VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wānanga o te Ūpoko o te Ika a Māui

DOCTORAL THESIS

A Novel Framework for Constructing Sport-Based Rating Systems

Author:
Ankit K. PATEL

Supervisor:
Dr. Paul J. BRACEWELL

A thesis submitted in fulfilment of the requirements
for the degree of *Doctor of Philosophy*
at Victoria University of Wellington, New Zealand

in the

School of Mathematics and Statistics
Te Kura Mātai Tatauranga

September 30, 2020

FORMAT STATEMENT

The format of this thesis is based on published works. It contains an unpublished introduction, three published papers provided throughout, Chapter Three, Chapter Four and Chapter Five, and a pre-published chapter presenting a novel evaluation metric. The published works are presented in their original format with permission from the Journal of Sport and Human Performance, Journal of Sports Analytics, International of Journal Sports & Coaching, Journal of Quantitative Analysis in Sport, and the 14th Australian Conference on Mathematics and Computers in Sport.

DECLARATION OF AUTHORSHIP

This thesis is presented to fulfil the requirements of a Doctor of Philosophy at Victoria University of Wellington. I hereby declare that this thesis and the work presented in it is entirely my own. No other person’s work has been used without due acknowledgement in this thesis. All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged.

Ankit Patel September 30, 2020

ABSTRACT

This doctoral thesis examines the multivariate nature of sporting performances, expressed as performance on context specific tasks, to develop a novel framework for constructing sport-based rating systems, also referred to as scoring models. The intent of this framework is to produce reliable, robust, intuitive, and transparent ratings, regarded as meaningful, for performance prevalent in the sport player and team evaluation environment. In this thesis, Bracewell's (2003) definition of a rating as an elegant form of dimension reduction is extended. Specifically, ratings are an elegant and excessive form of dimension reduction whereby a single numerical value provides an objective interpretation of performance.

The data, provided by numerous vendors, is a summary of actions and performances completed by an individual during the evaluation period. A literature review of rating systems to measure performance, revealed a set of common methodologies, which were applied to produce a set of rating systems that were used as pilot studies to garner a set of learnings and limitations surrounding the current literature.

By reviewing rating methodologies and developing rating systems a set of limitations and communalities surrounding the current literature were identified and used to develop a novel framework for constructing sport-based rating systems which output measures of both team and player-level performance. The proposed framework adopts a multi-objective ensembling strategy and implements five key communalities present within many rating methodologies. These communalities are the application of 1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling procedure to produce an overall rating.

An ensemble approach is adopted because it assumed that sporting performances are a function of the significant traits affecting performance. Therefore, performance is defined as $performance = f(trait_1, \ldots, trait_n)$. Moreover, the framework is a form of model stacking where information from multiple models is combined to generate a more informative model. Rating systems built using this approach provide a meaningful quantitative interpretation performance during an evaluation period. These ratings measure the quality of performance during a specific time-interval, known as the evaluation period.

The framework introduces a methodical approach for constructing rating systems within the sporting domain, which produce meaningful ratings. Meaningful ratings must 1) yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions and small sample sizes (robust), 2) be accurate and produce highly informative predictions which are well-calibrated and sharp (reliable), 3) be interpretable and easy to communicate and (transparent), and 4) relate to real-world observable outcomes (intuitive).

The framework is developed to construct meaningful rating systems within the sporting industry to evaluate team and player performances. The approach was tested and validated by constructing both team and individual player-based rating systems within the cricketing context. The results of these systems were found to be meaningful, in that, they produced reliable, robust, transparent, and intuitive ratings. This ratings framework is not restricted within the sport of cricket to evaluate players and teams' performances and is applicable in any sporting code where a summary of multivariate data is necessary to understand performance.

Common model evaluation metrics were found to be limited and lacked applicability when evaluating the effectiveness of meaningful ratings, therefore a novel evaluation metric was developed. The constructed metric applies a distance and magnitude-based metrics derived from the spherical scoring rule methodology. The distance and magnitude-based spherical (DMS) metric applies an analytical hierarchy process to assess the effectiveness of meaningful sport-based ratings and accounts for forecasting difficulty on a time basis. The DMS performance metric quantifies elements of the decision-making process by 1) evaluating the distance between ratings reported by the modeller and the actual outcome or the modellers 'true' beliefs, 2) providing an indication of "good" ratings, 3) accounting for the context and the forecasting difficulty to which the ratings are being applied, and 4) capturing the introduction of any subjective human bias within sport-based rating systems. The DMS metric is shown to outperform conventional model evaluation metrics such as the log-loss, in specific sporting scenarios of varying difficulty.

ACKNOWLEDGEMENTS

During this thesis I have received support, encouragement and advice from my supervisors, academics, sports people, fellow postgraduate students, and business leaders. I thank all these people whole-heartedly for their assistance, understanding and contributions during the research process.

I am especially grateful to the input provided by my supervisor, Adjunct Research Fellow Dr. Paul Bracewell, who inspired my pursuit of statistics, nurtured my passion for sport statistics, encouraged me to chase my ideas and provided me with the necessary direction and support required to complete this research. I cannot speak highly enough of Paul's wonderful influence upon my research and development as a student and a data scientist. I am forever grateful.

The encouragement and guidance provided by my co-supervisor, Associate Professor Ivy Liu, is also greatly appreciated.

The input of the data scientists and data engineers at DOT Loves Data was also extremely valuable. I am also grateful to Umano, Penny and the management and coaching staff at New Zealand Cricket for providing excellent support relating to the commercial feasibility of the models created. Additional to the academic and business contribution was the beneficial assistance provided by the numerous top-level coaches, selectors and managers who made themselves available for discussions and freely voiced their opinions.

My sincere thanks also go to Jason Wells, Mike Brough and Matt West, who provided me an opportunity to join their team as a budding young data scientist, and who gave access to their equipment and facilities. Without their support it would not have been possible to conduct this research.

I would like to thank DOT Loves Data for providing the funding for the uptake of this research and providing an excellent working environment.

Finally, I must express my gratitude to my parents, Suresh and Daxa Patel, and my wife, Madeleine, for their continued support and encouragement.

For Suresh and Daxa Patel

In Memoriam

Parbhubhai Patel
Laxmiben Patel
Lalbhai Patel
Shantiben Patel

Contents

Chapter One

AN INTRODUCTION TO RATING SYSTEMS

"Information is a source of learning. But unless it is organized, processed, and available to the right people in a format for decision making, it is a burden, not a benefit".

William Pollard, Physicist.

Pollard, C. W. (2011). The Soul of the Firm.
Illinois: The ServiceMaster Foundation.

1.0 INTRODUCTION

In the recent decade there has been a significant growth in the demand for data-driven rating systems. This growth in demand for such data-driven models to assess behaviour, expressed as performance on context specific tasks, has been experienced across many industries, although this effect is most evident across three major industries: 1) sport, 2) finance and 3) technology. Specifically, the following area within each industry have received considerable academic and commercial attention, respectively, the evaluation of team and player performance, the evaluation of an applicants' creditworthiness and repayment behaviour, and developer assessment – evaluating a developers' coding ability across three dimensions technical, procedural and behavioural. These modelling systems quantify the effectiveness of performance by producing a quantitative interpretation of performance, and consequently, are referred to as *rating systems*.

The modelling applications of such systems have an objective of evaluating, rating, and forecasting performance, such as player and team performance, a loan applicants' repayment behaviour and a developer's coding ability.

As evidence of this growing demand, DOT Loves Data was approached by three separate organisations, Umano[1], Penny[2] and New Zealand Cricket, to develop three bespoke rating systems. Consequently, DOT funded this research to develop a ratings framework to construct rating systems that can be commercially deployed across the sporting, credit-risk, and developer domains and systems that output meaningful ratings. Although this research was initially funded to develop a ratings framework to construct sports, credit-risk, and developer-based systems, this thesis purely focusses on the development of a novel framework to construct rating systems within the sporting context, also referred to as sport-based rating systems. This is due to the commercial sensitivity of credit-risk and developer data, intellectual property agreements and non-disclosure agreements.

This research extends Bracewell's (2003) definition of ratings, who stated that ratings are an elegant form of dimension reduction and enable the simplification of massive amounts of data into a single quantity. Specifically, ratings are an elegant and excessive form of dimension reduction whereby a numerical value provides a meaningful quantitative interpretation of performance. Meaningful ratings are defined as: 1) robust – the rating system must yield good performance where data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes. 2) Reliable – ratings produce accurate and highly informative predictions which are well-calibrated and sharp ratings. 3) Transparent – interpretable and easy to communicate. 4) Intuitive – ratings

[1] A software company which evaluates a developer's coding and programming ability.
[2] A peer-to-peer lending company.

must relate to real-world observable outcomes and the context to which the system is being applied.

This chapter introduces sport-based rating systems, specifically within the domain of credit risk application and sporting team and player-based evaluation, explains the research objectives, the methodologies and philosophies adopted in this thesis, and summarises the key methods and results surrounding rating systems.

1.0.1 Prologue

Invariably, when comparing attributes of players, teams, employees', products, or services, the conversation revolves around rating or ranking performance and ability, or the perception of performance and ability. This fascination with evaluating, rating, and ranking the outcomes of interest, such as performance, is inherent across many disciplines, and in recent decades the need to objectively quantify such ratings has garnered a large amount of academic and commercial attention.

The demand for quantitative methods to assess and measure performance has exponentially increased within the commercial industry due to the growth of big data, machine-learning, artificial intelligence, and data-driven business models. A few examples are outlined below.

Uber, a multinational peer-to-peer ride sharing company, implements a rating system which evaluates driver performance on four dimensions: conversation, vehicle cleanliness, timeliness, and safety, and assessing passenger '*performance*' on three dimensions: waiting times, courtesy and safety. This system allows both the driver and passenger to evaluate each other's performance during a trip. It enables the driver to understand the "type" of passenger/s they are picking up and enables the passenger to understand the "type" of driver. Effectively, the driver and passenger can evaluate "performance". A similar rating system is implemented by AirBnB, an online marketplace for arranging or offering lodgings, home stays or tourism experiences. Based on subjective inputs the AirBnB system allows the hostess and host to evaluate each other in a similar fashion.

Netflix, an American media services provider and production company, that implements a matching algorithm that evaluates the types of television series and movies that an individual has previously watched and based on past viewing behaviour recommends other TV shows and movies. Effectively, the rating system evaluates the quality of these matches using past actions.

FICO (FairIssac), a credit-scoring services company, is another example of an organisation that implements a rating system within their core services offering. FICO's rating systems rank-order consumers by how likely they are to pay their credit obligations as agreed (Smith, 2011). Effectively, FICO evaluates an applicant's ability to repay credit-obligation and evaluating repayment behaviour.

These are a few examples of rating systems applied within the commercial environment, there are many more organisation that implement data-driven rating or scoring systems within their core offering such as Numerix (https://numerix.com/), FindFace (https://findface.ru/), AlchemyAPI (https://www.ibm.com/watson/alchemy-api.html), Isograph (https://www.isograph.com/), FIFA player ratings (https://www.ea.com/games/fifa/fifa-20/ratings). This growth in demand for the implementation of *"rating systems"* has been experienced across many industries, however this effect is most evident in three major industries: 1) sports, 2) finance and 3) technology. Specifically, evaluating team and player performance, evaluating an applicants' creditworthiness and repayment behaviour, and developer assessment - evaluating a developer's coding and programming ability, respectively.

Specifically, this thesis focuses on the development of sport-based rating systems which output meaningful results. The primary reasons why rating systems within the sporting environment are chosen to construct the ratings framework is the growing demand of such systems, which assess player and team performance, increases the need to measure the performance and validity of the underlying model.

The fundamental philosophy adopted in thesis continues and extends the research explored at the masterate level by Patel (2016) which developed an optimised player rating and team selection algorithm for T20 cricket. This extension is two-fold: 1) developing a framework to construct sport-based rating systems, at both the team and player-level, and 2) developing a novel performance metric to evaluate the effectiveness of sport-based rating systems.

This thesis contains extracts from Patel (2016), specifically sections of the literature review, however full references and appropriate acknowledgements have been provided where necessary.

1.0.2 Research Motivation and Commercial Sensitivity

During the research process DOT Loves Data, a data science and statistical analysis agency, partnered with Umano, a software company, to construct a rating algorithm which dynamically measures the effectiveness of developers and programmers using their technical, process and behavioural capability.

Since March 2018, the Umano product has been operationalised and deployed across several developer teams across different organization, primarily within the banking and finance sector. The deployed models have been tested and validated across many scenarios and shown to provide managers with invaluable insight when evaluating employee performance in real-time. The technical details surround the underlying Umano models will not be discussed or disclosed in this thesis due to commercially sensitivity, and the intellectual property being owned by www.umano.tech. Although, the methodology used to produce the underlying rating models is outlined in Chapter Three. The deployed Umano models adopted an ensemble forecasting strategy applying a supervised hierarchical network-based approach to evaluate a developer's *technical*, *process* and *behavioural* traits. Moreover, the framework applied to develop the beta version has been peer-reviewed and published (please see Bracewell, P. J., Patel, A. K., Blackie, E. J., & Boys, C. (2017). Although the beta models do not adopt the ratings framework constructed in Chapter Three, there are communalities with the ensemble strategy outlined in Chapter Three and the work published in Bracewell, Patel, Blackie & Boys (2017).

During the research process DOT Loves Data partnered with Penny, a peer-to-peer lending service, to develop an application credit-risk scorecard to dynamically produce credit-scores for loan applicants. Again, given the commercial sensitivity of this work, this thesis does not disclose the technical details surrounding the development of the deployed scorecard.

Finally, in May of 2018 DOT Loves Data was approached by RugbyPass (https://index.rugbypass.com/), the premier online destination for global rugby fans, to build a revolutionary rugby rating system based on individual skill executed in real-time. This rating system applies a unique position-and-point-based approach which allocates players points based on their contribution to winning, during a rugby match. The individual and team-based rating system has been peer-reviewed and published (please see Moore. W. E., Rooney. S. J., Bracewell. P.J., & Stefani. R. (2018), and Bracewell. P.J., McIvor, J., Moore, W. E., & Stefani. R. (2018)).

Therefore, given the number of commercial entities showing an appetite for rating systems, and the prevalence and need for commercial rating systems, DOT Loves Data funded this research. Given the commercial sensitivity of this work, this thesis only develops sport-based rating systems. Therefore, the credit-risk rating system developed for Penny nor the developer rating system built for Umano are disclosed. The focus of this thesis is purely on the development of a ratings framework to construct meaningful rating systems applicable within the sporting context.

The motivation for this research is two pronged: 1) develop a quantitative ratings framework to construct sport-based rating systems that output meaningful ratings and 2) construct a performance metric to quantify the effectiveness of meaningful sport-based ratings (please see Section 1.9 for more details). Based on commercial needs, the research identified the need for a ratings framework that produced systems which produce meaningful ratings, defined as reliable, robust, intuitive, and interpretable. This definition was established by DOT Loves Data and the interested parties. Moreover, from a technical lens the secondary research motivation was to develop a novel performance metric that assess the effectiveness of meaningful sport-based rating systems.

1.0.3 Publications and Contribution to Knowledge

Throughout the research process several novel rating systems were published in academic journals and conference proceedings. Specifically, rating systems evaluating sporting teams and players, assessing a credit applicant's credit worthiness, and assessing a developer's coding ability were developed. A full list of peer-reviewed conference proceedings and journal publications are provided in Appendix A. These peer-reviewed publications provided the necessary exploratory research to identify the key elements required in a ratings framework.

This thesis resulted in 19 peer-reviewed publications, with 2 papers currently under review and 2 papers in preparation.

In performing an exploratory analysis and developing rating systems using current modelling methodologies, the communalities and limitations of the existing ratings literature are identified and addressed. The peer-reviewed rating systems that have been published as a result of this thesis have made meaningful contributions to the body of knowledge.

During the literature review of rating systems, specifically within sports and credit-risk (Chapter Two), the following limitations within the knowledge base were identified:

- *Lack of a sport-based ratings framework*– given the prevalence of sport-based ratings within the commercial and academic environment no modelling framework or approach currently exists in the literature to construct meaningful sport-based rating systems.
- *Lack of meaningful rating systems* – the literature echoed the sentiment expressed by Bracewell (2003); ratings are an elegant form of dimension reduction. Throughout this chapter it was shown that variable selection and dimension reduction are crucial elements of ratings methodologies. Although, given the loss of information during dimension reduction and the application of "black box" modelling techniques to produce ratings, the resultant ratings lack transparency and intuition, implying that results cannot be mapped to real-world observable outcomes.
- *No evaluation metric to assess the effectiveness of meaningful sport-based ratings* – to evaluate the predictive accuracy of the developed rating systems commonly applied evaluation metrics such as log-loss, root mean square error (RMSE) and mean absolute error (MAE), were used. Although, given the uniqueness of sport-based rating systems, it is necessary to construct a specific performance metric which quantifies the effectiveness of meaningful sport-based ratings.

Given the current limitations of the rating system knowledge base, the primary contribution of this thesis is the development of a modelling framework to construct rating systems to evaluate sports team and player performance. The framework is developed by constructing preliminary rating systems, specifically within the sporting context, identifying the communalities, distinctions, and limitations of these systems, and implementing and addressing these when developing the ratings framework, respectively. Specifically, sports-based rating systems were developed to identify these commonalities, distinctions, and limitations because data is not commercially sensitive, readily available and accounts for a range of sporting scenarios.

The proposed framework (Chapter Three) adopts a multi-objective ensembling strategy. An ensemble approach is adopted because it assumed that performance is a function of the individual traits that significantly affect performance. Therefore, performance is defined as $performance = f(trait_1, \ldots, trait_n)$. Moreover, the developed framework is a form of model stacking where information from multiple models is combined to generate a more informative model.

Given the lack of an evaluation metric to quantify the effectiveness of sport-based ratings, the secondary contribution of this thesis is the construction of a novel performance metric to quantify the effectiveness of sport-based ratings. This novel evaluation metric, which applies distance and magnitude-based measure associated with spherical scoring rule methodology (please see Chapter Three), is shown to be a better evaluator of sport-based ratings than commonly used evaluation metric, such as log-loss, in certain scenarios.

Throughout this research the author has significantly contributed to the body of knowledge. Specifically, the author has developed a novel ratings framework that produces robust, reliable, transparent, and intuitive ratings, and validates its worth within the sporting domains. Further, a novel performance metric which quantifies the effectiveness of meaning sport-based ratings is developed.

Furthermore, because of this research a substantial body of novel work has been generated during this research process. The work has been peer-reviewed and published and includes six journal articles and 13 peer-reviewed conference proceedings. In addition, the papers titled "*Estimating the expected total in the first innings of T20 cricket using gradient boosted learning*" (Patel, Bracewell & Bracewell (2018)) and "*A framework to quantify the impact of social engagement on data driven creative*" were awarded the Neville De Mestre Prize for best student paper and presentation, respectively, at the 14th Australian Conference on Mathematics and Computers in Sport (MathSport) conference[3].

1.0.4 Software and Hardware

Analyses and statistical programming were executed in R (R-GUI 64-bit v3.6; R Core Team, 2018). This is an S-PLUS statistical programming environment for statistical computing and graphics. The choice of software was determined by the extensibility for modelling packages and the need for flexible object-oriented data manipulation. By using R, which is free, open-

[3] MathSport is a special interest group of ANZIAM, the Australia New Zealand Industrial and Applied Mathematics organisation.

source and readily available over the Internet, all procedures carried out can be reviewed and replicated. All research was carried out on a desktop computer equipped with an Intel(R) Xeon(R)™ CPU 2.4HGHz (2 processors), 48GB RAM, running 64-bit Windows 7 Professional.

1.1 IDEOLOGY

Using objective match statistics from multiple sports, such as cricket, rugby and golf, and personal data from credit applicants, this thesis seeks to develop a framework for constructing rating systems within the sporting domain to evaluate team and player performances. The framework should construct sport-based rating systems which produce meaningful ratings, specifically reliable, robust, intuitive, and transparent ratings.

Techniques from multivariate analysis, ensemble forecasting strategies and machine learning techniques for regression and classification are implemented to develop a novel framework for constructing sport-based rating systems. The fundamental technique adopted in this thesis, to develop the ratings framework is an ensemble forecasting strategy. Birthed in meteorology, ensemble forecasting strategies are prevalent amongst meteorologist as they allow the use of many models and model uncertainties to understand a range of possibilities of future weather to evaluate the most likely outcomes. Moreover, atmospheric scientists have developed much of the underlying methodology of ensemble forecasting and the ensemble forecasting strategies adopted by such experts fall into one of two categories: 1) ensembles based on many different models and 2) ensembles based on many runs of one-computer model initialised from slightly different data (Kunst & Jumah, 2004).

Distance and magnitude-based measures associated with a proper scoring rule methodology, specifically a spherical rule, are used to develop a novel performance metric to quantify the effectiveness of meaningful sport-based performance ratings. The evaluation metric provides a unique way to compare ratings across different forecasting scenarios of varying forecasting difficulty.

As mentioned, this thesis develops a framework to construct meaningful rating systems by applying modelling methodologies prevalent within the credit risk, sporting, and the developer domains. Such rating systems produce a numerical interpretation of performance, which is defined as a function of the individual traits significantly affecting behaviour, expressed as performance on context specific tasks. Liu & Pentland (1999) developed an approach to model behaviour which considers the human as a device with many internal mental states, or traits, each with its own control behaviour and interstate transition probabilities. This approach is like that outlined in Chapter Three, whereby each trait is modelled individually, and these individual ‘trait-based’ ratings are ensembled to produce an overall rating representing a quantitative interpretation of performance.

A range of statistics relating to a developer's ability and an applicants' credit history are applied to construct Umano's and Penny's rating systems, respectively. Although, these are commercially sensitive and therefore details are not disclosed in this thesis.

As mentioned, although DOT funded this research to develop a rating framework to construct rating systems across multiple domains, the scope of this thesis is limited to the sporting context, specifically developing a framework to construct sport-based rating systems.

Chapter Two discusses and outlines the distinctions, communalities, and limitations between rating systems across the credit risk environment and the sports industry. Before developing rating systems and applying statistical techniques, the justification for implementing such procedures must be proven. Chapter Two reviews the large pool of literature supporting the application of statistics for assessing sports performance and evaluating a loan applicants' creditworthiness. Chapter Two also introduces the importance of rating systems across the credit-risk and sporting domains.

1.2 RESEARCH PROCESS

The research process was driven by the commercial needs of DOT Loves Data to develop their artificial intelligence capability. Dr. Paul Bracewell, Managing Director of DOT Loves Data, provided invaluable feedback regarding potential models and approaches from a commercial and statistical perspective.

Due to the commercial environment and the organisations industry connections, the access and quality of the data available was unparalleled. With input from some of New Zealand's top sporting minds, such as former Black Cap Grant Elliot and White Ferns cricket selector Jason Wells, credit risk expert, such as Dr. Paul Bracewell formerly a General Manager at Dun & Bradstreet (a company that provides commercial data, analytics and insights for business) and some of Australasia's top computer scientists and developers. The insight and feedback in creation of the data collection process and associated systems was invaluable. Furthermore, the commercial involvement led to direct discussions with other top-level experts and executives which was necessary to ensure the resultant ratings framework and statistics were suitable for the consumption by decision makers. These meetings proved crucial in establishing the justification of adopting the proposed ratings framework and in establishing how such a framework can be deployed within different environment.

In developing the initial Umano models and Penny's credit risk scorecard, the time pressure was immense, with the official beta launch of Umano and the deployment of an operationalised scorecard occurring simultaneously, in late March 2017, approximately four months after commencing analysis of the data. As mentioned, the operationalised and deployed Umano and Penny models are commercially sensitive, and their technical details have been excluded from this thesis. This initial work acted as a pilot study for the subsequent work. The knowledge gained through this pilot process proved crucial when constructing the novel ratings framework. Specifically, the feedback received from weekly stand-ups and fortnightly retrospective sessions, as part of the agile development methodology used by the software developers, with Umano and Penny provided the necessary insight for the requirements of a rating system aimed at the credit risk and developer environment.

Prior to commencing the research process DOT had an existing relationship with a San Francisco based start-up, a funding platform for athletes, and New Zealand Cricket (NZC). In early 2016, the start-up approached DOT to identify 'up and coming' or 'dark horse' golfers in the Professional Golfer Association (PGA) tour. Effectively, this organisation wanted to identify the golfers with untapped potential and those players who were on their way to becoming superstars but had not yet been identified and invest monetary resources into these players. Therefore, to address this question DOT developed a predictive PGA performance rating system, which was peer-reviewed and published (please see Patel, A. K., Rooney. S. J., Bracewell, P. J., & Wells. J. D. (2018)).

Moreover, DOT was commissioned to provide a player rating and team optimisation system, which measured a player's performance and selects the optimal team based on their performance. This work was commissioned leading up to the T20 Cricket world cup. The prototype of this rating and optimisation system was peer-reviewed and published (please see Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2017)).

The communalities and limitations of these exploratory rating systems were identified, and the thesis sought to create an expert ratings framework capable of mirroring the opinions and observations of 'unforgetful' human experts from observable statistics using an ensemble forecasting strategy. Throughout the research process it was found that a careful balancing act between pure statistical methodology and creative statistics was necessary to ensure the models produced meaningful ratings (i.e. reliable, robust, intuitive and transparent), providing an appropriate quantitative interpretation of performance.

At this point the bilateral theme that predominates this thesis becomes evident. To fully understand the entirety of this work, a background of ratings system across the credit risk and sporting domain is required, and an understanding of the key statistics and data required needs to be understood. The two separate themes run parallel throughout, with the emphasis shifting continuously. Therefore, the material within this thesis is compartmentalised. Specific sections deal predominately with current rating philosophies and prevalent statistical techniques (Chapter Two and Chapter Three). The implementation and validation chapters (Chapter Four and Chapter Five) provide common ground and show how to effectively utilise the information generated from both a ratings and statistical perspective.

1.3 RATINGS FRAMEWORK

The ratings framework developed in this thesis helps construct sport-based rating systems that produce reliable, robust, intuitive, and transparent ratings. The framework draws influence from two key fields in which rating systems have been very prevalent and experienced an increase in recent times: 1) credit risk scorecard and 2) sport ratings systems (Chapter Two).

The primary contribution of this thesis is the implementation of an ensemble forecasting strategy to develop an approach for constructing sport-based rating systems which produce meaningful ratings of behaviour, expressed as performance on context specific tasks. Meaningful ratings are defined as robust, reliable, transparent, and intuitive outputs. *Robust* ratings yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions and small sample sizes. *Reliable* ratings are accurate and provide highly informative predictions, implying they are well-calibrated and sharp. *Transparent* ratings are interpretable and easy to communicate. *Intuitive* ratings can be mapped to real-world observable outcomes; effectively incorporating forecasting contextuality.

In this thesis, performance is defined as a function of individual traits and can be notational represented as $performance = F(trait_1, \dots, trait_n)$. To effectively quantify performance, various statistical techniques capable of dimension reduction and feature selection are applied to extract the traits affecting performance and the features that significantly affect these traits, respectively.

To extract meaningful ratings that provide a numerical interpretation of performance, a multi-objective ensemble forecasting strategy has been adopted. Specifically, an ensemble forecasting strategy is applied to combine forecasts derived from statistical methods that differ substantially and draw from different sources of information leading to improved forecasting accuracy. Given performance (or behaviour) is a manifestation of different traits (van Strien, 1986; Argyle & Little, 1973; Halder, Roy & Chakraborty, 2017; Heinström, 2003) ensemble forecasting is an appealing modelling approach because instead of choosing a single method, a collection of "best" methods is selected to improve overall accuracy.

Given that the first task of rating systems is to identify the different traits that significantly affect performance expressed from the data, the problem lends itself to the field of dimension reduction. Assuming each dimension represents a specific trait and further, assuming there are multiple traits that define performance, a multi-objective approach is appropriate, where each modelling objective relates to a specific trait, and the outcome of each objective produces a trait-based rating. This trait-based rating provides a numerical representation of the trait. It is assumed that for each trait to account for a sufficient amount of uncertainty in performance, different feature-types (action, context and time) across varying levels of complexity must be used to derive the trait-based ratings.

After generating trait-based ratings for each significant trait, an ensembling forecasting method is used to combine the different trait ratings to produce an overall performance rating. This approach assumes that performance is a manifestation of individual traits, and these traits are measurable based on the observations of physical tasks performed in different conditions which are recorded as *action*, *context*, or *time*-based variables. This assumption has been heavily researched and validated within the academic literature (please see Bracewell 2003; Gonzalez, Mens, Colacioiu & Cazzola, 2013; Plomin, Owen & McGuffin, 1994; Delaney, Harmon & Ryan, 2013).

Ultimately, rating systems are a form of data scoring, also referred to as scoring models, a term commonly used within the data mining environment, which means filling in the outputs (Berry & Linoff, 2004). Scoring systems have been used in a range of academic fields, such as assessing an individual's repayment behaviour and calculates their risk of defaulting on a line of credit (please see Arsovski, Markoski, Pecev, Ratgeber & Petrov, 2014; Marikkannu & Shanmugapriya, 2011; Pedreschi, Giannotti, Guidotti, Monreale, Pappalardo, Ruggieri & Turini, 2018). A wide range of statistical and data mining techniques are applied to enable scoring to occur (please see Kitts, Freed & Vrieze, 2000; Langley, 1997; Grady, Schryver & Leuze, 1999). Effectively, different characteristics and dimensions are extracted from the data to provide interested parties suitable ratings from which dimensions and predictions can be derived (Berry & Linoff, 2004). Generally, scoring or rating is associated with dimension reduction, aligning with Bracewell's (2003) definition that ratings are an elegant form of dimension reduction; whereby high dimensional data is reduced to fewer, more manageable dimensions.

Although there is the danger of losing crucial information in the dimension reduction and feature selection process (Bracewell, 2003), given it is difficult for a human observer to detect patterns in multivariate situations (Grady *et al.*, 1999), it is often necessary to condense a large number of variables into a more manageable set. This enables the identification of trends and patterns within the data. The danger of information loss can be managed or minimised techniques that eliminate the redundancies in the data by identifying the true dimensionality (i.e. key performance traits) of the input data.

Most rating systems need to produce predictive and accuracy outputs which are robust in the sense that they are applicable to everyone and everything within the target population (Bracewell, 2003). Bracewell (2003) stated that the risks associated with scoring a model are dependent on the intended use of the obtained information. This research extends Bracewell's (2003) definition of ratings, who stated that ratings are an elegant form of dimension reduction and enable the simplification of massive amounts of data into a single quantity. Specifically, meaningful ratings are an elegant and excessive form of dimension reduction whereby a numerical value provides a meaningful quantitative interpretation of performance.

It is a common misconception that reliability and robustness are the only requirements for meaningful ratings, however, if ratings do not make sense to the people who consume them or cannot be easily communicated, the ratings will never be used. Therefore, transparency and intuitiveness are also necessary. Indeed, misinterpretation of the ratings is the greatest threat to the success of rating systems, provided the key issues are resolved.

Whilst *rating* is an overly broad term, fundamentally, it is the ensemble forecasting approach that is most closely related to the development of a novel framework for constructing sport-based rating systems assessing team and player performance. Further, it is the scoring rule methodology that is most closely related to the construction of a novel model evaluation metric to quantify the effectiveness of meaningful ratings.

1.4 RELEVANCE OF RESEARCH

Using the wrong 'type' of rating system is a common occurrence because not everyone can develop their own rating system. Commercial systems are difficult to assess because transparency and intuitiveness is usually absent, primarily because software suppliers want to maintain their competitive advantage and intellectual property.

This thesis develops an important novel framework for constructing rating systems, which can be applied to introduce transparency, intuition, reliability, and robustness to any rating system to assess performance and intends to produce a meaningful numerical interpretation of performance.

The application of analytics in the business environment has recently experienced tremendous growth (Henke, Bughin, Chui, Manyika, Wiseman & Sethupathy, 2016). Business analytics has transformed from a "nice-to-have" to a competitive advantage. "In the past few years, predictive analytics, has gone from a practice applied in a few niches to a competitive weapon with a rapidly expanding range of uses" (CGI: Predictive Analytics, 2013, p.1).

> *"By using real-time data on the merchants' transactions to build its own credit scoring system, Alibaba's finance arm was able to achieve better non-performing loan ratios than traditional banks"* (Henke, Bughin, Chui, Manyika, Saleh, Wiseman & Sethupathy, 2016, p. 26).

> *"Many companies have implemented rule-based lead scoring models to identify which leads get handed over to sales teams"* (Ericsson, Dansingani, O'Hair, Jackson & Edin, 2018, p. 6).

> *"Rating systems were developed to help us [Bioz] choose services and products, such as a hairdresser or a new car, ultimately guiding us in our evaluation of quality, relevance and performance…. Rating systems are based on data; they are usually displayed on a scale of 1 to 100. As these ratings rely on algorithms that are hard to*

> *manipulate, and on measurable usage data, these ratings are both objective and trusted"* (Lachmi, 2018, p.1).

A key factor for the rise in business analytics is the phenomenon of big data, and its acceptance by senior executives as an important business enabler. The goal of insight and information extraction or revealing hidden patterns within big data is achievable through the application of mathematical and statistical techniques. Sagiroglu & Sinanc (2013) stated that modern analytics, characterised by improvements in computing power, reduced cost in data storage, greater access to various data sources and cheaper commodity hardware, requires a revolutionary step forward, moving away from traditional data analysis. The Transforming Data with Intelligence (TDWI) survey revealed that the application of advanced analytics creates better aimed marketing, increased business insights, client-based segmentation and recognition of sales and market chances. Through analytics businesses have seen the benefits described above and have also been able to develop analytical techniques and models that are the core of their competitive advantage and offerings. Moreover, these advanced analytics tools put information in the hands of business analysts and business users, offering significant potential to create business value and competitive advantage.

Through big data analytics, not only have businesses seen the benefits described above but they have also been able to develop analytical techniques and that are the core of their competitive advantage and core offerings. Advanced business analytics tools enable deeper insights and discovery that will change business assumptions (Seddon, Constantinidis, Tamm & Dod, 2017). Moreover, these tools put information in the hands of business analysts and business users and offer significant potential to create business value ad competitive advantage (Pratt & White, 2018). There are many well-known organizations whose competitive advantage rely on powerful mathematical and statistical algorithms such as Google's Page Rank algorithm, DeepMinds (https://deepmind.com/) general purpose neural network algorithms, and Facebook facial recognition algorithm. Without these models the business would not be able to sustain a competitive advantage. This increase in demand for big data analytic teams has created a high demand for those specialising in mathematical, statistical sciences, software engineering and computer scientists and data scientists as organisations seek to develop their machine learning and analytics capabilities, in an ever evolving data-driven environment.

The objective of these are data-driven and modelling intensive applications is to evaluate, rank, rate, or predict the performance of an individual or collection of individuals. This common thread dictates that the results must be robust, transparent, and meaningful (Bracewell, 2003). The following section will provide brief context as to why the sporting evaluation and credit risk environments have received considerable commercial and academic success and why they are experiencing growth in demand and academic attention.

1.5 COMMERCIAL RATING SYSTEMS

1.5.1 Sporting Industry

Within the sporting world statistical ranking and rating methodology have been heavily applied in the past two decades at both the individual and team level. The growth of sports analytics and the need for meaningful sports related statistics has emerged in recent decades due to the large volume of monetary resources that is increasingly invested into teams and individual players. Moreover, the rise in player salary caps over the last 25 years provide ample evidence of the growth in sports analytics, with investors, franchises, clubs, and other stakeholders wanting to determine the value of their investment decisions. For example, in the National Football League (NFL) there has been an increase of approximately 950% in player salaries since 1980's, and an increase of 288% in salary cap since 1994 (Vrooman, 2012). With global sports revenue growing by U$145.3billion over the 2010-2015 period (Coopers, 2015), at an annual compound growth rate of 3.7%, and winning teams earning significantly larger revenue than that of losing teams, there is a strong incentive for coaches and managerial staff of sport teams to succeed. Additionally, "the regulated sports betting market is forecasted to reach $70 billion in 2016, representing a 20% increase from 2016" (Foley-Train, 2014).

Given the large investment of resources and stakes involved, coaches, managers and other stakeholders cannot solely rely on subjective views and personal beliefs to make team and player selections. Solutions must be augmented with objective approaches by implementing analytical techniques to rank, rate, evaluate and forecast selection decisions. This need to make informed data-driven decisions has given rise to the use of sport analytics by managers, coaches, athletes, and fans. Forbes (2015) claimed that the popularity of data-driven decision making in sports has trickled down to the fans, which are consuming more analytical content than ever.

To derive a deeper understanding of the requirements for a meaningful sports rating system, the author has undertaken significant, novel research and meaningfully contributed to the body of knowledge. These findings have been published and peer-reviewed publications: Patel, Bracewell & Rooney (2016); Patel, Bracewell & Rooney, 2017; Patel & Bracewell, 2018; Patel, Bracewell & Wells, 2017; Brown, Patel & Bracewell, 2017; Campbell, Bracewell, Blackie & Patel, 2018; Patel & Bracewell, 2017; Greer, Patel, Trowland & Bracewell, 2018; Mansell, Patel & Bracewell, 2018; Simmonds, Patel & Bracewell, 2018; McIvor, Patel, Hilder & Bracewell, 2018; Patel, Bracewell, Wells & Brown, 2018; Patel, Rooney, Bracewell & Wells, 2018; Campbell, Patel & Bracewell, 2018; Patel, Bracewell & Bracewell, 2018; (please see Appendix A for a full list of the published work). These rating systems were developed prior to developing a ratings framework to construct sport-based rating systems and were used to understand the communalities, distinctions, and limitations of commonly applied practices to develop sport-based rating systems.

1.5.1.1 De-regularisation of the sports betting industry

During the research process the United States de-regularised the sports betting market and opened online sports betting outside the state of Nevada. This de-regularisation has significant implications on this thesis and makes the development of sport rating systems within the commercial environment extremely relevant. Therefore, this section is dedicated to outlining the market effects of this de-regularisation and the impact such as event has on this research.

On May 14th, 2018 the United States Supreme court removed the Professional and Amateur Sports Protection Act of 1992 (PASPA), which federally prohibited sports gambling under state law, ruling that the federal ban was unconstitutional, and opened the door for all states to legalize sports betting. This legalisation of sports betting will impact the relationships between leagues, gambling institutions, data providers, the government and how fans interact with games. It is expected that by 2024, approximately 70% of the United States will offer legal sports betting. "Reports state that the removal of the PASPA will lead the United States to become the largest sports betting market in the world, a massive high-tech industry centred on professional and amateur athletes and fuelled by hundreds of billions of dollars" (Silverman, 2019, p. 1).

The overturned PASPA (1992) legislation also opens the doors for sports bookmakers, new betting agencies, diverse betting options and sports-based analytics companies primarily driven by the influx of new customers and gambling participation in the marketplace. With the rising number of participants on both the demand and supply of sporting odds, the type of odds and the level of participation needed to identify and provide appealing and diverse betting options simultaneously increasing in the demand for intuitive, transparent, robust and reliable sports statistics, and team and player rating systems that are easily digestible by sports fans.

New entrants such as bookmakers and betting agencies must offer exotic bets, lucrative odds and diverse betting options to capture a significant proportion of this growing and sophisticated market. The rapidly expanding sports betting market and the rise in mathematical and statistical models to inform decision making within the sporting industry (both academically and commercially) highlights a need to develop a novel ratings framework that can be deployed across multiple sports and domains. This leads to an increase in a more sophisticated marketplace, supplying and demanding more informative sports analytics and statistics detached from subjectivity, bias and tradition.

Much like market finance sports bookmaking examines the environment with sophisticated algorithmic trading systems, running and constantly adjusting prices and odds as players or in-game events occur. However, unlike the financial markets, sports are governed by a set of physical rules and are measurable and understood (Blume, 2019, p.

1). “In less than one year that bookmaking has been legal in New Jersey, a number of European companies have swooped into offer services to the racetracks and casinos licensed to books sports bets. These companies offer turnkey operation complete with quantitative analysts, software and modelling for profiling bettors and managing risks, access to data from sports leagues, and worldwide pools of liquidity” (Hill, 2019, p. 1).

The introduction of new entrants in the sports betting market has a profound effect on the expected growth in revenue, within the industry, with the global gambling market expected to reach revenue of over USD525 billion by 2023, growing at a compound annual growth rate (CAGR) of approximately 4% between 2017-2023 (Cision, 2018). Moreover, according Zion Market Research (2019), the global sports betting market was valued at approximately USD104.31 billion in 2017 and is expected to reach approximately USD155.49 billion by 2024, growing at a CAGR of 8.83% between 2018 to 2024. The expected growth in the global gambling market is driven by increasing penetration of online gambling and betting across the United States and European region. With prognosticators estimating that betting dollars could reach $287 billion, currently $4.9 billion, and that total sportsbook revenue could reach $4.6 billion, currently $800 million, by 2021, there is tremendous opportunity for stakeholders to capitalize on.

As mentioned, the deregulation of Americas sports gambling market creates an opportunity to monetize, many key stakeholders see opportunities to monetize, while others raise concerns about the impact legalized gambling could have on the integrity of the game, and federal and state governments consider their roles and legislative next steps. It has been reported that the key will be to form relationships between gambling institutions, governing bodies, and leagues to share the pie and ensure integrity. This growth also underscores the need to develop real-time data feeds in conjunction with leagues to support real-time and in-game betting, particularly on mobile platforms.

America’s fastest growing industry with a new breed of sports gamblers are known as “wall-street types” and are “adept at figuring percentage odds and statistical permutations” (Hill, 2019). According to Hill (2019) there are two ways to make money in this business: 1) fundamental analysis and 2) technical analysis – using finely tuned models to analyse the team, player ratings and odds. The growing need for tools and data to make informed sports betting decisions and setting odds for book makers and increase profitability for both parties. The benefit of model-based decision making is its lack of subjectivity and free from bias.

In 2009, the sports betting market was valued at $20 billion, however by 2016, it was valued at $40 billion. The American sports betting market has consecutively grown at a rate of approximately $10 billion per year, with a present market capitalization of between

$60-$73 billion. If this growth rate continues, Americas sports betting market will occupy an increasingly significant share of the worlds sports gambling market (Gary, 2019).

“Sports betting currently accounts for upwards of 40% of global gambling revenue around the world which is greater than any other section (inclusion of lotteries, casinos, poker etc.). According to the latest projections from market research firm Technavio, the CAGR is expected to increase by a whopping 8.62% from 2018-2022” (Chain, 2018, p. 1). In terms of participation, online sports betting has surpassed all other forms of gambling including lotteries, casinos, poker etc and currently accounts for upwards of 40% of the worlds global gambling revenue. “According to the latest projections from market research firm Technavio, the CAGR is expected to increase by 8.62% from 2018-2022” (Chain, 2018, p. 1).

Charlton (2013) states that betting on NFL football experienced the largest growth in the sports betting industry, growing by 69% between 2009-2012. This growth was primarily driven by the rise of in-play betting, while gambling via electronic gaming machines fell 20%, from 39% to 19%, between 2009-2011, while participation in sports betting increased by 13% over the same period.

With this rapid expansion in the global sports betting market and its continued surge in popularity the demand for accurate and predictive sports statistics applied to derive sporting odds has never been so high. The sports betting industry is one of the fastest growing sectors in the world, and the legalization of it within the states exponentially increase the growth and will intellectualize sports, strengthen team and player statistics and the type of data collected in each sport, amongst stakeholders such as bookmakers, fans and sports analytics companies. There has been a rise in the number of model-based sports betting systems and academic literature which analyse line movement and public betting data to identify “smart money” bets, and automatically identify where and when to place lucrative bets.

Moreover, given analytical strategies are shared between different teams largely due to the high turnover rate among coaches and managers, there is rapid progression and implementation of various statistical analyses, which is partially responsible for the boom in the sports analytics industry over the past decade. One example that illustrates this lies in the MIT Sloan Sports Analytics Conference, an annual event that discusses recent developments in sports data analytics. In 2007, there were 175 attendees. However, in 2013, there were over 2200 attendees; and 3500 attendees in 2019, this is an increase of over 1200%. Historically, predictive sports modelling has been accomplished through mathematical, theoretical models, based on human intuition and other primitive means. However, with the recent technological advances in modern analytics, opportunities have arisen for a transition into data-driven modelling.

These market and industry movements highlight the need and importance of sports analytics and the research within rating systems presented throughout this thesis. Given this recent development, it has never been more pertinent to produce accurate and predictive sport team and player rating systems to inform decision-making surrounding team selection, enticing odds and analytical outputs that offer in-depth insight relative to traditional outputs.

1.5.2 Financial Industry

Due to the 2007-2008 Global Financial Crisis (GFC), in 2010-2011, the Basel committee on Banking Supervision introduced the Basel III framework with the intention of strengthening capital requirements by increasing liquidity and decreasing leveraging. This regulation changed the way credit scoring systems were built and the type of applicant attributes that each system must incorporate, increasing financial institutions demand for scoring systems that detect subtle changes in an applicant's attributes, associated with probabilities of default. Most credit scoring systems are based on the 12-month view of historical applicants' behaviour and an assumption that a customer's future performance is like their past performance. However, during the GFC many applicants who were financially stable for many years ended up in financial difficulty. This revealed that the adopted scoring methodology was not necessarily reflective of an applicant's credit worthiness and highlighted flaws in the current scoring methodology. Hand & Henley (1997) stated that the most widely used techniques for building scorecards are linear discriminant analysis, logistic regression, probit analysis, non-parametric methods, Markov chain models, recursive partitioning, expert systems, genetic algorithms, artificial neural networks and conditional independence models. These techniques are used to predict the probability of default in the next 6, 9, 12 or 18 months (Peussa, 2016; Bolton, 2009).

There exist subtle nuances in the application of statistical methods within the financial services sector. Specifically, given data is not missing at random, this requires an approach called reject inference. Moreover, to maintain interpretability and minimise the impact of collinearity the data fed into the model in a manner satisfying commercial constraints by iteratively modelling on the residuals. This thesis has derived novel applications of reject inference techniques and outlines work on modelling residuals published in peer-reviewed journals. Please see Baez-Revueltas, 2009; Einarsson, 2008; Shad & Rehman 2012; Anagnostopoulos & Abedi 2016; Roy, 2016; Tabagari, 2015; Torosyan, 2017; Patel *et. al.* (2017).

These findings have been published in peer-reviewed academic journals (please see Bracewell, Coomes, Nash, Rooney, Patel & Meyer, 2017; Patel, Bracewell, Gazley & Bracewell, 2017; Patel, Bracewell & Coomes, 2018). Please see Appendix A and the Supporting Publications document for full copies of the published work.

1.6 SUMMARY OF COMMERCIAL RATING SYSTEMS

To successfully develop a ratings framework that is applicable within the sporting domain, the commonly used methodologies and statistical techniques, sector-specific terminology and underlying philosophies that shape these rating systems must be understood. In this section the key methods, the type of rating systems, important terminology and commonly used techniques within credit-risk and sports are summarised.

1.6.1 Credit Risk

Credit scoring is the term used to describe statistical methods used for classifying applicants for credit into good (low probability of loan defaults) and bad (high probability of loan defaults) classes. Such methods have become increasingly vital with the remarkable growth in consumer credit in recent years. Credit scoring has become one of the most successful application areas for statistical and operational research.

To measure risk, financial institutions apply statistical analysis called credit scoring to help make credit decisions. Credit scoring produces a numerical value known as a *credit score* measuring the likelihood of an individual ability to repay their debt sometime in the future. A high credit score indicates a lower likelihood of default, while a low credit score indicates a higher likelihood of default (i.e. an increased likelihood of not repaying the debt in the future). A customer with a high credit score is known as a *good customer* while a customer with a low credit score is known as a *bad customer*.

The financial industry has been utilizing statistical rating/ framework methodologies for many decades to evaluate consumer creditworthiness (i.e. credit risk scorecards), model corporate and developing scoring models for retail exposures. The financial industry regulates such development through the Basel Framework. “The Basel framework has three sets of banking regulations (Basel I, II and III) set by the Basel committee on Bank supervision, which provides recommendations on banking regulations regarding capital risk, market risk and operational risk” (Basel Committee, 2010, p.1). The framework aims to ensure that financial institutions have enough capital on account to meet obligations and absorb unexpected losses.

Such scoring models evaluate risk by applying statistical analysis so that the users can score an individual’s, group or businesses credit worthiness to help make decisions on the amount of risk to take, credit to provide, provide to promote etc.

Credit-scoring models are used by insurance companies, mobile phone companies, government departments, landlords, and their use continues to expand. Credit has existed in various formats for many years but in recent times consumer credit has increased in the form of credit cards, home loans, personal loans etc. This has resulted in a widespread use of credit scoring. However, there are many aspects of the methodology that have not received enough attention in the academic literature, due to the need for confidentiality resulting in a lack of availability of datasets for investigation purposes. Both application and behavioural scoring rely on the development of classification tools using statistical analysis (Hand, 1981; Johnson & Wichen, 1998).

There are two main types of credit risk scorecards widely used in the finance industry: 1) Behavioural scorecards, and 2) Application scorecards. Broadly speaking, banks apply application and behavioural scoring to deal with two different types of customers requiring different types of decisions: 1) *New customers'*– should the new applicants for credit be granted? And 2) *Existing customers'* – should the agency grant the request of an old customer to increase credit limit? How risky are the existing customers? What products to offer to the existing customers to maximize the profit? Application scoring is applied to determine the answer to the first question, while behavioural scoring is applied to answer the second questions.

1.6.1.1 Application scoring

Application scoring is more common in the literature to the point that when credit scoring is discussed, one automatically thinks of application scoring, however the literature on credit scoring is scarce due to the sensitivity in the data. Literature on behavioural scoring is almost non-existent. In this proposal credit scores refer to both application and behavioural scoring.

1.6.1.2 Behavioural scoring

Behavioural scoring has become an important task in the credit industry. Behavioural scoring has many benefits including closer monitoring of existing accounts, reductions in credit analysis costs, faster credit decisions and prioritizing credit collections (Brill, 1998). These types of scoring models aim to group customers that share similar behavioural patterns. Using these patterns, banks target different groups to promote new products, increase credit limits, target the group which will be encouraged to spend more and come up with strategies to manage recovery if a customer's repayments ability turns bad. In behavioural scoring models, historical transaction behaviour and payments are considered assuming that the customers' behaviour will be similar in the future. To model a customer's behaviour, behaviour scoring models establish an association between input variables and an output score, which measures the probability of default. Based on these associations, a score is assigned to each customer and customers are clustered into group for marketing purposes. The typical scoring method usually involves the steps shown in Figure 1.

The appropriate statistical framework is *'classification'*. This is an old modelling paradigm with mature literature review. Approaches to binary classification are legion within the credit scoring discipline, as credit card applicants are either *'good'* or *'bad'* defined by the probability of default or delinquency (number of days of missed payments).

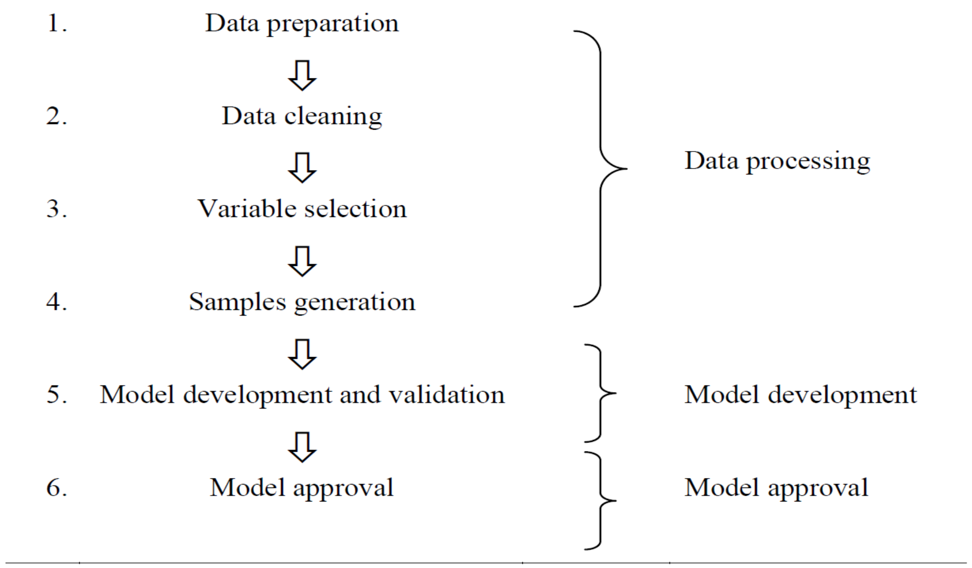


Figure 1: Process to create credit risk scorecards

1.6.2 Sport-based Rating Systems

Applying analytical techniques to quantitative sports data allow users to rank and evaluate player and team performances. A rank refers to ordinal placement of ratings, while "ratings come from a continuous scale such that the relative strength of team individual is directly reflected in the value of its rating" (Massey, 1997, p. 2). Early sports modelling work was based on ratings methodologies (Stefani, 2011).

In general, sport rating systems provide an objective evaluation of a team or individual based on prior performances and are implemented for player comparisons and improving player and team selection process. Generally, such systems are used by coaches, players, team managers and other key stakeholders.

Formally, a sport rating system assigns each team or individual a single numerical value representing a team or individual's strength relative to the rest of the league on some predetermined scale (Massey, 1997). These ratings are beneficial to numerous parties, especially athletes, coaches and managers who utilise such systems to track and predict form, progress and applied as a motivational and benchmarking tool. According to Leitner *et al*. (2010) sport ratings are typically derived by suitably aggregating a competitor's previous performances and provide predictive power in forecasting future performances. Many American sporting franchises, such as The Oakland A's (Baseball) and Dallas Mavericks (Basketball), adopt such an ideology.

Using a common framework, Stefani (1997) presented a survey of major world sport rating systems. The study stipulated that sport rating systems have 3 key steps: 1) weigh the observed results – this is the most important factor in determining points for a competitor, i, in any given competition n, 2) combine the competitive points to produce season value, and 3) Aggregate the seasonal value to produce a rating.

The most well-known sport rating methodologies are the Bradley-Terry (1952), Elo (1978) and Glicko (1999) models (please see section Chapter Two for more details).

1.6.2.1 Type of sport ranking systems

Sorensen (2000) claimed that sport ranking systems, in general, fall into one of the two following categories: 1) Earned ranking systems utilise past performances to provide a suitable method for selecting either a winner or a set of teams that should participate in a play-off (Sorensen, 2000). Earned ratings are assigned an ordinal rank to produce team rankings. Majority of international sports such as tennis, basketball and football adopt an earned ranking system to produce [conference] seedings to establish play-off matchups. 2) Predictive ranking systems utilise past performances to build a forecasting model to predict future match outcomes between two teams. No internationally recognised sport adopts this ranking approach to determine seedings, as in practise this would not make sense and be problematic to implement. However, betting agencies, sport networks and analyst use such systems to set odds, predict margin of victory and establish winning probabilities.

Stefani (1997) stated sport rating systems can be separated into three further distinct types depending on how new ratings are calculated for each rating system: (1) Adjustive systems (2) Accumulative systems and (3) Subjective systems.

1.6.2.2 Adjustive Systems

Adjustive systems, also known as adaptive systems, “provide the best predictors for future performance because each adjustment follows from a predictor correction action in which a rating for team i, can increase, decrease or stay the same, as each new result is compared to each prediction based on information available prior to the competition” (Stefani, 2011, p.8). Such systems cause ratings to fluctuate, depending on performances, and account for leapfrogging. This is a situation in which a player who cannot participate due to injury, is exposed to being overtaken by teammates who can play more games, and therefore can earn more points. Adaptive rating systems are adopted by sports such as golf, cricket, chess, football, and rugby. According to Stefani (2011) an adaptive system for competitor, i, has the following form:

$$r_i^n = r_i^{n-1} + k\left[w_i^n - P\left(r_i^{n-1}, r_j^{n-1}, W, O^{n-1}\right)\right] \tag{1}$$

Here, r_i^{n-1} represents the rating for competitor i after competition (i.e. match or game) n derived by adjusting the previous rating, r_i^{n-1}, for competition i, by a multiple k. As mentioned previously weighing the observed result is the most important factor in determining points, as a large value would make ratings respond aggressively to the error term in the square brackets, while a small k would make ratings unresponsive. The adjustment $k,$ depends on, $w_i^n,$ which represents the difference between the actual performance of competitor i in competition $n,$ (i.e. w_i^n), and the predicted performance $P(\dots)$ which is based on competitor $i's$ previous ratings. Competitor $i's$ and opponent $j's$ previous rating is affected by $W$and O^{n-1}, defined as weightings and other factors (i.e. money won, quality of entrants, number of skills used etc.) present in competition $n-1$, respectively. The weighting procedure, $W,$ converts performances to points and varies across sports. For example, FIBA basketball provides weightings ranging from 0.1-5 for various championships (i.e. Olympic and Worlds) over an eight-year window. The ATP [men's professional] and WTP [women's professional] tennis publishes a matrix where each row represents *final placement* points for a given championship and columns represent the placement for each championship (Stefani, 2011).

1.6.2.3 Accumulative Systems

Accumulative systems are 'running sums' rating methods that are non-decreasing over a defined time-frame. These systems are predominately adopted by athletic sports such as gymnastics, power lifting and cycling. According to Stefani (2011) an accumulative system for competitor i has the following form:

$$r_i^n = \sum_{k=1}^{n} f_i\left[w_i^k, W, A, O^k\right] \tag{2}$$

Here, r_i^n represents competitor $i's$ rating after competition n, based on past performances. "The function, f_i, for competition i operates on w_i^k which is the performance of i in competition k, using W, which is a weighting procedure used to convert performance to points" (Stefani, 1997, p.7). The performance points are adjusted by an 'ageing' factor, A, and other factors, O^K, for competition k. The factors W, A O^K and are sport dependent on the sport.

1.6.2.4 Subjective Systems

Subjective systems consist of a panel of experts (i.e. judges) who rank the competitors and then combine the individual ratings to produce the overall ratings. Subjective systems are formally adopted by sports such as kickboxing, mixed martial arts and boxing.

Although statistical models are utilised to evaluate many problems in the sporting industry, the focus of this study will purely centre on team and individual rating systems. An extensive review of individual and team-based rating systems can be found in Chapter Two (section 2.1, p.43-65).

1.7 RATING SYSTEMS

Rating systems produce a single real number ([0,1]) representing a team or player's ability to perform. This section provides, system definitions, common methods, and common modelling practices within credit-risk and sporting industry. A more comprehensive review of the sporting and credit-risk literature, common methods, and key modelling practices, are provided in Chapter Two.

1.7.1 Sport Rating Systems

Formally, a sports rating system assigns each team a single numerical value to represent team or player strength relative to the rest of the league on some predetermined scale (Massey, 1997). Stefani (1997) stated that sport rating systems have three steps: 1) Weigh the observed results to provide competition points - this is the most important factor in determining points for competition *i* for a given competition, 2) Combine the competition points to produce seasonal values, and 3) Aggregate the seasonal value to produce a rating. Generally, sport rating systems fall into two categories: 1) Earned ranking – These systems utilise past performance to provide a suitable method for selecting either a winner or a set if teams that should participate in a play-off, and 2) Predictive ranking – These systems utilises past performance to provide the best prediction of the outcome of future games between two teams. Additionally, Stefani (2011) stated that sport rating systems can be separated into three distinctive types depending on how new ratings are calculated for each rating system: 1) Adjustive, 2) Accumulative and 3) Subjective. A potential drawback of sport rating systems are small sample sizes due to a limited number of contested sporting events. To derive a deeper understanding of the requirements for a meaningful sports rating system, this research builds on work from: Patel, Bracewell & Rooney (2017); Patel, Bracewell & Wells (2017); McIvor, Patel, Hilder & Bracewell (2018), and Campbell, Patel & Bracewell (2018).

1.7.2 Credit Risk Scorecards

Application and behavioural scorecards incorporate a *binary* or *count* target variable (approval or non-approval, or a credit rating, respectively). However, unlike the target variable associated with sport rating systems, evaluating the actual 'creditworthiness' of an approved line of credit can take months to observe the true outcome. New scorecard regulations require more robust, dynamic, and flexible models capable of accurately measuring an applicant's credit worthiness using a smaller time window of transactional data. However, a smaller time window means a smaller sample size of transactional data, potentially leading to poorer, less predictive credit ratings. There are six key steps involved when developing a scoring method: 1. Data Preparation > 2. Data Cleaning > 3. Variable Selection > 4. Sample Generation > 5. Model Development and Validation > 6. Model Approval. The first 3 steps are data processing. These steps are essential in developing a scoring method; however, the literature predominately focuses on data preparation, model development and validation steps. These three steps have the potential for improving the performance of scorecards. Various model algorithms can be used with different input variables to see which gives the best result. The choice of modelling objective is the primary key to developing scorecards since it defines a full set of technical estimation procedures that are used to select the best model under the objective and defines how to assess its validity. Data preparation and variable selection steps are especially important in credit scoring, and it has been found that applying new and more predictive variables can improve the performance of scoring models (Hand & Henley, 1997). The 'model development and validation' step is used to discriminate between 'good' and 'bad' applicants. The better the classifier, the better the performance of the scoring method.

1.8 RESEARCH OBJECTIVES

Surprisingly, given the growing application of analytics in the business environment and the increasing demand for rating systems to evaluate sporting performances, and assessing an applicant's credit worthiness, there currently exists no known modelling framework for constructing rating systems within these two domains.

Although performance will differ across different sporting codes, it is hypothesised that some elementary traits exist within the data, identifiable through '*action*', '*context*' and '*time*' based attributes. A key question is: what methods are appropriate for extracting these elementary traits? Given the key is to identify the traits, or the latent dimensionality of performance, it is suggested that the most suitable techniques will involve dimension reduction and feature selection. Moreover, given that performance is a function of significant traits it is suggested that ensemble forecasting strategies are most suitable when combining 'trait-based' ratings to produce overall performance ratings. Specifically, given the complexity, high uncertainty, and difficulty of modelling performance within sports, adopting an ensemble approach is appropriate as it produces results whose probability law of error will rapidly decrease (Armstrong, 2001). Provided that an approach to develop meaningful ratings can be established, an additional question is broached, impacting the effectiveness of the proposed rating system.

As this thesis focusses on sport-based rating systems, it is important that a rating specific model evaluation metric is developed to quantify the effectiveness of ratings produced by sport-based rating systems, and therefore comparisons can be made between ratings from different rating scenarios and forecasting difficulty. There exists a gap in the literature for an evaluation metric that assesses the effectiveness of meaningful sport-based ratings.

Based on an extensive literature review (Chapter Two), peer-reviewed conference proceedings and journal publications (Appendix A) on sport-based rating systems, the thesis formulates potent, yet achievable, research objectives, which form the basis of this research. The three research objectives are:

i. Develop a quantitative ratings framework to construct sport-based rating systems that output meaningful performance ratings.

Specifically, identifying the communalities of good sport-based rating systems and convert these findings into a defensible operational framework relating to performance. Meaningful sport-based ratings are reliable, robust, interpretable, and intuitive. These characteristics are defined as 1) Robust – the rating system must yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes. 2) Reliable – Produce accurate and highly informative predictions (i.e. well-calibrated and sharp ratings). 3) Transparent – Interpretable, easy to communicate and break down. 4) Intuitive – must relate to real-world observable outcomes (i.e. contextuality).

ii. Develop a novel evaluation metric to quantify the effectiveness of meaningful sport-based ratings.

Specifically, techniques such as Gini coefficients, Area under the curve (AUC), K-S, classification accuracy and root mean square errors are limited. Performance metrics of a rating system need to quantitatively align with the attributes of a 'good' rating system. There are many systems across sports, which dynamically assess performance and calculates a single numerical representation of performance. The issue with sport-based rating systems are that the rating measure may not tangibly link to the event outcome. For example, if a player rating system produced a rating of 67 (out of 100) during the game, how can the accuracy of such a rating be evaluated? Can this rating be mapped to actual in-game events and actions? And is it representative of an intuitive outcome? Therefore, an evaluation metric which evaluates the effectiveness of sport-based ratings is necessary.

iii. Demonstrate the applicability of the developed ratings framework and novel performance metric within the sporting context.

Given this research evaluates performance the practical implications are crucial. By demonstrating in a real-world context, the characteristics of meaningful or 'good' (reliable, robust, transparent and intuitive) rating systems, selected through a novel performance metric (ii), developed as part of this thesis, this will prove the value of this body of work.

1.9 SUMMARY OF METHODOLOGIES

The primary technical challenge faced throughout this research was to build a ratings framework that output ratings of performance that adhere to the commercial requirements of meaningful ratings. That is, ratings that are accurate and highly informative (i.e. reliable), interpretable, and easy to communicate (interpretability), and ratings that relate to real-world observable outcomes and the context to which the rating system is being applied (intuitive), and ratings must be largely unaffected by outliers, small departures from model assumptions and small sample sizes. Existing rating system methodologies lack the ability to produce meaningful ratings, therefore the research needed to develop a framework that produced reliable, robust, transparent, and intuitive ratings. Most rating systems account for two to three of these characteristics but not all four.

Further, through the research process it was realised that commonly used performance metrics were not completely suitable to evaluate the effectiveness of sport-based ratings, in that various performance metrics are required to evaluate their effectiveness but no single evaluation index can be applied across all systems and none is universally regarded as the ‘gold-standard’ metric to assess ratings performance. Therefore, the secondary challenge faced was the lack of an evaluation metric to quantify the effectiveness of meaningful sport-based ratings.

To successfully address these challenges and appropriately quantify and evaluate performance, the methods applied and the underlying philosophy that shaped the use of these methods form an important partnership. In this section the key methods used are summarised before outlining the associated philosophies that affected the creation of the various rating systems developed throughout this thesis. Several techniques were used to quantify and evaluate performance using objective multivariate data.

1.9.1 Dimension Reduction and Feature Selection

The core methodology adopted by the framework is the application of an ensemble forecasting strategy, dimension reduction and feature selection techniques. Therefore, the ratings problem resides in the field of information theory.

Ratings are an elegant and excessive form of dimension reduction (Bracewell, 2003), therefore dimension reduction techniques are a core functionality of the sport-based rating systems. The core techniques that were used for dimension reduction are outlined in Chapter Two and applied in Chapter Four and Five. After dimension reduction, feature selection techniques are applied to automatically select metrics which significantly affect performance or specific traits. After applying dimension reduction and feature selection, models are applied to derive trait-based ratings representing a quantitative interpretation of a specific trait.

1.9.2 Feature Engineering

Feature engineering is an important strategy when constructing rating systems. Such strategies extract relevant, contextual, and highly informative features which are inherently available within the data, and therefore expert knowledge is required to create these latent features. These features provide an approach to account for a large amount of uncertainty within the ratings.

1.9.3 Ensemble Forecasting

The proposed ratings framework adopted a “multi-objective” ensemble forecasting strategy as it allows the evaluation of multiple traits (i.e. dimensions), at different layers. The different layers for each trait are regarded as different sources of information. Given there are multiple traits that significantly affect performance, the constructed framework incorporates different modelling objectives capturing information from the different traits which significantly affect performance.

Liu & Pentland (1999) developed an approach to model human behaviour which consider the human as a device with a large number of internal mental states, or traits, each with its own particular control behaviour and interstate transition probabilities. This approach is like the approach outlined in Chapter Three, whereby each trait is modelled individually, and the trait-based rating produced by each model is ensembled to produce an overall rating representing a numerical interpretation of performance.

The ratings framework is a multi-objective ensemble forecasting strategy. The methodology ensures that each trait rating utilises *action*, *context*, and *time*-based attributes to effectively account for the uncertainty within each trait. These trait-based ratings are combined using an ensemble strategy to output a rating which provides a numeric representation of performance.

1.9.4 Proper Scoring Rules

Commonly used performance metrics were applied to measure the effectiveness of meaningful sport rating systems and during the literature review process limitations associated with evaluation metrics were identified. A set of criteria were identified to construct a performance metric to assess the effectiveness of meaningful sport ratings. A proper scoring rule methodology was applied to construct such an evaluation metric. Specifically, the performance metric applies the distance and magnitude-based measures associated with the spherical scoring rule with an embedded Analytical Hierarchy Process which allows the user to incorporate expert-based knowledge.

1.10 KEY PHILOSOPHIES

To fully explore the theses research objectives, several different approaches were required to ensure suitable information was extracted from the data using the methods detailed previously. These are briefly described below.

1.10.1 Expert System Development

The ultimate goal of this thesis is to produce a statistical framework that allows a modeller to develop rating systems that mimic the opinion of an unbiased expert human observer, and produce meaningful ratings (reliable, robust, transparent and intuitive). Therefore, the philosophies discussed in this section help in the attainment of this goal.

1.10.2 Performance ~ $\mathcal{F}(\mathbf{trait_i})$

To produce ratings which accurately provide a numerical representation of performance the traits that influence performance need to be identified and ‘correctly’ scored. The underlying philosophy adopted to calculate this score (i.e. ratings) has been extracted from the sociology and psychology literature (Heinström, 2003; Kampe, Edman, Bader, Tagdae, Karlson, 1997; Scharli, Ducasse, Nierstrasz and Black, 2003; Silvia, 2008). That is, sporting performance is a manifestation of the significant traits that affect performances.

1.10.3 Specialisation vs. Generalisation

In the identification of significant traits affecting team and player performance, the number of traits applied needed to consider the trade-off between sport-specific (too many traits) and generalisation (too few traits).

1.10.4 Influence vs. Formulation

In creating a suitable model, it is the overall influence that is of interest rather than the actual model formula. Transparency and intuition are crucial for promoting the rating systems and for this to occur, the general influence of the features involved is more important than the specific coefficients required to calculate the ratings.

1.10.5 Optimal Solution

An optimal solution is not required. A good solution in reliable, robust, intuitive, and transparent terms will suffice (i.e. meaningful ratings). The accuracy and predictivity of the ratings are not of vital significance because ratings generally will be expressed as whole numbers.

1.10.6 Action, Context and Time

To ensure that each trait-based rating sufficiently accounts for the uncertainty surrounding performance each trait ratings must be derived using a combination of action, context, and time-based attributes. This ensures that the ensembled trait-based ratings are meaningful.

1.10.7 Conventional Features with Creative Complex Features

To obtain suitable models for a sport-based ratings framework which create rating systems that output meaningful results i.e. transparent, robust, intuitive, and reliable. To achieve such output characteristics conventional statistics needed to be supplemented and combined with creative and complex features.

1.11 SUMMARY OF RESULTS

This thesis shows that a dynamic multi-objective ensembling forecasting strategy is an advantageous methodology to implement when developing rating systems which produce meaningful ratings within a sporting context.

As an exploratory exercise a set of rating systems were developed to identify the key communalities amongst rating methodologies. These communalities included the application of

1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling mechanism to produce an ensembled rating of individual traits.

Using these findings, a ratings framework to construct sport-based rating systems was developed. The framework assumes that performance is a function of the traits that significantly affect it, and that these traits are a function of the feature-types (action, context, and time) that significantly affect the trait of interest. Dimension reduction identified the key traits (dimension) within the data, while feature selection identified the significant feature-types affecting each trait. Although the ratings framework is only applied within the sporting context, specifically within cricket, it can be applied across multiple sporting codes to evaluate both team and player performances. Such applications are considered outside the scope of this research.

The dynamic ratings approach is a form of model stacking where information from multiple trait-based models is combined to generate a more informative model. To address the issue of intuitive-results and transparency present within many rating systems, due to the application of "black-box" techniques, a manual approach is applied to ensure full autonomy and understanding of model inputs and input effects.

Adopting a multi-objective ensembling strategy, where each modelling objective represents a specific trait affecting performance, the applicability of the framework is tested against different sporting scenarios. These individual studies reveal that an ensembling forecasting strategy produces reliable, intuitive, transparent, and robust results and is an ideal strategy to implement when developing sport-based rating systems.

During the model evaluation process major problems were encountering with identifying a suitable performance metric to assess the effectiveness of meaningful sport rating systems. A novel model evaluation approach has been developed to address this problem.

The distance and magnitude-based spherical (DMS) performance metric was developed to assess the effectiveness of meaningful sport-based ratings. This approach is an evaluation index which accounts for forecasting difficulty, forecasting scenario and leverages expertise knowledge when determining the effectiveness of sport-based rating systems. The DMS performance metric applies distance and magnitude-based measures derived from the spherical scoring rule. A proper scoring rule methodology is applied because ensemble-based forecasts are generally assessed on two criteria: calibration and sharpness, and a metric which promotes such results was necessary. This resolves the issue of identifying reliable ratings, which is necessary to ensure rating systems are meaningful. Additionally, this method shows a great deal of promise as an evaluation tool for problems outside the sporting domain.

1.12 THESIS LAYOUT

This thesis is constructed in two parts. Part One reviews and applies rating systems in a range of novel contexts relating to performance. The communalities and limitations of these rating systems, within the context of rating sporting performance and credit-risk applicants', are then used to create a ratings framework with wide applicability and robustness. Validation of the proposed ratings framework occurs with the application to cricket at both the team and player-level. Importantly, use and creation of multivariate techniques to extract intuitive, robust, reliable, and transparent performance and trait-based features for individuals is the dominant theme throughout this thesis.

The intent of this thesis structure is to communicate the underlying limitations and assumptions of commonly used rating techniques. This provides the necessary background to understand and improve upon the applicability of objectively rating performance within the sporting context.

To aid the flow of the thesis and concentrate specifically on the development of the proposed ratings framework, supporting material, including peer reviewed conference proceedings and journal publications written as a direct consequence of this research are supplied in the appendix.

Rating systems (or scoring models) within credit-risk and the sports domain are used widely as is demonstrated in the literature review that forms the basis of Chapter Two. Chapter Two examines the application of various statistical methodologies to develop rating systems across credit risk and sports. Using the learnings (communalities, distinctions, and limitations) from the literature, this thesis develops rating systems adopting the key methodologies identified in the literature and applies these in novel settings. This is extended to derive research objectives based on literature gaps and the limitations identified when extending the existing ratings systems to wider use cases. Appendix A includes the substantial body of novel work generated during this research process that has been peer-reviewed and published. This includes six journal articles and 13 peer-reviewed conference proceedings.

To develop a novel evaluation metrics to quantify the effectiveness of meaningful sport-based rating systems (research objectives (ii)), a review of commonly applied performance metrics must be conducted. Therefore, Chapter Two reviews commonly applied model evaluation metrics, the field in which each metric is most applied and their limitations. Chapter Two identifies the five key criteria that a performance metric must adopt to assess the effectiveness sport-based ratings: 1) sensitivity to distance, 2) sensitivity to time-dependence, 3) evaluate the ratings based on the entire distribution, 4) provide an incentive for well-calibrated and sharp ratings and 5) adjust incentives based on forecasting difficulty.

Chapter Three focuses on developing a ratings framework for constructing sport-based rating systems. In this chapter Bracewell's (2003) definition of a rating is extended. Specifically, meaningful ratings are an elegant and excessive form of dimension reduction whereby a numerical value provides a meaningful quantitative interpretation of performance. The handling of the limitations and issues specific to rating systems are explored within the statistical framework. The core methodology adopted by the framework is the application of an ensemble forecasting strategy, dimension reduction and feature selection techniques. Therefore, the ratings problem resides in the field of information theory.

Given performance is defined as a function of individual traits that significantly affect performance, the framework develops a multi-objective approach, where each objective is dedicated to quantifying each individual trait, and ensembling these trait-based ratings produces an overall rating, representing a numerical interpretation of performance. The action, context and time-based attributes that significantly affect each trait are identified through dimension reduction and feature selection. In Chapter Three, it is stipulated that to derive meaningful trait-based ratings, defined as intuitive, reliable, robust, and transparent, 'action', 'context' and 'time' based attributes are necessary. Further, ensembling these trait ratings produce meaningful performance ratings.

Using the scoring rule methodology (Chapter Two), Chapter Three also develops a novel performance metric, known as the distance and magnitude-based spherical metric, to evaluate the effectiveness of meaningful sport-based ratings. The developed performance metric applies distance and magnitude-based statistics derived from the spherical scoring rule and adopts an Analytical Hierarchy Process which incorporates expertise knowledge. The distance and magnitude-based spherical (DMS) metric is applied to the rating systems developed in Chapter Four and Chapter Five and is shown to be a more appropriate measure of evaluating ratings than traditional evaluation metrics such as the log-Loss. Further, the metric is shown to perform better in certain forecasting scenarios and forecasting difficulty, than traditional metrics. Chapter Three addresses research objectives (i) and (ii).

Part Two applies and validates the ratings framework and DMS performance metric developed in Chapter Three. Specifically, Chapter Four and Five validates these findings by developing two unique ratings systems within the sporting context. Incorporating the lessons learnt from Part One, applicable data mining methods are explored from sporting and statistical perspectives. Key issues influencing the methodology for quantifying performance, reliability, robustness, transparency, and intuition, shape the techniques explored. Moreover, the effectiveness of the ratings is quantified and evaluated through commonly used performance metrics such as the log-loss. Chapter Four and Five test the validity of the novel performance metric by benchmarking it against evaluation metrics such as the log-loss and applying it to the ratings produced by sport-based rating systems built using the ratings framework. Chapter Four and Five address research objective (iii).

Chapter Six concludes the thesis with a detailed deliberation over the relevance and appropriateness of the methods and data involved, the work remaining, and answering the final question, namely, *“What is the relevance of this work for rating systems, in particular for rating systems in others domains other than sport performance evaluation?”*

1.13 DISCUSSION AND CONCLUSION

Bracewell (2003) stated that ratings are an elegant and excessive form of dimension reduction and that “good” ratings are reliable, contextual, and transparent. This chapter outlines the theses aim to extend Bracewell’s (2003), that is, meaningful ratings are an elegant and excessive form of dimension reduction whereby a value provides a meaningful quantitative interpretation of performance. Specifically, meaningful ratings should have the following characteristics: 1) robust – ratings must yield good performance where data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes. 2) Reliable – ratings must be accurate and highly informative predictions that are well-calibrated and sharp. 3) Transparent – ratings must be interpretable and easy to communicate. 4) Intuitive – ratings should relate to real-world observable outcomes and the context to which the system is being applied.

This chapter introduced performance-based sport evaluation systems, referred to as rating systems. Due to the growing application of ‘big-data’ and machine-learning within the commercial environment, the chapter outlines, the growing commercial demand for data-driven rating systems to evaluate performance and identifies limitations of current industry standards and methodologies, specifically within sports and credit-risk.

Before developing the ratings framework for constructing rating systems (Chapter Three), this thesis develops sport-based rating systems using current methodologies (please see Appendix A). This development and application process identified limitations of current industry standards, limitations of rating methodologies and formulates potent and achievable research objectives. Three objectives relating to sport-based rating systems, have been identified: (i) develop a ratings framework to construct meaningful sport-based rating systems that output meaningful sport-based rating systems (Chapter 3), (ii) develop a model evaluation metric to quantify the effectiveness of sport-based rating systems (Chapter 3) and (iii) demonstrate the applicability of the developed framework and novel performance metric within the sporting context (Chapter 4 and Chapter 5).

The following chapter provides an extensive literature review of credit-risk and sport-based rating systems, rating methodologies and model evaluation metrics, identifies major gaps in the literature pertaining to these rating systems, and develops a set of rating systems using commonly applied statistical methodologies.

PART ONE: REVIEWING THE RATINGS LITERATURE AND DEVELOPING RATING SYSTEMS

Chapter Two

A NOVEL LITERATURE REVIEW AND APPLICATION OF RATING SYSTEMS

"The math works. Over the course of a season, there is some predictability to baseball. When you play 162 games, you eliminate a lot of random outcomes. There's so much data that you can predict - individual players' performances and the odds that certain strategies will pay off".

Billy Beane, Moneyball (2008).

On the application of sports analytics in Baseball.

2.0 INTRODUCTION

To achieve the research objectives outlined in chapter one, there must be a thorough understanding of the ratings literature, model evaluation metrics, dimension reduction and feature selection techniques. Given the commercial requirements of this research, rating systems within the sporting and credit-risk environment are reviewed.

This chapter provides a comprehensive review of academic literature outlining the application of statistical techniques and current ratings methodologies to construct rating systems within the sporting and credit risk environment. This chapter also reviews commonly applied performance metrics within industry and academia, outlines the technical details and limitations of each metric, and explains why certain performance metrics work well with certain problems and lose information in other circumstances.

The primary objectives are to consolidate these findings to 1) identify the key elements required to construct a rating system and 2) identify the ideal set of criteria and the ideal methodology to construct a novel performance metric to evaluate the effectiveness of meaningful sport-based ratings.

Using the findings from the literature review and common ratings methodologies, a set of peer-reviewed conference proceedings and journal publications are written across various sports. The published papers are not provided in this chapter; however, the key findings are summarised in this chapter and chapter three. Further, a full list of these publications can be found in Appendix A.

Section one and two provide an overview of sports rating systems and credit risk models, respectively, an discusses the gaps in the ratings literature. Section Three reviews commonly used model evaluation metrics, also known as model performance metrics. The limitations of each performance metric and the areas in which their application is most prevalent is also discussed. Section Four describes the research objectives and the gaps in the literature. Section Five concludes with some closing remarks and discussions the outcomes of the review.

2.1 OVERVIEW OF SPORTS RATING SYSTEMS

Formally, sports analytics is defined as "the management of structural historical data, the application of predictive analytical models that utilise such data, and the information systems used to inform decision makers and enable them to help their organisations in gaining a competitive advantage on the field of play" (Alamar & Mehrotra, 2011, p. 1).

The distinction between quantitative data collection, within sports, and sport analytics exists within its application. Quantitative [sports] data collection is the measurement and storage of the performances or actions of a team or a player, while analytics is the use of data to inform decision makers.

The results generated from applying statistical techniques to sports related data are called sports statistics, which differs from sports analytics in the sense that sport statistics are the outcomes generated from the analytical techniques applied to the data. Bracewell (2003) claimed that sports statistics fall into one of two categories: (1) statistics that can be directly observed from a scoresheet, known as performance indicators, and (2) statistics that are not directly observed from a scoresheet, known as performance outputs. Sport statistics are utilised to make player selection decisions, develop training regimes, and determine optimal strategies.

There is a breadth of academic literature applying various statistical techniques to myriad sports. This chapter will review notable academic literature describing and developing sport rating systems, at both the team and individual level. Moreover, these analytical techniques allow users to rank and evaluate player and team performances. A rank refers to ordinal placement of ratings, while "ratings come from a continuous scale such that the relative strength of team individual is directly reflected in the value of its rating" (Massey, 1999, p. 2).

In general, sport rating systems provide an objective evaluation of a team or individual based on prior performances and are implemented for player comparisons and improving player and team selection process.

Formally, a sport rating system assigns each team or individual a single numerical value representing a team's or individuals strengthen relative to the rest of the league on some predetermined scale (Massey, 1999). These ratings are beneficial to numerous parties, especially athletes, coaches and managers who utilise such systems to track and predict form, progress and applied as a motivational and benchmarking tool. According to Leitner *et al*. (2010) sport ratings are typically derived by suitably aggregating a competitor's previous performances and provide predictive power in forecasting future performances. Many American sporting franchises, such as The Oakland A's (Baseball) and Dallas Mavericks (Basketball), adopt such a mentality and focus.

Using a common framework, Stefani (1997) presented a survey of major world sport rating systems. The study stipulated that sport rating systems have 3 key steps: 1) weigh the observed results – this is the most important factor in determining points for a competitor, i, in any given competition n, 2) combine the competitive points to produce season value; and 3) aggregate the seasonal value to produce a rating.

2.1.1 Sport Rating Systems

This section provides a brief review of academic literature outlining the application of statistical techniques to derive individual and team-level ratings for various sports. This section is divided into two sub-sections: 1) team-level sport rating systems and 2) individual-level sport ratings systems.

2.1.1.1 Team-based sport rating systems

Team-based sport rating systems evaluate team strength, performance and or ability using match level statistics or apply an aggregation function to individual level performances to produce team-based ratings.

West & Lamsal (2008) developed a predictive team-based ratings model using a simple linear regression technique to college football data. Regressing six predictors (scoring margin, offensive yards per game, defensive yards per game, strength of schedule, defensive touchdowns per game and turnover margins) on match outcome, West & Lamsal (2008) built a predictive model using previous season data and 'bowl' game outcomes to establish team ratings. The amount of rating points a team received was based on the 95% confidence interval (*c.i.*) for the expected outcome for a single game, and the team ratings were produced by aggregating these points across all games. Applying this model to college bowl competition it was found that model results agreed with actual outcomes in 59.4% (19/32) of games, and of the 13 incorrect predictions three of the confidence intervals included actual game outcomes.

Mease (2003) developed a penalized maximum likelihood approach for the ranking of college football teams independent of margin of victory. This ranking process attempted to reflect the opinion of human pollsters Applying the model to 1998 American College Football data and comparing the proposed model outcomes to computer-based outcomes, it was found that the penalized maximum likelihood approach outperformed two of the three [computer-based] models adopted by American College football. Moreover, the model produced rankings for college football teams which were highly correlated with expert rankings relative to BCS (bowl college series) models.

Dyte & Clarke (2000) developed a team-based rating method for predicting the distribution of scores in international soccer matches. Dyte & Clarke (2000) treated the number of goals scored by a team as an independent Poisson variable, dependent on FIFA team ratings and match venue. The Poisson regression model had two underlying assumptions: 1) the number of goals scored by a team in a soccer match is Poisson distributed and 2) it is independent of the number of goals scored by the opposing team. The model predictors were current team's FIFA ratings (TR), opponent's FIFA rating (OR) and a parameter (v) which changed according to venue. The expected number of goals scored per team, $m,$ was used to produce the marginal probabilities for each teams Poisson distribution of goals scored. Using the latest FIFA ratings to calculate the expected number of goals estimated through the regression analysis, it was possible to generate two Poisson random variables for every game and run a simulation for an entire tournament. Post simulation, the expected number of wins draws and losses for each team were calculated by aggregating the probabilities for each of the world cup matches. A Chi-squared test showed that there was no statistical difference between the expected and observed numbers, indicating that the form of Poisson model used for simulation was plausible.

Bracewell, Forbes, Jowett & Kitson (2009) developed a Rugby-based team rating system, providing outputs known as 'team Lodeings'. These outputs "measure the relative performance of sport teams and the competitive balance of competition" (Bracewell, Forbes, Jowett & Kitson, 2009, p.2). The ratings system enabled (Bracewell *et al.*, 2009) to measure a team relative performance to opponents within the same division, allowing meaningful comparisons and effectively evaluating competition competitiveness. Applying the framework to the 2004 New Zealand National Provincial Rugby Championship revealed that the ratings engine produced suitable comparisons of team performance across divisions. The results revealed that the standard deviation of the ratings provided good representation of the competitiveness of a given sports league. Moreover, it was found that a competitive league results in teams having similar winning percentages, and therefore a smaller standard deviation. Applying the ratings across 7 different sports it was found that soccer was the most competitive sport, followed by Basketball and American football, while Rugby was found to be the least competitive.

Bracewell, Downs & Sewell (2014) realized that the way limited overs cricket results are recorded complicates the ability to generate meaningful team ratings. Therefore, Bracewell *et al.* (2014) developed a method for creating performance-based team ratings for cricket utilizing a margin of victory that was solely runs based. This was achieved by developing a method for calculating the margin of victory for when the team batting second wins. The method estimated the number of runs that would have been scored had the team batting second continued until resources (i.e. balls and wickets) were exhausted. Using the Duckworth & Lewis (1998) framework, a score projection was carried out if both resources had been exhausted using $T_2 = \frac{C_2}{R_2}$, where C_2 is team two's actual score and R_2 (the DL resource remaining). An F-test found that the score projections did not produce margins of victory that were significantly different from those produced when the team batting first wins. Logarithmically transformed score ratios, were used in creating team ratings which were regressed against the winning percentages to deduce a linear transformation that would increase the spread of the ratings between 0 and 1. These score ratios were then input into the Team Lodeings algorithm developed by Bracewell *et al.* (2009) to quantify relative performance. A correlation of 0.91 between the ratings and the International Cricket Council ratings indicated that the team ratings generated by the proposed performance-based rating system was valid. Moreover, F-test results confirmed that the variance of the transformed margin of victory when the team batting first wins is not statistically different from the variance of the transformed margin of victory when the team batting second wins, indicating that the extrapolation did not introduce bias.

Similarly, to evaluate cricket team performance Clarke (1988) applied a dynamic programming model to one-day cricket to: 1) calculate the optimal scoring rate, 2) estimate the total number of runs to be scored in the first innings and 3) estimate the probability of winning in the second innings. The first innings formulation allowed the development an *'optimal scoring model'* outlining a team's optimal scoring rate (i.e. runs per over) to obtain a given expected total, for any given number of wickets lost and balls remaining. The second innings formulation enabled the development of a *'probability scoring table'* outlining the probability of the second innings batting team scoring the target total, for any given number of wickets lost and balls remaining. Results suggested that the scoring rates should be more uniform, and that the team batting second has an advantage.

Of all the team-based sports rating system, none is more famous and reputable than the Elo rating system. Although originally developed for rating chess players, the Elo rating system has been extended to many team has been adapted to a wide variety of sports, at both the individual (such as tennis [United, 2018; Raboin, 2013; Abstract, 2018] and golf [Broadie & Rendleman, 2013; Broadie, 2012; Levin, 2017]) and team-level (such as football [Curiel, 2018; Hvattum & Arntzen, 2009; Leitner & Hornik, 2009; Lasek, Szlávik & Bhulai, 2013; Goddard & Asimakopoulos, 2004]), American football (Silver, 2014), Basketball (Silver & Fischer-Baum, 2015), and among others. See (Aldous, 2017; Király & Qian, 2017; Stefani, 2009) for recent mathematical reviews.

An example of such an extension is Moore, Rooney, Bracewell & Stefani (2018). Moore, Rooney, Bracewell & Stefani (2018) measured team ratings for the 2017 super rugby season using the Elo model, and showed how to systematically determine all Elo model parameters, using an optimisation technique to achieve maximum power. This modified ratings model was applied to the 2017 super season rugby. The initial ratings are computed by fitting static ratings to the 2016 super rugby season. A logistic regression model was trained to predict win/ loss outcome from ratings differences with a latent parameter $B \approx 59$ and home ground advantage parameter $h \approx 49$ ensuring the win/ loss model was responsive to ratings differences. The model produced an accuracy of 77%. The learning parameter K was optimised for predictive performance using a three-fold cross validation approach with Lasso regularisation prevent over-fitting, and the inverse regularisation strength C was varied. The parameters K and C were simultaneously optimised for predictive power. A clear minimum loss was found at $k = 90$ and $C = 6 \times 10^{-3}$.

Glickman (1995) introduced an evolution to the Elo system, known as the Glicko model. The Glicko model is derived as an approximation to a Bayesian dynamic paired comparison model, where each player is given an initial prior rating described by a univariate normal distribution. Glicko breaks time into periods, during which skills are assumed to be constant, and overtime, these skills change according to a Markovian random walk.

The likelihood employed is the same as in Elo and under certain assumptions, Glicko recovers Elo as a special case (Glickman, 1995). Glicko extends the Elo system by computing two components: 1) rating, r_i, representing team $i's$ or player $i's$ strength and 2) rating deviation, RD_i, representing a standard deviation, which measures the uncertainty in a rating. The amount a players or teams rating changes depends on RD. This change in rating is small when player or team $i's$ RD is low, and the change in rating is high when their opponents' RD is high. The RD experiences a decrease after playing a game, but slowly increases over-time of inactivity, and therefore, a low RD indicates that a player competes frequently.

Since inception, the Glicko model has been applied to a variety of sports such as Chess (Vecek, Mernik & Crepinsek, 2014; Vecek, Mernik, Filipic & Crepinsek, 2016), basketball (DeLong, Terveen & Srivastava, 2013; Vaziri, Dabadghao, Yih & Morin, 2018), football (Kharrat, Pena & McHale, 2017; Lasek, Szlávik & Bhulai, 2013; Babic, 2017), tennis (Ingram, 2019) and volleyball (Glickman, Hennessy & Bent, 2017).

Herbrich, Minka & Graepel (2007) also introduced an evolution of the Elo system, known as TrueSkill. The TrueSkill rating system was developed by Microsoft for their Xbox Live gaming platform, which measures the skill level of players in multiplayer games. A player's true skill rating determines the team in which they will play for and the opponent they will play against. TrueSkill is a Bayesian rating system which can be viewed as a generalised system of the Elo system assessing a probabilistic generative model of match results. An intuitive random process is constructed to generate a player's skill and match results, and therefore, it is unnecessary to experiment with different formulae to update a player's skill rating and produce the 'right' player rating. The model is benchmarked against data and refined depending on discrepancies identified within the data. Once a 'good' model is found Bayesian inference is applied to identify the optimal algorithm for updating skill ratings (Herbrich, Minka & Graepel, 2007).

Bradley & Terry (1952) introduced a paired comparison probabilistic model predicting the outcome of paired comparisons. The model assumes that in a match-up between two players or two teams, i and j (Bradley & Terry, 1952), the odds that i beats j is $\frac{\alpha_i}{\alpha_j}$, where α_i and α_j are positive parameters which represent team or player 'ability' or strength. Assuming player and team 'abilities' are measured on a ratio scale, the Bradley-Terry approach can be applied to derive the probability of competitor i beating competitor j (Bradley & Terry, 1952). The Bradley-Terry model has been applied to a wide variety of sports, both individual, such as tennis (please see McHale & Morton (2011) and team, such as football (please see Leitner, Zeileis &Hornik (2010); basketball Koehler & Ridpath (1982); Katoh, Koyanagi, Ohnishi & Ibaraki (1992)).

2.1.1.2 Individual-based sport rating systems

Individual-based sport rating systems evaluate individual strength, performance and or ability using both individual and match level statistics produce individual player ratings.

Clarke (2011) applied multiple linear regression to rank tennis players using results from an Australian domestic doubles competition. Using indicator variables to tag individual players, Clarke (2011) fitted a regression model to 'games-up per set played' as a linear function of the two players involved and established model significance with an $R^2 =$ 0.074. Next, percentage of games won by opposition and set weaknesses were added to the regression model, producing a practically and statistically significant model with an $R^2 = 0.26$. Further, using separate player ratings, a larger regression model was developed, incorporating a constant for home advantage. The home advantage coefficient of 0.51 was significant with a *p-value* = 0.026. The two sets of ratings had an almost perfect linear relationship suggesting that the method of calculating ratings provide reasonable estimates of a players' relative ability. An exponential smoothing method was implemented to estimate a player's end-of-season rating. A correlation of 0.85 between the exponential smoothed ratings and regression ratings indicated that the smoothing method produced reasonable results. Moreover, this result indicated that the smoothing method could give reasonable ratings. The smoothing constant was optimised such that the best fit to the predicted set of results was produced. Each refinement in the method showed an increase in the correlation of the end of season exponentially smoothed ratings and the least squares regression ratings (Clarke, 2011), reinforcing the use of exponentially smoothed ratings to rank tennis players.

Similarly, Ingram (2019) developed a tennis ranking system using a dynamic paired comparison model with a Gaussian Process as a prior for the time dynamics rather than a Markovian process. The modified Gaussian Process was applied to ATP (Association of Tennis Professionals) tennis matches to evaluate player performance. Ingram (2019) stated that even though random walk is convenient to compute, it does not allow for mean reversion. Using the Gaussian process allows to evaluate other player skill evolutions. Using the kernel functions, K, a prior can be selected to evaluate how smooth the function should be, how quickly it varies and how much it varies. A player's skill was modelled as a combination of the radial basis function kernels with different length scales: 1) short-term variation (80 days), 2) medium-term variation (400 days) and 3) long-term variation (800 days). A Hamiltonian Monte Carlo sampler Stan was applied to fit the kernel hyperparameters. Once the hyperparameters for the kernel was fitted, a maximum a posteriori (mAP) point estimate was used. This model was fit to ATP matches from 2012 onwards and compared against a Glicko and Elo fit from the start of the open ERA (1969) using log-loss and accuracy measures. The Gaussian process model was found to have the lowest log-loss, despite using few years of data and only using the mAP estimates. However, the Gaussian process had a lower accuracy than the Elo and Glicko models. Ingram (2019) recommended using other kernels which adopt longer time scales or less smooth than the RBF kernel.

Bracewell, Farhadieh, Jowett, Forbes & Meyer (2009) applied time series clustering to map the test career progression of Australian cricketing legend Sir Don Bradman, acknowledged as the greatest batsman of all time with an unparalleled career batting average of 99.94, from 80 innings. However, part of his career was interrupted as international cricket was suspended during World War II. Given this 'disruption' in his test career Farinaz (2009) utilised time series clustering to characterise Bradman's test career and compared him to other 'great' batsmen to test if Bradman was denied his prime. The selected clustering method was based on global characteristics measures "as it does not require many conditions to be true before it can be utilised, relative to other clustering techniques" (Farinaz, 2009, p.3). Additionally, the approach clusters global features extracted from individual time series and can be applied on different length time series. The performance measure used to compare batsman was [scaled] average 'contribution' per innings. To estimate a batsmen's performance over their career, weighted least square regression is used to model scaled average contribution per calendar year for all test batsmen, who had careers spanning 17 years, participated in 70 innings and had averages > 40 runs. The average contribution was scaled by the range producing a minimum of 0 and maximum of 1. Results showed that Bradman's career progression was most like West Indian legend Brian Lara, indicating that Bradman's peak performance would have occurred in the 12th and 14th years of his career (1939-1941), coinciding with World War II. Imputing Bradman's likely performance (i.e. batting average) from 1939-1945 It was estimated his batting average to be 105.41, which was significantly higher at the 5% significance level than Bradman's actual average (i.e. 99.94).

Akhtar, Scarf & Rasool (2014) also derived cricket player and team ratings by fitting multinomial logistic regression models to session by session test match data to calculate match outcome probabilities given the match position at the end of each session *t*. The probabilities were used to measure the overall contribution each player had on match outcome, based on their individual contribution during each session. Additionally, a hypothetical position at the end-of session *t* was defined, in which bowlers had not taken any wickets, and match outcome probabilities were generated. A player's overall contribution during a given session was assessed by using the difference between the hypothetical match outcome probabilities and the actual match probabilities. The batting probability differences were observed with respect to 'not losing' and bowlers with respect to winning (Akhtar *et al*., 2014). The difference in probabilities were distributed to batters according to their share of the runs scores in session *t* and to bowlers according to their share of wickets taken the session, *t* (Akhtar *et al*. 2014). An individual i*'s* batting contribution in session *t* was evaluated via:

$$C_{i,t,bat} = C_{i,t,bat} \times \frac{r_{i,t}}{r_t} \tag{3}$$

Here, $r_{i,t}$ is the runs scored by player *i* in session *t* and r_t is the total runs scored by their team in session *t.* An individual, *i*, bowling contribution in session *t* was evaluated via:

$$C_{i,t,bowl} = C_{i,t,bowl} \times \frac{\sum_{j=1}^{n} Z_{itj}\alpha_j}{Z_t} \tag{4}$$

Here, Z_{itj} represents the total number of wicket taken by player i during session t for wicket taking contribution j, j = $\{1,2,3\}$, where $j = 1$ corresponds to a wicket taken by the bowler with no fielder involvement, $j = 2$ corresponds to catches taken by a fielder and $j = 3$ corresponds to run-outs. The α_j represents the share of points for a wicket awarded to the fielder. The net contribution of player i in the match is the aggregated of contributions from all sessions. However, it was found that the rating system took little account of contribution after a point when the win or draw probability of any team is close to unity. To overcome this problem Akhtar *et al*. (2014) adopted the contributions as one component of a weighted average rating system, while the other was raw runs and wickets in the match. Points gained were placed on a 'runs-like' scale by multiplying the net player contribution by the average runs per test match played 1877-2007. Team ratings for each nation were calculated by combining the individual player ratings, the final aggregated value represented the national teams overall rating.

Duckworth & Lewis (2005) developed real time player metrics, using the Duckworth-Lewis methodology, to evaluate player contribution at any given stage of an innings, producing context-based measures. The developed metrics were: 1) batsmen average run contribution per unit of resources consumed and 2) bowlers' average runs contribution per unit resources consumed. Applying these measures to the 2003 VB series final (Australia vs. England) it was shown that the Duckworth-Lewis based contribution measures were less susceptible to distortions compared to traditional performance metrics.

Individual-based sport rating systems also extends beyond players, for example, Scully (1994) applied survival analysis techniques to investigate manager retention rates in baseball, basketball, and football. Kaplan-Meier survival curves were fitted to evaluate managerial survival and compare differences in survival probabilities, across the three sports. The results from the test indicated that the survival curves for each sport were statistically different from each other. Further, Scully (1994) investigated which distribution was most appropriate to describe the survival probabilities. The Weibull distribution was initially implemented with the parameters (α, β) being estimated using maximum likelihood estimation. In each regression model, managerial efficiency was used as a covariate and regressed against managerial tenure. Managerial efficiency was calculated as a comparison of the manager's winning percentage with the manager's maximum win percentage. Results suggested that the Weibull distribution provided an accurate description of managerial survival rates. The results showed a highly significant positive relationship between managerial efficiency and managerial tenure across all three sports, suggesting that the higher the proportion of games won by a manager, the longer the manager will stay with the team.

Similarly, Ohkusa (2001) also investigated factors that affect the quit behaviour of professional baseball players in Japan. The author considered both pitchers and batters who played between 1977 and 1990 and applied Cox proportional hazard methodology. The dependent variable was defined as the time until the player quit. Duration was defined as the number of years since the player entered the baseball league. Ohkusa (2001) used wages, productivity, and their quadratic terms as explanatory variables. Batter productivity was measured as the *slugging rate*, pitcher productivity was defined using *hit rate* and *strike to walk rate*. The results found that higher income discouraged quitting among both batters and pitchers. Among batters, higher productivity was associated with a reduction in probability of quitting, while, among pitchers, higher productivity was associated with an increase in probability of quitting. This suggested that there may be other factors at play such as the impact on the body. For example, for batters, high productivity may put more strain on the body. As such, these results would suggest that higher body impact leads to greater retention.

McHale, Scarf & Folker (2012) outlined the Premier League player performance index for rating the performance of football players. The ratings index is a weighted ensemble of six sub-indices and constructed using different regression models. The six sub-indices model: match outcome, point-sharing, appearance, goal-scoring, assists and clean sheets. Match outcome was modelled as a Poisson function of goals for and goals against which was determined by the number of shots and shot effectiveness of the two teams, which are determined by the player actions of each team. The point sharing index is a linear function of the number of minutes played by a player, the total number of minutes played by all players on their team, and the number of points the team won in a given match. The appearance index divides the number of points won by all teams in the league among the players according to how many minutes they played. The assist index rewards points for each assist. Finally, the clean sheet index was modelled as a function of blocks, clearances, tackles won, interceptions and saves. The final index is a weighted sum of the points achieved in each sub-index.

Broadie & Rendleman (2013) investigated whether the Official World Golf Ranking (OWGR) system was prone to bias for the four major tours (PGA Tour, European tour, Japanese Tour and Asian Tour) by comparing the OWGR system with two unbiased methods for estimating golfer performance: 1) Score-based skill estimation (SBSE) method and 2) Sagarin method. The SBSE method provides a player's mean 18-hole score played on a neutral course, and statistically removes all intrinsic course difficulties such as course setup and weather. The Sagarin method uses a player's won-lost-tied record against other players when they play on the same course on the same day, and the stroke differential between those players, then links all players to one another based on common opponents. Highly correlated rankings were found between the three methods however a large difference depending on tour affiliation was found which illustrates the existence of bias. There was a clear tendency for OWGR/ SBSE ranking pairs to fall below the 45-degree line for non-PGA tour players and above the line for PGA tours. A similar result was found for OWGR/Sagarin relationship. Moreover, it was found that a golfer's primary tour affiliation is the PGA tour is penalised an average of 37 OWGR rankings positions relative to non-PGA Tour affiliated golfers (Broadie & Rendleman, 2013). The analysis revealed statistically significant tour bias in the OWGR against PGA tour affiliated golfers and was greater among less skilled players.

Jackson (2016) developed a novel metric for measuring the similarity between players in the Australian Football League (AFL). "A players involvement in games was measured as a combination of event type, the current state of the game, and the location of the event (Jackson, 2016, p.3). The similarity between two players, i and j, was calculated as a linear transformation of the vector angle between the players individual involvement vectors $\boldsymbol{w_i}$ and $\boldsymbol{w_j}$. A similarity of 0% is produced for players with a vector angle of $\frac{2}{\pi}$ (completely orthogonal). Applying this measure to the 2015 AFL season West Coasts Jason J. Kennedy and GWS Giants Jeremy Cameron were the two most similar players. The measure was also used to identify the most unique players within the AFL. The player similarity metric was used to compare player efficiency relative to the 5-most similar players using:

$$EFF_i = \frac{\bar{x}_i}{\frac{1}{5}\sum_{i=1}^{5} \bar{x}_{S_{i,j}}} - 100\% \qquad (5)$$

Here, $\bar{x}_i$, represents the average score per game of player i, calculated using the official AFL player ratings to measure player performance. $S_{i,j}$ is the jth most similar player for player i. The benefits of the similarity measure are that it measures player performance by examining player efficiency relative to similar player rather than raw average points per game. For example, Brisbane's Dayne Zorko was 42nd in the 2015 AFL competition for average points per game, but no. 1 for similarity relative to similar players.

Moore, Bracewell, McIvor & Stefani (2018) developed a result-driven system for rugby union. Initially several logistic regression models, applying random forest selection, were developed, with the match outcome (win or lose) as the target variable and player actions as covariates. These models did not sufficiently capture the signal to noise ratio, therefore a 'live odds', i.e. the estimated probability of victory $p(win)$ for a chosen reference team (the home team), was applied. An ensemble of time-based logistic regression model was developed to train the live odds model on a set of 600 matches. The model accuracy in terms of log-loss improved over the course of a match, with the greatest improvement occurring in the last 20 minutes. Using the probability of win, positional specific metrics are identified through regression analysis to derive statistically and practically significant attributes related to within game changes to the probability of win. To derive individual match ratings, a player's features are aggregated to a match-level and normalised by minutes played. These features were grouped by position, and for each position a linear transformation of the aggregated information is learnt. Next, a set of position-based quantile transformation are learnt to transform the match ratings into standard normal variables. A sigmoid function is applied to obtain match ratings in $[0, 1]$. An exponentially weighted moving average is applied to each player current rating to derive a players rating based on a series of matches. The odds provided by the New Zealand TAB were converted to probability of victory. These two statistics were compared against the ratio of victory. It was found that the individual-derived team ratings outperformed the TAB.

A note on sport ratings literature

During the research process Albert, Glickman, Swartz & Koning (2017) published a textbook titled "Handbook of Statistical Methods and Analyses in Sports". This work is a comprehensive review of commonly applied statistical techniques and methodologies used across a multitude of sports such as baseball, ice-hockey, basketball, American football, soccer, golf, and cricket. A handful of these studies and modelling approaches have been reviewed throughout this literature review and were applied during the development of the rating systems.

2.2 OVERVIEW OF CREDIT RISK MODELS

Credit scoring is the term used to describe statistical methods used for classifying applicants for credit into good (non-risky) and bad (risky) classes (Hand & Henley, 1997). Credit scoring has become increasingly vital with the remarkable growth in consumer credit in recent years and has become one of the most successful application areas for statistical and operational research.

There are two main types of credit risk scorecards widely used in the finance industry: 1) Behavioural scorecards, and 2) Application scorecards. Application scoring is applied to determine the answer to the first question, while behavioural scoring is applied to answer the second questions. Broadly speaking, banks apply application and behavioural scoring to deal with two different types of customers requiring different types of decisions: 1) *New customers'*– should the new applicants for credit be granted? and 2) *Existing customers'* – should the agency grant the request of an old customer to increase credit limit? How risky are the existing customers? What products to offer to the existing customers to maximize the profit?

Application scoring refers to the scoring an applicant's credit score using static data obtained from application forms and are used to decide whether to grant lines of credit for new applicants (Chen & Huang, 2003). Behavioural scorecards are used to analyse of existing customers (Setiono, Thong & Yap, 1998). The prerequisite for using a behavioural scorecard is that the financial institution observes and obtains data about payment behaviour on a month-by-month basis, so the scores are dynamic (i.e. change monthly).

Both application and behaviour scoring deal with classification analysis and their main objective is to classify customers into groups consisting of people with similar default risk (Lancher, Coats, Shanker, & Fant, 1995). In credit scoring, classification analysis is applied to categorize a new applicant as "accept "or "reject" by using characteristics such as age, income and marital status (Chen & Huang, 2003), whereas classification of behaviour scoring is used to describe the behaviour of existing customers, based on behaviour characteristics such as payment patterns and spending patterns, and to predict the future behaviour of existing customers (Setiono, *et al*. 1998). The standard techniques used in application scoring can be used for behavioural scoring. However, the data and the objective of behaviour scoring make it different from application scoring.

When credit scoring models were first introduced, the aim was to estimate future credit worthiness of applicants and to grant credit to those with low default risk. The underlying assumption for application scoring is that the creditworthiness of a customer is time dependent (Thomas, 2000). Application scoring models are typically built using a minimum of one year's credit performance of applicants. Generally, the data for application scoring is provided by credit bureaus.

The objective of application scoring is to classify a new applicant as good (non-risky) and bad (risky) based on characteristics such as age, income, marital status, number of dependents and employment type, whereas behavioural scoring classifies the behaviour of existing customers based on purchasing and payment patterns. Behavioural scoring models provide better information for setting credit limits, creating new products, and identifying risky customers.

Behavioural scoring models allow the user to understand their customers (i.e. debtors) spending and repayment patterns to minimize losses and estimates the probability that a customer's credit behaviour remains in, or returns to, a satisfactory condition in the future. Behavioural scores therefore make use of a customer's recent behaviour to predict if they are likely to default in the immediate future. A pure behavioural scoring system will only include variables dealing with the customers' performance and the current values of variables from monthly credit bureau reports. Other behavioural systems include personal characteristics such as age, time with banks, residential status, as well as pure behavioural characteristics.

A behavioural scoring model is developed using data for a sample of customers before and after a point in time, including all the characteristics which describe the performance of these customers over this period (Thomas *et al*., 2004). The period before the observation time point, usually 6-12 months, is called the performance period, while the period after the observation time is the outcome period, which is usually taken as 12 months.

Behaviour scores are not only used to identify risky customers, they are also used in assigning new credit limits for good customers, marketing new products to good customers, or managing recovery of debt if an account turns bad. The most widely used techniques for building scorecards are Linear Discriminate Analysis and Linear Regression. Other techniques which have been applied in the industry include logistic regression, probit analysis, non-parametric methods, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms, artificial neural networks, and conditional independence models.

Hand & Henley (1997) provides a brief introduction into the credit risk environment, summarizing the statistical classification methods found in the consumer credit scoring and outline the performance measurements predominately implemented to assess model accuracy.

Large datasets are not uncommon therefore statistically significant variables must be defined to produce a parsimonious model with over fitting effects. In credit scoring approaches to selecting characteristics (i.e. predictor variables) are commonly used: expert knowledge, stepwise statistical procedures, information value, discriminate analysis, regression, logistic regression, mathematical programming methods, recursive partitioning, expert systems, neural networks, smoothing parameter models and time-varying methods were mentioned as industry standard models. Hand & Henley (1997) identify various publications that have been implemented the techniques in the credit risk environment and outlines scenarios/ areas in which the methods have strong discriminatory power and areas of weak discriminatory power.

Hand & Henley (1997) concluded that the classification method is dependent on the details of the problem: the data structure, the characteristics used the extent to which it is possible to separate the classes by using the characteristics and the objective of the classification (i.e. overall misclassification rate, cost-weighted misclassification rate, bad risk rate among accepted applicants, probability metrics).

The performance of a credit risk scorecard is usually assessed using divergence statistics and information statistics. Industry standard metrics include the Receiver Operating Characteristic (ROC) curve in which the *true positive rate* (the proportion of the true good risks that are above the threshold) is plotted against the *false positive rate* (the proportion of true bad risks that are above the threshold).

2.2.1 Credit Risk Scoring Systems

Although statistical models are utilised to evaluate many problems in the financial industry, the focus of this section will purely centre on credit risk scorecards. The section provides a comprehensive review of academic literature outlining the application of statistical techniques to develop application and behavioural-based scoring systems. This section has been divided into two sub-sections: 1) Application scoring and 2) Behavioural scoring systems.

2.2.1.1 Application Scorecard

Application scoring refers to the scoring an applicant's credit score using static data obtained from application forms and are used to decide whether to grant lines of credit for new applicants (Chen & Huang, 2003). This section reviews of Application scorecards.

Banasik & Crook (2005) adopted an Accept-Reject (AR) augmentation model which used a set of predictors to determine if an applicant has been accepted or rejected and on the basis all applicants, accepted or not, are assigned a score. As there is a range of equivalent scores, applicants with similar scores were assigned to intervals or ranges, with each range of scores being presented by an interval wherein there are both accepted cases and rejected cases. In each interval the accept ratio was calculated and regarded as the probability of acceptance within a given interval. It was assumed that for a given interval that the probability of good repayment performance was equally likely among accepted and rejected applicants. This augmentation accept-reject technique was applied within a lean modelling framework, and was repeated for 23 models, one for each number of variables between 4 and 26. For each band and each number of variables, the total scope for reject inference was provided. An initial dataset containing 2540 applicants was used as a sample upon which an AR model was formulated and estimated. A good-bad model was built, using the remaining 9668 English and Welsh applicants, and adopted to assess the efficiency of reject inference. Two-thirds of the applicants were used to build training model parameters and choosing the cut-off probabilities. Bands were accumulated such that each included case from the preceding band. For each band, the variable coefficients are calculated using the bands training sample and a cut-off point that equalized the predicted number of bands with that observed. Band specific coefficients and cut-offs were used to score and classify the hold-out cases. The results showed that there was scope for reject inference to improve predictions. Overall, across the five bands the scope of reject inference to improve predictive performance would be 5.3%.

Recently, there been an increase in the use of ensemble strategies, neural networks, and hybrid-based modelling techniques in credit-risk scorecard development. To evaluate the modeling power of neural networks (NN) against traditional techniques, Alabi, Issa & Afolayan (2013) compared the predictive power of artificial neural network against a discriminant analysis. The techniques were applied to a dataset composed of 200 records (163 goods and 37 bads), and 15 variables (9 categorical and 6 numerical; 14 independent and 1 dependent). The error function that the network tries to minimize during training was *cross entropy error*. Discriminant Analysis results revealed a cross validation accuracy of 88.50%. The NN correctly classified 100% (47) of good customers and 88.9% of bad customers (8) and had an overall predictive accuracy of 98.2%. Given that the neural network model produced fewer *'bad accepted' (%, amount)* compared to the discriminant analysis models, the former model achieves a lower cost of misclassification. Moreover, the NN model produced an overall classification greater accuracy, and therefore it was found to be the superior model.

Similarly, Ince & Aktan (2009) explored the classification performance of credit scoring models using traditional methods (discriminant analysis and logistic regression) and artificial intelligence approaches (classification and regression trees, decision trees and neural networks). The discriminant function produced an average correct classification of 65.23% and 62%, across the training and testing samples, respectively. Of the misclassifications 31.9% were type I (good customers misclassified as bad customers) and 43.32% were type II (bad customers misclassified as good custom). The stepwise logistic regression produced a training accuracy of 66.37% and testing accuracy of 62.33%, with 42.86% type I error and 32.22% type II errors. The neural network applied a back-propagation algorithm and revealed a classification accuracy of 78.85% and 61.52%, across the training and testing sample, respectively, and produced 44.59% type I errors and 29.25% type II errors. The decision tree adopted a 1-SE pruning procedure and the optimal tree was selected by using the lowest cross-validated or testing set error criteria. Decision tree produced a 39.88% type I errors and 32.01% type II errors. Given that type II costs are significantly higher than those associated with type I, it was concluded that neural networks significantly reduced costs associated with misclassification, compared to the other 3 approaches. Overall, it was found that the CART produced the best average classification accuracy followed by logistic regression, discriminant analysis and neural networks.

West (2000) investigated the accuracy of five neural network architectures (MLP: Multilayer Perceptron; MOE: Mixture of Experts; RBF: Radial Basis Functions; LVQ: Learning Vector Quantization and FAR: Fuzzy Adaptive Resonance) for credit scoring application and benchmarked their performance against five traditional methods (LDA, LR, *k*-nearest neighbours, kernel density estimates and decision trees. The techniques were applied to two separate datasets, an Australian [credit] dataset and a German [credit] dataset. The results revealed logistic regression to have the lowest overall credit scoring error (0.2370), followed by MOE (0.2434), RBF (0.2540), MLP (0.2672), LDA (0.2667), LVQ (0.3163), CART (0.3044), kernel density (0.3080), k-NN (0.3240) and FAR (0.4039). For the Australian dataset, again, results showed logistic regression to have the lowest credit scoring error (0.1275) followed by RBF (0.1286), MOE (0.1332), LDA (0.1404), MLP (0.1416), KNN (0.1420), CART (0.1502), Kernel density (0.1660), LVQ (0.1703) and FAR (0.2461). For the German dataset, the [top 5] model results showed MOE (0.2243) to be the most accurate neural network model followed by logistic regression (0.2370), RBF (0.2437) and ML (0.2496). Overall, the results, across both datasets, showed MOE, RBF, MLP and logistic regression to be superior in terms of overall errors, while LVQ, LDA, KNN, Kernel density and FAR and CART were labelled as inferior models. Results suggested that MOE and RBF networks produced fractional improvements in credit scoring accuracy ranging from 0.5% up to 3%, this was due to their ability to partition the input subspace. Moreover, it was claimed that traditional methods suffer the curse of dimensionality producing inferior results relative to the MOE neural network models.

Jensen (1992) applied a neural network using back propagation to 125 credit applicants to predict loan outcomes. The neural network consisted of 24 input neurons, each representing an applicant’s characteristics obtained from their application form, 2 hidden layers each consisting of 14 neurons. The output layer consisted of three neurons, one for each possible outcome. The models’ predictive power was subjected to two individual tests. The first test utilised 75 applicants to train the network, while the remaining 50 applicants were used to evaluate the model. The network correctly classified 0%, 28.5% and 94.6% of delinquent loans, charged-off loans, and paid-off loans, respectively. Evaluating the credit scoring scheme revealed 76% of applicants were correctly classified. The results of the study indicated the commercial benefits and strong predictive power of building neural networks to evaluate credit worthiness of loan applicants.

Similarly, Pacelli & Azzollini (2011) compared the predictive power of two feed-forward neural networks, that differ in activation function and model parameters, applied to two separate datasets containing a set of Italian manufacturing companies and financial industries associated with each company. Pacelli & Azzollini (2011) objective was to analyse the ability of neural networks to forecast the credit risk of each Italian manufacturing company (*i.e. safe, vulnerable, and risky*). Both neural network models were trained through back propagation and consisting of an input layer, containing 24 neurons, two hidden layers, and 1 output layer. Neural Network A was build using a sample of 273 Italian companies and their associated financial variables. After 101,470 cycles the network produced an error revealed a classification accuracy of 84.2% among companies labelled as safe, a classification accuracy of 73.9% among companies labelled as vulnerable. However only 34.8% of risk companies were correctly labelled. Neural Network B was build using a sample of 507 Italian companies and their associated financial variables. After 10,000 iterations the network produced an error rate = 0.3308. Results revealed that the model was unable to correctly classify companies, classifying all 148 companies into the first class, and producing a validation error rate = 0.3311.

Len & Chen (2005) extended the application of neural networks by developing a two-stage hybrid credit scoring model using multivariate adaptive regression splines (MARS) to identify significant variables and using these variables as the input node of a neural network model. Len & Chen (2005) hypothesized that by adopting MARS to identify the significant variables to input into the model: (1) the training time to build the optimal neural network would significantly decrease and (2) the predictive power of the neural network would significantly increase. The proposed model was applied to a dataset containing 510 housing loan customers (459 good customers and 51 bad customers). A 5-fold cross validation (CV) scheme was adopted to evaluate the capability of the built model. The result of the hybrid model was compared against discriminant analysis, logistic regression, MARS, and neural network results. It was found that the two-stage hybrid method produced the highest classification accuracy (84.7%), followed by neural networks (84%), MARS (81%), logistic regression (76%) and discriminant analysis (75.5%). Moreover, the proposed model had fewer type I errors (classify good customers as bad) and type II errors (classify bad customers as good), therefore a lower expected cost of misclassification. The optimal neural network typology consisted of an input layer made up of 5 nodes, a single layer made up of 20 nodes and an output layer containing a single node. The top 5 significant variables using MARS were: Monthly instalment/ monthly income, number of guarantors, loan types, loan amount/ house appraisal value and marital status.

Another example of the application of hybrid modelling strategies in application scoring in Bahrammirzaee (2011). Bahrammirzaee (2011) developed a hybrid intelligent system to produce credit scores using reasoning-transformational models. The hybrid intelligence system was created due to three key reasons: (1) Hybrid systems overcome the limitation of each individual technique, (2) A single technique is not applicable to many sub-problems that a given application may have and (3) Hybrid systems encapsulate multiple information processing capabilities within a single architecture. The proposed model was trained using 100 loan applicants (50 personal and 50 corporate) and tested on a dataset with the same composition. The first module contained a knowledge base of personal loan applicants and corporate loan applicants. The knowledge was extracted from a credit ranking model developed by several banking experts (Bahrammirzaee, 2011). The inference engine applied Aristolean logic composed of 136 rules utilizing backward inference methodology. The output of the expert system is the score for each criterion and a final score which is the sum of all individual criteria (50 personal scores and 50 corporate scores). The scores produced by the expert system, was used as inputs to the neural network which implemented a back-propagation algorithm. The target scores for the neural networks were produced by several banking experts for each of the 100 applicants (i.e. expert scores). Applying the model to 50 personal and 50 corporate loan applicants, a statistically significant difference was established between expert system's scores and hybrid credit rating system scores. Moreover, it was found that the errors (i.e. MSE, RMSE and MAD) produced by the hybrid system was significantly less than those produced by the expert system, demonstrating that the hybrid intelligence systems provide greater accuracy and power in credit ranking compared to expert systems.

Similarly, Chuang & Huang (2011) developed a hybrid credit scoring model with the capability of enhancing classification accuracy and reducing misclassification. The proposed model incorporated three key techniques, rough set theory (RST), artificial neural networks (ANN) and cased based reasoning (CBR). The model first integrates the RST and ANN model to identify the accepted and rejected applicants, CBR is then applied to detect type I errors i.e. rejected applicants that should have been accepted. The hybrid model adopts RST due to its ability to handle noise and isolate relevant attributes, reducing model-training time and increasing classification accuracy. The hybrid model was trained, tested, and validated using credit card applicant data. The neural network implemented a back-propagation algorithm, consisting of an input layer (9 nodes), a single hidden layer (5 nodes) and an output layer (1 node). The network was trained using learning rates ranging from 0.01 - 0.2, momentum rates from 0.7 - 0.93 and varying lengths from 1,000 – 10,000. The optimal network architecture had a classification accuracy of 81.5%. Moreover, benchmarking the hybrid model against traditional scoring models revealed that the RST-ANN-CBR model was the optimal model in terms of accuracy rate, least number of Type I and II errors, and reducing cost of misclassification.

Another example of the application of ensemble modelling in application scoring in Bahrammirzaee Bao, Lianju & Yue (2019). Bao, Lianju & Yue (2019) proposed an ensemble strategy integrating unsupervised learning with supervised learning at different stages, for credit risk assessment. Unsupervised learning techniques were applied at two different stages: 1) the data clustering and 2) the consensus stage. The ensemble strategy determines consensus classification decisions based on the predictive outcomes produced by individual machine learning models. The strategy was applied to a combination of the German and Australian credit datasets from the UCI machine learning repository and a Chinese P2P enterprise. The results suggest that the cluster-based consensus model could obtain a more accurate and reliable classification as it achieves the best MCC of 0.542. Moreover, it was inferred that the strategy of combining unsupervised and supervised machine learning at multiple stage proved to be an effective strategy. To prove the effectiveness of the consensus, model the authors compared the consensus and cluster-based models. It was found the consensus strategy helps consensus models to outperform the individual models.

Khandani, Kim & Lo (2010) applied generalized classification and regression trees (CART) to construct a model which forecasts credit delinquencies and defaults. To improve predictive power an adaptive boosting technique was adopted to address the issue of highly skewed proportion of good and bad realisations. The models' predictive power was evaluated by assessing its ability to forecast '90-day-or-more' delinquent customers during a 6-month period. The results showed that the average CScore among 90-day-or-more delinquent customers (2.4% of account) was 61.2 across the 10 calibrations (model applied 10-fold CV) and testing periods, while those accounts that were not delinquent (97.6%) averaged a CScore of 1.0, indicating strong discriminatory power. The results showed that the average forecast among customers who were current and did not reach delinquency was 0.7, while the average forecast for straight roller was 10.3. Khandani *et al.* (2010) rebuilt the model by dividing the data into equally sized sets separated by the availability of features and performing a 10-fold cross validation (CV) on each one. The results revealed a significant improvement in both model precision and recall between groups 1 (accounts with the most missing feature) and group 2 (accounts with the fewest missing features). It was shown that regressing forecasted delinquencies on realized delinquencies produced and R^2 of 85% for a 6 month and 12 month forecast horizon indicating that the model can generate leading indicators of deterioration in consumer credit worthiness. Overall, it was concluded that the regression tree model produced accurate credit forecasts 3-12 months in advance and yielded costs saving between 6% - 23%.

2.2.1.2 Behavioural Scorecard

Behavioural scorecards are used to analyse of existing customers (Setiono, Thong & Yap, 1998). Such scoring techniques allow financial institutions to determine for example whether a customer's credit-limit should be extended, what financial products should be served to various customer segments, etc. Behavioural scorecard allows for deeper analysis and allows creditor to mine a customer's payment history, credit utilisation over time and type of financial products purchased. This section reviews Behavioural scorecards adopting both traditional and non-traditional techniques.

Given behavioural scorecards form the backbone for a lot of financial institutions, the amount of published literature surrounding the development of such systems is limited and scarce. Like application scorecard, behavioural scorecards have also experienced an increase in use of ensemble strategies, neural networks, and hybrid modelling strategies.

Previously, such modelling strategies were non-existent within the credit risk industry given hybrid ensemble strategies suffered from interpretability and transparency issues (please see: Kim, Lee, Shin, Yang, Cho, Nam, Song, Yoon & Kim, 2019; Zhang, He & Zhang, 2018; Bao, Lianju & Yue, 2019; Shen, Zhao, Li, Li & Meng, 2019; Papouskova & Hajek, 2019). Specifically, the most common ensemble strategy within the credit risk assessment is integrating different machine learning models for credit scoring, and one of the mainstream ensemble strategies is to make consensus classification decisions based on predictive outcomes of individual machine learning models. There are different approaches to perform ensemble strategy in terms of using different base learners (single classifiers or models) and different consensus techniques. The difference between the traditional credit risk research and modern more recent credit risk research focuses on implementing unsupervised learning techniques at different stages.

Fadaei-Noghani & Moatter (2017) proposed a hybrid data-mining methodology, which considered feature selection and the decision cost, to increase the accuracy of detecting fraudulent credit-card behaviour. The developed methodology adopted a feature selection approach which incorporated prior feature filtering and a wrapper approach using C4.5 decision tree, and an ensemble classification is performed using cost sensitive decisions trees in a decision forest framework. The ensemble classifier yielded a performance improvement of 33% compared to Naïve Bayes, Bayesian network, ID3 and J48 classifiers.

Zieba and Swiatek (2012) proposed an ensemble classification method based on switching class labels, which switches the class of an observation according to an estimated probability, $p(i|j)$, which represents the probability that an object in the jth class will be switched to the ith class, for credit assignment. The switching techniques addresses imbalanced dataset and issues with asymmetric cost matrices. The proposed method was found to have the lowest false negative (FN) ratio and experimental risk index (ERI) when benchmarked against C.45, K-NN, MLP, LR and NB classifiers.

Feng, Xiao, Zhong, Qiu and Dong (2018) proposed an ensemble classification method. The classifiers are initially selected based on classification ability and the relative costs of type I and type II error in the validation set. With the selected classifiers, different classifiers were combined for the samples in the testing set based on their classification results to get an interval probability of default by using soft probability. The proposed method was compared with some well-known individual classifiers and ensemble classification methods for credit scoring, including five selective ensembles, by using ten real-world data sets and seven performance indicators. Through these analyses and statistical tests, the experimental results demonstrated the ability and efficiency of the proposed method to improve prediction performance against the benchmark models.

Kennedy *et al*. (2013) examined the contrasting effects of altering the performance period and outcome period on the stability of predictions produced by behavioural scoring methods. The study evaluated the efficacy of varying performance and outcome window sizes on the classification accuracy of a logistic regression model. Kennedy *et al. (2013)* compared performance window sizes by classifying loans over a range of fixed outcome window sizes and varying performance windows. These performance window sizes are: 6, 12, and 18 months, while fixed outcome window sizes are: 3, 6, 12, 18, and 24 months. The study assessed classification accuracy across two behavioural scoring approaches: 1) *Current status* - this approach assigns either a *'good'* or *'bad'* status to consumers based on their account status at the end of the outcome window. 2) *Worst status* - This approach assigns either a *'good'* or *'bad'* status to consumer based on the account status during the outcome window. Comparing classification performance of varying outcome window size and a fixed performance window of 12 months, using a *'worst status'* and '*current status'* approach, revealed that in the '*worst status'* scenarios a clear separation existed between shorter outcome windows (3, 6 and 12-months) and longer outcome windows (18 months and 24 months). A Kruskal-Wallis test revealed at least one significant difference between the results. The '*current status'* approach revealed that a logistic regression classifier using "a 3-month outcome window consistently achieved the highest average class accuracy" (Kennedy *et al*. 2013, p.), followed by 6, 12, 18 and 24-month outcome windows. It was found that when using the *worst status* approach, the shorter outcome windows produce relatively superior average class accuracy for a 12-month performance window. The results found that a 3-month or 6-month outcome window produced the highest average class accuracy in conjunction with a performance window of 12 months using a logistic regression classifier. Overall Kennedy *et al*. (2013) revealed that a "classification task based on worst status approach and a longer outcome window size achieves a higher average class accuracy" (Kennedy *et al*., 2013, p. 9), using a 12-month performance window.

Unlike Kennedy (2013), Yobas & Ross (2000) conducted a comparison study evaluating the predictive power of linear discriminant analysis (LDA), neural networks

(NN), genetic algorithms (GA) and decision tress (DTs) in the classification of credit cards customers. The four techniques were applied to a dataset containing 1001 credit cards consumers. A case was declared ‘bad’ if the individual has missed at least one payment in the sample period and ‘good’ otherwise. The optimal neural network typology was identified by testing various learning and momentum rates, activation function and epoch numbers. A classification accuracy of 64.2% was achieved by the neural network. The decision tree model achieved an average accuracy of 62.3% across the 10 trees. The GA model achieved an accuracy of 64.5%. LDA results revealed that the model correctly classified 68.4% of the cases. Although results indicated LDA to be superior of the 3 investigated models, Yobas *et al.* (2000) stated that further analysis is required as these results were inconsistent with results presented in other studies, and stated that factors such as differences in the types of individuals in the samples, differences in sample sizes, differences in the transformation applied to the data, could explain these inconsistencies.

Similarly, Hsieh (2004) applied a self-organising neural network to identify profitable customers and segments based on repayment behaviour, recency, frequency and monetary (RFM) behaviour. Using account and transition data Hsieh (2004) applied self-organising neural network methodology to identify customer segments based on repayment behaviour (i.e. transaction users, convenience users or revolver users – target scores), and RFM behaviour scoring predictors. Hsieh (2004) constructed a 4 × 4 SOM (self-organising map) to identify profitable customer segments based on previous repayment behaviour and RFM behavioural scoring predictors. Hsieh (2004) developed customer profiles using neural network (32-20-3) sensitivity analysis and an aprior association inducer. The SOM results showed that customers fall into three major profitability groups dispersed over 16 clusters. Moreover, it was found that customers with values tending towards R↓ F↑ M↑ can be targeted with greater accuracy. The clusters were then profiled by feature attributes determined using the apriori association inducer. Overall Hsieh (2004) presented a behavioural scoring model that enabled the user to deduce profitable and non-profitable customer segments from credit data.

Wang, Jiang, Ding & Liu (2018) proposed a novel behavioural scoring model based on a mixture survival analysis framework to dynamically predict the probability of default over time. ‘Cured’ borrows are those that never default and uncured are those that will eventually default at some point during the loan-term. A random forest modelling technique was utilised to determine whether a borrow defaults and a random survival forecast is introduced to model the time to default. The proposed ensemble mixture random forest (EMRF) model, using an averaging ensemble method, was compared against the mixture cure model (MCM) and the Cox proportional hazard model (Cox PH) which predicted the probability of default over time. The EMRF model predicted whether a borrow will default, through the random forest component, and predict when they are most likely to default, through the random survival forecast. The model was trained on 60% of the data and tested on 40% was held for training. A repeated 10-fold cross validation was applied across the three models, probability of default was predicted for a 12-month loan and the time interval was one month. Across 10 individual time intervals the proposed EMRF model outperformed the MCM and Cox PH model 7 out of 10 times using the AUC performance metric, while it outperformed all models across all time interval using the K-S statistics.

Thomas (2000) proposed two extensions to behavioural credit scoring models, producing more robust, highly improved and focussed scorecards. First, it was suggested incorporating current economic conditions into scoring methodologies, as an individual's financial situation is dedicated by economic conditions. Given the several years' time lag between transactional data collected and its use in scorecards, the model scores for each consumer may not be indicative of their current financial situation. It was established that economic variables such as unemployment claims had a major impact on default. It was suggested that one way to incorporate various economic conditions into a consumer's credit score would be to have two scores, one for prosperous economic conditions and the other for failing economic conditions. However, to build a model that incorporates all the stages of economic conditions one would have to use old data. Second, it was suggested changing the overall objective of credit scoring model from *'minimising the risk of a default customer'* to *'maximising the profit a customer brings'*. To build a profit scoring model Thomas (2000) suggested three approaches: 1) Build on existing scoring models which estimate default rates, attributes, and acceptance, and demographically segment the population according to their score and these measures. Finally establish the profitability of the various segments. 2) Describe profit as a linear function of categorical variables obtained from the application form using the regression of credit scoring. A drawback to this method is that almost all the data will be censored in that total profit is not known. 3) "Build on Markov chain approaches to behavioural scoring to develop more precise stochastic models of customer behaviour" (Thomas, 2000, p.166).

Papouskova & Hajek (2019) proposed a model which modelled the overall credit risk of a consumer's loan using expected loss (EL). To model EL, three key credit parameters required estimation: 1) probability of default (PD), 2) loss given default (LGD) and 3) exposure at default (EAD). Papouskova & Hajek (2019) proposed a two-stage credit risk modelling approach integrating 1) class imbalance ensemble learning for predicting PD and 2) an EAD prediction using regression ensembles. A stacking method was applied to combine multiple predictive models and consisted of two steps: 1) generating a set of base predictions and 2) these predictions are used to train the meta-classifier or meta-regressor.

The results show that stacking with RF as the meta-learning algorithm outperformed the other classifiers on all evaluation metrics. Therefore, this method was chosen for modelling PD in the two-stage method. Regression modelling of EAD for the sub-population of default loans was applied and included the single regressors (regression tree, random forest, linear regression, SVR, deep neural network), homogenous (RF, Rotation Forest, Additive Regression, Bagging and Random Subspace) and heterogenous stacking ensemble methods. Overall staking with Linear regression performed best across all the evaluation metrics. This was preferred method for modelling EAD in the two-stage EL model, while Stacking with RF was selected for the second stage as the best performance. Finally, the two stage EL modelling in an integrated framework including stacking with RF to model PD and Stacking with LR to model EAD. The results showed that misclassification cost can decreased by using a heterogenous method with RF as a meta-classifier in modelling PD.

Similarly, Bakoben, Adams & Bellotti (2019) proposed a two-stage approach for determining dissimilarity between pairs of time series objects. Bakoben, Adams & Bellotti (2017) applied cluster analysis to the behaviour of credit-card accounts to help assess credit risk level. The first stage fitted a multivariate time series model 1) to characteristics the dynamic nature of an account and 2) to reduce data dimensionality. Stage two computes the dissimilarity between confidence region of the model parameters identified in stage one. The accounts were clustered using Euclidean distance clustering and an uncertainty-aware clustering approach. A logistic regression model was developed to predict defaults and evaluates clustering performance. The default status is predicted based on cluster assignment and the analysis is performed on the first 2/3 of each account and forecast default is measured over the last 1/3 period. The logistic regression prediction and forecasting models based on ellipsoid clustering showed good performance in comparison to the models based on the outcomes of the clustering analysis which uses Euclidean distance and outperformed models based on the aggregated behaviours.

2.3 AN EXPLORATION OF MODEL EVALUATION METRICS

The secondary objective of this research is to develop a novel evaluation metric to quantify the effectiveness of sport-based ratings. Here, a comprehensive review of commonly used evaluation metrics, also known as model performance metrics, is provided. An exhaustive review has not been conducted as the non-reviewed performance metrics are not suitable or applicable when evaluating sport-based rating systems.

Specifically, this chapter reviews commonly applied performance metrics within industry and academia, outlines the technical details and limitations of each metric, and explains why certain performance metrics perform well within certain forecasting scenarios and lose information in other circumstances. Rating systems apply different objectives depending on the forecasting scenario, and therefore adopt different evaluation metrics to assess model outputs. Throughout the literature review it was found that there exists no universal evaluation technique or evaluation strategy to measure the validity of meaningful sport-based rating systems. Therefore, the objective is to consolidate these findings to identify the ideal set of criteria and the ideal methodology to construct a novel performance metric to evaluate the effectiveness of ratings.

The rating systems developed and reviewed throughout this chapter have applied a myriad of performance metrics such as coefficient of determination, correlation, accuracy, root mean squared error (RMSE), Symmetric mean absolute percent error (SMAPE), mean absolute error (MAE), 10) Wilcoxon *p-value*, Spearman's rank coefficient, area under the curve (AUC), Kolmogorov-Smirnov test, leave-one-out cross validation (LOOCV), logarithmic-loss, calibration plot – empirical probabilities vs. predicted probability, and Hosmer-Lemeshow test statistics.

The section is divided into two parts: regression and classification. This division is because the evaluation criteria applied to any given modelling exercise is dependent on the model type and the outcome variable.

2.3.1 Regression-Based Evaluation Metrics

Predictive models can either produce continuous or classification outputs, and the measuring criteria to establish the predictive power associated with these models is dependent on the type of outcome variable.

Hyndman & Koehler (2006) stated that regression-based performance metrics can be classified into 4 groups: 1) scale dependent, 2) percentage error, 3) relative error and 4) scale-free error.

The mean squared error (MSE) is a scale dependent measuring metric which means there can be major variation in the scale of observations between series such that a few series with large values can dominate the comparisons (Chatfield, 1988). MSE can be inappropriate for comparing predictive accuracy on different variables or different time intervals because it is a scale-dependent measure. Another major drawback is that MSE is heavily influenced by large errors compared to small errors, because of the "squared error" effect. This squared error effect within the numerator "over-inflates" the mean squared error for predictions further away from the actual outcome. The most common shortcomings of the MSE are: 1) quadratic loss may not correspond to the modeller's loss function, 2) scale dependent, implying that it depends on the measurement unit for the outcome of interest) and 3) vulnerable to outliers (i.e. farther distance from the actual outcome). MSE performs well for forecasting procedures that avoid large forecast failures (Armstrong, 2001).

Another scale dependent metric is the root mean square error (RMSE) which is vulnerable to outliers in errors. Although, the measure is expressed in the same unit as the outcome variable, making for easier interpretations, than the MSE. While, the MSE illustrates the variance, the RMSE illustrates the standard deviation. The measure is the square root of the average squared errors, therefore, larger errors have a disproportionately larger effect on the root mean square error because the effect of each error on RMSE is proportional to the size of the squared error (Chai & Draxler, 2014). “The RMSE is an inappropriate and misinterpreted measure as it is a function of 3 characteristics of a set of errors, rather than of one (average error) (Willmott & Matsuura, 2005, p.1)”. The RMSE and MSE measure is commonly applied in the climatic and environmental literature.

The mean absolute error (MAE) is also scale dependent, and therefore cannot be compared across different data series. MAE measures the prediction accuracy in absolute terms and is easy to understand and compute. Compared to RMSE and MSE this measure is less vulnerable to large errors because of the absolute value characteristic. The absolute values prevent negative and positive errors from offsetting each other (Hyndman, 2006). The scale deficiency issue can be solved through percentage error measures such as mean absolute percentage error. Another drawback is that the MAE assumes that the mean is stable over time (Choi, Hui & Yu, 2013; Hyndman, 2011). Due to the absolute loss function, MAE is more sensitive to small deviations from 0 and less sensitive to large deviations compared to that squared loss function in MSE. MAE performs well for forecasting procedures that produce occasional large forecast failures, while performing reasonably well on average.

An example of scale independent metric is the mean absolute percentage error (MAPE) metric which allows for meaningful forecast comparisons between two models. The MAPE measure is a strictly positive value between $[0, 100]$ and is a unit free measurement.

It suffers from two major drawbacks: First, it is infinite or undefined if there are zero values in the data series. Hyndman (2006) and Makridakis & Hibon (2000) stated that percentage errors have an extremely skewed distribution when actual values are close to zero. Second, MAPE penalises positive errors heavier than the penalties placed on negative errors. It can be shown that the MAPE is asymmetric where equal errors above the actual value result in a greater absolute percentage error than those below the actual value.

The MAPE is bounded on the low side by an error of 100% but there is no bound on the high side (Goodwin & Lawton, 1999), to resolve this issue Makridakis & Hibon (2000) proposed a symmetric MAPE (SMAPE) measure involving dividing the absolute error by the average of the actual observation and the forecast. Makridakis’s resolution is expressed as follows:

$$SMAPE(t) = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \qquad (6)$$

Here, A_t is the actual observation at time *t*, F_t is the forecasted (i.e. predicted) value at time *t* and *n* represents time period *n*. The forecast, F_t, is likely to be zero if the actual value, A_t, is zero. Therefore, the SMAPE involves division by a number close to zero. Moreover, the SMAPE may produce negative values indicating an unintuitive interpretation. An advantage of the SMAPE is their scale independence, and therefore are frequently applied when comparing forecast performance across different datasets and forecasting scenarios (Goodwin & Lawton).

The median relative absolute error (MdRAE) is also a scale independent metric which is advantageous over mean-error metrics such as MSE, RMSE and MAE, and due to the absolute error, it is more resilient to outliers. MdRAE is found by ordering the RAE from the smallest to the largest, and using their middle value, or the average of the middle values, as the median. However, a major drawback of MdRAE is the denominator in the presence of small or zero errors, because applying the benchmark method is no longer possible as it involves division by zero, therefore leading to extremely large or infinite relative errors.

Hyndman and Koehler (2006) proposed the mean absolute scaled error (MASE) as a generally applicable measurement of forecast accuracy without the problems such as scale dependence and outlier vulnerability found in the other measurements. Hyndman & Koehler (2006) proposed scaling the errors based on the in-sample MAE. A one-period-ahead forecasts from each data point in the sample is made. The result of MASE is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average one-step forecast computed in-sample. “The in-sample MAE is used in the denominator because it is always available and effectively scales the errors” (Hyndman & Koehler, 2006). The MASE can be used to compare forecast methods on a single series, because it is scale-free, to compare forecast accuracy across series. Moreover, MASE produces interpretable results for example, values greater than one indicate worse forecasts, on average, than the in-sample one-step forecasting of the naïve method (Hyndman & Koehler, 2006). It was claimed that in situations where there are vastly different scales including data which are close to zero or negative and intermittent demand studies, the MASE is the best available metric for forecasting accuracy. Although, Chen, Twycross & Garibaldi (2017) showed that the MASE can be dominated by a single large error, while Franses (2016) illustrated scenarios and criteria where the MASE did not ‘fit’ as these criteria did not imply the relevant moment properties.

2.3.2 Classification-Based Evaluation Metrics

The number of classification-based evaluation metrics out number that of regression-based metrics. This is primarily due to the ease of computing or communicating the error between the actual [continuous] outcome and predicted outcome in regression models. However, in classification models the problem of interpretation validity is slightly more challenging as the distance between actual and observed outcomes is categorical.

There are many ways to assess the validity of a classification model and the selected method of assessment is dependent on the modelling objective, the outcome of interest and forecasting scenario. In the case of regression, the goal is to decrease the distance (i.e. error) between actual and predicted outcome, while the goal of classification-based models is dependent on the number of outcomes and overall classifier objective such as increasing overall accuracy, increasing accuracy across the individual classes, or increasing precision or recall etc. In this section classification-based evaluation metrics are reviewed, beginning with confusion matrix-based metrics.

A confusion matrix, also known as an error matrix, consists of information about the actual and predicted classifications created by a classification system. “A confusion matrix is a clean and unambiguous way to present the prediction results of a classifier” (Brownlee, 2014, p.1). Typically, the rows represent the predicted classes, while the columns represent the actual classes. The confusion matrix is widely used to measure the accuracy of classification-type models by applying statistical measures derived from the matrix (Table 1).

A confusion table is a 2×2 matrix that reports the number of true positives (i.e. power), false positives (Type I error), false negatives (Type II error) and true negative. Here a brief explanation of the performance metrics that can be derived through the confusion matrix is provided.

	A**ctual condition**		
Predicted condition	**Total population**	**Positive class**	**Negative class**
	Predicted positive	True positive	False positive (Type I)
	Predicted negative	False negative (Type II)	True negative

Table 1: Confusion Matrix (i.e. Error Matrix)

Accuracy is a metric which can be derived using the confusion matrix. It is expressed as a percentage of the number of correct classifications divided by the total number of predictions. Accuracy is calculated from a tally of the correctness of the classification generated by sampling the classified data and expressed in the form of an error matrix (Story & Congalton, 2006). Generally, accuracy is an *unreliable* evaluation metric in the presence of unbalanced data (i.e. classes are not represented equally), this is also known as the *accuracy paradox*. It suffers from the issue of imbalance dataset or class imbalance because a classifier built using this data will be geared towards classifying majority of the observations. This issue of imbalanced data can also lead to model overfitting. The accuracy measure can be applied across multiple disciplines and its application can be found in many fields and is the one of most common classification evaluation metrics (Story & Congalton, 1986).

To mitigate the issue of class imbalance the Index of Balanced Accuracy (IBA) can be applied. Imbalanced classes occur when the ratios of prior probabilities between classes are significantly skewed. A two-class dataset is imbalanced when one of the classes is heavily under-represented relative to the other classes.

Another confusion matrix metric is Sensitivity, also known as *recall* or the *true positive rate*. Sensitivity is the number of true positive cases divided by the number of positives conditions (i.e. true positives + false negatives) and measures the true positive rate (Halligan, Altman & Mallet, 2015). It measures the classifiers ability to identify actual positives and label those observations as positives. Sensitivity measures the model's ability to correctly classifies the number of true positives that have '*the condition*' of interest (effectively quantifies the avoidance of false negatives). It is most suited for scenarios where the true negative cases are not of interest and having '*the condition*' is of importance (Trevethan, 2017), and therefore is also known as the '*detection rate*' in the clinical literature (please see Sano, Quarrancino, Aguas, Gonzalez, Harada, Krupitzki & Mordoh, 2008; Colin, Lanoir, Touzet, Meyaud-Kraemer, Bailly & Trepo, 2003; Rocco, Cobelli, Leon, Ferruti, Mastropasqua, Matei, Gazzano, Verweji, Scardino, Musi, & Djavan, 2006).

The 'inverse' of Sensitivity is Specificity, also known as the *true negative rate*. Specificity is the number of true negative cases divided by the number of negative conditions, i.e. false positive + true negative, (Halligan, Altman & Mallet, 2015). It measures the proportion of actual negative that are correctly identified, effectively quantifying the avoidance of false positives (Trevethan, 2017). Specificity measures a model's ability to correctly classify the number of true negatives who have the condition (for example, detecting the proportion of patients classified as <u>*not*</u> having cancer). The specificity measure is most suited for scenarios where the true positives cases are not of interest and <u>*not*</u> having '*the condition*' is of importance. Like sensitivity, specificity is commonly used within medical diagnostic testing (please see Altman & Bland, 1994; Akobeng, 2007; Altman & Bland, 1994(a)).

Sensitivity and Specificity individually apply *true positive rates* and *true negative rates*; respectively, however, Prevalence is the number of positive conditions (i.e. true positive + false negative) divided by the total population (Noordzij, Dekker, Zoccali & Jager, 2010).

Prevalence measures the proportion of positive conditions from the population. Effectively, measuring the frequency of positive conditions within the population. It is commonly applied in Epidemiology, health care providers, insurers and toxicologists, when measuring the proportion of the population affected by a particular medical condition, such as a disease or a risk factor, i.e. smoking or obesity (please see Ellickson, Bird, Orlando, Klein & McCaffrey, 2003; Linder, Rigotti, Brawarsky, Kontos, Park, Klinger, Marinacci, Li, Haas, 2013; O'Neil, 2015; Bolton-Smith, Woodward, Tunstall-Pedoe & Morrison, 2000; Chen, Rennie, Cormier & Dosman, 2005).

Another classification metric, like Prevalence, which applies true positives is Precision, also known as the *positive predictive value*. Precision is the number of true positives divided by the number of predicted condition positive (i.e. *true positives* + *false positives* = *total predicted positives*). "Precision expresses the proportion of the data points that the classifier says were actually relevant" (Koehrsen, 2018, p.1). The precision represents the model bias evaluating the model's tendency to output positive classes. Therefore, the precision and recall measures should be used when the correct identification of negative cases is unnecessary. Precision is commonly applied in pattern recognition (please see Lanz, Marti & Thormann, 2003; Pettersson, 2005; Brodersen, Ong, Stephan & Buhmann, 2010), information retrieval (Carlberger, Dalianis, Duneld & Knutsson; Cormack & Lynam, 2006; Holmes & McCabe, 2002) and machine learning classification problems (Loh, 2009; Landgrebe, 2000 & Cano, Herrera & Lozano, 2007).

The disadvantage of the precision and recall measure is that neither capture information surrounding the model's ability to handle negative cases (Davis & Goadrich, 2006, p. 234). Recall relates to positive conditions (i.e. true positives and false negatives), while precision relates to predicted positive conditions (i.e. true positives and false positives). These metrics can produce misleading results when all observations are classified as positives.

Averaging Precision can be viewed as finding the area under the precision-recall graph (Su, Yuan & Zhu, 2015). "Average precision is a measure that combines recall and precision for ranked retrieval results. Specifically, the average precision is the mean of the precision score after each relevant document is retrieved" (Su, Yuan & Zhu, 2015, p. 350). Like the precision measure, the average precision score is commonly applied in information and image retrieval and document analysis (i.e. topic modelling and sentiment analysis). It is useful when comparing how well different models are ordering or ranking predictions.

An alternative to the average precision score is the F_1 score which combines the recall and precision using a harmonic mean. The F_1 measure, also known as the balanced *F*-score, is a single value metric based on two parameters (recall and precision). The *F*-score combines the recall and precision metric using the harmonic mean. "F_1 score is needed when you want to seek a balance between precision and recall" (Shung, 2018, p.1). F_1 score is a better measure to identify a balance between Precision and Recall because it accounts for both false positives and false negatives predictions avoiding the possibility of being deceived by very poor precision and very high recall. The F_1 score is optimal when there is perfect precision and recall (i.e. F_1 score = 1) and worst when F_1 score = 0, implying that the F_1 score can not be greater than precision (Shung, 2018). Like precision and recall the F_1 score fails to capture information surrounding the model's ability to handle negative cases.

Like prevalence, the F_1 score is commonly applied in information retrieval for assessing document classification and query classification performance (please see Jiang, 2009; Buttcher, Clarke, Yeung & Soboroff, 2007; Cao, Hu, Shen, Jiang, Sun, Chen & Yang, 2009; Li, Zhong, Xu & Kitsuregawa, 2012; Beitzel, Jensen, Chowdhury, Frieder, 2008) and has widely used in Natural Language Processing literature (please see Collobert, Weston, Bottou, Karlen, Kavukcuoglu & Kuksa, 2011; Maarouf, Bradbury, Baisa & Hanks, 2014; Yao, Zweig, Hwang, Shi & Yu, 2013; Wang, Liu, Afzal, Rastegar-Mojarad, Wang, Shen & Liu, 2018).

Another way to evaluate the effectiveness of a binary classifier is the Receiver Operator Curve (ROC). ROC is a two-dimensional graph in which true positive rate (*Y*), sensitivity, is plotted against the false positive rate (*X*). “It depicts the relative trade-offs between true positives and false positives” (Vuk & Curk, 2006, p. 90).

The area under the curve (AUC) is a value between 0 and 1, where a value of 1 corresponds to a classifier that can perfectly separate observations across the two classes, while an AUC of 0.5 corresponds to a classifier that cannot distinguish between the two classes. This area represents the probability that a randomly selected case will have a higher result than a randomly selected control (Fawcett, 2006). A disadvantage of AUC is that it does not account for prevalence or different misclassifications costs. Further criticisms of the AUC are heavily cited in clinical and medical literature, noting its lack of relevance and un-interpretability of small magnitude changes. This technique is commonly applied in financial sector, specifically credit risk, for assessing risk discrimination (please see Zhou, Lai & Yen, 2009; Joao, 2007; Brown & Mues, 2012; Abdelmoula, 2015).

As discussed, each confusion matrix metric is geared towards measuring the classifiers ability to assesses various aspects of the model’s predictive power. Although, there is no universal method to assess classifier “accuracy”. The choice of confusion matrix metrics is dependent on the modelling objective, what is needed from the overall classifier, the forecasting scenario, and the outcome of interest.

2.4 BEYOND THE CONFUSION MATRIX

Although the confusion matrix outlines many evaluation metrics to assess the accuracy of classification-based models, it does not encompass an exhaustive list of such metrics. Therefore, this section outlines validation metrics that are outside the confusion matrix and are heavily used within industry and academia.

The Log-loss metric is an example of a classification evaluation metric which resides outside the confusion matrix and heavily used in machine learning tasks. Minimising the log-loss is equivalent to maximising accuracy of the classifier, therefore a lower log-loss value implies better predictions. The log-loss heavily penalises classifiers that confidently produce incorrect classification (i.e. producing a high probability for incorrectly classified observations). Log-losses closer to 0 indicate high accuracy, whereas if the log-loss is away from 0 indicates lower accuracy, therefore, the log-loss increases as the predicted probabilities diverge from the observed class. The log-loss is a logarithmic proper scoring rule (please see section 2.12) and is applied in many machine learning problems to identify the optimal solution, specifically binary classification models, such as energy (please see Esser, Appuswamy, Merolla, Arthur & Modha, 2015; Belanger & McCallum, 2016; LeCun, Chopra, Hadsell, Ranzato & Huang, 2006), credit risk (please see De Fontnouvelle, Jesus-Rueff, Jordan & Rosengren, 2003; Bielecki, Cousin, Crepey & Herbertsson, 2014; Chen, 2007; Zhang, 2009) and pattern recognition (please see Almeida, Backovic, Cliché, Lee & Perelstein, 2015; Masood, Ellis, Nagaraja, Tappen, LaViola & Sukthankar, 2011; Cheng, Zhang, Shao & Zhou, 2016; Ding, Chen, Lui & Huang, 2016).

The Brier score is another example of a proper score function, specifically a quadratic function, measuring the accuracy of probabilistic predictions. It is heavily used in medical research and meteorological forecasting to assess and compare the accuracy of binary classifiers (Lix, 2010; Wilks, 2010; Ferro, Richardson & Weigel, 2008). The most common Brier score is:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \qquad (7)$$

In effect this is the mean squared error (MSE) error of the forecast. The Brier score simultaneously addresses calibration, the statistical consistency between the predicted probabilities and the observations, and sharpness, i.e. the concentration of the predictions, (Rufibach, 2010). The more concentrated the predictions, the sharper the forecasts, and the sharper the better, subject to calibration (Gneiting, Balabdaoui & Raftery, 2007). A lower Brier score implies better model calibration and classification accuracy. The Brier score is widely reported in the meteorology literature and survival analysis (please see Ferro, 2013; Jewson, 2004; Young, 2010; Hersbach, 2000; Prasad, Dash & Mohanty, 2010).

A gain and lift chart measures classifier effectiveness through the ratio between the results obtained with and without the model (Jaffery & Liu, 2009). It measures the improvement in results when applying the classifier compared to the results when the classifier is not applied. The chart represents the cumulative percentage ‘correct’ and cumulative population (Figure 2(a) and Figure 2(b)). Figure 2(a) illustrates a gains chart, showing the percentage of the total number of observations, for example, the first observation is at (10%, 30%) implying that when contacting 10% of the customer base 30% of these contact customers will have positive responses, and contacting 50% of the customer based 85% will have a positive responses. The diagonal line represents the baseline curve (i.e. without model), for example if 10% of observations are randomly selected from the ‘scored’ dataset, it is expected that approximately 10% of the observations are classified as a ‘positive response’.

The lift chart can be obtained from the gain chart by identifying the values on the $y - axis$ corresponding to the ratio of the cumulative gain for each curve to the baseline (Vuk & Curk, 2006). Figure 2(b) shows the corresponding lift chart for the gains chart in Figure 2(a).

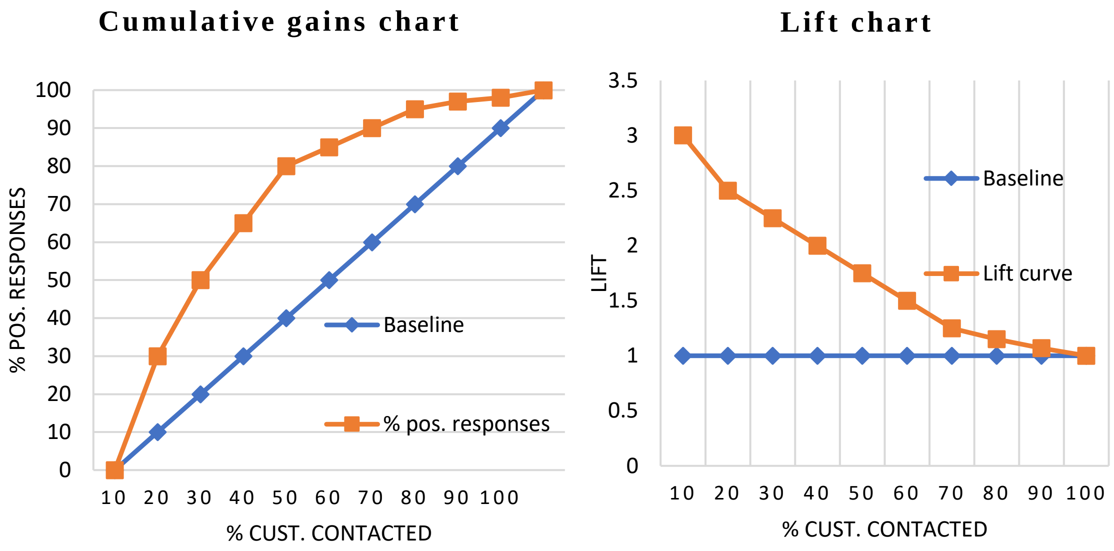


Figure 2(a): Gain chart and Figure 2(b) Lift chart

These techniques are commonly applied in marketing to evaluate the success of marketing campaigns (i.e. customer response before prediction model vs. customer response after prediction model), establish the acquisition of new customers after the release of a new product/s or launch of a marketing strategy (please see Amini, Rezaeenour & Hadavandi, 2015; Rosset, Neumann, Eick, Vatnik & Idan, 2001; Sing'oei & Wang, 2013; Surma & Furmanek, 2010, Kim, 2009; Vuk & Curk, 2006).

The Kolmogorov-Smirnov (KS) is another non-parametric goodness-of-fit test which measures whether a sample is drawn from a population with some known distribution and that two populations have the same distribution (Dodge, 2008). Like AUC, this technique is heavily used in financial sector for assessing risk discrimination within the credit risk environment. The KS statistic quantifies the distance between the empirical distribution of the sample and the cumulative distribution associated with the null hypothesis (i.e. reference distribution) or quantifies the distance between the empirical distribution functions of two samples (Lopes, 2011).

Calibration measures how close the predicted probabilities are to the actual probabilities. A model is "well-calibrated" if the probabilistic effectively reflect the true likelihood for the event of interest. For linear regression models, a calibration plot is a simple scatter plot, while for binary outcomes smoothing techniques such as loess can be used to estimate the observed probabilities for the outcome in relation to the predicted probabilities. Calibration is a common characteristic required by many probabilistic model outputs as it measures the statistical consistency between the predictive distribution and the observations (Gneiting, Balabdaoui & Raftery, 2007).

Proposed by Youden (1950) the Informedness statistic combines both sensitivity and specificity into a single measure. The measure is a value between 0 and 1, where 1 represents a perfect test and 0 represents an imperfect test. It measures how consistently the predictor predicts the outcome of interest by combining surface measures about what proportion of outcomes are correctly predicted (Powers, 2007). Effectively, the Informedness metric measures the likelihood of observations (i.e. subjects) with a given condition to test positive against those observations (or subjects) without the condition to test positive. It is commonly applied in Clinical literature (please see Smits, 2010; Nonhoff, Rottiers & Struelens, 2005; Li, Shen, Yin, Peng, Chen, 2013, Youden, 1950; Ruopp, Perkins, Whitcomb & Schisterman, 2008).

2.4.1 Model Selection

This section provides a brief overview of four commonly used model selection and evaluation techniques: Akaike Information Criteria (AIC), Coefficient of Determination (R^2), Persons Correlation Coefficient, Hosmer-Lemeshow test, Stepwise regression, and Cross-validation. Model selection is an important phase of developing the ratings framework as under-fitting a model leads to insufficient information being captured surrounding the true nature of variability in the dependent variable, while an over-fitted model leads to selecting a model that loses generality.

The Akaike Information Criteria (*AIC*) is a model selection and comparison technique. AIC is a penalized likelihood and requires the likelihood to be maximised before it is calculated. The AIC algorithm tests varying combinations of the independent variables regressed on the dependent variables, estimating several candidate regression-models, and selecting the model that produces the lowest AIC value, which represents the optimal model. "The AIC attempts to balance the trade-offs between complexity of a given model and its goodness of fit" (Mohammed, Naugler & Far, 2015).

The coefficient of determination, also known as the R^2, measures the correlation between the dependent variable and the independent variables jointly (Zhang, 2017). It measures the proportion of variation in the dependent variable explained by the independent variables included in the model. Within linear regression models, the R^2 is most used as a measure of goodness-of-fit of the underlying models, as it measures the proportion of variation explained by the model predictors.

Pearson's correlation coefficient measures the strength and direction of the relationship between two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. It has a value between +1 and -1, where 1 is a completely strong positive correlation, 0 is no correlation and -1 is a completely string negative correlation. A disadvantage of the correlation coefficient is that it only measures linear relationships between $\boldsymbol{X}$ and $\boldsymbol{Y}$. A disadvantage of the correlation measure is that it is meaningless when applied to categorical data and leads to inaccurate results when applied to non-linear relationships. A fundamental assumption when using the R^2 measure in linear regression is that all the observations and exploratory variables are correctly observed. However, occasionally, the variables may not be correctly observed and measurement errors creep into the data (Cheng, Shalabh & Garg, 2014).

The Hosmer-Lemeshow test is used measure the *goodness of fit* for logistic regression models. The test statistic assesses whether the observed event rates match expected event rates in subgroups of the model population (Hosmer, Lemeshow & Sturdivant, 2013). A model is considered *well-calibrated* if the observed and expected outcome rates in subgroups are similar.

The Hosmer-Lemeshow goodness of fit test is based on dividing the sample up according to their predicted probabilities (Bartlett, 2014). Effectively, it measures *how well the logistic regression model fits the data*. A small *p-value* associated with the Hosmer-Lemeshow test statistics indicates that the model does not fit the model well. The Hosmer-Lemeshow test is commonly applied in risk prediction models across many disciplines such as actuarial science (please see Mendoza, Rose, Geiger, & Cash, 2016; Klotz, Vesprini, Sethukavalan, Jethava, Zhang, Jain, Yamamota, Mamedov & Loblaw, 2014; Klugman & Parsa, 1999), credit risk (Altman & Sabato, 2008; Soureshjani & Kimiagari, 2012; Sun & Guo, 2015), health care (Homser, Taber & Lemeshow, 2011; Saunders, Krause, Acuna 2012; Rello, Lujan, Gallego, Valles, Belmonte, Fontanals, Diaz, Lisboa, 2010), insurance (Paefgen, Staake & Fleisch, 2014; Kim, 2016 & Kelz, Gimotty, Polsky, Norman & DeMichele, 2004) and marketing (Jensen, 2006; Jensen & Jepsen 2008; DesJardins, 2002).

The stepwise regression technique is a model selection technique which implements a combination of the forward selection and backward elimination variable selection techniques. The forward selection technique is initialized with the null model and inputs predictors to the model with the lowest $p - value$ less than α_{crit} and stops when all $p - values$ are less than α_{crit}. The backward elimination procedure starts with all the predictors in the model, $Y = B_0 + B_1 + \cdots + B_{r-1}X_{r-1} + \epsilon$, and removes the predictors with $p - values$ greater than α_{crit} and then refits the model. This process is continued until all $p - values$ are less than α_{crit}. Stepwise regression is a modification of the forward selection method in that after each step in which a variable is added, all the candidate variables in the model are checked for statistical significance, if significance is achieved the variable is retained in the model and removed otherwise. The final optimal model minimises the Akaike Information Criteria or maximises the adjusted R^2 value. A drawback of the stepwise regression is the stopping criteria which only produces a single model while there may be a variety of models with a similar goodness-of-fit (Seber & Lee, 2012). An additional drawback is that the forward selection method only selects independent variables that maximises the partial correlation coefficient with the dependent variable (Bendel & Afifi, 1977).

Cross-validation evaluates a model's ability to predict data that was unobserved during the model training process to identify over-fitting or selection bias issues and is the most widely used method for estimating prediction error. Therefore, the objective of cross-validation is assessing a model by measuring how well it will generalise to an independent data set.

The basic idea behind cross-validation partitions sample data into *k* complementary subsets. These complementary subsets, *k*, are used to derive model estimates, while the remaining *k* parts are used to identify the quality of the model estimates. However, model results greatly depend on the random splitting procedure, for example if the training set does not reflect the structure of the original data, then the developed model will not generalise well to the validation set.

There are two types of cross-validation: *exhaustive and non-exhaustive*. Exhaustive cross validation methods train and test on all possible partitions of the original data into a training and validation set, while non-exhaustive methods do not train and test on all possible partitions of the original data.

Cross-validation has become an attractive modelling approach and has several advantages: 1) *uses all data* – predictions can be made on all data using k different models, 2) *availability of more metrics* – building k different models allow k different evaluations on the test set and reduces bias, 3) *dependent/ grouped data* – by performing random train-test split on the data, it is assumed that the examples are independent, 4) *parameter tuning* – training and testing k different models allows the modeller to identify the best or optimal model parameters, and 5) *model stacking* – meta-ensemble to combine information from k different predictive models.

The most used cross-validation techniques will be discussed briefly: 1) *leave-p-out*, 2) *leave-one-out*, 3) *k-fold*, 4) *holdout method* and 5) *Repeated random sub-sampling validation*.
An example of exhaustive cross-validation is *leave-p-out* cross validation. *Leave-p-out* cross validation $N - p$ subsamples are used as the training set, while the remaining *p* subsamples are used as the validation set. This process is repeated for all combinations in which the original sample can be separated into a validation set of *p* observations and a training set of $N - p$. This process can be computationally intensive and time consuming as a model is trained and tested *N* times.

Leave-one-out cross validation (LOOCV) is a case of *leave-p-out* where $p = 1$. In LOOCV the data is partitioned into $k = N$ equal size subsamples; therefore, LOOCV is *k*-fold cross validation. At each step, a model is fitted to *N-1* subsamples used as the training set and the remaining subsample (i.e. one observation) is used as a validation case to calculate the prediction error of the fitted model. This cross-validation process is repeated N times (i.e.

number of observations in the data) with each observation being used once as the validation set. The N results are averaged or combined to a produce a single observation.

An example of exhaustive cross-validation is *hold-out* cross validation. In *Holdout* cross validation the data is randomly partitioned into two subsamples: 1) training set, and 2) testing set. First, a model is built using the training set to estimate a model, second, the model estimates are assessed against the validation set to see how well the model generalises to an unseen data source. Holdout cross-validation process can be costly for small data sources as it excludes a proportion of the data for training and testing.

In k-fold cross validation the data is randomly partitioned into k equal size subsamples. At each step, $k-1$ subsamples are used as the training set, while the remaining k subsample is used as a validation set. The prediction error of the fitted model is calculated when predicting the k^{th} part. This cross-validation process is repeated k times, with each subsample being used once as the validation data. The k results are either averaged or combined to a produce a single estimation. "The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once" (Hastie, Tibshirani & Friedman, 2009, p. 258). As k increases, the difference in size between the training set and the validation sets decreases, and as this difference decreases so does the bias of the technique. Therefore, there is a bias-variance trade-off associated with the choice of k

In *repeated random sub-sampling* cross validation, sampling is run over k iterations, with each iterations procedure randomly selecting a fixed number of samples, S, without replacement. At each iteration, S, samples are used as the training set, while the remaining n-S samples are used as the validation set. This cross-validation procedure is repeated k times.

2.4.2 Scoring Rules

A scoring rule can be any function of the predictions and the observations, which measures the informativeness or valuableness of specific probability predictions by providing a score, to each probabilistic prediction, based on the prediction and on the occurrence event. Measure of the quality of a probabilistic forecast. The use of a proper scoring rule encourages the forecaster to be honest to maximise the expected reward. A score can be thought of as either a measure of the "calibration" of a set of probabilistic predictions, or as a "cost function" or "loss function". Scoring rules were developed to provide the modelling system or the forecaster with an incentive to honestly report probabilistic predictions and encourage the assessor to reveal their 'true' beliefs.

In an ex-ante (i.e. results based on forecasts) sense, strictly proper scoring rules provide an incentive for careful and honest forecasting by the modeller i.e. well-calibrated, while in the ex-post (i.e. results based on actual outcomes) they reward accurate forecasts, i.e. *sharp* probabilities, and penalise inferior forecasts (Winkler, 1996). "The ex-ante incentive aspect of scoring rules, however, suggests that we should restrict our attention to strictly proper scoring rules, for which the assessor can maximise his or her expected score only by reporting probabilities honestly" (Winkler, 1996, p.3).

An example of probabilistic forecasting is in meteorology where a weather forecaster may give the probability of rain on the next day. One could note the number of times that a 25% probability was quoted, over a long period, and compares this with the actual proportion of times that rain fell. If the actual percentage was substantially different from the stated probability, we say that the forecaster is poorly calibrated. A poorly calibrated forecaster might be encouraged to do better by a bonus system. A bonus system designed around a proper scoring rule will incentivise the forecaster to report probabilities equal to their personal beliefs. In addition to the simple case of a binary decision, such as "rain" or "no-rain", scoring rules may be used for multiple classes, such as 'rain', 'snow' or 'clear'.

The modeller desires to maximise the expected score from a strictly proper scoring rules, which requires well calibrated and sharp probabilities. Calibration assesses how well model predictions align with observed probabilities. It is crucial when developing insightful predictive models, especially for decision-making. A commonly used technique for calibration is Hosmer-Lemeshow test statistic which assesses a model's goodness-of-fit by comparing observed probabilities against predicted probabilities at quantiles of predicted probabilities. Predicted probabilities that match the expected probability distribution for each class or quantiles are referred to as calibrated. For the model to be well-calibrated, the probabilities must effectively reflect the true likelihood of the event of interest. Calibrated probabilities can result in an improved calibration on a reliability diagram, plotting the relative frequency of observed probabilities vs relative frequency of predicted probabilities. Better calibrated probabilities may or may not lead to better class-based or probability-based predictions. This depends on the specific metrics used to evaluate predictions.

The scoring rule motivates the modeller to report well calibrated and sharp probabilities, as there is a desire to maximise the expected score from a strictly proper scoring rule. Many have noted that ensemble-based forecasts are generally assessed on two statistical criteria: calibration and sharpness, also known as reliability and resolution, respectively (please see Gneiting, Stanberry, Grimit & Held, 2008; Gneiting, Balabdaoui & Raftery, 2009; Gneiting, Raftery, Westveld and Goldman, 2005; Wilks, 2018; Feldmann, 2012; Hudson, 2017). Given scoring rules promote the reporting of well-calibrated and sharp probabilities, the scoring rule methodology lends itself well to the construction of a novel performance metric to quantify the effectiveness of sport-based rating systems. "Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialize. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only (Gneiting, Balabdaoui & Raftery, 2007, p.2). Sharp predictions refer to prediction concentration and can be measured using entropy and variance.

Savage (1971) stated that scoring rules induce an assessor to reveal their true opinion, expressed through probabilities, associated with events or, more generally, their personal expectations of random quantities. Winkler (1996) made a similar statement, stating that situations arise where experts maybe self-interested and, consequently, are not necessarily honest when reporting their beliefs. For example, experts with a reputation to protect might report forecasts near the most likely group consensus, whereas experts who have a reputation to build might overstate the probabilities of outcomes they believe will be understated in a possible consensus. Therefore, it is necessary to distinguish between an expert's belief $\boldsymbol{p},$ and the expert's reported forecast $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$. It is desirable that an expert's forecast equals the expert's belief $\boldsymbol{p} = \boldsymbol{r}$, and that the expert's forecast is accurate.

The elicitation of "true" beliefs and probabilities can be observed when gambling. The gambler is incentivised to place bets based on their belief on event outcome and is rewarded with cash if the subjective reported probability (r) equals the actual probability (p). The situation where $r = p$ is known as perfect calibration. The gambler is penalized by sustaining financial losses, if $r \neq p$ or if the reported probability (r) is "far-away" from the actual probability (p). This example illustrates how scoring rules can be used to capture an assessor's or a model's "true" opinion and how they can be used to self-evaluate "vagueness" or "uncertainty". Vagueness or uncertainty is a major obstacle when forecasting, however after the initial predictions, the predictions that follow are modified when reflecting upon the implications of the initial predictions (Maskey, 2004). Effectively, calibrating and sharpening the probabilities considering new information.

Scoring rules are commonly applied in finance (please see Figini & Maggi, 2014; Offerman, Sonnemans, Van de Kuilen & Wakker, 2009; Hanson, 2012; Tsaih, Liu, Liu & Lien, 2004; Henley, 1995), meteorology (please see Winkler & Murphy, 1970; Ferro, 2013; Murphy, 1973; Stanski, Wilson & Burrows, 1989) and pattern recognition (please see Malley, Kruppa, Dasgupta, Malley & Ziegler, 2012; Lakshminarayanan, Pritzel & Blundell, 2017; Nock, Ali, D'Ambrosio, Nielsen & Barlaud, 2014; Kauppi, Kamarainen, Lensu, Kalesnykiene, Sorri, Kalviainen, Uusitalo & Pietila, 2009).

2.4.2.1 Strictly Proper Scoring Rules

A scoring rule, *S*, is strictly proper if:

$$E_p[S(p)] > E_r[S(r)] \qquad for\, r \neq p \tag{8}$$

Here, *p* denotes the modeller's subjective probability that A will occur, and *r* denotes the actual objective probability of A. It is evident from (8) that an assessor maximises their expected score when the reported forecast, r, equals the assessors' true beliefs (i.e. $\boldsymbol{r} = \boldsymbol{p}$; perfect calibration). Therefore, strictly proper scoring rules incentivise the assessor for careful probabilistic assessment and for gathering information to 'improve' probabilities by recalibrating the probabilities by using more informative attributes.

The difference between the user's subjective and model's objective probabilities allows the modeller to 1) effectively calibrate the predictions and 2) build in any subjective or expertise knowledge that may be missed by the model. All strictly proper scoring rules encourage honest and careful assessment ex-ante and reward calibration and sharpness ex-post (Wrinkler, 1996).

Building in such subjective, domain expertise and individual interpretation to the model are important as the model is more likely to produce outcomes that are reflective of reality or nature. Therefore, to maximise expected score, the modeller should set $r = p$ (i.e. the modeller should be honest).

A scoring rule, S, gives the modeller a score of $S_1(r)$ if A occurs and $S_2(r)$ if A does not occur. The expected score is: $E_p[S(r)] = pS_1(r) + (1-p)S_2(r)$. Alternatively, the choice to report $\boldsymbol{R}$ can be analysed in terms of expected loss: $L(r,p) = E_p[S(p)] - E_p[S(r)] = \sum p_i S_i(p) - \sum p_i S_i(r)$.

Three frequently used scoring rules for dichotomous response variables s are the quadratic, logarithmic and spherical rules.

Quadratic	**Logarithmic**	**Spherical**
$S_1(r) = -(1-r)^2$	$S_1(r) = \log r$	$S_1(r) = \dfrac{r}{[r^2 + (1-r)^2]^{\frac{1}{2}}}$
$S_2(r) = -r^2$	$S_2(r) = \log(1-r)$	$S_2(r) = \dfrac{1-r}{[r^2 + (1-r)^2]^{\frac{1}{2}}}$

In the presence of an event with more than a dichotomous outcome, the Quadratic, Logarithmic and Spherical scoring rules take the following form:

Quadratic	**Logarithmic**	**Spherical**
$S_j(r) = 2r_j - \sum_{i \in I} r_i^2$	$S_j(r) = \log r_j$	$S_j(r) = \dfrac{r_j}{\left(\sum_{i \in I} r_i^2\right)^{\frac{1}{2}}}$

In the presence of continuous random variables, the Quadratic, Logarithmic and Spherical scoring rules take the following form:

Quadratic	Logarithmic	Spherical
$2r(x) - \int_{-\infty}^{\infty} r^2(x)\, dx$	$S_x(r) = \log r(x)$	$S_x(r) = \dfrac{r(x)}{\left(\int_{-\infty}^{\infty} r^2(x) dx\right)^{1/2}}$

2.5 KEY LITERATURE REVIEW OUTCOMES

Reviewing the sports ratings and credit risk literature several important elements to construct rating systems were revealed. It was found that many rating systems apply ensemble modelling strategies, such as model stacking, and use feature selection techniques and dimension reduction mechanisms. These approaches combined trait or feature-based ratings, and extract as much information as possible to reduce ratings uncertainty, respectively.

Applying these statistical techniques and modelling methodologies a set of rating systems were developed. These systems have been peer-reviewed and published in numerous academic journals and conference proceedings. A list of the most notable rating systems, constructed because of this literature review, are provided below and in Appendix A.

Moreover, as a result of the literature review this chapter has identified several gaps in the academic literature and outlined a set of key research objectives. An additional consequence is the development of sport rating systems which have been published in peer-reviewed academic journals and presented at peer-reviewed academic conferences. These systems were developed using some of the key methodologies identified in the literature, which enabled the identification of key elements and limitations that must be considered when developing the ratings framework.

As mentioned, based on this literature review, the peer-reviewed conference proceedings, journal publications and limitations within the ratings literature, this thesis formulates three potent, yet achievable, research objectives which form the basis of this research. (i) Develop a quantitative framework to construct sport-based ratings systems that output meaningful ratings. (ii) Develop a novel evaluation metric to quantify the effectiveness of meaningful sport-based ratings. (iii) Demonstrate the applicability of the developed ratings framework and novel evaluation metric within the sporting context.

2.5.1 Limitations of Rating Systems

During the literature review several rating systems were developed using modelling techniques identified throughout the literature (please see section 2.5.4 and section 2.5.5). During this development process, gaps within the literature, and the limitations and communalities of these systems were identified. These are the limitations of the current knowledge base[4]:

[4] The limitations within the credit-risk and sport-based ratings literature led to the research objectives outlined in Chapter One.

- *Lack of a sport-based ratings framework*– given the prevalence of sport-based ratings within the commercial and academic environment no modelling framework or approach currently exists in the literature to construct meaningful sport-based rating systems.
- *Lack of meaningful rating systems* – the literature echoed the sentiment expressed by Bracewell (2003); ratings are an elegant form of dimension reduction. Throughout this chapter it was shown that variable selection and dimension reduction are crucial elements of ratings methodologies. Although, given the loss of information during dimension reduction and the application of "black box" modelling techniques to produce ratings, the resultant ratings lack transparency and intuition, implying that results cannot be mapped to real-world observable outcomes.
- *No evaluation metric to assess the effectiveness of meaningful sport-based ratings* – to evaluate the predictive accuracy of the developed rating systems commonly applied evaluation metrics such as log-loss, root mean square error (RMSE) and mean absolute error (MAE), were used. Although, given the uniqueness of sport-based rating systems, it is necessary to construct a specific performance metric which quantifies the effectiveness of meaningful sport-based ratings.

2.5.2 Ensembling Forecasting

Throughout the literature review [of rating systems] it was found that many credit-risk and sport-based systems apply ensemble forecasting strategies, such as model stacking, to produce highly predictive and reliable outputs.

Formally, "an ensemble consists of a collection of two or more forecasts that try to realise the possible uncertainties in a numerical forecast (Cheung, 2001, p. 315). Birthed in meteorology, ensemble forecasting strategies are prevalent amongst meteorologist as they allow the use of many models with varying initial atmospheric conditions and model uncertainties to understand the range of possibilities of future weather to evaluate the most likely outcomes. Much of the underlying methodology of ensemble forecasting and ensemble forecasting strategies have been developed by atmospheric scientists. These strategies fall into one of two categories or a combination: 1) Ensembles based on many different models and 2) Ensemble based on many runs of one-computer model initialised from slightly different data (Kunst & Jumah, 2004).

Ensemble forecasting is an appealing modelling approach because instead of choosing a single method, a collection of the most appropriate methods is selected to improve accuracy, assuming each method statistically and practically contributes to the modelling objectives. Ensembling forecasts are advantageous 1) when the modeller is uncertain as to which method is best or which variables or combination of variables are best, and 2) when it is assumed that each modelling method has some validity, but no single method provides perfect forecasts.

Armstrong (2001) stated that an ensemble forecasting strategy produce results whose probability law of error rapidly decreases. “Combining forecasts improves accuracy to the extent that the component forecasts contain useful and independent information” (Armstrong, 2001, p.1). There are two ways to generate independent forecasts: 1) analysing different data sources and dimensions within the data and 2) applying different forecasting methods. The greater the difference between the data dimensions and modelling methods, the greater the expected improvement in the accuracy over the average of the individual forecasts. Compared with single forecasts, ensembled forecasts provide more complete information on the possible future states of the modelling system, because they allow forecasters to estimate, in an objective and reliable way, the range of possible future scenarios. Moreover, a competent ensemble algorithm is a key requirement for a successful ensemble forecasting strategy.

In recent years ensembling forecasting strategies have seen a surge in implementation and applicability within different fields such as weather forecasting (Wu & Lin; 2017; Yu, Nakakita & Jung, 2016; Dahl, Brun, Kirsebom & Andresen, 2018; Ye, Deng, Ma, Duan, Zhou & Du, 2019; Craparo, Karatas & Singham, 2017), industrial economics (Qin, Xie, He, Li, Chu, Wei & Wu, 2019; Sun, Wang & Wei, 2018; Hao & Tian, 2019), credit risk scorecards (Bao, Lianju & Yue, 2019; Papouskova & Hajek, 2019; Yu, Wang & Lai, 2019), sports analysis (Lessmann, Sung, Johnson & Ma, 2012; Baboota & Kaur, 2019; Cai, Yu, Wu, Du & Zhou, 2019; Gjoreski, Kaluza, Gams, Milic & Lustrek, 2015), and energy generation (Tang, Wu & Yu, 2018; Zhou & Chen, 2009; Kim & Hun, 2018; Ziliani, Ghostine, Ait-El-Fquih, McCabe & Hoteit, 2019).

Peimankar, Weddell, Jalal & Lapthorn (2018) presented an ensemble time series forecasting algorithm using multi-objective optimisation to predict dissolved gas contents of power transformers. The results showed that the proposed algorithm produced higher accuracy and reliable forecasts for one day, two day and three-day forecasts compared with other statistical methods.

Sun, Wang & Wei (2019) designed a multiscale decomposition ensemble approach for forecasting exchange rates, which was found to produce promising forecasting foreign exchange rates. Allison, Crocker, Tran & Carrieres (2014) developed an ensemble forecasting model adopting a Monte Carlo approach to embedded random variability to the model parameters to predict the drift of icebergs. Although the proposed method did not improve the skill of the forecast, relative to non-ensemble forecasts methods, it was shown to be consistent and the statistical properties of the model provide useful information on the uncertainty inherent in the forecasts.

Liu, Zhan & Bai (2019) investigated the effects of PV solar power variability and proposes a data-driven recursive ensembling modelling technique, implementing SVM (support vector machines), MLP (multilayer perceptron) and MARS (multivariate adaptive regression splines) methods, to improve the predictive accuracy of PV power generation. In general, the ensembling strategy demonstrated higher accuracy compared to a stand-alone forecasting model.

Liu, Xu & Chen (2019) developed a multi-step stacking ensemble forecasting method with Empirical wavelet transform for urban fine particle concentration. The proposed model was shown to have better accuracy and wider applicability compared to existing models.

Furthermore, ensemble forecasting strategies are becoming increasingly prevalent in the credit risk environment (for more examples please see Bao, Lianju & Yue, 2019; Papouskova & Hajek, 2019; Ala & Abbod, 2016a; Ala & Abbod, 2016b; Beque & Lessmann, 2017; Dahiya, 2017; Wang, Hao, Ma & Jiang, 2011; Noghani & Moattar, 2015; Bouaguel & Limam, 2015, Zhu, Zhou, Xie, Wang & Nguyen, 2019; Abellan & Castellano, 2019; Yu, Wang & Lai, 2019; Thomson, Pollock, Onkal & Gonul, 2019; He, Zhang & Zhang, 2018). In recent years, classification ensemble strategies or multiple classifier systems have been widely applied to credit scoring and have been found to achieve significantly better performance than individual classifiers (Feng, Xiao, Zhong, Qiu & Dong, 2018).

2.5.3 Dimension Reduction and Feature Selection Techniques

Along with ensemble forecasting strategies, many rating systems apply a combination of dimension reduction and feature selection techniques to reduction dimensions and increase model parsimony, respectively, to produce transparent and robust.

Dimension reduction and feature selection techniques are vital elements in rating systems due to the issues of multicollinearity, interrelationships between various attributes-types and the high dimensionality of datasets. Therefore, dimension reduction and feature selection techniques are necessary to handle such issues and identify features that significantly affect human traits and to construct trait-based ratings which significantly account for uncertainty within sporting performances. To ensure that statistically significant and important features are identified two areas of feature selection are considered: (1) classical parametric techniques, such as principal component analysis, linear discriminant analysis, stepwise regression and hierarchical variable clustering, and (2) non-parametric techniques, such as regression trees, random forests and gradient boosted machines. While a preliminary regression analysis can produce an assessment of variable significance by evaluating statistical significance and effect size, such analyses can generate unreliable and inaccurate results, due to the inability to handle multicollinearity and interaction effects. Additionally, given the multitude of features and the research require to produce meaningful sport-based ratings, an accurate means of assessing feature significance was paramount to research success.

Common variable assessment strategies in credit scoring include automated logistic regression variable selection (e.g. stepwise, forward and backward), Factor Analysis, Principal Component Analysis (PCA), and field-specific statistics such as ‘Weight of Evidence’ and

‘Information Value’ (Lin, 2013). Common feature selection and dimension reduction strategies in sports ratings include a combination of traditional and machine learning techniques, such as random forest, linear regression, stepwise regression, PCA and regression trees (Patel, 2016).

Feature selection is a process whereby a heuristic or algorithm identifies the variables that best accomplish a given modelling objective (e.g. explanatory value, prediction accuracy). The current apex of literature is ‘feature selection’ in genomics. Gene expression microarray data sets range from 20,000 to 60,000 variables (Dziuda, 2010, p.100; Guyon & Elisseeff, 2003, p.1158), often numeric, coded (i.e. not intuitively interpretable), and with only cursory theoretical knowledge available to guide selection.

The three main strategies in feature selection are ‘wrappers’, ‘filters’ and ‘embedded methods’ (Guyon & Elisseeff, 2003). With wrappers, the feature importance measures of a supervised learning machine, trained to predict a response variable, are used to determine variable selection in subsequent models. Filters, in contrast, assess importance during a ‘pre-processing step’ separate from the response variable, while embedded methods are automated and self-contained within the model (e.g. stepwise selection in regression).

Due to the multicollinearity and high dimensionality of the data, various selection and reduction techniques should be applied within the ratings framework to minimise the presence of such effects and reduce the number of features that are implemented when rating sporting performance.

2.5.4 Evaluating Meaningful Sport-Based Ratings

Given meaningful sport-based rating systems must be reliable and intuitive (along with transparent and robust), it is necessary to apply an evaluation metric which is able to capture expert knowledge and account for sporting context. The ideal metric should reward or favourably weight ratings based on honest reporting and the system’s ability to output ratings which align with a tolerable level of intuition. This notion of intuition is important to sport-based rating systems as the outputs derived from such systems must incorporate an element of intuition. Therefore, the ideal performance metric when evaluating sport-based rating systems should elicit an adequate level of information from experts to make informed, objective and intuitive judgements, and ensure ratings encompass a tolerable level of intuition; therefore the results can be mapped to real-world outcomes.

Therefore, it is proposed that a novel performance metric be developed which meets the following criteria: 1) sensitivity to distance, 2) sensitivity to time-dependence, 3) evaluate ratings on the entire probability distribution (i.e. non-local metric), 4) incentivisation for well-calibrated and sharp ratings, and 5) adjusts incentives based on forecasting difficulty. The justification for each of these criteria are provided in Chapter Three (section 3.7).

A metric that accounts for these five criteria can determine the effectiveness of meaningful sport-based rating systems (i.e. reliable, robust, transparent, and intuitive). Given these criteria, and the limitations of current evaluation metrics, and that ensembled ratings are generally assessed on calibration and sharpness (Gneiting, Raftery, Westveld & Goldman, 2005), a proper scoring rule methodology will be applied to develop the novel performance metric to evaluate the effectiveness of meaningful sport-based rating systems (research objective (iii)).

2.5.5 Sport-Based Rating Systems

To derive a deeper understanding of the requirements for meaningful sport rating systems, the author has undertaken significant, novel research and meaningfully contributed to the body of knowledge. These findings have been peer-reviewed and published as outlined below:

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2017). An Individual-Based Team Rating Method for T20 Cricket. *Journal of Sports and Human Performance 5(1): 1-17.*

Patel, A. K., Bracewell, P. J., & Wells, J. D. (2017, June 23). Real-time measurement of individual influence in T20 cricket. Paper Presented at The Proceedings of the 17th MathSport International 2017 Conference Proceedings. (pp. 61-70). Padua, Italy. ISBN: 978-88-6938-058-7.

Patel, A. K., Bracewell, P. J., Gazley, A. J., Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. *Journal of Sports Science and Coaching 12(3): 328-338.*

Brown, P., Patel, A. K. & Bracewell, P. J. (2017, June 23). Optimising a Batting Order in Limited Overs Cricket using Survival Analysis. Paper Presented at The Proceedings of the 17th MathSport International 2017 Conference Proceedings. (pp. 71-80). Padua, Italy. ISBN: 978-88-6938-058-7.

Simmonds, P., Patel, A. K., & Bracewell, P. J. (2018). Using network analysis to determine optimal batting partnership in T20 cricket. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

McIvor, J. T, Patel, A. K., Hilder, T.A., & Bracewell, P. J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Rooney. S. J., Bracewell, P. J., & Wells. J. D. (2018). Constructing a predictive PGA performance rating using hierarchical variable clustering. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

2.5.5 Credit risk rating systems

Patel, Bracewell, Gazley & Bracewell (2017) applied methodologies and techniques heavily used within the credit risk environment to calculate the probability of bowler success and rank-order each bowler in terms of international success. Specifically, residual logistic regression, a two-stage regression method which is commonly used in credit risk, is applied to estimate an applicant's probability of default (please see Baez-Revueltas, 2009, for more details). Given the confidential and commercially sensitive nature of credit risk scorecards the credit risk system developed because of this research is not disclose (embargoed research).

Patel, A. K., Bracewell, P. J., Gazley, A. J., Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. *Journal of Sports Science and Coaching 12(3): 328-338.*

Patel, A. K., Bracewell, P. J., & Coomes, M. (2020). Inferring bowling strike rate in T20 cricket. *Journal of Sports Analytics (under review).*

These rating systems reinforced Bracewell's (2003) definition of ratings being an elegant form of dimension reduction. To adequately measure performance using a single numerical value, which provides an interpretation of performance, the key dimensions affecting the performance must be sufficiently accounted for in the rating system. Further, to predict this sporting performances appropriate modelling techniques must be applied to derive these measures of performance. Various modelling techniques must be used, and the ratings associated with different traits must be ensembled to produce a meaningful rating. Given these findings, Chapter Three develops a ratings framework for constructing sport-based rating systems that produce robust, reliable, transparent, and intuitive ratings, also known as meaningful ratings, of performance. Further, Chapter Three develops a novel performance metric to assess the effectiveness of meaningful rating systems.

2.6 DISCUSSION AND CONCLUSION

This chapter provided a comprehensive review of statistical methodologies and techniques commonly applied when constructing rating systems within the sporting (team and player-based) and credit risk (application and behaviour-based) environments. Based on this literature review, several modelling techniques were applied to develop rating systems across these three domains. Throughout this development process limitations were identified, and three research objectives were formulated. Addressing these limitations and answering the research questions will plug the gap in the literature and provide a novel solution for constructing sport-based rating systems.

The literature review revealed that many rating systems apply an ensemble strategy and use machine learning techniques as feature selection and dimension reduction mechanisms. These two approaches combine multiple feature specific ratings or trait-based ratings to extract as much information as possible and reduce uncertainty. Consequently, it was identified that the most suitable and applicable method for use within the ratings environment, for the problems defined in this thesis, is a multi-objective ensemble forecasting strategy.

Applying commonly used statistical methodologies and techniques revealed key communalities when constructing rating systems. These communalities were the application of 1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling procedure to produce an overall rating. The following chapter develops a ratings framework which implements these five communalities to construct sport-based ratings that output meaningful ratings of performance.

Through the rating systems literature review, it has become apparent that there exists no unique performance metric when assessing the effectiveness of sport-based ratings. This is because commonly applied performance metrics are plagued with limitations and are not suitable when assessing sport-based ratings.

Chapter Two also provided a comprehensive review of commonly applied model evaluation metrics to assess the predictive accuracy of regression and classification-based models. In this chapter it become apparent that there exists no universal model evaluation technique to measure the validity of sport-based ratings that output meaningful ratings of performance. Given this review and the limitations of the performance metrics that were identified, a set of criteria are identified to construct a performance metric to quantify the effectiveness of sport-based ratings. These criteria include: 1) sensitivity to distance, 2) sensitivity to time-dependence, 3) evaluates the ratings on the entire probability of distribution, 4) provides an incentive for calibration and sharp ratings and 5) adjusts incentives based on forecasting difficulty. Moreover, the proper scoring rule methodology is identified as the most appropriate method to construct this evaluation metric as it allows the incorporation of these five criteria.

The literature review revealed that ensemble forecasts are generally assessed through two key statistics: reliability and resolution (i.e. calibration and sharpness, respectively). The reliability, or calibration, of a forecast indicates how confident the assessor can be in their predictions and can be evaluated by comparing the standard deviation of the error in the ensemble mean with the forecast spread (Gneiting, Balabdaoui & Raftery, 2007). The resolution, or sharpness, of a forecast indicates how much the forecasts deviates from the climatological event frequency, given that the ensemble is reliable, increasing this deviation will increase the usefulness of the forecast.

Given these ideal criteria and the need for calibration and sharpness to assess ensembled ratings, the proper scoring rules methodology is identified as the most suitable methodology to construct an evaluation metric that quantifies the effectiveness of sport-based ratings.

The following chapter provides a technical overview of scoring rules, constructs a novel performance metric quantifies the effectiveness of meaningful sport-based ratings and describes how the constructed metric accounts for the five criteria mentioned above. Therefore, Chapter Three answers research objectives (i) Develop a quantitative framework to construct sport-based ratings systems that output meaningful ratings; and (ii) Develop a novel evaluation metric to quantify the effectiveness of meaningful sport-based ratings.

Chapter Three

A NOVEL RATINGS FRAMEWORK AND EVALUATION METRIC

"This black box problem of artificial intelligence is not new, and its relevance has grown with modern, more powerful machine learning solutions and more sophisticated models. Meanwhile, models can outperform humans in complex tasks like the classification of images, transcription of speech, or translations from one language to another. And the more sophisticated the model, the lower its explainability level".

Dr. Markius Noja, The Digitialst.

On bringing transparency into AI.

3.0 OVERVIEW

Given the lack of a ratings framework for constructing rating systems, the first half of this chapter develops a novel framework for constructing sport-based rating systems which produce intuitive, robust, reliable, and transparent outputs, also known as meaningful performance ratings. Specifically, the framework is developed to construct rating systems within sporting industry to evaluate team and player performances. The framework adopts a multi-objective ensembling strategy and implements five key communalities present within many rating methodologies. These communalities are the application of 1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling procedure to produce an overall rating. Key elements of the ratings framework are the application of feature engineering, feature selection and dimension reduction techniques; therefore, the ratings problem resides in the field of information theory.

An ensemble approach is applied because it is assumed that performance is a function of the individual traits significantly affecting overall performance. Therefore, performance is defined as $performance = f(trait_1, \dots, trait_n)$. Moreover, the ratings framework is a form of model stacking where information from multiple models is combined to generate a more informative model. A key part of the proposed framework is feature selection and dimension reduction; an ideology held throughout credit scoring literature. The development of this ratings framework to construct sport-based rating systems addresses research objective (i).

Applied to sport-based ratings, current model evaluation metrics such as RMSE, MAE and SMAPE are limited as they do not evaluate forecasting difficulty, capture the introduction of bias and cannot assess the effectiveness of meaningful sport-based rating systems. Therefore, the second half of this chapter develops an evaluation metric which quantifies the effectiveness of meaningful ratings systems applicable within the sporting context. The proposed metric resolves these issues and quantifies elements of the human decision-making process by 1) evaluating the distance between reported ratings, actual outcomes and averaged forecasts, 2) measuring the distance between ratings across different time-frame, 3) providing an incentive for well-calibrated and sharp ratings, 4) accounting for the context and the difficulty of the forecasting scenario and 5) evaluating ratings on the entire probability distribution. A proper scoring rule is the underlying methodology, specifically distance and magnitude-based measures derived through the spherical scoring are applied. The development of the novel performance metric addresses research objective (ii).

3.1 BACKGROUND

The application of analytics in the business environment has experienced tremendous growth (Analytics, 2016). Business analytics has transformed from a “nice-to-have” to a competitive advantage. “In the past few years, predictive analytics, has gone from a practice applied in a few niches to a competitive weapon with a rapidly expanding range of uses” (CGI: Predictive Analytics, 2013, p.1).

A key factor for the rise in business analytics is the phenomenon of “big data” and the data science revolution, and its acceptance by senior executives as an important business enabler. The goal of insight and information extraction or revealing hidden patterns within big data is achievable through the application of mathematical and statistical techniques. Importantly, these insights need to be relayed appropriately to the intended audience. Sagiroglu & Sinanc (2013) stated that modern analytics, characterised by improvements in computing power, reduced cost in data storage, greater access to various data sources and cheaper commodity hardware, requires a revolutionary step forward, moving away from traditional data analysis. The Transforming Data with Intelligence survey revealed that the application of advanced analytics creates better aimed marketing, informed decision-making, client-based segmentation, and recognition of sales opportunities. This information offers significant potential to generate business value and competitive advantage.

This growth in demand for analytics and data capture has been experienced across many industries, resulting in considerable academic and commercial attention (e.g. Stefani, 1997; Siddiqi, 2012; Bracewell *et. al.*, 2017, respectively). The consequence is the development of data-driven and modelling intensive applications with an objective of evaluating, rating, and forecasting the performance of an individual or collection of individuals (i.e. team).

Such data-driven models must produce robust, transparent, and contextual results (Bracewell, 2003) to generate trust leading to implementation of the insights. In this chapter, this definition is extended, specifically, ratings are an elegant and excessive form of dimension reduction whereby a value provides a meaningful quantitative interpretation of sporting performance.

It is stated that rating systems must produce meaningful results, which have the following characteristics: 1) Robust – ratings must yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes. 2) Reliable – ratings must be accurate and produce highly informative predictions which are well-calibrated and sharp. 3) Transparent – ratings must be interpretable and easy to communicate. 4) Intuitive – ratings must relate to real-world observable outcomes.

Given the limitations found within the ratings literature, this chapter addresses these limitations by 1) developing a quantitative ratings framework to construct sport-based rating systems that output meaningful ratings, applicable within the sporting industry to evaluate team and player performance and 2) develop a novel evaluation metric to quantify the effectiveness of meaningful sport-based rating systems built using the ratings framework.

3.2 KEY ELEMENTS OF A RATING SYSTEM

Although there is high demand for rating systems within the commercial environment and the academic literature is highly active, currently there is no known modelling approach for constructing sport-based rating systems. Moreover, a limitation of the academic literature is the lack of methodologies to output robust, reliable, transparent, *and* intuitive ratings. Through the literature review it became apparent that rating systems lack at least one of these four output characteristics. Specifically, many systems lack transparency and intuition due to the application of machine-learning techniques such as neutral-networks, random forest, gradient boosted machine etc, which although produce highly predictive and accurate results, lack transparency and intuition due to their "black-box" syndrome. Such systems produce outputs which are difficult to communicate and map to observable real-world outcomes. Further, commercially deployable rating systems are difficult to assess because transparency and intuitiveness is usually absent because software suppliers want to maintain their competitive advantage and intellectual property. The need for transparency and intuition in sport-based ratings is necessary as such ratings must align with observable real-world outcomes and must be interpretable.

Given the commercial prevalence of sport-based rating systems and the gaps in the literature (Chapter Two, section 2.5), this chapter develops a ratings framework which output reliable, robust, transparent, and intuitive ratings, also known as meaningful ratings. Specifically, the framework constructs sport-based rating systems which produce meaningful ratings that quantify sporting performances, at both a team and player-level, and are comparable across different forecasting scenarios. The proposed framework adopts multi-objective ensembling forecasting strategy. Specifically, the ratings framework is a form of model stacking where information from multiple models is combined to generate a more informative model. Further, the proposed framework implements five key elements: 1) dimension reduction and feature selection techniques, 2) feature engineering strategies, 3) multi-objective modelling framework, 4) time-based variables and 5) ensembling forecasting. These five elements were determined during the development of the sport-based rating systems outlined in Chapter Two (section 2.5.5). The contribution of each published paper to these five elements is outlined below.

Patel, Rooney, Bracewell & Wells (2018) applied hierarchical clustering as a dimension reduction tool and randomForest as a variable section technique to identify meaningful clusters and significant attributes, respectively, to construct a predictive PGA performance rating system. Using a hierarchical clustering technique, four meaningful and expected clusters were found to influence a player's earnings: 1. Short game, 2. Putting, 3. Accuracy and 4. Driving. The most important attributes within each cluster was identified by applying a random forest technique. A simple linear model was created with these metrics for each of the 2010-2017 seasons. The models were applied to the following season, explaining no less than 64% of variation associated with player earnings for each season.

Similarly, Patel, Bracewell, Gazley & Bracewell (2017) applied a classification tree technique to account for the collinearity and complex interactions amongst player metrics and identified the key dimensions influencing the selection of fast bowler to play test cricket. In this research Patel *et al.* (2017) develop a methodology for determining individuals with a greater propensity to play test cricket for New Zealand, based solely on youth performances. The framework enables the probability of playing test cricket for each player to be determined by fitting a regression model to the regression tree residuals. This framework serves as a useful ranking system.

Simmonds, Patel & Bracewell (2018) developed a framework to assess the influence any batting partnership has on a T20 match. Match and player attributes were analysed using exploratory data techniques and random forests to identify the most important attributes influencing the number of runs scored, within a partnership. The important partnership attributes were: (1) partnership strike rate, (2) change in expected total, (3) proportion of resources consumed and (4) partnership contribution. These attributes were aggregated to create a partnership match influence (PMI) metric that quantifies the strength of a partnership. Applying the PMI metric as edge weights for network analysis allows visualisation of partnership strength within a team. It was found that the PMI in the second innings of T20 cricket was more indicative of match outcome relative to individual player influence, showing that building strong partnerships must be built to successfully reach the target total. Simmonds, Patel & Bracewell (2018) outlined the importance of dimension reduction and feature selection techniques, and the need to utilise intuitive and transparent metrics when developing a rating system to assess the influence of batting partnerships in T20 Cricket.

Similarly, Brown, Patel & Bracewell (2018) applied Cox proportional hazard modelling to develop a partnership rating system for one-day batters. The paper successfully formulated models capable of calculating how likely a partnership is to survive each ball, for different partnerships based on within-game events. Based on these survival ratings the optimal batting order for the New Zealand black Caps is identified using a boot-strapping optimisation approach. A stepwise regression survival analysis is applied to identify the significant batting features. Brown, Patel & Bracewell (2018) illustrated the importance of feature selection techniques and time-based metric when building a rating system.

McIvor, Patel, Hilder & Bracewell (2018) applied feature engineering strategies to extract player specific metric from [cricketing] commentary data to predict in-game events in T20 Cricket. McIvor *et al.* (2018) outline a four-stage process (i.e. phase extraction, player identification, performance analysis and future predictions) for extracting and using key components of to produce player ratings based on commentary sentiment. This research outlined the need to apply feature engineering techniques to derive intuitive and reliable sentiment metrics when constructing rating systems. Moreover, McIvor *et al.* (2018) applies a linear multi-objective framework to ensure the commentary-based player metrics are robust. Additionally, it was hypothesised that high sentiment leads to improved future outcomes, therefore an autoregressive distributed lag model using time-lagged sentiment metrics was applied to predict future match state. The results found that time-lagged metrics include additional information which helped predict future match state.

Patel, Bracewell & Rooney (2018) constructed an adaptive roster-based optimisation rating system for limited overs cricket by deriving team ratings as an ensemble of individual ratings. The attributes used to define the individual rating were based on the statistical and practical contribution to winning. An adaptive system was sued to create the individual ratings using a modified version of a product weighted model. It was shown that when developing team rating systems, the individual player components must be appropriately constructed and measured. Moreover, ensembling these individual components produce superior ratings relative to those based on summary statistics of team performance. Therefore, it was recommended that an ensemble mechanism should be applied when building rating systems.

Similarly, Patel, Bracewell & Wells (2017) developed a framework to evaluate the influence of individual players in T20 Cricket. It is hypothesised that three key dimensions are needed to accurately measure player influence: 1) volume of contribution, 2) efficiency of contribution and 3) contributions made under pressure. The ratings framework applies a multi-objective ensembling strategy to derive player ratings. Moreover Patel, Bracewell & Wells (2017) This outlined the importance of dimension reduction and feature selection techniques when constructing rating systems and highlighted the importance of ensembling forecasting strategies when building rating systems.

Patel, Bracewell, Blackie & Boys (2017)[5] developed a predictive rating system to assess the effectiveness of computer programmers to optimise recruitment. The paper constructs a modelling framework for building a developer-based rating system. It was found that a developer's performance score is a combination of *'accuracy'*, *'timeliness'* and *'difficulty'* based features. These three traits form a meaningful predictive measure of performance by using a non-linear optimisation routine which ensembles these three 'trait-ratings'. Patel, Bracewell, Blackie & Boys (2017) identified the key traits that significantly impact developer performance and the key building blocks necessary to construct a rating system. Moreover, they highlighted the advantage of ensembling individual traits to produce an overall rating of performance.

[5] This paper was assessed and selected by IGI Global's executive editorial board as a reprinted chapter in IGI Global research anthology titled *Human Performance Technology.* Bracewell, P. J., Patel, A. K., Blackie, E. J., & Boys, C. (2019). Using a Predictive Rating System for Computer Programmers to Optimise Recruitment: Using Ratings to Optimise Programmer Recruitment. In Management Association, I. (Ed.), *Human Performance Technology: Concepts, Methodologies, Tools, and Applications* (pp. 397-412). IGI Global. http://doi:10.4018/978-1-5225-8356-1.ch020.

Through the development of these sport-based rating systems it is abundantly clear that to output meaningful ratings of sporting performance the [ratings] framework must apply dimension reduction and feature selection techniques, feature engineering strategies, multi-objectives, time-based attributes, and ensemble forecasting strategies. Table 2 outlines the reason for each element.

Element	Reason
Dimension reduction and feature selection	Identifies the traits that significantly affect performance and identify the attributes-types (of varying complexity) that significantly affect each trait, respectively.
Feature engineering	These strategies extract the latent traits affecting performance.
Multi-objective modelling	Derive trait-based ratings.
Time-based variables	Allows the dynamic evaluate ratings. This is necessary as sports are regarded as a dynamic system.
Ensemble strategies	These procedures combine the trait-based ratings and produce results that have better predictive performance relative to single predictions and are more stable.

Table 2: Key elements of sport-based ratings framework

Moreover, given the complexity of modelling performance and the different traits needed to construct a numerical representation of the performance, it is assumed that multiple dimensions, corresponding to different traits that significantly affect performance are necessary. Effectively, the ratings assigned to each significant trait is ensembled to produce intuitive, robust, reliable, and transparent ratings. This “meaningful” combination of trait-based ratings is achieved using an ensemble strategy and produces a performance-based rating, referred to as a sport-based rating. Therefore, performance is defined as a function of individual traits, specifically $performance = \mathcal{F}(trait_1, trait_2, \ldots, trait_3)$.

An ensemble forecast strategy, using different sources, dimensions and modelling methods, is applied because error reductions are larger when ensembling is based on different methods and each method is applied to different dimensions within the data (Batchelor & Dua, 2011). Batchelor & Dua (2011) found that combining forecasts based on diverse assumptions reduces error more than when combining forecasts based on similar assumptions. In Chapter Two it was shown that such ensemble strategies are heavily applied within the credit risk environment and the sporting industry to evaluate team and player performances. Therefore, a multi-objective ensemble forecasting strategy has been applied to construct the ratings framework, because ensembling different trait-based ratings derived from methods that differ substantially and draw from different dimensions and sources of information lead to improved forecasting accuracy (Lichtendahl Jr. & Winkler, 2019).

Effectively, the ensembling strategy will assign ratings to each significant trait, known as trait-based ratings, that affect performance, and ensemble these ratings to produce a rating indicative of performance, during the "evaluation period". Formally the "evaluation period" is defined as the period in which a modeller evaluates human performance for a given task against some standard, depending on the ratings scenario, and assign trait-based ratings. During this period, the trait-based ratings are ensembled to produce an overall human rating, representing a numerical interpretation of performance. For example, when calculating credit-ratings, for a loan applicant, which quantifies an applicants' ability to make timely repayments, on their line of credit, different repayment traits must be considered, such as the applicants' *current lines of credit, existing limits, default history, etc.* Each repayment trait is scored using statistical analysis to produce a trait rating. These "trait" ratings are then ensembled to produce a resultant credit-rating.

To produce meaningful ratings, trait-based ratings must also be meaningful, and therefore, must be reliable, robust, transparent, and intuitive. Key elements of the ratings framework are the application of feature engineering, feature selection and dimension reduction techniques. This reveals the sport-based ratings problem resides within the field of feature selection and dimension reduction, because to effectively measure performance the traits significantly affecting performance must be identified, using feature selection, and quantitatively understood, using dimension reduction.

3.3 MULTI-OBJECTIVE ENSEMBLING

Given the complexity and difficulty of modelling performance it is assumed that performance cannot be evaluated with a single predictive run due to the inherent uncertainty and dynamic nature of the 'subject' during the evaluation period[6]. Therefore, a multi-objective ensemble forecasting strategy is adopted to construct the ratings framework. Such an approach is necessary "to improve forecasting strategy, combine forecasts derived from methods that differ substantially and draw from different sources of information" (Armstrong, 2001, p.1). Lichtendahl Jr. & Winkler (2019) expressed similar sentiments stating that, it is generally

[6] The "evaluation period" is defined as the period in which a modeller evaluates performance for a given task against some standard, depending on the ratings scenario, and assigns trait-based ratings.

accepted that using a combination of forecasting methods instead of a single forecasting method can lead to improvements in forecast accuracy. Moreover, such an approach is appropriate because it is hypothesised that performance is a weighted average or a function of the traits significantly affecting performance (Heinstrom, 2003; Kampe, Edman, Bader, Tagdae, Karlson, 1997; Scharli, Ducasse, Nierstrasz and Black, 2003; Weiten, 2007). The "multi-objective" element of the forecasting strategy will enable the framework to evaluate individual human traits at different layers, either *shallow* or *deep*. The greater the number of traits that significantly affect performance, the more complex the performance.

Given there are different traits that significantly affect performance, the constructed framework should incorporate different modelling objectives that quantify these traits. Effectively, the multi-objective ensembling forecasting strategy is an approach to establish a quantitative understanding of each significant trait, known as trait-based ratings. These individual trait ratings must also be intuitive, transparent, reliable, and robust. Moreover, these individual trait ratings must be unique and not highly correlated, this is because Lichtendal Jr & Winkler (2019) stated that including a poorer method that has forecasting errors which are highly correlated with those of a better method is redundant and can degrade the performance of a combination, while on the contrary, applying a poorer method that has forecasting errors which are negatively correlated with those of a better method can improve a combination. These trait-based ratings are ensembled to produce a meaningful rating of sporting performances.

As sport-based rating systems require a high level of intuition and interpretability, there is an inherent trade-off between the predictive accuracy and the ability to produce such results. Therefore, the ratings framework applies a manual feature selection and engineering, dimension reduction and model selection, based on each trait-based objective, ensuring meaningful outputs are produced.

The trait-based modelling objectives are determined based on the context of the problem and the decision makers criteria for success. To reduce the uncertainty within different traits, multiple (trait-based) objectives are defined and methods that involve fundamentally different approaches are applied to reduce the dependence. Each trait has a unique modelling objective designed to produce a rating which provides a numerical interpretation for the specific trait, also known as a trait-based rating. The trait-based objectives correspond to individual traits that significantly affect performance. For example, within the cricketing context, to successfully measure a player's match performance, their contribution across both innings, with the bat and ball must be examined. Moreover, their contribution at both the match-level and ball-by-ball level must also be measured. By evaluating a player's performance across different dimensions (i.e. *batting*, *bowling*, *match*, and *ball-level*), using multiple objectives it is possible to produce a numerical representation of performance, known as a player rating. Effectively, the "multi-objective" element ensures that all relevant traits pertaining to performance are captured and the uncertainty surrounding the numerical interpretation of performance (i.e. human rating) is sufficiently reduced. Given that performance are difficult to measure, it is hypothesised that multiple layers for each trait must be evaluated to ascertain a significant proportion of uncertainty within each trait, and thus producing a rating which is an ensemble of the trait-based ratings.

It is hypothesised that each trait-based rating must incorporate information from different layers of each trait. The different layers within each trait account for different levels of uncertainty. Therefore, it is necessary to extract performance information layer-by-layer for each trait. Further, the multi-objective ensemble approach ensures that the individual trait-based ratings are sufficiently diverse. Specifically, no two traits account for the uncertainty within the same layer *for a given trait*.

An additional benefit of multi-objective ensembling forecasts is it extends the period of skilful forecasts. This effectively increases the amount of information and reduces uncertainty to produce highly predictive forecasts. For example, the evaluation period for a cricketer is the length of a match (i.e. 20 overs, 50 overs and 5-days). However, trait ratings from varying lengths need to be combined to generate an overall view. A players "batting" evaluation period is the number of balls faced till dismissed, while their "bowling" evaluation period is the number of balls bowled. Therefore, a player's rating is an ensemble of their individual trait-based ratings, specifically, batting and bowling ratings. Moreover, ensembling forecasts from different time periods extends the evaluation period, which increases the skilfulness of the forecasts (Pardowitz, Osinski, Kruschke & Ulbrich, 2016).

3.4 CONSISTENCIES OF RATING SYSTEMS

Through the literature review and the process of developing rating systems (Chapter Two and Appendix A), key communalities were identified within the rating methodologies. These included: 1) features-types, 2) sample size issues, 3) lack of transparency and intuitive ratings, 4) lack of an evaluation metric to assess the effectiveness of ratings systems and 5) a quantitative representation of performance.

3.4.1 Feature Types - Action, Context and Time

A communality found across sport and credit-risk rating systems was the application of three feature types: 1) action, 2) context and 3) time. Patel, Bracewell, Blackie & Boys (2017) found that a developer's performance score is a combination of *'accuracy'*, *'timeliness'* and *'difficulty'* based features. These three developer-trait ratings were derived using three significant feature-types, specifically action, context, and time, that affect each trait. These three traits form a meaningful predictive measure of performance by using a non-linear optimisation routine which ensembles these three 'trait-ratings'.

Given these findings, it is assumed that trait-based ratings must use these three feature-types to account for a significant proportion of uncertainty. These feature types are classified at varying levels of complexity (i.e. *traditional*, *environment* or *time*). This implies trait-based ratings are an ensemble of three feature-types of varying complexity and performance ratings are an ensemble of trait ratings, which are derived using action, context, and time-based features. Figure 3 illustrates the relationship between feature, traits, and performance.

The literature review showed that when developing rating systems, a combination of action, context and time-based features are necessary to produce highly predictive trait-based ratings. For example, a discriminatory credit risk scorecard uses a combination of 1) traditional applicant characteristics such as *age*, *gender* and *occupation*, 2) environmental characteristics such as *living area* and *deprivation score*, and 3) time-based characteristics such as *time at current residence* and *length of tenure*. To develop a highly discriminatory scorecard, a credit risk rating system should apply features from varying levels to generate ratings for repayment 'traits' that significantly affect performance. Therefore, to significantly determine trait-based ratings which accurately quantify a given trait a combination of action, context, and time-based feature-types at varying levels of complexity must be applied.

Further, in the cricketing context, to accurately quantify a player's batting trait the rating system must use a combination of *action* (runs scored), *context* (strike rate) and *time* based (balls faced) features at varying levels of complexity.

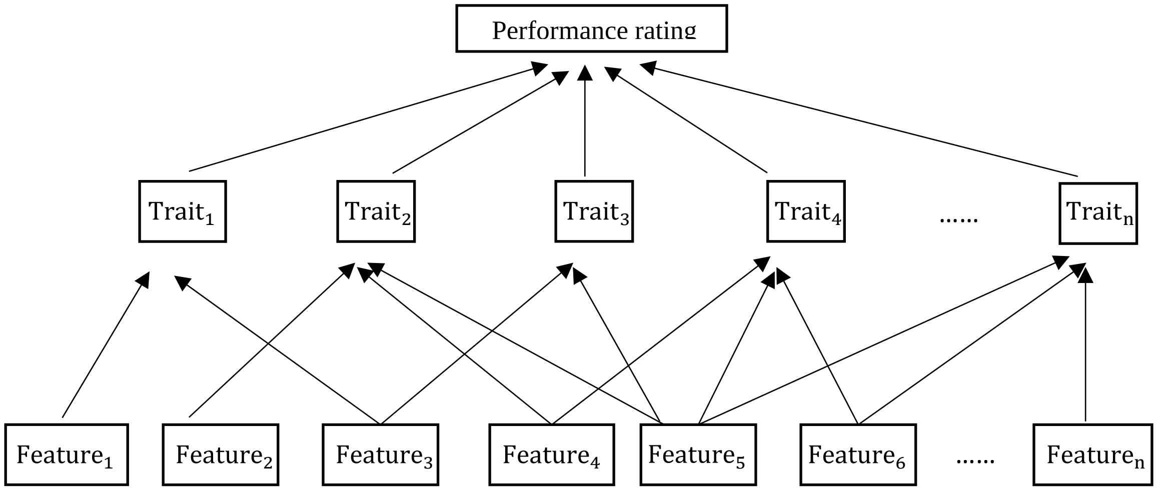


Figure 3: Relationship between feature, traits, and performance

3.4.1.1 Definitions

Action-based features are associated with an observation's physical contribution or demographic characteristic, such as *gender*, *age*, *runs conceded*, *coding language* etc. These attributes summarise an observations static ability without recognising the context of the performed action.

Context-based features contextualise an observation's action for a given condition. A context-based feature provides insight into an individual's action. For example, if '*age'* is an action-based feature, then '*occupation*' or '*relationship status*' are context-based feature. If 'runs scored' is an action-based feature, then '*strike rate*' is context-based. The context features provide information on how and why an action was conducted, efficiency of an action and supplements the action with '*reduced uncertainty*'. For example, in the credit risk environment, if the scorecard only applies an applicant's 'age' and assigns a credit rating of 520. Now suppose the applicant's occupation, a context-based feature, is applied in the scorecard, increasing the applicant's credit-score to 660. The addition of a context-based feature allows the model to account for an additional level of certainty for an applicant's credit rating, overall reducing the uncertainty surrounding their repayment behaviour. By introducing more information, the system becomes increasingly representative of the applicant's ability to make 'good' on a line of credit.

Now suppose instead of introducing '*occupation'*, an additional action-based feature, such as *'gender'* is applied. The credit rating may experience an increase; however, this increase would not be as significant as when *occupation* was introduced. Therefore, combining action and context features reduces uncertainty concerning creditworthiness, producing a more predictive rating.

Time-based features introduce an element of time to an observations action or contextual ability. A time-based feature provides information on the length of time elapsed for an observation to perform an action. Time features are important when constructing rating systems because it informs the decision maker about the length of time taken to perform an action. Effectively such attributes inform the model of the humans' longevity when performing an action within a given context. For example, in the cricketing context '*runs scored'* and '*strike rate'*, are action and context-based attributes respectively, while '*resources remaining*' and '*balls remaining*' are time-based features, because they measure the amount of time remaining until all batting resources have been exhausted, leading to match completion. Moreover, in the credit risk context, time-based features are more subtle. For example, '*time at current residential address*', '*length of tenure at current job*' or '*length of current relationship*' are time-based features providing information on the longevity when performing an action. If '*relationship status*' is considered a context-based feature, the corresponding time-based feature would be '*time spent in current relationship*'.

As mentioned, all three feature-types are required to build meaningful sport-based rating systems. Each feature-type must be applied to produce highly predictive ratings. Moreover, the following relationship of feature "informativeness" should exist: $time \geq context > action$. Specifically, time-based features account for an equal or greater proportion of uncertainty than context-based features, while context-based features account for a greater proportion of uncertainty than action-based features.

3.4.2 Sample Size Issues

Given performance-based ratings are an ensemble of trait-based ratings, an issue was the lack of data-points for certain observations during the earlier stages of an evaluation period. This sample size issue leads to an inability to significantly reduce uncertainty, ultimately producing ratings unindicative of performance. The sample size issue occurs during the earlier stages of the evaluation period, when insufficient amount of information is available. Although, this issue dissipates as the evaluation period matures and more data becomes available. Due to this "time-dependent" nature of rating systems, a dynamic ratings framework is developed. Moreover, an additional benefit of multi-objective ensembling forecasts is that it extends the period of skilful forecasts, leading to an increase in sample size over time.

3.4.3 Lack of Transparency and Intuition

The most prevalent issue amongst sport-based rating systems is the lack of transparency and intuition. Although ratings can be highly predictive and accurate, many suffer from the "black-box" syndrome, implying that ratings are difficult to map to observable real-world outcomes. This leads to communication and interpretation issues of model outputs in terms of observable performance-based attributes. As mentioned, this lack of transparency and intuition is due to the application of black box modelling techniques such as support vector machines (SVM), neural networks (NN), multilayer perceptron (MLP), gradient boosted machines (GBM). To account for this lack of transparency and intuition the ratings framework ensures complete autonomy over model features, feature selection, dimension reduction and modelling techniques.

3.4.4 Numerical Representation of Performance

The most common characteristic shared across many rating systems is the output of a numerical value representing a quantifiable interpretation of sporting performances. Therefore, the constructed ratings framework must output a numerical interpretation of performance when performing a given task, for example repayment behaviour for credit risk. This credit '*ratings*' or '*score*' evaluates how well an applicant performed while completing [credit] tasks over the course of the loan period (i.e. evaluation period).

3.5 RATINGS FRAMEWORK: A DYNAMIC HIERARCHICAL MULTI-OBJECTIVE ENSEMBLE FORECASTING STRATEGY

Given the complexity and difficulty of modelling performance it is assumed that performance cannot be evaluated and described with a single predictive run due to the inherent uncertainty and dynamic nature of the 'subject' during the evaluation period. Therefore, a dynamic multi-objective ensemble forecasting strategy is adopted when constructing a sport-based ratings framework.

Liu & Pentland (1999) developed an approach to model performance which considers the human as a device with many internal mental states, or traits, each with its own control behaviour and interstate transition probabilities. This approach is like the ratings framework approach outlined in this section, whereby each performance-trait is modelled individually, and the trait-based ratings are ensembled to produce an overall rating representing a numerical interpretation of performance.

Therefore, it is hypothesised that ratings perform better the deeper a dimension is explored. The intuition is that engineering features at deeper layers, of a given trait, capture more information and account for a greater level of uncertainty, relative to shallow layer features [of the same trait]. Therefore, the ratings framework incrementally builds more complex features, reducing the uncertainty surrounding performance and producing a rating indicative of performance.

Figure 4 shows a pictorial representation of the ratings framework and represents the different dimensions and multiple layers that exist within each dimension. The rectangular boxes represent the different layers within each trait dimension, and the different layers represent the 'depth' of each trait. The deeper layers contain 'complex' features, which explain a greater proportion of variation within a given trait, relative to shallow layer features. These complex features are constructed by applying modelling or transformation techniques to traditional or simple features. To generate these trait-based ratings action, context and time-based features must be applied. The *edges* linking the layers represent the methods applied to access deeper layers leading to deeper understanding of specific traits, effectively reducing the level of uncertainty.

The shallow layers of each trait produce weak forecasts as they apply simple features that explain a smaller proportion of uncertainty in performance compared to deep layer features. It is assumed that to completely understand the complexities of each trait, first the trait must be understood at shallow levels, and these shallow-level [trait] "understandings" is functionally ensembled to construct increasingly complex features that explain a greater proportion of uncertainty within the trait. As mentioned, these complex features are constructed within deeper layers using the features constructed in the shallow layers.

To ‘delve deeper’ into each trait relevant features must be engineered to construct ‘informative’ trait-based ratings. Therefore, feature engineering, feature selection and dimension reduction techniques ensure that a significant proportion of uncertainty has been removed from each trait rating, and the correct features are selected when building a model to predict each trait-based rating.

Specifically, the deeper layers combine shallow layer forecasts and features, and account for interactions to develop complex features which incorporate information corresponding to the individual performance. Therefore, the deeper layers for each trait contain greater amounts of information pertaining to performance than the weaker forecasts that were built in the shallower layers. These deeper layers are synonymous with “*getting to know*” the performance on a deeper level, to better understand and evaluate and establish the depth of the performance (or performance). When the trait-based rating explains an “appropriate” level of uncertainty, the modeller progresses to the succeeding trait, and constructs the trait-based rating. Once the significant traits are assigned a rating a modelling function is applied to ensemble these trait-based ratings and produce an overall rating representing a numerical interpretation of performance, at a given point during the evaluation period.

Chapter Four and Five apply the ratings framework within the sporting context to build novel rating systems, at both the team and individual level, to evaluate both team and player performance, respectively.

The proceeding sections of this chapter are dedicated to the development of a performance metric which quantifies the effectiveness of meaningful sport-based rating systems (research objectives (ii)).

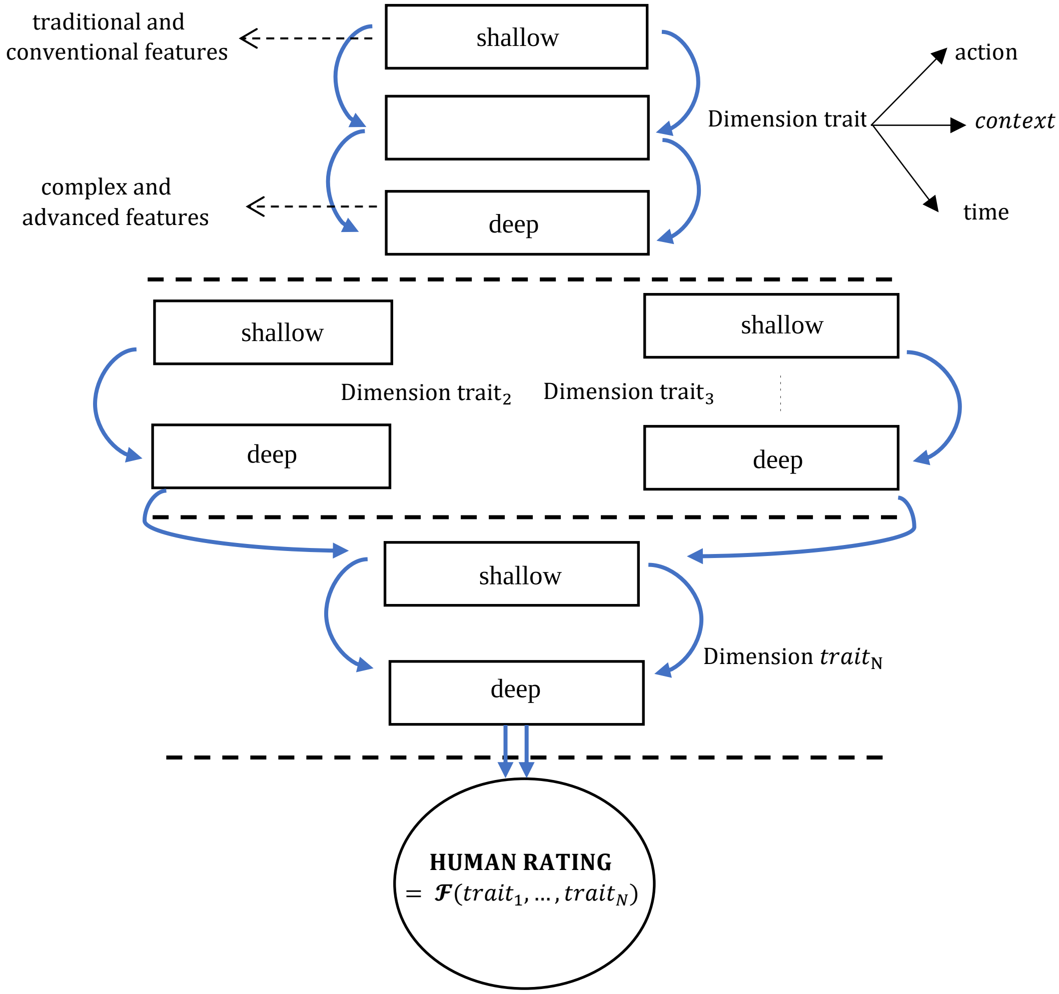


Figure 4: Ratings framework

3.6 METRIC INTRODUCTION

Sport-based rating systems developed using the ratings framework produce intuitive, robust, reliable, and transparent outputs, or more simply meaningful ratings. In the literature review it was found that current model evaluation metrics are limited and lack applicability in certain ratings or forecasting scenarios. Moreover, it was found that traditional performance metrics lack the ability to utilise domain and forecasting specific knowledge. Therefore, it is necessary to develop an evaluation metric which appropriately quantifies the effectiveness of meaningful sport-based rating systems, accurately distinguishes between ‘good’ and ‘bad’ ratings, measures the distance between the reported ratings and the actual outcome, and appropriately accounts for the sporting context and forecasting difficulty. Such a metric will measure the ‘meaningfulness’ and subsequent effectiveness of sport-based ratings systems.

Until now, performance metrics applied to assess the effectiveness of sport-based rating systems are considered insensitive to the subtle nuances of performance ratings and do not sufficiently account for specific attributes that are important to sport performance-based ratings. It is stipulated that such ratings require an evaluation metric which sufficiently accounts for information from different dimensions (i.e. traits), such that model outputs are rewarded for sufficiently incorporating this information. This section is dedicated to the development of a novel metric to quantify the effectiveness of meaningful sport-based rating systems; therefore, addressing research objective (ii).

Throughout this section the terms assessor, expert, forecaster, modeller, and rating systems will be used interchangeably, and are defined as model-based approaches to derive ratings. When forecasting, an assessor can choose one or two of the following approaches: 1) A *model-based approach* which depends on statistical models built on historical data to forecast outcomes of future events and 2) An *expert-driven approach* which, given the current situation, depends on a forecaster's expert beliefs and knowledge to forecast the occurrence of future events. The following section defines and outlines a set of ideal criteria for a performance metric to assess the effectiveness of sport rating systems and describes the importance of each criteria when evaluating the effectiveness of meaningful sport-based rating systems.

3.7 PERFORMANCE METRIC CRITERIA

Given meaningful sport-based performance ratings must be reliable and intuitive [transparent and robust], it is necessary that an evaluation metric incorporates on element of expert knowledge to account for reliability and intuitiveness of sport-based rating systems. Effectively, it is vital that ratings align with a tolerable level of intuition and reality. The performance metric should reward ratings based on honest reporting and provides motivation to report ratings equal to their beliefs. This notion of honest reporting is important to sport-based rating systems because the ratings derived from such systems incorporate an element of intuition. An ideal performance metric should elicit adequate information from experts to make informed, objective, and intuitive judgements.

Chapter Two revealed a set of ideal criteria for constructing a novel performance metric. These metric criteria are: 1) sensitivity to distance, 2) sensitivity to time-dependence, 3) evaluate ratings on the entire probability distribution (i.e. non-local metric), 4) incentivisation for well-calibrated and sharp ratings, and 5) adjusts incentives based on forecasting difficulty.

A metric that accounts for these five criteria can determine the effectiveness of meaningful sport-based rating systems (i.e. reliable, robust, transparent, and intuitive). Given these criteria, and the limitations of current evaluation metrics, and that ensembled ratings are generally assessed on calibration and sharpness (Gneiting, Raftery, Westveld & Goldman, 2005), a proper scoring rule methodology, specifically a spherical scoring methodology is applied to develop the novel metric. Distance and magnitude measures associated with the spherical scoring rule are used to develop the novel performance metric to assess the effectiveness of meaningful sport-based rating systems.

3.7.1 Sensitivity to Distance

Ratings are sensitive to distance, that is, ratings that are closer to the actual outcome, during early stages of the evaluation period, should receive a greater expected score than equivalent ratings outputted during latter stages of the evaluation period. Therefore, it is important that the performance metric considers the distance between the reported probability ($\boldsymbol{r}$) and the actual probability ($\boldsymbol{p}$) and consequently, is sensitive to the size of the deviation from the truth. Specifically, a metric that rewards ratings which are closer to the actual outcome during the early stages of an evaluation period than the latter stages is necessary, as the assessor who produces ratings closer to what actually happened during earlier stages of the evaluation period has done more with less (i.e. produce more informative predictions with less information). Specifically, during the early stages of the evaluation period, the rating system, and therefore the ratings are less reliable about the outcome of interest compared to ratings outputted during the latter stages, where more reliable information is available. Therefore, it is much easier to make predictions about actuality relative to the early stages.

This notion of sensitivity to distance involves probabilities assigned to values close to the observed values compared with probabilities assigned to values farther from the truth (Winkler, 1996). For example, suppose two rating systems, A and B, report probabilities, $\boldsymbol{p}$, the probability it rains on Sunday, given today is Tuesday. The possible values are 0 ($\overline{\text{rain}}$) and 1 (rain). These probabilities, as reported by rating system A and B, are updated daily until Saturday evening. The vector $\boldsymbol{p}$ for the five-day forecast (Tuesday, Wednesday, Thursday, Friday and Saturday) is $[0.64, 0.73, 0.82, 0.85, 0.87]$ for system A and $[0.5, 0.6, 0.82, 0.88, 0.91]$ for system B. Now suppose that it rains on Sunday, under the quadratic (Brier) scoring rule, system A receives a score of 0.89 and system B receives a score 0.81. This example illustrates the notion of "sensitivity to distance" which involves the comparison of the probabilities assigned to values close to the observed outcome with the probabilities assigned to the values farther from the outcome. Even though system B reported a probability on Saturday closer to the actual event relative to system A, it is observed that system A reported probabilities closer to the actual event throughout the entire evaluation period[7] and reported 'closer' probabilities to the actual outcomes further from Sunday (i.e. better prediction during the earlier stages of the evaluation period). Therefore, system A is said to be more consistent.

[7] The evaluation period is defined as the period in which an assessor evaluates a for a given task against some standard, depending on the ratings scenario, and assign ratings.

To evaluate the effectiveness of spot-based rating systems a metric that is sensitive to distance is required. Therefore, if it is believed that a set of probabilities closer to the true value reflect greater predictive power and ability to incorporate sporting context, a performance metric that is sensitive to distance is reasonable.

3.7.2 Sensitivity to Time-Dependence

Ratings are generated at equal time-intervals during the evaluation period, and the length of an evaluation period depends on the sporting context to which the ratings framework is being applied. For example, when rating batting performance, the evaluation period is the batting-innings. As the batting-innings matures and nears completion, so too does the information surrounding the outcome of interest, and the uncertainty surrounding match outcome decreases. Given this information asymmetry, with less information available during the early stages of the evaluation period and more information available during latter stages, it is important that the performance metric weights the early-stage ratings differently than later-stage ratings. For example, suppose two sport pundits: pundit A and pundit B, predict the outcome of a rugby game for team i at 5 different time intervals. The vectors of assessed probabilities for the 5-time intervals for pundit A and B are $[0.67, 0.72, 0.75, 0.82, 0.88]$ and $[0.56, 0.65, 0.72, 0.85, 0.93]$, respectively. Now suppose that team i won the match, even though pundit B's final prediction of match outcome (0.93) was more accurate than pundit A's (0.88) prediction, at the end of the evaluation period, pundit A's predictions for the first three intervals ($[0.67, 0.72, 0.75]$), were significantly more predictive than pundit B's initial three predictions $[0.56, 0.65, 0.72]$. Moreover, there was more uncertainty in pundit B's initial two predictions (0.56 and 0.65) compared to pundit A's initial two predictions (0.67 and 0.72). This reveals that pundit A was able to extract more information about the match outcome from fewer data-points and lesser information than pundit B, and therefore pundit A should be rewarded more.

By the time pundit A and pundit B reveal their predictions, for time-intervals 4 and 5, majority of the uncertainty surrounding match outcome has dissipated and majority of the information surrounding *what will happen* has been established. Therefore, one could say that the value of expert opinion diminishes as the evaluation period matures. Further, it is argued more accurate expert opinions with less information should be rewarded (or weighted) more favourably than accurate expert opinions with more information. Therefore, it is suggested that a rating system which is able to do "more with less" (i.e. extracts greater amount of information from lesser data) or a rating system which is "most right earliest" should be weighted more favourably than a system which is able to do "more with more" or "least right earliest".

3.7.3 Evaluate Ratings on The Entire Probability Distribution

This criterion is a direct consequence of sensitivity to distance and sensitivity to time-dependence. Again, consider the example of pundit A and pundit B (see section 3.7.2), even though it is a dichotomous problem, the ideal performance metric should assess the entire probability distribution, rather than a single probability. The pundits should be evaluated on their reported forecasts throughout the entire evaluation period. This criterion is closely linked to sensitivity to distance and sensitivity to time-dependence because the entire evaluation period accounts for information asymmetry and the assessors' ability to produce predictive ratings under scenarios of asymmetric information by analysing the assessor's entire probability vector. Therefore, the novel metric should be non-local as the effectiveness of rating systems should to be evaluated on the entire probability distribution. "A scoring rule is local if it the score depends only on the probability or density assigned to what is observed" (Wrinkler, 1996, p.15).

The logarithmic rule is a local scoring rule as the score depends *only* on the probability or density assigned to the observed outcome. The logarithmic score assigns a score that is independent of normalization of the quoted probability distribution (Parry, 2016).

Moreover, an additional challenge with sport-based rating systems is the lack of events for the outcome of interest (Patel & Bracewell, 2018), therefore, a performance metric which assesses the entire probability distribution rather than a single probability is appropriate for sport-based rating problems.

3.7.4 Incentive for Well-Calibrated and Sharp Ratings

Calibration and sharpness are criteria shared by most predictive models. These properties are highly sort-after when producing probabilistic predictions (please see Gneiting, Balabdaoui & Raftery, 2007). Gneiting, Balabdaoui & Raftery (2007) stated that calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the events that materialise. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. The more concentrated the predictions, the sharper the forecasts, and the sharper the better, subject to calibration.

Winkler (1996) stated that "good" probability forecasts should be well-calibrated and sharp. Moreover, Naeini, Cooper & Hauskrecht (2015) stated that predictive models that are well-calibrated and sharp are critical for many prediction and decision-making tasks in artificial intelligence; and Niculescu-Mizil & Caruana (2005) mentioned that in many applications it is important to predict well-calibrated and sharp probabilities.

The performance metric must reward well-calibrated ratings by penalising deviations from perfect calibration, that is, $\boldsymbol{r} = \boldsymbol{p}$, where $\boldsymbol{r}$ represents the output vector reported by the rating system, and $\boldsymbol{p}$ represents the actual outcome vector. The metric must also reward sharpness by penalising ratings as they move from zero or one and towards one-half. Effectively, moving towards greater amounts of uncertainty in the ratings. To maximise expected score, the system should set $\boldsymbol{r} = \boldsymbol{p}$. Therefore, the ratings should be assessed using calibration (or reliability) and sharpness (or resolution) statistics.

3.7.5 Adjusts Incentives Based on Forecasting Difficulty

Specific strategies must be developed for tailoring proper scoring rules to specific ratings scenario and aligning the interests of the expert and the rating system. Therefore, the evaluation metric should heavily weight systems for reporting accurate forecasts when other systems perform badly relative to when most systems perform well and when performing well under extremely difficult forecasting situations.

Given, sport-based rating systems can be built for different sporting scenarios, the developed performance metric should be capable of adjusting for various forecasting difficulty and sporting contexts. Moreover, the metric should maximise the expected reward for the "most skilful" forecast (a probability of zero or one being true for a single event) and minimise the expected reward for the "least skilful" forecast. "If one extreme forecast is viewed as more skilful (for example perhaps because it is more difficult to forecast precipitation perfectly than to forecast no precipitation perfectly), then the expected score should be maximised at that extreme" (Winkler, 1996, p.17).

A forecast is regarded as "most skilful" during the early stages of the evaluation period, when the forecasted rating is close to the observed outcome. Extreme ratings (i.e. 0 or 1) that are closer to the actual outcome during the early stages of the evaluation period are viewed as "more skilful", because it is more difficult to forecast 'accurate' ratings earlier than forecasting 'accurate' ratings during latter stages of the evaluation period. Therefore, the expected reward should be maximised at the extremes.

It is important that the performance metric rewards ratings derived during the early stages of the evaluation period, differently than similar ratings derived during the latter stages of the evaluation period. This difference in incentives, based on the time at which ratings are produced, is due to the information asymmetry phenomenon present within the ratings system. Therefore, the skilfulness of ratings and adjusting ratings for forecasting difficulty is required depending on their distance from *what is actually observed* and the time during the evaluation period at which the rating was outputted. Therefore, time adjusting the ratings accounts for scenarios where the outputted ratings perfectly or closely align with the actual outcome (i.e. perfect calibration or well-calibrated, respectively), however perfect calibration or well-calibrated ratings occur during the latter stages of the evaluation period, where prior knowledge is redundant and contributes insignificant information. During these situations, the performance metric should not maximise the expected reward as most of the information surrounding the outcome of interest is available.

3.8 PROPER SCORING RULES

It is hypothesised that a performance metric which encompasses the five criteria (section 3.7) can determine the effectiveness of meaningful sport-ratings by measuring the distance between ratings reported by the rating system and the actual outcome, accounting for the context and difficulty of the forecasting tasks, accounting for time-dependence, measuring ratings on the entire probability distribution, and incentivising for well calibrated and sharp output. As mentioned, a spherical scoring methodology is used to develop a novel performance metric to assess the effectiveness of meaningful sporting performances.

Specifically, distance and magnitude measures derived through the spherical scoring rules are applied to construct the novel performance metric, known as the DMS (distance and magnitude-based 'spherical') metric, to assess the effectiveness of meaningful sport-based rating systems and measure the accuracy of ratings for inducing honest reporting within a certain context.

3.8.1 Application of Scoring Rules

Carvalho (2016) stated that there has been a tremendous increase in the number of published articles applying proper scoring rules to evaluate probabilistic forecasts. The areas which have experienced the largest growth in the application of proper scoring rules are: meteorology (Rasp & Lerch, 2018; Jordan, Kruger & Lerch, 2017; Sillman, Thorarinsdottir, Keenlyside, Schaller, Alexander, Hegerl, Seneviratne, Vautard, Zhang & Zwiers, 2017; Katz & Murphy, 2005; Murphy & Winkler, 1973; Murphy & Winkler, 1982; Murphy & Winkler, 1984; Murphy & Winkler, 1992), prediction markets (Witkowski, Atanasov, Ungar & Krause, 2017; Cummings, Pennock & Vaughan, 2016; Carvalho, 2016), sport analytics (Jackson, 2016; Patel & Bracewell, 2018), psychology (Hollard, Massoni & Vergnaud, 2016; Mellers, Stone, Atanasov, Rohrbaugh, Metz, Ungar, Bishop, Horowitz, Merkle & Tetlock, 2015; Shah, Zhou & Peres, 2015; Bolger & Rowe, 2015; Schlag, Tremewan & Van der Weele, 2015; Roughgarden & Schrijvers, 2017) and energy markets (Robu, Chalkiadakis, Kota, Rogers & Jennings, 2016; Scheuerer & Hamil, 2015; Chen, Jiang, Yu, Liao, Xie & Wu, 2017; Papakonstantinou & Pinson, 2016).

Machete (2013) suggested that the chosen proper scoring rule should depend on its application and consider future decisions associated with high impact, low probability events. The choice of the most appropriate scoring rule is dependent on the desired properties, which depends on the underlying context. Given the forecasting context and the level of forecasting difficulty, choosing the most appropriate scoring rule indicates that: 1) properness is not the only important property, 2) the scoring function should evaluate different forecasts in terms of penalising errors and 3) incorporate information about the decision maker/s who will use the forecast and incorporate new information or attributes as the evaluation period matures and uncertainty or vagueness surrounding the outcome of interest diminishes. Wrinkler (1969) expressed similar sentiments stating that "it is insufficient to use a scoring rule simply because it is strictly proper; instead, it is beneficial to consider the specific way in which the scoring rule rewards and penalizes forecasts" (Wrinkler, 1969, p.753).

An example of probabilistic forecasting is in meteorology where a weather forecaster may give the probability of rain for the next day. One could note the number of times that a 25% probability was quoted, over a long period, and comparing this with the actual proportion of times that rain fell. If the actual percentage was substantially different from the stated probability, here it is stated that the forecaster is poorly calibrated. A poorly calibrated forecaster might be encouraged to do better by a reward-based system. A reward system designed around a proper scoring rule will incentivize the forecaster to report probabilities equal to their personal beliefs. In addition to the simple case of a binary decision, such as "rain" or "no-rain", scoring rules may be used for multiple classes, such as 'rain', 'snow' or 'clear'.

The modeller desires to maximise the expected score from a strictly proper scoring rules, which requires well calibrated and sharp probabilities. Here, calibration assesses how well model predictions align with observed probabilities. A commonly used technique for calibration is the Hosmer-Lemeshow test statistic (Hosmer & Lemeshow, 2013) which assesses a model's goodness-of-fit by comparing observed probabilities against predicted probabilities at quantiles of predicted probabilities. Predicted probabilities that align with the expected probability distribution are known as calibrated. Furthermore, the model is well-calibrated, if the probabilities effectively reflect the true likelihood of the event of interest.

3.9 SPHERICAL SCORING DISTANCE AND MAGNITUDE MEASURES

Given the DMS metric utilises distance and magnitude statistics derived from the spherical scoring methodology, this section discusses the characteristics and properties associated with the spherical scoring rule, how the methodology lends itself to evaluate the effectiveness of rating systems and how these characteristics meet the five ideal criteria for an evaluation metric to assess the effectiveness of meaningful sport-based rating systems.

The quadratic score is not sensitive to distance, while the logarithmic score is a local scoring rule and does not consider the entire probability distribution. A spherical scoring rule is non-local and considers the entire probability distribution, therefore, distance and magnitude measures derived from the spherical scoring rule are applied to develop the novel performance metric to evaluate the effectiveness of meaningful sporting performance (i.e. ratings).

3.9.1 The Spherical Scoring Rule

The spherical scoring rule, first introduced by Roby (1965) in the context of psychological testing, has several geometric properties and a strong connection to the statistical notion of surprise (Good, 1971). The notion of statistical surprise or information content is the amount of information gained when a random variable or signal is sampled. Surprisal represents the surprise or unexpectedness of observing an outcome, for example, the occurrence of a highly improbable outcome is very surprising. The identification of surprise makes a forecaster reconsider the validity of their modelling assumptions, the selected features, the data, and applied techniques. It can provoke the forecaster to change their subjective assessment of previous hypotheses and generate hypotheses that had not been previously entertained.

The expected score, $E_p[S(r)]$, and spherical scoring rules for the occurrence, $S_1(r)$, and non-occurrence, $S_2(r)$, of an event, are expressed as follows:

$$E_p[S(r)] = pS_1(r) + (1-p)S_2(r)$$

$$S_1(r) = \frac{r}{\sqrt{r^2 + (1-r)^2}} \qquad \text{and} \qquad S_2(r) = \frac{1-r}{\sqrt{r^2 + (1-r)^2}}$$

Jose (2007) expressed the expected spherical scoring rule as eqn. 9 and the score for reported probabilities, $\boldsymbol{r}$**,** for, i, as eqn. 10:

$$E_p(p) = \sum p_i \frac{p_i}{\|\boldsymbol{p}\|} = \|\boldsymbol{p}\|, \qquad \text{where} \quad \|p\| = \left(\sum_{i=1}^{n} p_i^2\right)^{\frac{1}{2}} \tag{9}$$

$$S(\boldsymbol{r}, i) = \frac{r_i}{\|\boldsymbol{r}\|} = \frac{r_i}{\sqrt{r_i^2 + \cdots + r_n^2}} \tag{10}$$

Equation (9) shows that the expected spherical score for reporting honestly is the Euclidean length of the vector associated with point p. Here, $\boldsymbol{p}$ represents the true outcome and $\boldsymbol{r}$ represents the reported or modelled probability. The expected score of the spherical scoring rule for reporting honestly is the Euclidean length of the vector associated with the point p representing the assessment $P \epsilon \wp$ (Jose, 2007). Given $p, r \in \Delta n$, $E_p[S(r)]$ can be expressed as follows:

$$E_p[S(r)] = \sum_i \frac{p_i r_i}{\|\boldsymbol{r}\|} = \|\boldsymbol{p}\| \sum_i \frac{p_i r_i}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|}$$

$$= \|\boldsymbol{p}\| \cos\theta \tag{11}$$

Here, θ represents the angle between vectors $\boldsymbol{r}$ and $\boldsymbol{p}$. This implies that the expected score is related to the angle of deviation and the spherical rule is strictly proper since the score is maximized only if the angle between the two vector is equal to zero, that is $\boldsymbol{r} = \boldsymbol{p}$ (Jose, 2007). This shows that the spherical scoring rule motivates the assessor to forecast well-calibrated and sharp probabilities and maximises the expected score when $\boldsymbol{r} = \boldsymbol{p}$**.** The comparison of losses considers the informativeness of the assessments. This is because the norm of a vector $\boldsymbol{p}$ measures to some extent the sharpness of this probabilistic assessment, as $\boldsymbol{p}$ moves away from the centre of the simplex towards the edges the value of the norm increases, and a prediction is viewed to be more informative. Practically, it seems reasonable that when faced with n events where an assessor has complete ignorance over, most people by the principle of insufficient reason would tend to defer to a uniform distribution rather than making a categorical forecast. Therefore, to prevent the possibility of such scenarios developing and appropriately incentivising the assessors is to *provide an incentive for well-calibrated and sharp ratings*. Eqn. 11 illustrates that the spherical scoring rule is *sensitive to distance*. Since the expected loss can be measured from the angle between vector $\boldsymbol{r}$ from the vector $\boldsymbol{p}$, this angular deviation is a measure of the 'distance' between the reported rating ($\boldsymbol{r}$) and the true outcome ($\boldsymbol{p}$).

Figure 5 is a geometric representation of the spherical scoring rule. It can be seen that the expected score (i.e. rating) of $\boldsymbol{r}$ given the true belief $\boldsymbol{p}$ can be interpreted as the norm of an orthogonal projection of a vector of the same size as that of $\boldsymbol{p}$ but in the direction of $\boldsymbol{r}$. This supports the notion that the angular deviation is a sufficient statistic, when taken together with $\boldsymbol{p}$, for the expected score $E_{\boldsymbol{p}}[S(\boldsymbol{r})]$. Effectively, this measures how well probability distribution $\boldsymbol{r}$ represents probability distribution $\boldsymbol{p}$, and measures the dispersion in the forecaster's true probability assignment.

Given the spherical scoring function is sensitive to distance and non-local, and therefore assesses the entire probability distribution when generating scores, it is evident that the spherical scoring rule is 1) sensitive to distance, 2) evaluates the entire distribution (i.e. non-local) and 3) provides incentives for well-calibrated and sharp ratings.

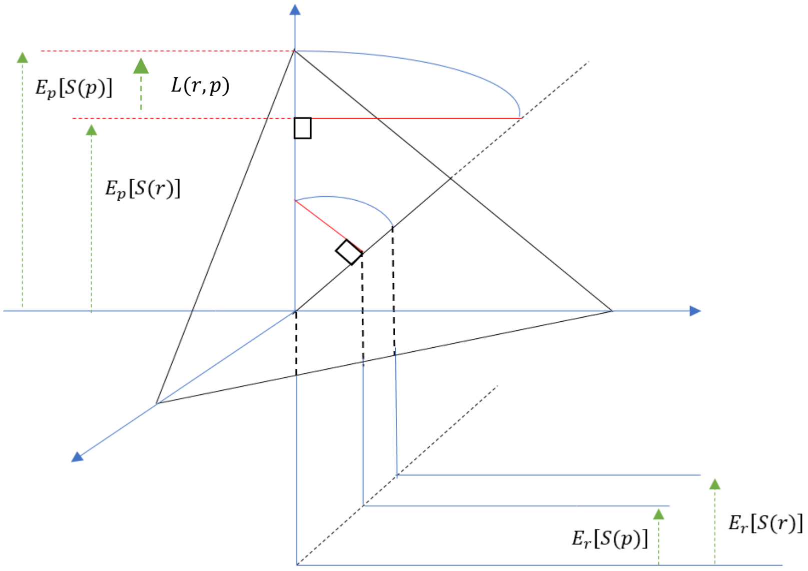


Figure 5: Geometric representation of the spherical scoring rule

3.9.2 Extending the Spherical Scoring Rule

In the previous section it was shown that the spherical scoring rule meets the ideal criteria of *sensitivity of distance*, *provides an incentive for well-calibrated and sharp rating* and *evaluates ratings on the entire distribution* (i.e. non-local). To ensure the developed performance metric sufficiently meets the last two criteria of: 1) *sensitivity to time-dependence* and 2) *adjusts incentives based on forecasting difficulty*, there is the need to extend the spherical scoring rule such that it is applicable to rating problems. As a consequence, this section draws from Winkler (1994) and Saaty (1971) to introduce a novel adaptation of the spherical scoring rule to construct a performance metric which evaluates the predictive power of sport-based ratings and simultaneously meet the last two criteria (time-dependence and adjusting incentives based on forecasting difficulty).

3.9.2.1 Time-dependence

Ratings close to the outcome of interest during the early stages of the evaluation period should be given a higher expected score, than ratings closer to the outcome of interest, during the latter stages of the evaluation period. Therefore, the constructed performance metric should heavily weight (i.e. reward) outputs that are more accurate "the earliest" (i.e. early stages of the evaluation period) relative to outputs that are more accurate "the latest" (i.e. early stages of the evaluation period). Therefore, the performance metric should be non-linear, as the 'predictive accuracy gains' assigned to the ratings during different stages of the evaluation period are not linearly related to the expected score. Given, the ratings are time dependent, the scores assigned to early ratings of the evaluation period are not equivalent to the scores assigned to latter ratings of the evaluation period. Effectively, a rating system that provides outputs during the earlier stages of the evaluation period that are 'close' to the outcome of interest should receive a higher expected score than rating systems that provide 'close' outputs, during the latter stages of the evaluation period. This is because early ratings close to the actual outcome better utilise lesser information and lesser reliable information to produce accurate forecasts, while close ratings during the latter stages should receive relatively smaller scores as more reliable information is available. The first system was able to extract greater information using fewer data points (i.e. the model was able to do "more with less"). Therefore, an asymmetric weight adjustment is implemented due to the time-dependent information asymmetry property of rating systems.

A possible method of adapting the distance and magnitude statistics derived from the spherical scoring rule for time-dependent scenarios is applying a positive affine transformation to dynamically evolve as the evaluation period matures. Such transformations are useful in converting the range of possible values for these scoring rules into other ranges which may be appropriate to the decision context. This time dependent "evolution" of the positive affine transformation can be decision context specific, such that the scoring rule "rewards" an assessor depending on the context of the forecasting situation[8]. For example, the scoring rule can be transformed such that α accounts for the time and β accounts for the context. This serves as a useful transformation given the way scores are allocated simultaneously depend on the time and sporting context.

3.9.2.2 Adjusting incentives based on forecasting difficulty

Suppose there are two assessors', A and B. Assessor A evaluates the probability that a credit applicant defaults on their line of credit, i.e. cannot make monthly repayments over a six month period, and in-light of new monthly information, updates these probabilities on a monthly basis and has the following probability vector $[0.34, 0.32, 0.31, 0.29, 0.21, 0.19]$. Assessor B evaluates the probability that a cricketing batter will be dismissed, i.e. losses their wicket over a 6-ball evaluation period, and in-light of new information, updates these probabilities on a ball-by-ball basis and has the following probability vector $[0.34, 0.32, 0.31, 0.29, 0.21, 0.19]$. Now suppose a performance metric that only accounts for the first four criteria (i.e. sensitive to distance, sensitive to time dependence, evaluate ratings on the entire probability distribution and provides an incentive for well-calibrated and sharp ratings) was used to evaluate the forecasting ability of the two assessors. Here both assessors would receive the same score, and assessor A and assessor B are considered equal in terms of forecasting ability. However, a performance metric that considers the forecasting context and difficulty needs to be applied. This is because the 'score' a rating system receives should depend not only on their forecasting ability, but also on the nature of the forecasting scenarios. Given the performance metric is specific to the sporting context, it is necessary to apply a mechanism which accounts for forecasting difficulty and provides incentives depending on each scenario. Therefore, a metric which adjusts incentives based on the forecasting difficulty is a necessary criterion. As mentioned, a positive affine transformation will be applied to adjust incentives based on forecasting difficulty and ratings context.

3.9.2.3 Asymmetric scoring rules

Given the complexity of sporting performances (at both the team and player level), the difficulty in assigning meaningful ratings that assess performance and the variations of the performance across different sporting codes, it is necessary to adjust for forecasting difficulty and tailor the scoring rule for certain scenarios. An expected-score function that

[8] An interesting property of proper scoring rules is that any positive affine transformation of a proper scoring rule is still proper, therefore any positive linear transformation of a strictly proper scoring rule is itself strictly proper. Effectively, if S is a strictly proper scoring rule then so is $\alpha S + \beta$, for $\alpha > 0$.

minimises the expected score for probabilities deemed to be the “least skilful” prediction is required. The expected score should be measured for probabilities deemed to be most “most skilful” predictions. The “most skilful” forecast is a perfect forecast (i.e. a rating of 0 or 1) in the assessment if a probability for a single event. “If an extreme forecast is viewed as more skilful then the expected score should be maximised at the extreme point” (Winkler, 1996, p.17).

Strictly proper scoring rules that attempt to evaluate the skill of probabilistic forecasts will almost always be asymmetric and their precise form for a given situation can be based on an evaluator’s judgments concerning the relative skill of different probability values (Wrinkler, 1994, p.1405). Winkler (1994) found that the outputs from asymmetric scoring rules correspond to an intuitive notion of a good forecast and are preferable and lead to good scores relative to skill scores and symmetric scoring rules.

Therefore, to develop a performance metric which evaluates the effectiveness of sport-based ratings an asymmetric spherical scoring rule is appropriate, and it should be tailored to reflect forecasting skill in different rating (i.e. forecasting) scenarios.

Skill scores are necessary when comparing two different forecasting systems, because different forecasters generally do not deal with the same situations. Further, different forecasting systems provide forecasts on different time intervals across different domains. For example, when forecasting the probability of precipitation, in extremely dry climates it is easy to give a probability close to zero on many days. However, in areas with higher frequency of precipitation due to highly unstable weather conditions, it may be difficult to give extreme forecasts near zero or one. Both these scenarios will lead to different distributions, $g(r)$, which represents the proportion of time the value r occurs in a forecast series and consequently to different average scores.

In scenarios where two forecasts or forecasting systems receive different average scores, the difference in scores, across the two systems, could reflect differences in forecasting ability or in the forecasting situation. To address this issue, Murphy (1974) developed “skill scores” to produce average scores that reflect the relative ability of forecasters rather than a combination of the forecasters’ ability and the level of difficulty of the forecasting task. The skill scores attempt to neutralise the contribution of ‘situational’ effect by comparing a forecasters average score to the average scores obtained from an unsophisticated forecasting strategy for the same set of forecasting situations. “The typical forecasting strategy chosen when a base rate (a relative frequency of occurrence of the event based on past data) is available is simply a forecast equal to that base rate” (Winkler, 1994, p.1398). For example, in meteorology the base rate, $\boldsymbol{c}$, is known from climatology. Suppose it has rained for 30% of the days in April at a given location, the climatology forecasts would be a probability of 0.3 for each day in April.

In scenarios where $c = 0.5$ (i.e. base rate), the scoring rule is symmetric and asymmetric otherwise (i.e. $c \neq 0.5$). “The intuition supporting this result is that the uncertainty about whether the event E will occur or not is greatest when the probability of E is 0.5” (Winkler, 1994, p. 1398), and therefore, the average skill scores will be lowest for the least skilful forecast. Here, the “least skilful” forecast are 1) those near or close to all information or data needed to access the outcome is readily available, or 2) situations where no real expert opinion is required to assess the outcome of interest. Generally, such forecasts occur during the latter stages of the evaluation period or when nearing the end of the evaluation period where all data and information to evaluate the system is available. For example, within the cricketing context reasonableness is defined as benchmark, c, which evolves as an innings matures. Suppose a forecasting system outputs the probability that the second innings batting team wins the match, on a ball-by-ball basis. This benchmark, c, is the proportion of times the team batting second went onto to win the match, given the number of balls, runs scored, wickets lost and resources remaining in the innings. Here, the average score is minimised for forecasts identical to the benchmark rate, i.e. c. For example, suppose a model outputs a probability of winning of 0.78 for a given ball, with a benchmark value $c = 0.6$, and the team batting second wins the match. Then the prediction of 0.78 is considered ‘reasonable’ as the forecasted probability is closely aligned with the actual outcome relative to the benchmark (i.e. $c = 0.6$). Therefore c can be viewed as an appropriate benchmark probability and the skill score represents the system’s ability to be more discriminatory than c.

For sport-based rating systems, the benchmark vector, $\boldsymbol{c}$, is different from the climatology example as $\boldsymbol{c}$ dynamically changes over the course of the evaluation period.

3.10 A NOVEL PERFORMANCE METRIC USING SPHERICAL SCORING: THE DISTANCE AND MAGNITUDE-BASED SPHERICAL METRIC

The objective of the novel evaluation metric is to measure the predictive accuracy of a sports-based rating systems’ reported rating, $\boldsymbol{r}$**,** against the actual ratings (i.e. outcome) vector, $\boldsymbol{p}$, and an average rating vector, $\boldsymbol{c}$, during different time intervals of the evaluation period. Effectively, $\boldsymbol{c}$ (i.e. benchmark vector) is the average observed outcome given the current conditions. The novel metric utilises distance and magnitude measures (calibration and sharpness statistics) associated with the spherical scoring method (such as the *rate of change in the difference in vector magnitude* and *the rate of change in vector angles*), and therefore will be known as the DMS (Distance and Magnitude-based Spherical) metric. These “similarity” measures are calculated using the three vectors, $\boldsymbol{c}, \boldsymbol{r}$ and $\boldsymbol{p}$, across different time intervals during the evaluation period. Vector $\boldsymbol{r}$ are ratings derived from the rating systems, vector $\boldsymbol{p}$ is the actual rating and vector $\boldsymbol{c}$ is the benchmarked rating based on historical data.

The DMS metric is derived using an algorithmic process such that the reported probability vector, $\boldsymbol{r}$, is weight-adjusted based on the *rate of change in the difference in vector magnitude* and the *rate of change in vector angle* measures between $\boldsymbol{r}$, $\boldsymbol{p}$, and $\boldsymbol{c}$. A weight-adjustment is applied to vector $\boldsymbol{r}$ at each time interval, t, of the evaluation period, and these adjustments weight (i.e. reward and penalise) $\boldsymbol{r}$ based on the rate of change measures. These weights are derived using the Analytical Hierarchy Process (Technical details provided in Appendix B).

Figure 6 illustrates the rate change in the difference between vector magnitude and the rate change of in vector angle measures.

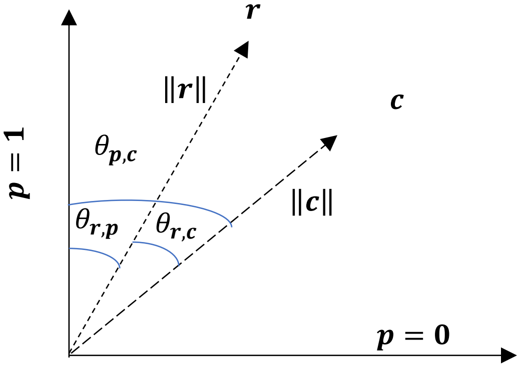

$\boldsymbol{r}$
$\|\boldsymbol{r}\|$
$\boldsymbol{c}$
$\theta_{\boldsymbol{p},\boldsymbol{c}}$
$\boldsymbol{p}=\boldsymbol{1}$
$\|\boldsymbol{c}\|$
$\theta_{\boldsymbol{r},\boldsymbol{p}}$
$\theta_{\boldsymbol{r},\boldsymbol{c}}$
$\boldsymbol{p}=\boldsymbol{0}$

Figure 6: Vector representation

In Figure 6, $\theta_{\boldsymbol{p},\boldsymbol{c}}$, represents the angular distance between vector $\boldsymbol{p}$ and $\boldsymbol{c}$, $\theta_{\boldsymbol{r},\boldsymbol{p}}$ represents the angular distance between vector $\boldsymbol{c}$ and $\boldsymbol{p}$, and $\theta_{\boldsymbol{r},\boldsymbol{c}}$ represents the angular distance between vector $\boldsymbol{r}$ and $\boldsymbol{c}$. The number of elements within $\boldsymbol{r}$ and $\boldsymbol{c}$ increases as the evaluation period matures and as more information becomes available (i.e. time interval, t, increases). The vector $\boldsymbol{p}$ represents the actual outcome of interest and is realised once the evaluation period is completed (i.e. fully matured).

3.10.1 Rate of Change in Difference in Vector Magnitude

As the evaluation period matures the magnitude of $\boldsymbol{r}$, $\|\boldsymbol{r}\|$, should tend towards $\|\boldsymbol{p}\|$. $\|\boldsymbol{c}\|$ should also tend towards $\|\boldsymbol{p}\|$, but at a slower rate. Although the averaged forecasts, $\boldsymbol{c}$, improves as the evaluation period matures, it uses match information less "efficiently" than the ensembled ratings system used to derive $\boldsymbol{r}$. Further, given $\boldsymbol{r}$ is assumed more informative than $\boldsymbol{c}$ as the evaluation period matures, the difference between $\|\boldsymbol{r}\|$ and $\|\boldsymbol{c}\|$ increases over time, while the difference between $\|\boldsymbol{r}\|$ and $\|\boldsymbol{p}\|$ decreases. Therefore, $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$ faster than $\|\boldsymbol{p}\| - \|\boldsymbol{c}\|$ as the evaluation period matures.

3.10.2 Rate of Change in Vector Angles

As the evaluation period matures the angle between $\boldsymbol{r}$ and $\boldsymbol{p}$, $\theta_{\boldsymbol{r},\boldsymbol{p}}$, decreases. As the evaluation period matures and more information becomes available the rating systems reported rating, $\boldsymbol{r}$, should converge faster to the actual outcome, $\boldsymbol{p}$, relative to $\boldsymbol{c}$. Therefore, $\theta_{\boldsymbol{r},\boldsymbol{p}} \to 0$ faster than $\theta_{\boldsymbol{c},\boldsymbol{p}}$ as the evaluation period nears completion.

Further, as time matures the angle between $\boldsymbol{r}$ and $\boldsymbol{c}$ (i.e. $\theta_{\boldsymbol{r},\boldsymbol{c}}$) should slowly increase. As stated, the system's reported ratings, $\boldsymbol{r}$, should converge faster to $\boldsymbol{p}$, relative to $\boldsymbol{c}$, over time as more information becomes available. Although $\theta_{\boldsymbol{c},\boldsymbol{p}} \to 0$, the rating system produces increasingly informative ratings than the average forecast, $\boldsymbol{c}$. Therefore, $\theta_{\boldsymbol{r},\boldsymbol{p}} \to 0$ faster relative to $\theta_{\boldsymbol{c},\boldsymbol{p}}$. Although both $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{c},\boldsymbol{p}}$ converge to zero, $\theta_{\boldsymbol{r},\boldsymbol{p}}$ converges at a faster rate, therefore the rate of angular change, $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{c},\boldsymbol{p}}$, diverges over time, and the angular difference between $\boldsymbol{r}$ and $\boldsymbol{c}$ (i.e. $\theta_{\boldsymbol{r},\boldsymbol{c}}$) increases over time. Specifically, this difference in rate of angular change between $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{c},\boldsymbol{p}}$, it is assumed that $\theta_{\boldsymbol{r},\boldsymbol{c}}$ will increase as the evaluation period matures.

Therefore, over time $\theta_{\boldsymbol{p},\boldsymbol{c}}$ decreases and tends towards zero, however the rate of angular change between $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{p},\boldsymbol{c}}$ should not be the same. As mentioned, it is assumed that $\boldsymbol{r}$ is more informative than $\boldsymbol{c}$ at any given point, especially during the latter stages of the evaluation period where $\boldsymbol{r}$ increases at a greater rate than $\boldsymbol{c}$. The rate at which $\theta_{\boldsymbol{r},\boldsymbol{p}} \to 0$ must be greater than the rate at which $\theta_{\boldsymbol{p},\boldsymbol{c}} \to 0$, therefore, although $\theta_{\boldsymbol{c},\boldsymbol{p}} \to 0$, it tends towards zero at a slower rate than $\theta_{\boldsymbol{r},\boldsymbol{p}}$, during the evaluation period.

These vector-based rate of change distance and magnitude measures can be used to quantify sport-based ratings effectiveness, at different time intervals of the evaluation period. The magnitude-based measures evaluate the similarity between the vectors, while the angle-based measures evaluate the distance between the vectors. The time remaining until the evaluation period fully matures measures length till completion.

3.10.3 Mathematical Notation

- Rate of change in vector angle between $\boldsymbol{r}$ and $\boldsymbol{p}$, during time t to $t+1$, is represented by $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}(t, t+1)$.
- Rate of change in vector angle between $\boldsymbol{r}$ and $\boldsymbol{c}$, during time t to $t+1$, is represented by $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}(t, t+1)$.
- Rate of change in vector angle between $\boldsymbol{p}$ and $\boldsymbol{c}$, during time t to $t+1$, is represented by $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t, t+1)$.
- Difference in vector magnitude between $\boldsymbol{p}$ and $\boldsymbol{r}$ at time t is represented by $(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_t$.
- Difference in vector length between $\boldsymbol{p}$ and $\boldsymbol{c}$ at time t is represented by $(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_t$.

- Difference in vector length between $\boldsymbol{r}$ and $\boldsymbol{c}$ at time t is represented by $(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_t$.
- Rate of change in the difference in vector length between $\boldsymbol{p}$ and $\boldsymbol{r}$ from time t to $t+1$ is represented by $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$.
- Rate of change in the difference in vector length between $\boldsymbol{p}$ and $\boldsymbol{c}$ from time t to $t+1$ is represented by $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$.
- Rate change in the difference in vector length between $\boldsymbol{r}$ and $\boldsymbol{c}$ from time t to $t+1$ is represented by $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$.

3.10.4 DMS Weight Adjustments

The weight adjustment strategy for the DMS metric is based on the rate of change in vector angle and the rate of change in the difference in vector magnitude, from time t to $t+1$ (at different time intervals of the evaluation period). These weight adjustments are calculated using the AHP (please see Appendix B for technical notes).

Higher weight adjustments are applied to $\|\boldsymbol{r}\|$ when it is closer to $\|\boldsymbol{p}\|$ relative to $\|\boldsymbol{c}\|$, and the rate of change in the difference in vector magnitude between $\boldsymbol{r}$ and $\boldsymbol{p}$ → 0 (i. e $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1} \rightarrow 0$) faster than the rate of change in the difference in vector magnitude between $\boldsymbol{c}$ and $\boldsymbol{p}$ (i. e $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1} \rightarrow 0$).

Lower weight adjustments are applied when the rate of change in the difference in vector magnitude does not significantly decrease over time or when $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$ tends towards 0 faster than $(\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$. These weight adjustments are initialised using expert knowledge and are updated at each time interval of the evaluation period. This updating process is dependent on the forecasting context and the time at which ratings are derived.

Based on the *rate of change in vector angles* and *rate of change in difference in vector magnitude* rules. Table 3 and 4 outlines a distance, magnitude, and time-based weight adjustment schema for the DMS performance metric.

3.10.4.1 High weight adjustments

Weight adjustments	Vector movement	Rule
Maximum reward	If $\|\boldsymbol{p}\| = \|\boldsymbol{r}\|$	
High reward	If $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During early stages of the evaluation period and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'very close' to zero.
Medium reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During latter stages of the evaluation period and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'close' to zero.
Small reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During early stages of the evaluation period and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'far' from zero.

Smallest reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During latter stages of the evaluation period and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'far' from zero.

Table 3: High weight adjustment schema for the DMS performance metric

3.10.4.2 Low weight adjustments

Weight adjustments	Vector movement	Rule
High reward	If $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During latter stages of the evaluation period, and where $\theta_{\boldsymbol{c},\boldsymbol{p}}$ is 'closer' to zero, relative to $\theta_{\boldsymbol{r},\boldsymbol{p}}$.
Medium reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During latter stages of the evaluation period, and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'further' from zero, relative to $\theta_{\boldsymbol{c},\boldsymbol{p}}$.
Small reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During early stages of the evaluation period, and where $\theta_{\boldsymbol{c},\boldsymbol{p}}$ is 'closer' to zero, relative to $\theta_{\boldsymbol{r},\boldsymbol{p}}$.
Smallest reward	if $\|\boldsymbol{p}\| - \|\boldsymbol{r}\| \rightarrow 0$	During early stages of the evaluation period, and where $\theta_{\boldsymbol{r},\boldsymbol{p}}$ is 'further' from zero, relative to $\theta_{\boldsymbol{c},\boldsymbol{p}}$.

Table 4: High weight adjustment schema for the DMS performance metric

3.10.5 Calculating the Distance and Magnitude Spherical metric

Based on the sporting context the modeller defines the importance for each of the *rate of change in difference in vector magnitude* measures (i.e. $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$ and $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$), and the *rate of change in vector angle* measures (i.e. $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$, $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$). The importance of each rate of change measure is calculated through the Analytical Hierarchy Process (AHP) between each time interval, t. Please see Appendix B for more details.

The comparison matrix, as defined by the AHP, establishes the importance for each of the rate of change measures and defines how these measures are updated over time. The AHP is applied to each comparison matrix for each time interval, t.

The rate of change metric $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t+t+1}$ is the most important magnitude-based measures, followed by $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t+t+1}$ and $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t+t+1}$. The weight assigned to $(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t+t+1}$ is significantly different than the weights assigned to $(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t+t+1}$ and $(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t+t+1}$. These weights are applied to each difference in vector length measure to output a result which is used to scale the vector $\boldsymbol{r}$, and create a new vector $\boldsymbol{r'}$.

The rate of change metric $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}$ is the most important distance-based measures, followed by $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}$. As vector $\boldsymbol{p}$ represents the actual outcome, the angular measures $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{p},\boldsymbol{c}}$ are more important than $\theta_{\boldsymbol{r},\boldsymbol{c}}$. Weights are applied to each of these angular distance measures to produce weighted vectors, which are applied to $\boldsymbol{r}$ to create a new vector $\boldsymbol{r'}$**.** Finally, vector $\boldsymbol{r'}$ is applied to the spherical scoring rule to evaluate the effectiveness of ratings. In this section a simple worked example of the distance and magnitude-based spherical (DMS) performance metric is provided. Chapter Five applies the DMS metric within the cricketing context. Specifically, to assess the predictive accuracy of a probability of win model and quantify the effectiveness of a player ratings model.

The following algorithmic process generates the DMS metric:

1) Establish the time intervals in which the data will be split. For example, an evaluation period partitioned into 20%-time intervals and a vector with 20 elements will be split into 10-time intervals of length 2. The size of these 'time' blocks are set by the modeller using expert knowledge.

2) Establish the importance for each of the three *difference in rate of change in vector length* (i.e. magnitude) measures (i.e. $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$ and $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$) and *rate of change in vector angle* (i.e. distance) measures (i.e. $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$, $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$). The relative importance of each distance and magnitude measure between each time interval, t to $t+1$, is established using the AHP.

3) Calculate the *vector lengt*h (i.e. magnitude) and *vector angle* (i.e. distance) for each time interval, t: (i.e. $\|\boldsymbol{r}\|_t$, $\|\boldsymbol{c}\|_t$, $\|\boldsymbol{p}\|_t$, $\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t}$, $\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t}$ and $\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t}$).

4) Calculate the *difference in vector length* and *vector angle* for time t between each vector (i.e. $(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_t$, $(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_t$ **,** $(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_t$, $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{\boldsymbol{t}}$, $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t}$).

5) Calculate the *rate of change in the difference in vector length* and *rate of change in vector angle* between each time interval (i.e. $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{\boldsymbol{t,t+1}}$, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{\boldsymbol{t,t+1}}$, $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{\boldsymbol{t,t+1}}$, $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$, $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$**,** and $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$).

6) Apply the AHP importance weightings to the *rate of difference in vector length* and the *rate of change in vector angle* measures across each time-interval. A multiplicative approach is applied to derive the equations for the rate of change in the difference in vector magnitude:

$$\omega^{\Delta(\|\mathbf{p}\|-\|\mathbf{r}\|)_{\mathrm{t,t+1}}} = \left(\Delta(\|\mathbf{p}\| - \|\mathbf{r}\|)_{\mathrm{t,t+1}} \times \omega_{\mathrm{t,t+1}}^{\Delta(\|\mathbf{p}\|-\|\mathbf{r}\|)_{\mathbf{t,t+1}}}\right)$$

$$\omega^{\Delta(\|\mathbf{p}\|-\|\mathbf{c}\|)_{\mathrm{t,t+1}}} = \left(\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{t,t+1} \times \omega_{t,t+1}^{\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{\boldsymbol{t,t+1}}}\right)$$

$$\omega^{\Delta(r-\|\mathbf{c}\|)_{\mathrm{t,t+1}}} = \left(\Delta(\|\mathbf{r}\|-\|\mathbf{c}\|)_{\mathbf{t,t+1}} \times \omega_{\mathrm{t,t+1}}^{\Delta(\|\mathbf{r}\|-\|\mathbf{c}\|)_{\mathbf{t,t+1}}}\right)$$

A multiplicative approach is also applied to derive the equations for the rate of change in vector angle (i.e. distance):

$$\omega^{\Delta\theta_{r,c}{}^{t,t+1}} = \left(\Delta\theta_{\boldsymbol{r},\boldsymbol{c}} \times \omega_{t,t+1}^{\Delta\theta_{r,c}{}^{t,t+1}}\right)$$

$$\omega^{\Delta\theta_{\boldsymbol{p},c}{}^{t,t+1}} = \left(\Delta\theta_{\boldsymbol{p},\boldsymbol{c}} \times \omega_{t,t+1}^{\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}}\right)$$

$$\omega^{\Delta\theta_{r,\boldsymbol{p}}{}^{t,t+1}} = \left(\Delta\theta_{r,p} \times \omega_{t,t+1}^{\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}}\right)$$

7) Generate an additive scalar and apply it to the vector $\boldsymbol{r}$ to produce $\boldsymbol{r'}$ for each time interval, t.

8) Calculate the spherical score and expected spherical score for $\boldsymbol{r'}$.

$$S(\boldsymbol{r'}, i) = \frac{r'_i}{\|\boldsymbol{r'}\|} = \frac{r'_i}{\sqrt{r_i^2 + \cdots + r_n^2}}$$

$$E_p[S(\boldsymbol{r'})] = \|\boldsymbol{p}\| \cos\theta$$

9) Compare the spherical score and, $E_p[S(\boldsymbol{r})]$ against $E_p[S(\boldsymbol{r'})]$. The higher the expected score the better the predictive accuracy and 'meaningful' of the rating systems.

3.10.6 Distance and Magnitude Spherical Metric Example

Calculate a spherical score for the prediction vector, $\boldsymbol{r}$, for each time interval t. This is achieved by partitioning the evaluation period into t equal time intervals which are user defined. This allows the user to define '*close*' or '*far*' from evaluation completion. For example, if the evaluation period is partitioned into 20%-time intervals, a prediction vector $\boldsymbol{r}$ of ten elements is partitioned as follows:

$$\boldsymbol{r} = [0.81, 0.83 \mid 0.85, 0.86, \mid 0.87, 0.89 \mid 0.91, 0.95 \mid 0.97, 1]^{\mathrm{T}}$$

Based on the sport (i.e. forecasting scenario) and sporting context (i.e. forecasting difficulty) the decision-maker defines the importance of each of the rate of change in difference in vector length (i.e. magnitude) measures: 1) $\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{t,t+1}$, 2) $\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{t,t+1}$ and 3)

$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{t,t+1}$, and the rate of change in vector angle-based (i.e. distance) measures: 1) $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}(t, t+1)$, 2) $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t, t+1)$ and 3) $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}(t, t+1)$.

The importance assigned to each distance and magnitude measure is derived using the AHP between each time interval, t to $t+1$. Table 5 and 6 illustrates an example of comparison matrices between each time intervals, t to $t+1$, for a vector, $\boldsymbol{r}$.

Time 1 - 2	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{\mathbf{1,2}}$	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{1,2}}$	$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{1,2}}$
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{\mathbf{1,2}}$	1	3	4
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{1,2}}$	0.33	1	4
$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{1,2}}$	0.25	0.25	1
Time 2 - 3	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{2,3}$	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{2,3}$	$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{2,3}$
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{\mathbf{2,3}}$	1	3	4
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{2,3}}$	0.33	1	4
$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{2,3}}$	0.25	0.25	1
Time 3 - 4	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{3,4}$	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{3,4}$	$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{3,4}$
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{\mathbf{3,4}}$	1	3	4
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{3,4}}$	0.33	1	4
$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{3,4}}$	0.25	1/4	1
Time 4 - 5	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{4,5}$	$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{4,5}$	$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{4,5}$
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{r}\rVert)_{\mathbf{4,5}}$	1	3	4
$\Delta(\lVert\boldsymbol{p}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{4,5}}$	0.33	1	4
$\Delta(\lVert\boldsymbol{r}\rVert - \lVert\boldsymbol{c}\rVert)_{\mathbf{4,5}}$	0.25	1/4	1

Table 5: Comparison matrix for the rate of change in vector length, between time t and t+1

Time 1 - 2	$\boldsymbol{\Delta\theta_{r,p}}^{(1,2)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(1,2)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(1,2)}$
$\boldsymbol{\Delta\theta_{r,p}}^{(1,2)}$	1	6	7
$\boldsymbol{\Delta\theta_{p,c}}^{(1,2)}$	1/6	1	6
$\boldsymbol{\Delta\theta_{r,c}}^{(1,2)}$	1/7	1/6	1
Time 2 - 3	$\boldsymbol{\Delta\theta_{r,p}}^{(2,3)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(2,3)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(2,3)}$
$\boldsymbol{\Delta\theta_{r,p}}^{(2,3)}$	1	7	7
$\boldsymbol{\Delta\theta_{p,c}}^{(2,3)}$	1/7	1	6
$\boldsymbol{\Delta\theta_{r,c}}^{(2,3)}$	1/7	1/6	1
Time 3 – 4	$\boldsymbol{\Delta\theta_{r,p}}^{(3,4)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(3,4)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(3,4)}$
$\boldsymbol{\Delta\theta_{r,p}}^{(3,4)}$	1	8	8
$\boldsymbol{\Delta\theta_{p,c}}^{(3,4)}$	1/8	1	5
$\boldsymbol{\Delta\theta_{r,c}}^{(3,4)}$	1/8	1/5	1
Time 4 - 5	$\boldsymbol{\Delta\theta_{r,p}}^{(4,5)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(4,5)}$	$\boldsymbol{\Delta\theta_{r,p}}^{(4,5)}$
$\boldsymbol{\Delta\theta_{r,p}}^{(4,5)}$	1	9	8
$\boldsymbol{\Delta\theta_{p,c}}^{(4,5)}$	1/9	1	5
$\boldsymbol{\Delta\theta_{r,c}}^{(4,5)}$	1/8	1/5	1

Table 6: Comparison matrix for the rate of change in vector angles, between time t and t+1

The comparison matrix defines the importance of each of the rate of change measures and shows how this importance evolves over time. Applying the AHP over each comparison matrix for each time interval, t. Table 7 and 8 outline the weights for the rate of change in the difference in vector length and the rate of change in vector angle.

Time	$\boldsymbol{\omega}^{(\Delta\lVert \boldsymbol{p}\rVert-\lVert \boldsymbol{r}\rVert)_{t,t+1}}$	$\boldsymbol{\omega}^{(\Delta\lVert \boldsymbol{p}\rVert-\lVert \boldsymbol{c}\rVert)_{t,t+1}}$	$\boldsymbol{\omega}^{(\Delta\lVert \boldsymbol{r}\rVert-\lVert \boldsymbol{c}\rVert)_{t+t+1}}$
$\mathbf{time\ 1 - time\ 2}$	0.60	0.30	0.10
$\mathbf{time\ 2 - time\ 3}$	0.69	0.23	0.08
$\mathbf{time\ 3 - time\ 4}$	0.70	0.23	0.01
$\mathbf{time\ 4 - time\ 5}$	0.73	0.21	0.06

Table 7: AHP weights for each rate change in difference of vector length element, between time t and t+1

Here, $(\Delta\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t+t+1}$ is the most important magnitude-based measure, it describes the rate change in the difference in vector length between vector $\boldsymbol{p}$ and $\boldsymbol{r},$ for time t to $t+1$. This is followed by $(\Delta\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t+t+1}$ and $(\Delta\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t+t+1}$. Here, the reported probability vector, $\boldsymbol{r},$ is the most important magnitude vector throughout the evaluation period. Moreover, the weight assigned to $(\Delta\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t+t+1}$ is significantly different than the weights assigned to $(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t+t+1}$ and $(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t+t+1}$. These weights are applied to each difference in vector length element to produce a value which is used to scale the vector $\boldsymbol{r}$ to create a new vector $\boldsymbol{r}'$.

Time	$\omega^{\Delta\theta_{r,p}}$	$\omega^{\Delta\theta_{p,c}}$	$\omega^{\Delta\theta_{r,c}}$
time 1 − time 2	0.73	0.21	0.06
time 2 − time 3	0.74	0.20	0.06
time 3 − time 4	0.78	0.17	0.05
time 4 − time 5	0.79	0.16	0.04

Table 8: AHP weights for each rate change in vector angle element, between time t and t+1

The rate of change in vector angle $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$ is the most important angle-based distance measure, it describes the rate of angular change between $\boldsymbol{r}$ and $\boldsymbol{p},$ for time t to $t+1$. This is followed by $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$. As vector $\boldsymbol{p}$ represents the actual outcome, the angular measures $\theta_{\boldsymbol{r},\boldsymbol{p}}$ and $\theta_{\boldsymbol{p},\boldsymbol{c}}$ are more important than $\theta_{\boldsymbol{r},\boldsymbol{c}}$. These weights are applied to each rate of change in vector angle measure to produce a value which is applied to scale the vector $\boldsymbol{r}$ to create a new vector $\boldsymbol{r}'$.

This modified vector $\boldsymbol{r}'$ is used within the spherical scoring rule to measures the effectiveness (i.e. predictive accuracy) of the forecasted ratings. The following section outlines a simple example of the distance and magnitude-based metric algorithm, followed by a more complex example applied within the cricketing context.

3.10.6.1 The Distance and Magnitude Spherical Metric Algorithm

Using the previous example, the following steps are necessary to calculate the distance and magnitude-based spherical metric:

1) Establish the time intervals in which the data will be split. Here, the evaluation period is partitioned into 20%-time intervals; for example, a vector with 10 elements will be split into 5-time intervals.

2) Establish the importance of the three difference in rate of change in vector length measures, for each time interval (i.e. $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$ and

$\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$) and rate of change in distance measures (i.e. $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$, $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$ and $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$). The relative importance of each element between each time interval is established using the AHP.

3) Calculate the *vector length* and *vector angle* for each time interval, t. For example, the vector $\boldsymbol{r}$ at time interval $t = 1$, $\boldsymbol{r}^1$, is (0.81,0.83), with a length, $\|\boldsymbol{r}\|_t$, of 1.16, and an angle, $\theta_{r,c}$, of 0.21 degrees (0.0037 radians). The vector length and vector angle for each time interval in the evaluation period for vectors, $\boldsymbol{r}, \boldsymbol{p}$ and $\boldsymbol{c}$ are:

Time	**Vector length**			**Angles**		
t	$\|\boldsymbol{r}\|$	$\|\boldsymbol{c}\|$	$\|\boldsymbol{p}\|$	$\theta_{\boldsymbol{r},\boldsymbol{c}}$	$\theta_{\boldsymbol{p},\boldsymbol{c}}$	$\theta_{\boldsymbol{r},\boldsymbol{p}}$
1	1.16	0.89	1.41	0.21	0.91	0.70
2	1.68	1.37	2.00	3.20	4.45	1.31
3	2.09	1.75	2.45	3.30	4.97	1.75
4	2.47	2.10	2.83	3.04	5.50	2.75
5	2.83	2.42	3.16	2.67	5.80	3.80

4) Calculate the *difference in vector length* and *vector angle* between each vector, $\boldsymbol{r}, \boldsymbol{p}$ and $\boldsymbol{c}$ for time t.

Time	**Vector length**			**Angles**		
t	$\|\boldsymbol{p}\| - \|\boldsymbol{r}\|$	$\|\boldsymbol{p}\| - \|\boldsymbol{c}\|$	$\|\boldsymbol{r}\| - \|\boldsymbol{c}\|$	$\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$	$\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$	$\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$
1	0.25	0.52	0.72	0.49	0.91	0.21
2	0.32	0.63	0.31	1.89	1.25	3.20
3	0.36	0.70	0.34	1.55	4.97	3.30
4	0.36	0.73	0.37	0.29	2.46	3.04
5	-0.33	0.74	0.41	-1.13	3.13	2.67

5) Calculate the *rate of change in the difference in vector length* and the *rate of change in vector angle* between time t and $t + 1$.

Time	**Vector length**			**Angles**		
t	$\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$	$\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$	$\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$	$\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}{}^{t,t+1}$	$\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}{}^{t,t+1}$	$\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}{}^{t,t+1}$
1-2	-0.07	-0.11	0.41	-2.99	-3.54	-0.61
2-3	-0.04	-0.07	-0.03	-0.10	-0.52	-0.44
3-4	0.06	-0.03	-0.03	0.26	-0.53	-1.00
4-5	-0.69	-0.01	-0.04	0.37	-0.30	-1.05

6) Apply the corresponding AHP weights to each element to the distance and magnitude measures between time-interval, t to $t + 1$. Table 7 and 8 outlines the weights for the *rate of change in difference in vector length* and *rate of change in vector angle* measures, respectively. The equations for the rate of change in the difference in vector length: 1) $\left(\Delta(\|\mathbf{p}\| - \|\mathbf{r}\|)_{\mathrm{t,t+1}} \times \omega_{\mathrm{t,t+1}}^{\Delta(\|\mathbf{p}\|-\|\mathbf{r}\|)_{\mathbf{t,t+1}}}\right)$, 2) $\left(\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1} \times \omega_{t,t+1}^{\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{t,t+1}}\right)$ and 3) $\left(\Delta(\|\mathbf{r}\| - \|\mathbf{c}\|)_{\mathrm{t,t+1}} \times \mathrm{W}_{\mathrm{t,t+1}}^{\Delta(\|\mathbf{r}\|-\|\mathbf{c}\|)_{\mathbf{t,t+1}}}\right)$. The equations for the rate of change in vector angle: 1) $\left(\Delta\theta_{\boldsymbol{r,c}} \times \omega_{t,t+1}^{\Delta\theta_{r,c}}\right)$, 2) $\left(\Delta\theta_{\boldsymbol{p,c}} \times \omega_{t,t+1}^{\Delta\theta_{p,c}}\right)$ and 3) $\left(\Delta\theta_{\boldsymbol{r,p}} \times \omega_{t,t+1}^{\Delta\theta_{r,p}}\right)$. The following table outline the multiplicative scalars for the *rate of change in the difference in vect*or *length* between each time interval: $time\ 1 - 2 = -0.034$, $time\ 2 - 3 = -0.0416$, $time\ 3 - 4 = 0.0347$ and $time\ 4 - 5 = -0.5085$.

	Vector length		
$\boldsymbol{t}$	$\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{\boldsymbol{t,t+1}} \times \omega_{t,t+1}^{\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{t,t+1}}$	$\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{\boldsymbol{t,t+1}} \times \omega_{t,t+1}^{\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{t,t+1}}$	$\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{\boldsymbol{t,t+1}} \times \omega_{t,t+1}^{\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{t,t+1}}$
1	(-0.07×0.60) = -0.042	(-0.11×0.30) = -0.033	(0.41×0.10) = 0.041
2	(-0.04×0.69) = -0.028	(-0.07×0.23) = -0.016	(-0.03×0.08) = 0.0024
3	(0.06×0.70) = 0.042	(-0.03×0.23) = -0.007	(-0.03×0.01) = -0.0003
4	(-0.69×0.73) = -0.504	(-0.01×0.21) = -0.0021	(-0.04×0.06) = -0.0024

The following table outlines the multiplicative scalars for the *rate of change in vector angle* between each time interval: $time\ 1 - 2 = -2.96$, $time\ 2 - 3 = -0.2044$, $time\ 3 - 4 = 0.0629$ and $time\ 4 - 5 = 0.2023$.

	Vector angles		
$\boldsymbol{t}$	$\Delta\theta_{\boldsymbol{r,c}}{}^{t,t+1} \times \omega_{t,t+1}^{\Delta\boldsymbol{\theta}_{r,c}}$	$\Delta\theta_{\boldsymbol{p,c}}{}^{t,t+1} \times \omega_{t,t+1}^{\Delta\boldsymbol{\theta}_{p,c}}$	$\Delta\theta_{\boldsymbol{r,p}}{}^{t,t+1} \times \omega_{t,t+1}^{\Delta\boldsymbol{\theta}_{r,p}}$
1	(-2.99×0.73) = -2.183	(-3.54×0.21) = -0.7434	(-0.61×0.06) = -0.037
2	(-0.10×0.74) = -0.074	(-0.52×0.20) = -0.104	(-0.44×0.06) = -0.0264
3	(0.26×0.78) = 0.203	(-0.53×0.17) = -0.0901	(-1.00×0.05) = -0.05
4	(0.37×0.79) = 0.2923	(-0.30×0.16) = -0.048	(-1.05×0.04) = -0.042

Time intervals	Magnitude scalar	Angular scalar	Multiplicative scalar
1-2	-0.034	-2.96	0.10064
2-3	-0.0416	-0.2044	0.0085
3-4	0.0347	0.0629	0.0022
4-5	-0.5085	0.2023	-0.103

7) Generate an additive scalar and apply it to vector $\boldsymbol{r}$ to produce $\boldsymbol{r'}$.

$$\boldsymbol{r'} = [0.81, 0.83, 0.94, 0.96, 0.94, 0.97, 0.91, 0.95, 0.88, 0.90]^{\boldsymbol{T}}$$

8) Calculate the spherical score and expected spherical score for $\boldsymbol{r'}$.

$$S(\boldsymbol{r'}, i) = \frac{r'_i}{\|\boldsymbol{r'}\|} = \frac{r'_i}{\sqrt{r_i^2 + \cdots + r_n^2}}$$

$$\|\boldsymbol{r'}\| = \sqrt{0.81^2 + 0.83^2 + 0.94^2 + 0.96^2 + 0.94^2 + 0.97^2 + 0.91^2 + 0.95^2 + 0.88^2 + 0.90^2}$$

$$\|\boldsymbol{r'}\| = 2.88$$

$$S(\boldsymbol{r'}, i) = [0.281, 0.288, 0.326, 0.333, 0.326, 0.336, 0.316, 0.330, 0.306, 0.313]^T$$

$$E_p[S(r')] = \|\boldsymbol{p}\| \cos\theta$$

9) Compare the spherical score and expected score of $\boldsymbol{r}$ against $\boldsymbol{r'}$. The spherical score for $\boldsymbol{r}$ and $\boldsymbol{r'}$, respectively:

$$S(\boldsymbol{r}, i) = [0.286, 0.293, 0.300, 0.303, 0.307, 0.314, 0.321, 0.335, 0.342, 0.353]^T$$

$$S(\boldsymbol{r'}, i) = [0.281, 0.288, 0.326, 0.333, 0.326, 0.336, 0.316, 0.330, 0.306, 0.313]^T$$

The expected spherical score for $\boldsymbol{r'}$:

$$\sum_i \frac{p_i r_i}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|} = \frac{(0.81 * 1)}{9.10} + \frac{(0.83 * 1)}{9.10} + \frac{(0.94 * 1)}{9.10} + \frac{(0.96 * 1)}{9.10} + \frac{(0.94 * 1)}{9.10} + \frac{(0.97 * 1)}{9.10} + \frac{(0.91 * 1)}{9.10} + \frac{(0.95 * 1)}{9.10} + \frac{(0.88 * 1)}{9.10} + \frac{(0.90 * 1)}{9.10}$$

$$\sum_i \frac{p_i r_i}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|} = 0.999$$

Next, the angle θ is calculated:

$$\cos\theta = \sum_i \frac{p_i r_i}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|}$$

$$\cos\theta = \frac{\boldsymbol{p}.\boldsymbol{r}}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|}$$

$$\theta = arccosine\left(\frac{\boldsymbol{p}.\boldsymbol{r}}{\|\boldsymbol{p}\| \cdot \|\boldsymbol{r}\|}\right)$$

$$\theta = arccosine\left(\frac{9.09}{9.10}\right) = 0.047$$

Therefore, the expected score, $E_p[S(\boldsymbol{r'})]$ is:

$$E_p[S(\boldsymbol{r})] = 3.15$$

$$E_p[S(\boldsymbol{r'})] = 3.16$$

The vector $\boldsymbol{r'}$ is shown to have a higher score across the evaluation period and a higher expected spherical score. This shows that relative to $\boldsymbol{c}$, in terms of angular distance and magnitude, $\boldsymbol{r'}$ was closer to the actual outcome over the evaluation period. This implies that $\boldsymbol{r'}$ should be rewarded for having better prediction than the benchmark during certain time intervals of the evaluation period. Figure 7 illustrates the evaluation of the spherical score for $\boldsymbol{c}, \boldsymbol{r'}, \boldsymbol{r}$ and $\boldsymbol{p}$.

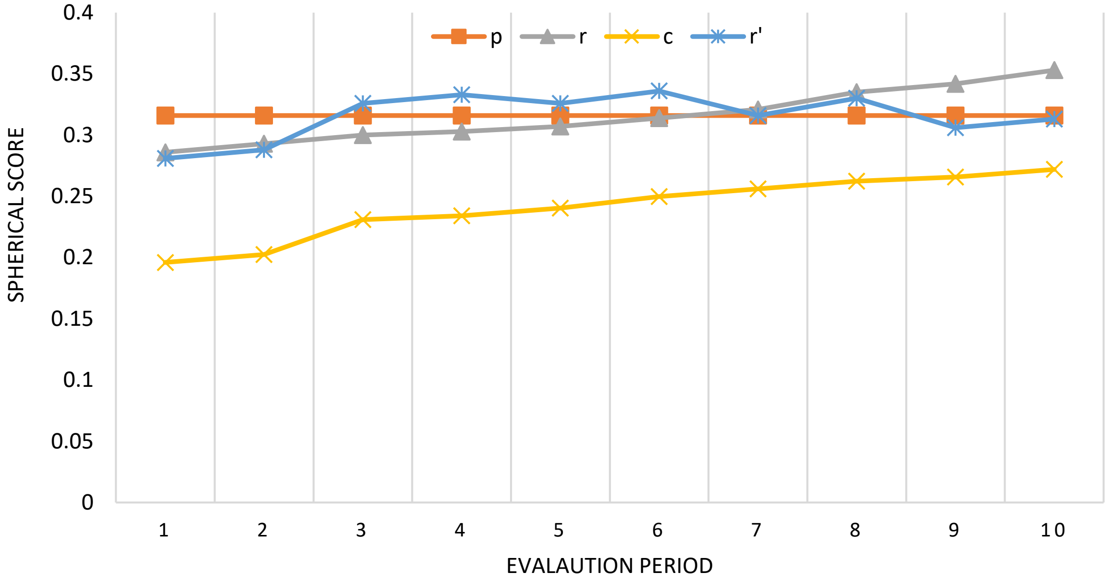


Figure 7: Evaluation of the spherical scoring rule **c, r**, $\boldsymbol{r}'$and **p**

3.11 VALIDATION THROUGH APPLICATION

Given the DMS metric has been shown to work on a simple example. In this section, the metric is applied to the probability of win model developed in Chapter Five (Patel, Bracewell & Wells, 2018) and benchmarked against the general spherical scoring rule and the log loss method. Specifically, the 2019 Big Bash semi-final between the Hobart Hurricanes and Melbourne stars is analysis. For more details on the probability of winning model, please see Chapter Five (section 5.5.4).

There are 120 elements $\boldsymbol{r}$, $\boldsymbol{c}$ and $\boldsymbol{p}$, as there are 120 balls in T20 cricket, and the evaluation period is partitioned into equally spaced intervals of 2 balls. Such a specific split is produced as each vector for each time-split will contain 2 elements ($(2/120)$ ×100= 1.67% of the evaluation period). Therefore, the evaluation period has been split into 60-time intervals with 59 between time intervals.

As mentioned, the vector $\boldsymbol{r}$ will be adjusted based on three magnitude-based rate of change distance measures and three angle-based rate of change measures over the 59-time intervals. Instead of constructing 118 comparison matrices, a 'matrices update' procedure is adopted. For example, assume for any given time interval, t to $t+1$, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{t,t+1}$ is required to be more important than $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$, and $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1}$ is required to be more important than $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$. Therefore, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1} > \Delta(\|\boldsymbol{p}\| - \|\boldsymbol{c}\|)_{t,t+1} > \Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{t,t+1}$ in terms of importance.

Further, for any given time interval t to $t+1$, $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}(t,t+1)$ is more important than $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t,t+1)$, and $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t,t+1)$ is more important than $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}(t,t+1)$. Therefore, $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}(t,t+1)$ > $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t,t+1)$ > $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}(t,t+1)$ in terms of importance.

Finally, assume it is known that in the first time interval the most important rate of change-based metrics, i.e. $\Delta\theta_{\boldsymbol{r},\boldsymbol{p}}(t,t+1)$ and $\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{t,t+1}$, are *absolutely more important* than $\Delta\theta_{\boldsymbol{r},\boldsymbol{c}}(t,t+1)$ and $\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{t,t+1}$, are *strongly more important* than $\Delta\theta_{\boldsymbol{p},\boldsymbol{c}}(t,t+1)$ and $\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{t,t+1}$, respectively. Therefore, the AHP pairwise comparison matrix for time interval 59-60 is:

Time 59-60	$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{\mathbf{59,60}}$	$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{\mathbf{59,60}}$	$\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{\mathbf{59,60}}$
$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{\mathbf{59,60}}$	1	9	10
$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{\mathbf{59,60}}$	1/9	1	8
$\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{\mathbf{59,60}}$	1/10	1/8	1

Time 59-60	$\Delta\boldsymbol{\theta}_{\boldsymbol{r},\boldsymbol{p}}{}^{(\mathbf{59,60})}$	$\Delta\boldsymbol{\theta}_{\boldsymbol{p},\boldsymbol{c}}{}^{(\mathbf{59,60})}$	$\Delta\boldsymbol{\theta}_{\boldsymbol{r},\boldsymbol{c}}{}^{(\mathbf{59,60})}$
$\Delta\boldsymbol{\theta}_{\boldsymbol{r},\boldsymbol{p}}{}^{(\mathbf{59,60})}$	1	8	10
$\Delta\boldsymbol{\theta}_{\boldsymbol{p},\boldsymbol{c}}{}^{(\mathbf{59,60})}$	1/8	1	7
$\Delta\boldsymbol{\theta}_{\boldsymbol{r},\boldsymbol{c}}{}^{(\mathbf{59,60})}$	1/10	1/7	1

Given the pairwise comparison matrix weights for each distance and magnitude measure, for time interval 1-2 and time interval 59-60, are known, therefore the AHP weight for each measure is known, at the beginning and end of the evaluation period. Therefore, at each time interval, the step change applied to each rate of change metric is also known.

A weighting update approach is dynamically applied to adjust the weights for each rate of change measure by a pre-defined value until match completion and the weights have 'converged' to the pre-defined weight as of time 59-60.

Given the small partitioning (i.e. 1.67%) of the innings, there are fifty-nine comparison matrices and therefore the importance value assigned to each element of the pairwise matrix is also small and must incrementally increase as the match progresses. The AHP comparison matrices for *the rate of change in the difference in vector length* and *rate of change in vector angle* for time interval 1-2 are:

Time 1-2	$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{\mathbf{1,2}}$	$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{\mathbf{1,2}}$	$\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{\mathbf{1,2}}$
$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{r}\|)_{\mathbf{1,2}}$	1	3	6
$\Delta(\|\boldsymbol{p}\|-\|\boldsymbol{c}\|)_{\mathbf{1,2}}$	1/3	1	7
$\Delta(\|\boldsymbol{r}\|-\|\boldsymbol{c}\|)_{\mathbf{1,2}}$	1/6	1/7	1

Time 1-2	$\Delta\boldsymbol{\theta}_{\boldsymbol{r,p}}^{(1,2)}$	$\Delta\boldsymbol{\theta}_{\boldsymbol{p,c}}^{(1,2)}$	$\Delta\boldsymbol{\theta}_{\boldsymbol{r,c}}^{(1,2)}$
$\Delta\boldsymbol{\theta}_{\boldsymbol{r,p}}^{(1,2)}$	1	3	5
$\Delta\boldsymbol{\theta}_{\boldsymbol{p,c}}^{(1,2)}$	1/3	1	6
$\Delta\boldsymbol{\theta}_{\boldsymbol{r,c}}^{(1,2)}$	1/6	1/6	1

Given the number of splits within the evaluation period and the comparison matrices for time interval 1-2 and 59-60 are known, the increments needing to be applied to each distance and magnitude measure at each time-interval are also known. For example, $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{1,2}$ is *more important* than $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{1,2}$, with an AHP scale of 6, and $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{59,60}$ is *absolutely more important* than $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{59,60}$, with an AHP scale of 10. Therefore, given there is a difference of 4 scales (10-6) between these two rates of change measures, over the 59 time-intervals, the pre-defined step change applied to the AHP [for $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)$ and $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)$] is 4/59 = 0.068. At each time interval of the evaluation the AHP will experience an incremental increase of 0.068, and at the end of the evaluation period it converges to an AHP scale importance of 10 (i.e. $\Delta(\|\boldsymbol{p}\| - \|\boldsymbol{r}\|)_{59,60}$ vs. $\Delta(\|\boldsymbol{r}\| - \|\boldsymbol{c}\|)_{59,60}$). This linear increment is pre-defined and applied to each rate of change measure and at each time interval the comparison matrices are updated accordingly.

These changes are pre-defined and applied to each metric as the evaluation period matures. Accordingly, as these changes are applied a spherical score is recalculated for each time-interval. Figure 31(a) and 31(b) compares the log-loss metric against the DMS metric to assess the predictive power of the probability winning model (please Patel, Bracewell & Wells (2018)).

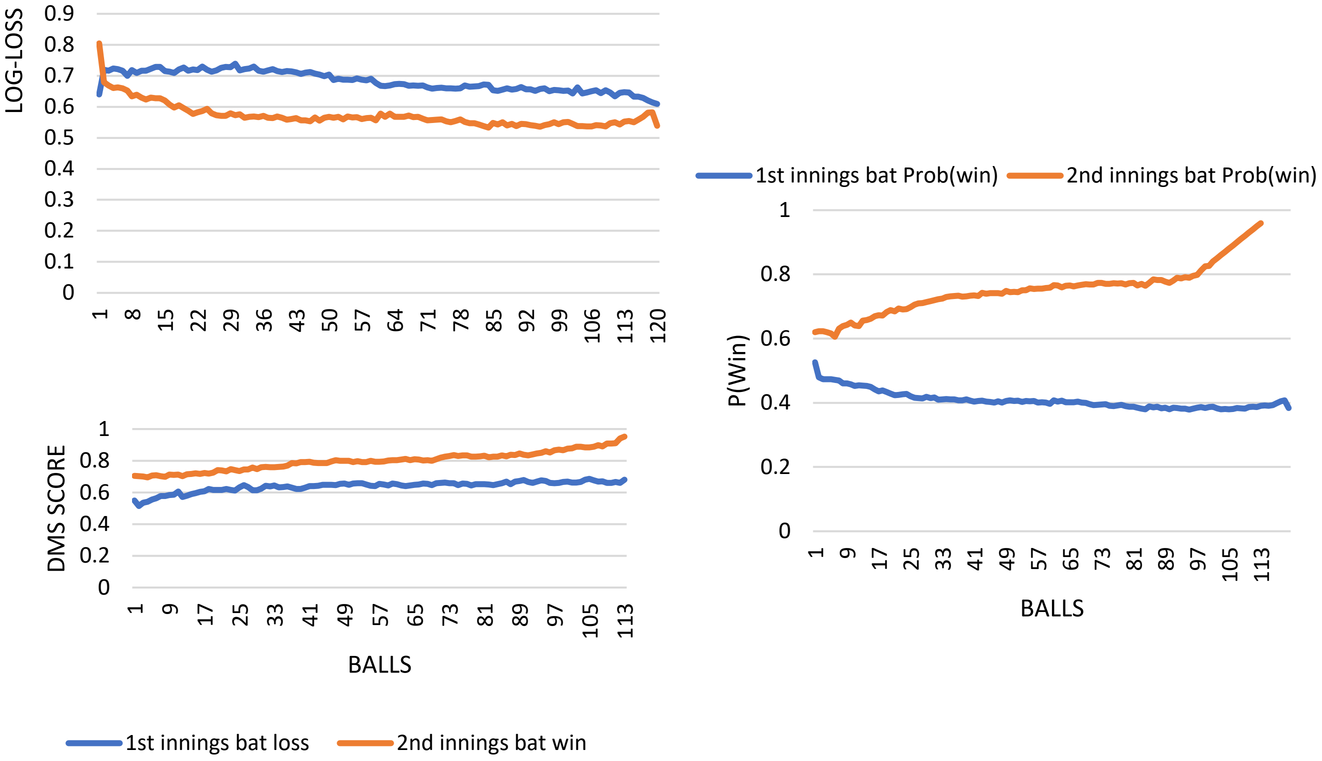


Figure 8(a), 8(b) and 8(c): spherical scoring vs. log-loss and probability of winning for first and second innings batting team, respectively.

To demonstrate the effectiveness of the DMS metric, a case study is examined. Here, the first semi-final from the Australian Big Bash 2018 between Hobart Hurricanes and Melbourne Stars on 14th February 2019 is explored (https://www.espncricinfo.com/series/).

The Hobart Hurricanes batted first and got off to a relatively poor start losing two wickets in the second over, leaving them 5 for 2 after 2. The third wicket put on an additional 37 runs prior to the major contribution coming from the 4th wickets, where the score advanced to 117 for 4, after 16 overs. The final total of 148 is arguably below par, given a $p(win)$ for the team batting first of 0.40. Intuitively at this stage the team batting second, the Melbourne Stars, would be picked as the favourites, aligning with the model.

Reflecting on the insights, observable in Figure 8(b), by the spherical scoring rule, there is relatively rapid movement in the first two overs of the first innings with the metric going from 0.51 after ball 2 to 0.59 after ball 12 and fall of the second wicket. Conversely, the log-loss score does not begin to move after the fifth over, with a gradual decline from 0.74 to 0.68 after 56 balls (end of eight overs). This indicates slight improvements in score from an interpretation of the log-loss as oppose to the early and meaningful escalation of the spherical score which provides greater insight into model performance relative to match context.

Similarly, in the second innings the spherical prediction is approximately monotonically increasing from ball one, going from 0.7 to 0.95. Conversely, the log-loss drops only within the power-play indicating an improvement in model prediction from 0.68 to 0.58 at the end of over 4. It continues to hover around 0.55 for the last 10 overs. However, with the Melbourne Stars chasing down the total only 4 down within 19 overs indicating they were comfortable throughout the chase. Consequently, its expected probability is high, this demonstrates the modified spherical scoring metric rewards the expected runs model based on its ability to predict accurately from long-range. This indicates the methodology is providing reliable, intuitive, robust, and transparent outputs.

Figure 8(a) and 8(b) shows that the spherical scoring rule outperforms the log-loss during the middle and latter stages of the evaluation period, given the meaningful change in the metric relative to the match context. Between overs 1-4 a power-play phase is conducted and during this time of the match, a lot of uncertainty, especially in the first innings, is present due to scoring rate which affects match volatility. The figures clearly show both metrics indicate model performance is improving as more information is obtained as the game advances.

This is represented in the latter stages of the spherical score, Figure 8(a), as the trend converges faster to what happened in the second innings, while the first innings trend remains relatively flat. Further. Compared to the log-loss predictions across both innings the modified metric converges at a faster rate to the actual outcome, revealing a better use of information.

In both instances the log-loss and spherical metric predictions are best in the 2nd innings, which is expected as more reliable or informed data is available because the first innings has passed. Figure 8(a) reveals that the modified metric is better at utilising new information and converges faster to actuality relative to the log-loss metric, this is due to the assigned weights representing the forecasting scenario.

These results reveal that the modified spherical scoring rule is an appropriate metric to assesses the effectiveness of human-based ratings and outperforms well-known performance metric of the log-loss.

3.12 DISCUSSION AND CONCLUSION

Given the lack of a modelling framework to construct sport-based rating systems, this chapter endeavoured to plug this gap with a robust methodology that produces reliable, robust, transparent, and intuitive ratings; also referred to as 'meaningful' ratings.

Through the literature review and the findings established during the development of multiple sport-based rating systems (please see Appendix A), it was identified that sport-based rating systems implement five key elements: 1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling procedure to produce an overall rating of performance. Given these findings, a framework which applies these elements was developed.

Specifically, the framework is a form of model stacking where information from multiple models is combined to generate a more informative model and applies a dynamic multi-objective ensembling forecasting strategy. An ensemble approach was adopted as it is assumed that sporting performances are a function of the individual traits that significantly affect performance.

The framework does not necessarily produce the 'optimal' rating system; however, it outlines the process to construct sport-based rating systems which produce meaningful ratings. It is hypothesised that a meaningful rating system identifies the important attributes for each significant trait, with respect to performance, and applies an ensemble strategy to construct meaningful ratings. This hypothesis is tested in Chapter Four and Chapter Five by applying the ratings framework within the sporting context to assess its applicability and validity. These sport-based rating systems attempt to show that the framework leads to meaningful outputs. The rating systems developed in Chapter Four and Five have been published in a peer-reviewed journal and conference proceedings, respectively.

This chapter also developed a novel performance metric, known as the DMS metric, which applies distance and magnitude measures derived from the spherical scoring rule to assess the effectiveness of rating systems. The DMS metric accounts for five criteria which leads to effective evaluation of meaningful ratings. This is achieved by 1) evaluating the distance between reported ratings, actual outcomes and averaged forecasts, 2) measuring the distance between ratings across different time-frame, 3) providing an incentive for well-calibrated and sharp ratings, 4) accounting for the context and the difficulty of the forecasting scenario and 5) evaluating ratings on the entire probability distribution. Based on these criteria and given that ensemble forecasts are generally assessed on calibration and sharpness, a proper scoring rule methodology is applied to construct the DMS metric. Specifically, distance and magnitude-based measures derived through the spherical scoring are applied to develop the DMS metric.

The DMS was developed using three rating vectors $\boldsymbol{r}$ (modelled ratings), $\boldsymbol{c}$ (benchmarked ratings) and $\boldsymbol{p}$ (actual ratings) and their corresponding rate of change metrics based on magnitude and angular difference.

An AHP was applied to assign importance weightings to each vector magnitude and angular distance measure. These weights were linearly updated using a pre-defined step parameter value until the evaluation period fully matures.

Using the probability of win model presented in Patel, Bracewell & Wells (2018), the DMS metric was benchmarked against the log-loss scoring rule. The DMS metric was shown to outperform the log-loss during the middle and latter stages of the evaluation period, however it performed equivalently during the earlier stages. This is mostly likely to due to the weightings assigned to the rate of change magnitude and angular metrics. Chapter Five extends this validation beyond a single cricket match and applies it to 400 matches.

Chapter Four and Five applies the ratings framework within the sporting context to build novel rating systems, at both the team and individual level, to evaluate team and player performance within a cricketing context, respectively. To demonstrate the effectiveness of the DMS performance metric, a cricket-based case study is also examined in Chapter Five. The effectiveness of the rating system (presented in Chapter Five) to output meaningful sport-based ratings of performance is validated using the DMS metric. Finally, the validity of the DMS metric is measured by comparing it against the log-loss when assessing the probability of win and player rating models (research objective (iii)).

PART TWO: APPLYING THE RATINGS FRAMEWORK AND DMS METRIC

Chapter Four

ESTIMATING EXPECTED TOTAL IN THE FIRST INNINGS OF T20 CRICKET USING A NOVEL RATINGS FRAMEWORK

> *"Essentially, all models are wrong, but some models are useful".*
>
> George Box (1976).

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

4.0 INTRODUCTION

The primary goal of this chapter is to evaluate the sport-based ratings framework and its ability to produce meaningful, i.e. robust, reliable, transparent, and intuitive, ratings, specifically to assess team performances within the cricketing context. It is shown that systems that ensemble features of varying complexities from different dimensions of the ratings scenario produce superior predictions than systems which only apply traditional statistics. Specifically, this chapter applies the ratings framework within the cricketing context to build a run prediction model which outputs meaningful ratings of team performances (i.e. expected total). These team ratings are interpreted in terms of runs the first innings batting team is expected to score, at any stage of the innings.

Current run prediction systems utilised within limited overs cricket suffer from two model issues: 1) overly broad match representation metrics and 2) inability to account for contextual match factors. In this chapter a gradient boosted model (GBM) ensembling strategy is developed to account for these two issues. The model outputs are benchmarked against a popular media tool, dynamic programming model (DPM), and actual first innings runs scored. The results show that the developed model converged to actual first innings total faster than the DPM and the winning and score prediction system (WASP). Importantly, the GBM model outperformed the DPM and WASP across several statistical accuracy metrics. The proposed run prediction model utilises traditional metrics, such as current runs and wickets, match-level metrics, such as runs remaining, resource-based metrics, such as resources remaining, and team (or innings) specific metrics relating to the batting and bowling teams, such as percentage dot balls and percentage boundaries. Conceptually, these attributes begin to account for environmental and team specific factors. The improvement in accuracy whilst maintaining a simplicity of deployment suggests that maintaining contextual information and *intuition* within an estimated runs model is appropriate for limited overs cricket. The results show that ensembling metrics or predictions from different match dimensions, such as player and environmental, produce more predictive and *interpretable* run estimates than models that only consider macro-perspectives such as Duckworth-Lewis model and the dynamic programming model. This indicates model *robustness* and *reliability*.

The ratings framework outlined in Chapter Three is applied to develop a run prediction system which implements different attributes from various dimensions of a cricket match to derive ball-by-ball run predictions, indicative of team performance (i.e. team-based ratings). Using the ratings framework, it is assumed that a model which ensembles traditional, resource-based and match-level metrics result in better predictions than models that either 1) do not incorporate ensembling techniques and 2) only utilise environmental-level or traditional (or shallow) metrics such as wickets, current total and balls. Similar ensemble strategies have been applied in the baseball sabermetrics literature focussing on run prediction. The literature states that models which utilise both traditional and complex metrics generate better run predictions; for example, "A mixture of conventional independent variables and sabermetrics independent variables would be broad enough to find models to correlate highly with run production and run preventions" (Benevenatno, Berger & Weinberg, 2012, p. 67).

The primary goal of this chapter to use the ratings framework to develop a real-time team-based rating system that predicts the expected total of the first innings batting team, at each stage of an innings, and to demonstrate the applicability of the ratings framework to construct sport-based systems that produce meaningful ratings of team-based performances. By demonstrating the applicability of the framework, this chapter addresses research objective (iii). Specifically, it demonstrates the need to apply 1) dimension reduction and feature selection techniques, 2) feature engineering strategies, 3) multi-objectives, 4) team-based variables and 5) ensemble forecasting strategies, to construct meaningful sport-based rating systems. Overall, the aim is to show that the ratings framework produces highly predictive expected runs, indicative of team performance[9,10].

4.1 RUN PREDICTION SYSTEMS IN CRICKET

T20 cricket is a dynamic and fast paced game where the team's prospects of winning can change within a few balls. This allows players to significantly influence match result off fewer deliveries relative to longer formats. Consequently, each ball carries more weight as it represents a greater proportion of the match. Although, this introduces a greater level of uncertainty when predicting results as only a small number of balls are necessary to the change match situation. An area of considerable uncertainty is the number of runs the batting team is expected to score in the first innings. It is hypothesised that ball-by-ball predictions of the first innings total can be improved by using match level metrics, such as resources remaining, or shallow metrics such as current total, with team inning metrics, such as percentage dots to produce better first inning run predictions than a model that only considers shallow metrics. The rationale is that some of these within game descriptive actions will encapsulate information about playing conditions.

Traditional run prediction models do not consider the complex interactions existing between resource-based, traditional, match level and team-specific (or inning specific) metrics, due to data inaccessibility, implementation issues, inability to produce intuitive results that are understandable by players and coaches and inability to implement offline.

[9] Patel. A. K., Bracewell. P.J., & Bracewell, M.G. (2018). Estimating expected total in the first innings of T20 cricket using gradient boosted learning. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

[10] This paper was awarded the Neville De Mestre Prize for best student paper at the 14th Australian Conference on Mathematics and Computers in Sport (MathSport) conference.

Refining estimates using data that is descriptive of actions within the innings is useful for applications in coaching, strategy, and entertainment. Although, they are not suited to adjusting totals for defining the formal outcome of a match, where the Duckworth-Lewis-Stern is used (Stern, 2016). The now defunct Indian Cricket League used the VJD method, developed by Jayadevan (2002). The outputs from targeting setting and readjusting models are subject to scrutiny and can have a bearing on match and tournament outcomes, thus tremendous rigour must be applied to ensure fair results.

Consequently, forecasting totals is an area that has received considerable attention. Notable research in run prediction in limited overs cricket include: Duckworth and Lewis (1998), Stern, (2016), Jayadevan (2002), Ovens and O'Riley (2006), Brooker & Hogan (2011), Clarke (2000), Scarf, Akhtar & Shi (2010), Kaluarachchi & Varde (2010), Bailey & Clarke (2006), Bandulasiri (2004), Jhawar & Pudi (2016), Asif & McHale (2016), Davis, Perera & Swartz (2015) and Shah, Jha & Vyas (2016).

The Duckworth-Lewis-Stern (DLS) system is the most famous of this research (Duckworth and Lewis, 1998; Stern, 2016) with the primary function is to reset the target total during interrupted matches of limited overs cricket. Importantly, the DLS system can also be used to produce first innings run predictions for uninterrupted matches, with the output embedded in live scorecard publication tools and websites like crichq.com and nzc.nz. This elegant method, which is well entrenched in club, domestic and international cricket due to simplicity of use, is well described in both academic and popular literature (e.g. espncricinfo.com). The premise of the method is that batting teams have two resources to produce runs: balls and wickets. This two-factor relationship is then used to calculate the average number of runs that can be scored given the remaining resources.

Clarke (1988) applied a dynamic programming model to one-day cricket to: 1) calculate the optimal scoring rate, 2) estimate the total number of runs to be scored in the first innings and 3) estimate the probability of winning in the second winnings. The first innings formulation generated a team's optimal scoring rate to obtain a given total, given the number of wickets lost and balls. The second innings formulation generated a probability scoring table outlining the probability of the second innings batting team achieving the target total, given the number of wickets lost and balls remaining. Ovens and O'Riley (2006) evaluated the ball-by-ball run prediction ability of four models: Average Run Rate, PARAB, Duckworth Lewis (D/L) and Jayadevan. Results showed that the D/L method had the strongest predictive power, predicting 4.50 runs below the actual total, followed by ARR with prediction 17.29 runs below the actual total, Jayadevan with 31.13 runs below the actual total and PARAB 41.60 runs below the actual total. Similarly, Brooker and Hogan (2011) utilised a dynamic programming model to develop a Winning and Score Prediction (WASP) system for limited overs cricket. The system produces predictions using factors such as pitch conditions, weather, boundary size and the quality of the batting team and bowling attack. The WASP works backwards to solve inning specific models. The first innings model produces ball-by-ball prediction of the runs scored, while the second innings model calculates the probability of the batting team reaching the target total and therefore winning.

Swartz, Gill and Muthukumarana (2009) developed a discrete generator simulator, as there is finite no. of outcomes that can occur for any given delivery, for one-day cricket. Applying a Bayesian Latent model, ball-by-ball outcome probabilities were estimated using historical ODI data and were dependent on batter, bowler, total wickets lost, total balls bowled and current match score. It was found that the proposed simulator produced reasonably realistic results, with actual runs and simulated runs revealed an excellent agreement. Ovens & Bukiet (2006) developed a Markov chain approach to predict the runs scored for a given batting line-up. Realising that the interaction between bowler and batter is the primary factor dictating the dynamics of run production, a match was modelled as a sequence of one-on-one interactions, through a multi-dimensional matrix, M, with entries (b, r, w, b_1, b_2) representing the number of balls, runs scored, wickets lost, and the striking and non-string batter, respectively. The probability of being in any given state was calculated, for any given number of balls, by multiplying M, representing the set of probabilities after $b-1$ balls, by the probability of each event (i.e. number of runs scored off any given ball). Simulation resulted in a runs distribution table and "summing the product of each possible number of runs and its probability of being the result for the match gives the expected number of runs for the batting order considered" (Ovens *et al.*, 2009, pg. 497).

Jayadevan (2004) developed a method for resetting the target total during an interrupted limited overs cricket match. A normal score represented a team's general scoring pattern, while the target score represented a team's ideal scoring pattern too achieve the target score. Regressing cumulative percentage runs on cumulative percentage overs it was found that a cubic polynomial equation of order 1 represented a team's scoring pattern (a similar approach was adopted by Mansell, Patel, McIvor, and Bracewell, 2018). Moreover, the effect of a wicket was incorporated into the model by examining the pattern of wickets fallen. Applying the model produces a "target runs" percentage table that allocates a proportion of runs that needed to be scored by the batting team during any stage of the second innings.

Swartz, Gill, Beaudoin & deSilva (2004) used simulated annealing to conduct a search over a space of permutation of batting orders to find the optimal or near optimal first innings order. A first innings run simulator was built using a Bayesian log-linear model to generate ball-by-ball outcomes. The model was applied to the 2003 India World cup squad and posterior estimates of the parameters were obtained by averaging output from a Markov chain. Simulating 71,000 first innings runs using India's 2003 World cup final batting order a good fit between actual runs and simulated runs was found. Overall, it was found that the optimised batting order produced 6 more runs than that of the actual batting order.

Singh, Singla and Bhatia (2015) developed a first innings run prediction model and a second innings match outcome probability model for one-day cricket by applying linear regression and Naïve Bayes classifiers for each innings applied in 5 over intervals. The first innings model used current run rate and wickets fallen, while the second innings used current run rate, wickets fallen and target score. The error produced by the linear regression classifier were less than a current run rate projection method and the Naïve Bayes classifier had an accuracy of 68% in the 0-5th overs, increasing to 91% between the 40-45th over.

Bracewell *et. al*. (2014) generated team ratings where margin of victory was represented in terms of runs only. Like the approaches outlined previously, the resources available at the end of the second innings were used to determine a likely final total if the innings continued until all resources were consumed. This was used to generate team ratings that outperformed popular opinion for result prediction.

The hypothesis of combining conventional and advanced metrics builds on sabermetrics literature stating that run prediction models that utilise both conventional and advanced metrics generate better predictions.

4.2 METHODS

Using the ratings framework, the intent is to show that ensembling traditional (i.e. current total and balls etc.), resource-based (i.e. resource remaining), match level and team-based metrics (i.e. percentage dots and percentage boundaries) produce better first inning run predictions in T20 cricket than models that only consider traditional (or shallow) metrics. It is anticipated that team-based specific metrics inherently include information relating to environmental, situational, and competitive factors. For example, high percentage boundaries could indicate either poor bowling, good batting, favourable batting conditions or any combination of these factors. Due to the large variations between balls in T20 cricket, ball-by-ball metrics that capture this variation must be utilised to produce predictive outputs. The traditional and less informative metrics produce less informative model outputs. Therefore, advanced, and more informative metrics must be adopted to significantly explain the underlying variation to produce accurate predictions. In effect resulting in non-meaningful outputs. Therefore, a modelling framework, such as ensembles, that consider these subtle nuisances and capture complex interactions must be applied to output meaningful ball-by-ball run predictions.

It is assumed that models which ensemble traditional, resource-based match level, and team specific metrics produce better run predictions than models that only utilise match level or traditional metrics, such as wicket, current total, and balls. This assumption has been extracted from baseball sabermetrics literature focusing on run prediction, which is a well-researched and documented problem within baseball. Effectively, the literature states that models which utilise both conventional and advanced metrics generate better run predictions. For example, "A mixture of conventional independent variables and sabermetrics independent variables would be broad enough to find models to correlate highly with run production and run preventions" (Benevenatno, Berger & Weinberg, 2012, p. 67). For more readings on baseball run prediction please see Bukiet, Harold & Palacios (1997); Freeze (1974); Beneventano, Berger & Weinberg (2012); Cserepy, Ostrow & Weems (2013). Only first inning run prediction models are considered as there is less information available regarding what is expected to be a winning total (which is known in the second innings).

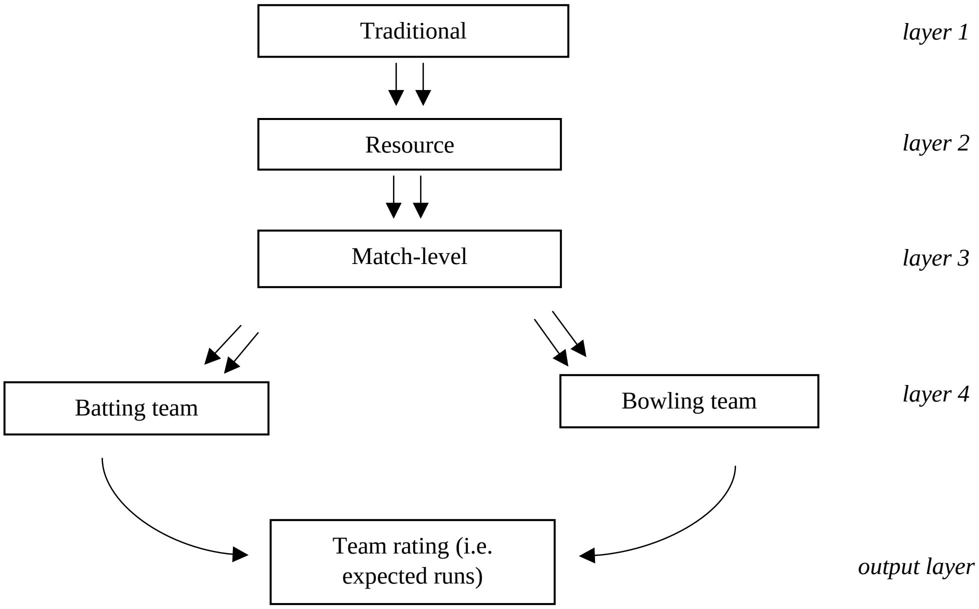


Figure 8: Run prediction framework

Figure 8 outlines the adapted ratings framework to develop a run (i.e. team-based) performance prediction system that accurately forecasts the expected number of runs in first innings of a T20 match, at any stage of the first innings. The run prediction framework applies various metrics from different dimensions of a cricketing match, these metrics are categorised into five dimensions: resources-based, traditional, match-level, team-based (i.e. bating and bowling dimensions). Each of these dimensions represent different elements of a cricket match, and within each dimension there are multiple metrics of varying complexity.

Ensembling metrics of varying levels of complexity from each dimension produces better predictions of expected run than traditional models such as the WASP system which only utilises match level, such as wicket, current total, and balls. Moreover, the framework will account for complex interactions between metrics across the five dimensions of interest to produce highly accurate predictions.

The objective of each layer of the run prediction framework is as follows: 1) the traditional layer measures a team's ability to score runs based on traditional statistics such run rate, wickets, and current total. 2) The resource layer derives resource-based metrics which measures the proportion of resources at the batting team's disposal from which to accumulate runs, at each stage of an innings. 3) The match layer considers match-level statistics, such as percentage dots and percentage boundaries to derive expected runs. 4) The team-level layer (i.e. batting and bowling team) using metrics such as projected total (i.e. *current runs/ resources remaining*), strike rate, and team economy rate (i.e. *current runs/total balls*).

The shallow (i.e. initial) layers of the framework consider match-level, traditional and resource-based metrics to build a better understanding of the team's run scoring ability and the complex interactions metrics which explain state of play. Team-based metrics are derived from resource metrics, for example, projected total is calculated using the resources-remaining feature. Finally, ensembling these metrics using a modelling function which accounts for interactions is applied to reduce variation in predicted run and produce highly predictive and meaningful run predictions.

Each dimension of the run prediction framework quantifies the state of play on a ball-by-ball basis and explains how team-level contributions affect the expected total. At each layer of the ratings framework action, context and time-based attributes are applied. All features are time-based as they are derived on a ball-by-ball basis. Table 9 outlines the metrics that will be used in the ensembled model, and the attribute-type and layer to which they relate.

Performance metrics	**Associated layer**	**Attribute group**
Projected total	Match-level	Action/ Time
Team adj. strike rate	Batting	Context/ Time
Current run rate	Match-level	Action/ Time
Wickets	Traditional	Action/ Time
Percentage dots	Bowling	Context/ Time
Percentage boundaries	Batting	Context/ Time
Resource remaining	Resource	Context/ Time
Balls bowled	Traditional	Action/ Time
Current runs	Traditional	Action/ Time
Percentage extras	Bowling	Context/ Time

Table 9: Metrics used in the ratings model by layer and attribute-type

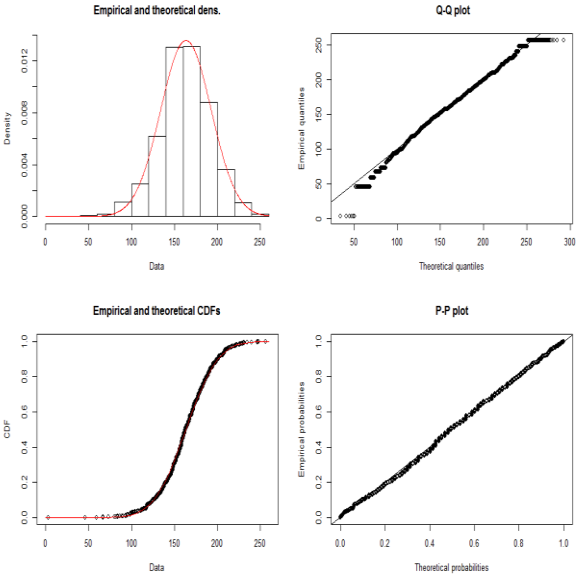
4.3 DATA

Model development utilised ball-by-ball observations from the following T20 competitions: Indian premier league (IPL; 2015, 2016, 2017 and 2018), Australian Big Bash League (BBL; 2016-2017 and 2017-2018), English NatWest T20 league (2015, 2016), South Africa Ram slam (2016 and 2017), Caribbean Premier league (CPL; 2014, 2015, 2016 and 2017) and New Zealand Super Smash League (2017-2018).

A process was developed to programmatically extract and parse ball-by-ball observations from ESPNCricinfo (http://www.espncricinfo.com/) commentary logs and provide a more convenient data structure (using the R programming language). The process extracted relevant data on a ball-by-ball basis and stored the data in a tabular form for easy access.

Overall, the dataset contained 85,700 1st inning ball-by-ball observations from 704 matches. Model development utilized 50% of the data for training, 25% for testing and 25% for validation.

Figure 9 shows that the underlying distribution for first innings total can be well approximated by a normal distribution. Given this finding a normal distribution will be used to during the modelling process. Exploratory analysis found the average first innings runs scored = 160, while the average first innings winning total = 168. Figure 10 illustrates the evolution of first innings total since 2014. There is an upward trend, with runs experiencing an average yearly increase of 4.75.



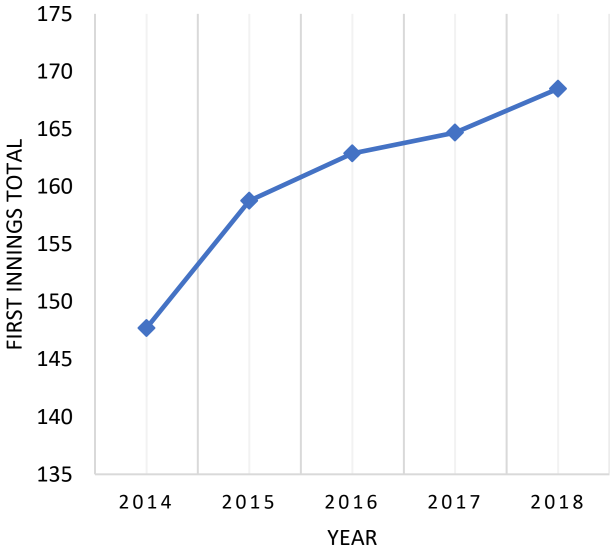


Figure 9 and 10: Distribution of first innings total and Avg. 1st innings total (2014 - 2018)

4.4 GENERALISED BOOSTED REGRESSION MODELING

The proposed model uses a gradient boosted regression technique (GBM) to account for the complex interactions between match and inning level metrics by taking a sequence of weak leaners to construct a complex leaner – increasing model complexity. The initial learners fit simple model and then the weighted combinations can grow more and more complex as learners are added. This produces regression models consisting of a collection of regressors. Learners do so sequentially with earlier stages fitting simple models to the data and analysing the errors. Latter models focus on trying to account for as much error as possible. The models are given weightings and the different models are combined into an overall predictor. Moreover, the gradient boosted method serves as a dimension reduction technique to identify the relative importance of each performance metric, allowing the evaluation of metric importance and elimination of uninformative metrics.

The proposed model was built in *R* using the *gbm* package and incorporates the following parameters: 1) distribution = Poisson – runs scored is a count outcome, 2) n.trees = 20,000 - optimal number of tree for out-of-bag variance, 3) interaction.depth = 5 – 5-way interaction to capture complex variable relationships and 4) shrinkage = 0.0001 – step-size learning rate. The combination of weak-leaners, to build a complex learner, that incorporate a 5-way interaction effect will slowly start to reduce the error in first innings total. Ultimately, the new complex learner will account for greater variation and understand the complex interaction between match and inning-specific metrics. The metrics included in the model: projected total $(i.e. current\ total\ /\ resources\ remaing)$, team strike rate, run rate, current runs, wickets, percentage dots, percentage boundaries, resources remaining and balls. These metrics contain match level, batting, and bowling-level performance information. Team specific metrics included in the model are percentage dots, percentage boundaries, percentage extras, strike rate and economy rate. These metrics inherently store information about the interactions between the bowling and batting environment. The gradient boosted technique considers the interactions across these ‘meta-information’ rich metrics to gauge match-level understanding.

4.5 RESULTS

The GBM model was benchmarked against the Dynamic Programming Model (DPM) outlined in Clarke (1988), and the model predictions were evaluated against actual runs scored. The ball-by-ball predictions and were aggregated to an over. A relative importance analysis revealed projected total, team strike rate, percentage boundaries and run rate as the 3 most important metrics. These results show that in T20 cricket the first inning total is heavily dependent on efficient run production. Specifically, run production is dependent on the volume of runs scored per percentage of resources used. Moreover, the analysis reveals the metrics that are utilised by the dynamic model: i.e. current total, balls and wickets are contained within the top 3 important metrics: 1) balls influence team strike rate, 2) current total influences team strike rate and project total and 3) wickets influence resources remaining, which is also present in projected total. This indicates that although current total, balls and wickets are important to evaluate the expected total in the first innings they are more informative when combined with other metrics to explain a greater proportion of variation. Model performance was evaluated using two measures: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). RMSE has the benefit of penalising large errors, while MAE has interpretative power and only describes error. Figure 11 illustrates the over-by-over predictive accuracy of the two models, measured against RMSE and MAE.

These run predictions (i.e. team ratings) are intuitive and transparent, as the results can be mapped to real-world observable outcomes and the context to which the system is being applied, and are interpretable and easily communicated, respectively. Moreover, the predictions are reliable and robust, as they yield good performance during different stages of the first innings, and are well-calibrated and sharp, respectively.

Figure 11 illustrates that on average the proposed model out-performs the dynamic program on an over-by-over basis across both RMSE and MAE. A bootstrapped sample of the over-by-over performance measures created confidence intervals. A statistically significant difference was found between the performance metrics for the two models: GBM and DPM (α == 5%). This statistically significant difference between the models for the performance measures existed up until the 13th over (~78 balls) suggesting that for 65% of the first innings of a T20 match the GBM model produced statistically better results than the DPM. Although, examining the performance measures on a match-by-match basis revealed instances where the DPM produced better predictions. On average, the DPM produced better predictions for low scoring matches (i.e. $\leq$ 158). This could be because in low scoring matches metrics such as current total, balls and wickets metrics have a greater impact on expected total, while efficiency metrics (i.e. percentage dots and percentage boundaries) are of lesser importance. A 30% increase in predictive accuracy is observed between the 5th and 10th over. Although, the prediction error during this period is large given that the data does not contain sufficient match information. Surprisingly, model accuracy experiences a decrease in the 12th over. It was found that between overs 6 -10 the batting team the run rate, percentage boundaries and innings strike rate are relatively constant. Although, in overs 11-13 these metrics begin to experience a steady increase, indicating that the batters are starting to pick-up and increase aggression. It is assumed that the GBM and DPM fail to effectively account for this sudden increase in batting intensity. The 12th over is where the difference between RMSE and MAE becomes statistically insignificant, indicating that after the 12th-13th over enough match information is known, therefore both models are producing similar results and both models are extracting similar information from the metrics relating to the first innings expected total.

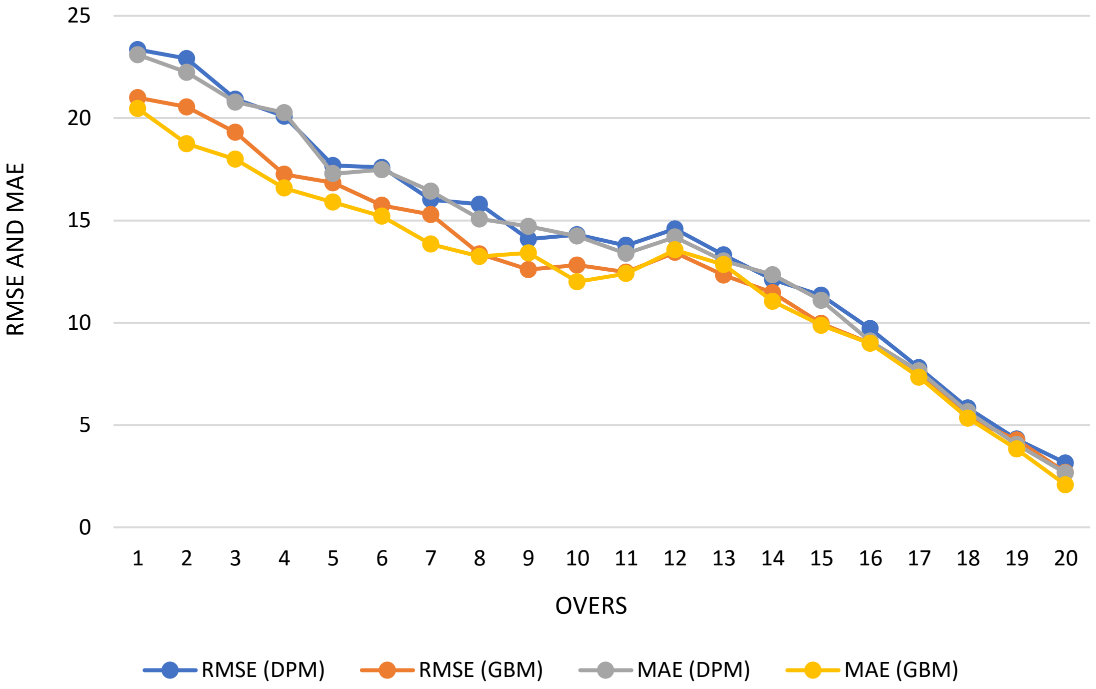


Figure 11: GBM vs. DPM performance measurements

4.6 HYBRID MODEL: GBM using DPM

Given the dynamic model generates predictive results and produces better prediction for low scoring first innings, the GBM model was updated using the DPM predictions as an input metric. This hybrid model did not produce prediction improvements as the metrics that are present in the DPM are already present in the GBM. As stated, current total, balls and wickets are included in the GBM model in a meaningful manner, such that more information regarding match-state is incorporated. Therefore, the DPM metric is not introducing any new information into the proposed model and introducing confounding issues. An importance analysis revealed DPM metric was the most important metric. This is expected as the DPM combines three conventional metrics in a meaningful way to produce a more informative metric. This suggests that combining weak predictors, in a meaningful manner, creates a stronger predictor that explains a greater proportion of variation.

Implementing a GBM ensembling approach into the run predictions framework produces highly predictive and meaningful team ratings, which is represented by their expected total at each stage of the innings. The run predictions are 1) robust - the prediction yield good performance where data is drawn from a wide range of distributions that are largely unaffected by model assumptions. 2) Reliable – the predictions produce accurate and highly informative forecasts which are well-calibrated and sharp. 3) Transparent – the predictions are easy to interpret and communicate. Each ball-by-ball prediction illustrates the number of total numbers of runs the batting team is expected to score. 4) Intuitive - each ball-by-ball prediction can be mapped back to a cricketing context, for example, suppose at $ball_i$ a batter hits a boundary four, increasing the expected total from 125 to 132, between $ball_i$ and $ball_{i+1}$, it can be said that the value of that boundary on the first innings expected total was 7 runs (i.e. 132-125).

4.7 DISCUSSION AND CONCLUSIONS

Although the proposed model produced better results, there are scenarios where the dynamic programming model produced better predictions. The DPM produced better outputs in low scoring matches (i.e. ≤ 163 runs) where the batting team had a ‘slow’ start. This scenario arises because the proposed model considers metrics that are relatively more important in high scoring matches, such as percentage dots and percentage boundaries and therefore is more sensitive to performances that significantly affect or deviate the slope of the expected total. Figure 10 illustrates the evolution of first innings total since 2014. Overall, there is an upward trend across time, although this seems to flatten out. Although recently (2017-2018) there has been a small increase in gradient. Assuming the trend continues it is assumed that the proposed model will continue to outperform the DPM as scores will continue to rise, meaning the latter model will continue to produce more varied predictions over time as it fails to accommodate for highly sensitive metrics that significantly affect expected total and matches where more than 163 are scored in the first innings.

The results confirmed the hypothesis that a model that utilises both meta (i.e. match-level) and shallow metrics, and advanced metrics will output better predictions than a model that only utilise meta and shallow metrics. This shows that advanced metrics store additional information, and combining shallow metrics reveal features that account for additional proportion of variation. Moreover, ensembling weak predictors creates stronger, more complex predictors that explain a greater proportion of variation than its individual counterparts. At each layer of the ratings framework action, context and time-based attributes are implemented to produce highly predictive run predictions. These three attribute-types are necessary to produce meaningful output because combining action, context and time-based attributes create trait-based ratings which result in meaningful performance-based ratings (please see Chapter 3 section 3.4.1).

Current models do not consider complex interactions existing between innings specific and match level metrics. Therefore, a modelling technique that incorporates these subtle nuisances and interactions is expected to produce more accurate predictions. The literature relies heavily on meta-level metrics such as pitch conditions, boundary size, and shallow team metrics such as batting and bowling characteristics that fail to incorporate inning dynamics and capture the interaction effects between players and team metrics. This novel methodology attempts to address these issues to dynamically predict the first innings total. The hypothesis that first innings predictions could be improved by using match-level, team and inning specific data was found to hold true in the first 12 overs over an innings.

Using the framework (Chapter Three), this chapter developed a novel team-based rating system, within the cricketing context, predicting the number of runs the first innings batting team is expected to score. Specifically, the results of this system are a meaningful representation of how a team is performing on a ball-by-ball basis. It is revealed that meaningful ratings of team performance (i.e. run predictions) are generated by 1) identifying the important dimensions of a cricket match using dimension reduction, 2) establishing the key feature within each of these dimensions, 3) apply time-based (i.e. ball-by-ball) features to understand match context and team performance on a ball-by-ball level, 4) considering the different objectives (or dimensions) of a cricket match and 5) applying an ensembling forecasting strategy (i.e. gradient boosted modelling) to effectively account for the complex interaction present within cricket and team performances. This chapter shows that rating systems must implement these key communalities to output meaningful ratings of team performance. Moreover, it was shown that different match dimensions must be applied to produce meaningful predictions.

It has been shown that the framework can be applied within the sporting context to produce meaningful ratings i.e. intuitive, robust, reliable, and transparent. It is concluded that the ratings framework allows the construction of rating systems which produce meaningful team-based ratings of performance within the sporting context.

The following chapter applies the ratings framework within the cricketing context, to develop a novel player-based rating system to evaluate the amount of influence a player exerts at each stage during a match of twenty-twenty (T20) cricket. Moreover, to demonstrate the effectiveness of the DMS performance metric, chapter five applies the DMS metric to a *probability of win* and *player rating model*, and benchmarks it against the log-loss metric (research objective (iii)) to prove that it better accounts for match context. The effectiveness of the player-based rating systems to output meaningful performance ratings is validated using the DMS metric.

From a team selection and commercial point of view, more accurate estimations of a first innings total provide interested parties with useful information for both strategic and entertainment purposes highlighting the value of deploying the GBM model in real-time.

Future research will benchmark the GBM system against the WASP model (Brooker *et. al.,* 2011, Shah *et. al.,* 2016*).* Although the WASP utilises team specific metrics such as the average team score, opposition's bowling performance, ground average score, these shallow metrics fail to capture match and inning information at a deeper level. Although metrics such as climate, pitch conditions and boundaries are important when predicting runs, this meta (i.e. match-level) information can be captured and stored in team and inning-specific metrics; for example, a high percentage of dots and quick depletion of resources indicating strong bowling attack, weak batting performance and/ or poor batting conditions. In addition, reviewing tournament specific model performance will also provide greater insight into the applicability of various models.

Finally, given that the DPM falls-over for high scoring matches (i.e. ≤ 158) and the first innings total is experiencing a 4.75 runs increase year-on-year, it is suggested that future research benchmark the two models year-on-year and observe the period in which the DPM outperforms the GBM. It is assumed that earlier seasons (2014 and 2015) the DPM would outperform GBM due to the low scoring first innings.

Chapter Five

DYNAMICALLY EVALUATING PLAYER INFLUENCE IN T20 CRICKET USING THE RATINGS FRAMEWORK

“Human behaviours can be accurately described as a set of dynamic models sequenced together by a mathematical or statistical function”

Alex Pentland (1999).

Pentland, A., & Liu, A. (1999). Modeling and prediction of human behavior. *Neural computation*, *11*(1), 229-242.

5.0 INTRODUCTION

Limited overs cricket is an ideal sport to isolate individual team member contribution. This is due to the availability and volume of machine-readable data, combined with the relatively isolated nature of the batter versus bowler contest observed per ball.

Cricket is a team sport based on the balance of two key resources: 1) *balls* and 2) *wickets*. Simply, the batting team that utilises these two resources most effectively wins the match. As an inning progresses, the total number of resources allocated to the batting team decreases. During this time the batting team aims to score as many runs as possible given an allocated number of resources, while the bowling team aims to restrict the total number of runs conceded, by taking wickets. The bowling teams' overall goal is to deplete the batting team resources as quickly as possible for the least number of runs.

The first innings batting team is assigned the task of maximising the total number of runs scored given two resource constraints: 1) *balls* and 2) *wickets*; while the second innings batting team is assigned the task of outscoring the first innings batting team, given the allocated resources. The first innings reaches completion when all resources are depleted, while the second innings reaches completion when all resources are depleted, or the batting team has achieved the target score. "The optimisation exercise in either team's task involves choosing some compromise between scoring fast and hence taking higher risks of losing wickets and playing carefully and hence risking making insufficient runs" (Duckworth & Lewis, 1998, pg. 220).

Cricket is intertwined with numerical values that ultimately translate to a match result. Although, given its numerical depth academic and commercial literature regarding the application of analytical techniques within cricket is limited. Historically this has been due to accessing data. Although, with rich online data sources such as ESPNCricinfo (http://www.espncricinfo.com/) this is rapidly evolving. The most notable application of analytics within cricket is the Duckworth Lewis (1998) resource allocation method. Duckworth and Lewis (1998) developed a framework which mathematically allocates resources to appropriately reset or recalculate target scores during interrupted one-day cricket matches. This system is currently implemented by the International Cricketing Council (ICC) as the primary method to recalculate the target score during an interrupted limited overs cricket match.

The primary goal of this chapter to use the ratings framework to develop a real-time individual-player rating system that accurately measures the amount of influence a player exerts on a T20 match of cricket, at each stage of an innings[11]. Overall, the aim is to show that the ratings framework produces meaningful ratings for player performance. The secondary goal of

[11] Patel, A. K., & Bracewell, P. J. (2019). Dynamic evaluation of player performance in T20 cricket. *Journal of Quantitative Analysis in Sport.*

this chapter is to apply the DMS performance metric developed in Chapter Three to the probability of win and player rating model, and benchmark it against the log-loss metric to prove that it outperforms the log-loss and better accounts for match context. Specifically, this chapter demonstrates the applicability of the developed ratings framework and novel performance metric within the sporting context. Therefore, addressing research objective (iii).

As the emphasis for this method is in real time, novel ball-by-ball metrics are built. To derive meaningful metrics, the ratings framework is applied. The ratings framework applies multi-modelling objectives to derive player specific metrics which are used to produce player ratings, quantifying the amount of influence a player exerts on a match of T20 cricket. Each model outputs ball-by-ball metrics which aim to explain a player's in-game contribution in terms three key dimensions: *volume of contribution*, *efficiency of contribution*, and *contributions made under pressure*.

To accurately measure a player's influence three key dimensions of a player's game must be considered: 1) volume of contribution, 2) efficiency of contribution and 3) contributions made under pressure. Many conventional performance metrics are inapplicable when evaluating in-play influence due to indefinability i.e. batting average is undefined until match completion, and event dependence i.e. bowling average and strike rate are undefined until a bowler takes a wicket (i.e. event).

Moreover, traditional metrics do not account for many match factors, such as opposition strength, venue, or pitch conditions. Therefore, before developing an influence model, performance metrics that are measurable on a ball-by-ball basis are engineered. These engineered metrics must incorporate the different dimensions of a player's game and capture their ability to affect match outcome. An influential player is one which can significantly shift the probability of winning in their teams' favour and contribute to team victory with quantity, efficiency and in pressure situations.

No such approach to evaluate real-time player influence, using a combination of volume, efficiency and pressure-based metrics, and match outcome, while the game is in progress, was identified in the academic literature. Although, it should be noted that real time tracking, and estimation is investigated in Akhtar & Scarf (2012), Bailey & Clarke (2006), Scarf & Shi (2005) and Bracewell (2015). Moreover, the idea of player efficiency was first introduced by Beaudoin & Swartz (2003) and the by Lewis (2005).

Before presenting the model, section 5.1 outlines the most notable research into the application of analytical systems in cricket. Section 5.2 describes the research methodology is described, followed by a description of the data used for model development and the covariates that will be experimented within the modelling. Section 5.4 provides an in-depth description for each of the models and assesses model accuracy and predictive power. Section 5.5 evaluates model fit diagnostics, provides a comparison of predicted probabilities with actual outcomes, and applies the DMS performance metric to the probability of win and player rating models. Section 5.6 concludes with closing remarks and discusses potential future work.

5.1 APPLICATION OF ANALYTICAL SYSTEMS IN CRICKET

As stated in Asif & McHale (2013) previous work in cricket has focussed largely on the problem of resetting target in the limited overs format following interruptions to play (please see McHale & Asif (2013); Duckworth & Lewis (2004); Duckworth & Lewis (1998); Bhattacharya *et al.,* (2011); Jayadevan (2002); Preston & Thomas (2002); Stern (2016)).

Although there are a small number of published articles on the application of analytics within cricket, there is increasing analytical literature and the adoption of predictive methodologies at the professional level. It has been noted that "during the past decade many academic papers have been published on cricket performance measures and predictive methods" (Lemmer, 2011, pg. 1). Moreover, there is increasing commercial demand for data-driven decision-making regarding topics such as player selection, in-game strategies. For example, Trent Woodhill claimed that the problem [with cricket] is, for all its obsession with numbers, the sport has yet to move onto a data obsession… Data is not something cricket has invested greatly in yet (ESPNCricinfo, 2017).

Critically, there remains an academic and commercial gap surrounding real-time, player rating systems. Proceeding is a review of the most notable academic literature outlining the application of analytical techniques to ball-by-ball cricketing data:

Clarke (1988) applied a dynamic programming model to one-day cricket to: 1) calculate the optimal scoring rate, 2) estimate the total number of runs to be scored in the first innings and 3) estimate the probability of winning in the second innings. The first innings formulation allowed the development an *'optimal scoring model'* outlining a team's optimal scoring rate (i.e. runs per over) to obtain a given expected total, for any given number of wickets lost and balls remaining. The second innings formulation enabled the development of a *'probability scoring table'* outlining the probability of the second innings batting team scoring the target total, for any given number of wickets lost and balls remaining.

Similarly, Davis, Perera and Swartz (2015) developed a T20 simulator that calculated the probability of a first-innings batting outcomes dependent on batsmen, bowler, and number of overs consumed and total wickets lost. These probabilities were based on an amalgamation of standard classical estimation techniques and a hierarchical empirical Bayes approach, where the probabilities of batting outcomes borrow information from related scenarios (Davis *et al.*, 2015). Simulation suggested that batting teams were not incrementally increasing aggressiveness when falling behind the required run rate.

Swartz, Gill and Muthukumarana (2009) developed a discrete generator simulator, as there are an infinite number of outcomes that can occur for any given delivery, for one-day cricket.

Applying a Bayesian Latent model, ball-by-ball outcome probabilities were estimated using historical ODI data and were dependent on batsmen, bowler, total wickets lost, total balls bowled and current match score. It was found that the proposed simulator produced reasonably realistic results, with actual runs and simulated runs revealing an excellent agreement. Moreover, comparing wickets taken, the actual results compared favourably with simulated results.

Bukiet & Ovens (2006) developed a Markov chain approach to predict the expected runs scored of a batting line-up. Realising that the interaction between bowler and batsman is the primary factor dictating the dynamics of run production, a match was modelled as a sequence of one-one interactions, through a multidimensional matrix, $M,$ with entries (b, r, w, b_1, b_2) representing the number of balls, runs scored, wickets lost, and the striking and non-striking batsmen, respectively. The probability of being in any given state was calculated, for any given number of balls, by multiplying $M,$ representing the set of probabilities after $b - 1$ balls, by the probability for each event (i.e. number of runs scored off any given ball). "Summing the product of each possible number of runs and its probability of being the result for the match gives the expected number of runs for the batting order considered" (Ovens et al., 2009, pg. 497). The results indicated that the optimal batting line-up had a minimum and maximum expected number of runs approximately 219 and 235, respectively, and on average, the optimal batting order produced at least 70 runs more than the worst batting order.

Duckworth & Lewis (2005) developed real time player metrics, using the Duckworth-Lewis methodology, to evaluate player contribution at any given stage of an innings, producing context-based measures. The developed metrics were: 1) batsmen average run contribution per unit of resources consumed and 2) bowlers' average runs contribution per unit resources consumed. Applying these measures to the 2003 VB series final (Australia vs. England) it was shown that the Duckworth-Lewis based contribution measures were less susceptible to distortions compared to traditional performance metrics.

Brown, Patel, and Bracewell (2016) investigated the likelihood of an opening batsman surviving (i.e. not being dismissed) each ball faced over the course of an innings. Using various model formulation and selection techniques, Brown et al. (2016) developed a contextually and statistically significant Cox proportional hazard model that predicted the probability of survival for any opening batsmen, given certain model conditions. Practically and statistically significant predictors were: 1) cumulative number of runs scored, 2) cumulative number of consecutive dot balls faced and 3) cumulative number of balls faced in which less than two runs in four balls had been scored. The results illustrated that as the magnitude of the three predictors increased for an opening batsman, the associated survival probabilities for the batsman either remained constant or decreased on a ball-by-ball basis. Applying the model to opening batsmen from 68 ODI matches, played between 8th December 2014 and 8th February 2014, it was found that

Kumar Sangakkara was the most effective batsman at occupying the crease. Moreover, calculating the area under the survival curve served as a unique validation method of rating batsmen contribution in real-time.

Bhattacharya *et al.* (2011) applied a Gibbs sampling scheme relating to isotonic regression to observed scoring rates to produce a non-parametric ball-by-ball resource table. The desired resource table required non-decreasing elements along the rows and down the columns. To accommodate these requirements Bhattacharya *et al.* (2011) implemented an isotonic regression optimization problem subject to monotonicity constraints applied to the rows and columns. Recognizing that the problem arises from a normal likelihood, a Bayesian model using a flat default prior subject to monotonicity constraints was adopted. Consequently, a Gibbs sampling was carried out via sampling from the full distribution. Sampling was carried out using a normal generator and reject sampling.

A comprehensive review of the application of statistical methods analyses in cricket is provided in Albert, Glickman, Swartz & Koning (2017).

5.2 RESEARCH OBJECTIVES

The primary aim of this chapter is to develop a real-time rating system that accurately measures the amount of influence a player exerts on a T20 match of cricket, at each stage of an innings. Such a system has considerable implications for a variety of stakeholders, such as players, coaches, managers, and franchise owners. The system will enable players and coaches to isolate specific match situations where performances increase or decrease and identify the key metrics leading to performance fluctuations. Moreover, coaches and managers can utilise the model results to build effective, player specific training regimes, determine optimal batting line-ups, evaluate player-selection decisions, and develop in-game strategies.

The secondary aim of this chapter is to assess the predictive accuracy of the probability of win model and the player rating model using the DMS metric, and benchmarking it against the log-loss metric to show that it outperforms and better accounts for match context relative.

5.3 DATA PREPARATION

The model development process implemented ball-by-ball observations from the following T20 competitions: Indian Premier League (IPL; 2014, 2015, 2016), Australian Big Bash League (BBL; 2014, 2015), English NatWest T20 League (2015, 2016) and Caribbean Premier League (CPL; 2014/2015, 2015/2016).

A process was developed to programmatically extract and parse ball-by-ball observations from ESPNCricinfo (http://www.espncricinfo.com/) commentary logs and provide a more convenient data structure (using the R programming language). The process extracted relevant data on a ball-by-ball basis and stored the data in a tabular form for easy access.

The data contains approximately 95,000 observations across 400 matches, excluding rain interrupted and abandoned matches.

The performance metrics fall into three categories: *pre-match* metrics, which are measured pre-play, *in-play* metrics, which are measured during play, and *post-match* metrics, which are measured post-match.

5.4 RESEARCH METHODOLOGIES

To measure a player's match influence, during any stage of an innings, performance metrics that significantly affect match outcome are necessary. These metrics must capture match, inning and player-specific information to accurately measure a player's influence, therefore it is necessary to implement metrics that capture 3 key dimensions: *volume of contribution*, *efficiency of contribution* and *contributions made under pressure*. Appendix C lists the metrics that are categorised under each playing dimension, and the corresponding attribute type (i.e. action, context, or time).

To effectively measure an individual's in-play influence many conventional performance metrics are inapplicable for two reasons: 1) *Indefinability* – for example batting average is undefined until inning completion; 2) *Event dependence* – for example bowling strike rate and bowling average are undefined, until the bowler takes a wicket. Therefore, given this lack of definable and event independent player metrics, engineering meaningful ball-by-ball metrics that are measurable was paramount to research success.

Generally, cricketing statistics are regarded as traditional or out-of-date as many are unobserved and immeasurable in real-time and have not evolved since the sports conception. Duckworth & Lewis (2005) expressed a similar sentiment stating that traditional performance metrics are susceptible to greater distortion relative to context-based metrics. Therefore, engineering measurable and definable metrics which captures volume, efficiency and pressure metrics are paramount to research success. These metrics will account for a greater proportion of variation in a player's influence and demonstrate how their individual performances affect match outcome.

The ratings framework is applied to extract meaningful metrics capturing information pertaining to the three dimensions of player influence and evaluate player performance (i.e. player-based ratings). The ratings framework engineer features encompassing information relating to different dimensions and layers within the data. For example, batting strike rate and runs scored are metrics that capture volume and efficiency information, respectively. The number of dots faced, and percentage boundaries also capture volume and efficiency information, however these metrics capture a deeper level of information as they provide insight into a batter's volume and efficiency with context, specifically relating to volume of non-scoring balls and efficiency of scoring boundaries.

In Chapter Three it was stipulated that ensembling traditional metrics, (i.e. metrics accounting for a small proportion of variation), with complex metrics, (i.e. metrics accounting for a large proportion of variation), resulting in meaningful ratings (i.e. reliable, robust, intuition and transparent). As mentioned, the five key elements of the ratings framework are dimension reduction, feature engineering strategies, feature selection techniques, multi-objectives (accounting for different dimensions within a sport), time-based variables and ensemble forecasting strategies. Here, the ratings framework has been adapted to output meaningful player-based ratings and implements these five key elements. Appendix C outlines all the action, context and time-based features applied within the player ratings framework. All features are time-based as they are derived on a ball-by-ball basis.

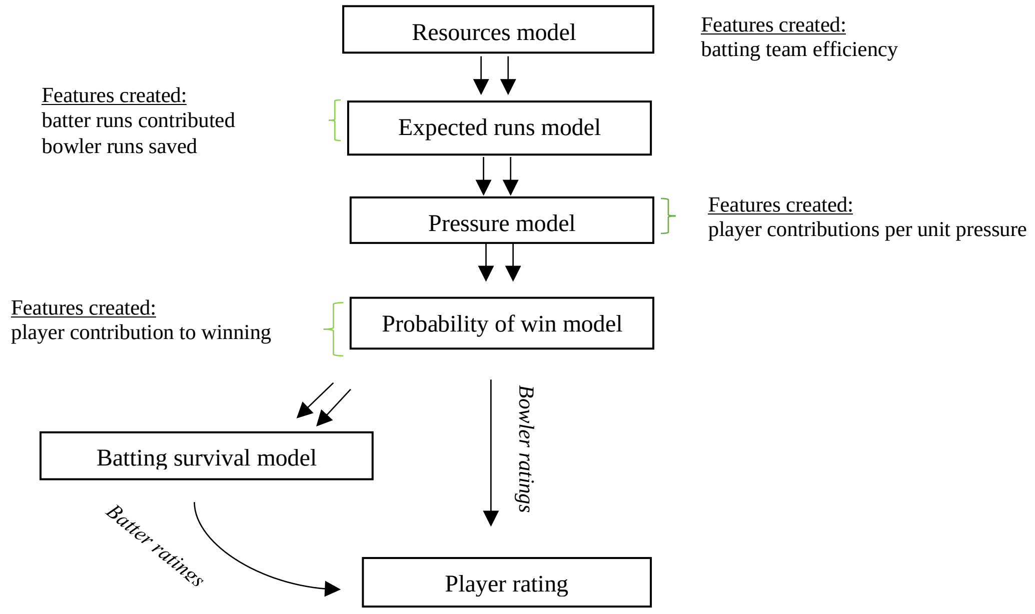


Figure 12: Player ratings framework

Figure 12 outlines the adapted framework applied to develop a real-time player rating system that accurately measures the amount of influence a player exerts on a T20 match of cricket. The player ratings framework applies multi-objectives with each objective modelling key dimensions of a T20 cricket match and engineer important features within each dimension. Specifically, five models are built to derive ball-by-ball (i.e. time-based) player specific metrics that are used to calculate player ratings: 1) resources model, 2) expected runs model, 3) pressure model 4) batting survival model and 5) probability of win model. Each model captures a specific dimension of a T20 cricket match and encompasses a significant proportion of variation in player influence.

The objective of each model is as follows: 1) The *resource model* measures the proportion of resources at the batting team's disposal from which to accumulate runs. The proportion of resources remaining at each stage of an innings indicates the rate at which the batting team are losing wickets and the rate at which the bowling team is accumulating wickets. The rate of resource decay measures which team is performing 'better' and describes innings progress from a team perspective. 2) The *expected-runs model* calculates the total number of runs the batting is expected to accumulate at the conclusion of their innings based on team-level batting metrics and the proportion of resources remaining. This measure shows how the batting team is performing and whether they will accumulate runs above par in the first innings or reach the target score in the second innings. 3) The *pressure model* measures the amount of pressure the batting and bowling is under given the amount of runs the batting team is expected to accumulate. The pressure metric is a value between 0-100, where a higher value indicates greater pressure. For example, if the second innings batting team is expected to accumulate more runs than the target total, the pressure exerted on the bowling team is greater than the pressure exerted on the batting team. 4) The *probability of win model* derives the batting team's probability of winning the match based on a set of team, match, and player metrics. This model utilises a mixture of individual and match-level metrics, such as resources, expected runs, batting team strike rate etc., because combining features from different match dimensions captures a greater level of information than the individual counterparts. 5) The *batting survival model* calculates the probability of a batter being dismissed given current batting performance, the amount of resources remaining and the level of pressure. This survival model is an extension of the work presented in Brown, Patel & Bracewell (2018) and measures the current batter's ability to occupy the crease (i.e. not lose their wicket). 6) The *player ratings model* quantifies the amount of influence a player exerts on a match of T20 cricket, on a ball-by-ball basis. The player rating models implement logistic regression to evaluate the amount of influence a bowler and batter exerts as the inning progresses. Consequently, the impact of a player's action relative to the current match state is evaluated, which enables dynamic player tracking. This methodology allows the quantification of an individual's match influence and can be applied to measure the importance of individual players to a team's probability of winning.

A rating system like that presented in this chapter could be used for several purposes; for example, team coaches may use in-play player rating to assess the merits of various strategies or analyses overall team performance. Further, the media could use the model to identify key moments in a match and further enhance television converge. The ratings framework is a dimension reduction, feature engineering and feature selection exercise, with a goal to reduce model complexity, increase variation explained and decrease the number of match metrics needed to explain the amount of influence a player exerts on a T20 match of cricket on a ball-by-ball basis.

The shallow layers of the framework adopt match-level metrics to engineer increasingly complex metrics which explain state of play, including resources, expected runs, pressure and probability of win. The deeper layer, derives player specific metrics using these match-level (i.e. complex metrics); for example, the number of runs contributed and saved by a batter and bowler, respectively, are calculated using expected runs feature (please see section 5.5.2). The features engineered within each layer are fed to succeeding layers to engineer more complex and informative features. Finally, ensembling these features provides a more informative understanding match situation and player influence (i.e. player ratings). Therefore, quantifying the state of play on a ball-by-ball basis allows accurate measurement of a player's actions and how their actions affect match outcome. The resources, expected runs, pressure and probability of win models enable the engineering of features associated with state of play and the derivation of player specific metrics. These player specific metrics are applied and ensembled in the survival and player rating models, respectively, to generate meaningful player ratings.

The following section details the methodology used to extract ball-by-ball features and outlines the features engineered to develop the dynamic player ratings model.

5.5 MODEL DEVELOPMENT

5.5.1 Resources Model

In-play resources remaining measures the proportion of resources at the batting team's disposal from which to accumulate runs, at each stage of an innings. Resources remaining represents the time till inning completion and quantifies the amount of time to accumulate runs. It allows the evaluation of whether the bating team is effectively utilising their time at the crease.

The original Duckworth & Lewis (1998) resource allocation method was designed for one-day cricket. Consequently, this study calculated proportion of resources remaining at any given stage of an inning uses a modified Duckworth-Lewis system developed by McHale & Asif (2013). McHale & Asif (2013) illustrates that the function $F(w)$ a positive decreasing step function with $F(0) = 1$, interpreted as the proportion of runs that are scored with *w* wickets lost compared with that of no wicket lost – results in unintuitive consequences on the value assigned to wicket partnership and produces erratic patterns for the value assigned to wicket partnerships. To solve this problem McHale & Asif (2013) smoothed $F(w)$ producing a survival function based on a truncated normal distribution. $F(w)$ was smoothed using equation (12):

$$F(w) = \frac{\phi(10;\mu_1,\theta_1)-\phi(w;\mu_1,\theta_1)}{\phi(10;\mu_1,\theta_1)-(0;\mu_1,\theta_1)} - \infty < \mu_1 < \infty, \theta_1 > 0 \qquad (12)$$

Here $F(w)$ is a survival function based on a truncated normal distribution, and ϕ is the normal cumulative distribution function and μ_1, θ_1, and parameters to be estimated.

Further, McHale & Asif (2013) suggested a truncated-Cauchy distribution which introduces slower decay towards the asymptotes, a heavier tail and produces more acceptable reset targets, compared to the exponential function outlined in Duckworth & Lewis (1998), which decays rapidly towards the asymptotes, $Z_0F(w)$, leading to situations where the D/L model under compensates. Therefore McHale & Asif (2013) suggested a truncated-Cauchy distribution which introduces slower decay towards the asymptotes, a heavier tail and produces more acceptable reset targets. The average number of runs scored in the remaining u overs when w wickets have been lost was given by:

$$Z(u,w) = Z_0F(w)\left\{\frac{tan^{-1}\left(\frac{u-\mu}{\theta_0F(w)}\right)-tan^{-1}\left(\frac{-\mu}{\theta_0F(w)}\right)}{\frac{\pi}{2}-tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\} \tag{13}$$

$F(w)$ is defined as above. Eqn. (13) was found to produce more intuitive results for Z_0 than the D/L method. A λ parameters was introduced to (11) to account for high and low scoring matches, as in Duckworth & Lewis (2004). It was assumed, in high scoring matches Z_u tends to become linear and hypothetically, the value of each wicket tends to zero. To account for effect θ_0 and Z_0 are scaled; Eqn. (11) is transformed to the following model:

$$Z(u,w|\lambda) = Z_0\lambda^{n(w)+1}F(w)\left\{\frac{tan^{-1}\left(\frac{u-\mu}{\theta_0\lambda^{n(w)}F(w)}\right)-tan^{-1}\left(\frac{-\mu}{\theta_0\lambda^{n(w)}F(w)}\right)}{\frac{\pi}{2}-tan^{-1}\left(\frac{-\mu}{\theta_0}\right)}\right\}$$

This modification produced more intuitive results for Z_0 than the D/L method. McHale & Asif (2013) scaled the existent parameters and introduced an additional parameter, λ, to account for high and low scoring matches, as in Duckworth & Lewis (2004). Therefore, the resources available are given by:

$$R_i = 1 - \sum_{j=1}^{n_i}(P_N\left(u_{1,j}, w_j\middle|\lambda\right) - P_N\left(u_{2,j}, w_j\middle|\lambda\right)) \tag{14}$$

This proposed method was found to be superior to the methods presented in Bhattacharya, Gill & Swartz (2011), Jayadevan (2002) and Stern (2009). Figure 13 illustrates resource decay during a T20 cricket match. It is revealed that resources decay linearly in T20 cricket.

5.5.1.1 Resources-based features

The resources remaining metric is used to assess a team's scoring efficiency and an individual's scoring efficiency (eqn. 15 and 16, respectively). This scoring efficiency metric measures the number of runs scored per unit of resources consumed.

$$runs\ per\ \%\ resource_i = \frac{current\ total_i}{(1-resources\ remaining_i)} \tag{15}$$

$$batters\ run\ efficiency_k = \frac{runs\ scored_k}{\sum resources\ consumed_k} \tag{16}$$

Here, $runs\ scored_k$ represents the number of runs scored by batter k, and $resources\ consumed_k$ represents the proportion of resources consumed by batter k; calculated using:

$$resources\ consumed_i = resources\ remaining_i - resources\ remaining_{i-1}$$

For example, if batter k hits a boundary four off $ball_i$, and resources remaining decreases from 0.86 to 0.84, between $ball_{i-1}$ and $ball_i$, then batter k has consumed 0.02 resources. This idea to assess the performance of batters and bowlers in one-day cricket was first developed by Beaudoin & Swartz (2006) and later by Lewis (2005).

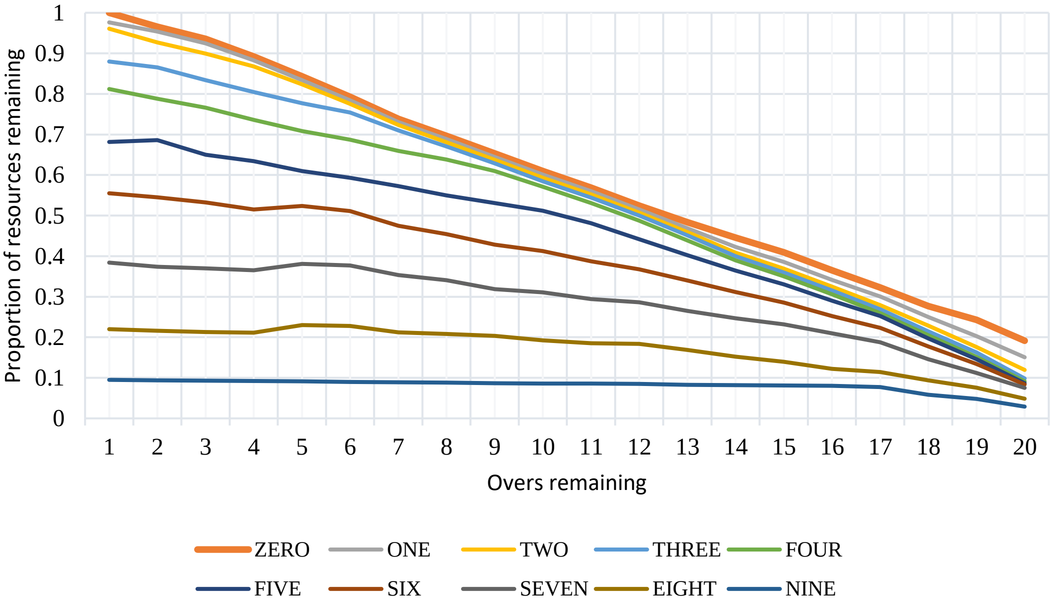


Figure 13: T20 Resources remaining

5.5.2 Expected Runs Model

In-play expected total predictions were generated using the methodology outlined in Patel, Bracewell & Bracewell (2018). The method applies a gradient boosted machine (gbm) technique with a Poisson distribution, 20000 iterations, a 5-way interaction depth and a step-size learning rate of 0.0001. A gbm technique was implemented because it accounts for the complex interactions between match and inning-level metrics by a taking a sequence of weak learners to construct a complex learner and increasing model complexity. The combination of weak-learners to create an increasingly complex learner that incorporates a 5-way interaction effect will reduce the error in expected total.

First and second innings models predicting ball-by-ball expected total was developed. The models used innings total as the dependent variable, while the model covariates were current total, wickets, balls bowled, run rate, projected total, percentage boundaries, percentage dots, resources remaining, runs per percentage resources, runs remaining and required run rate. The number of runs the batting team is expected to score depends on the team's batting performance, opposition strength and meta-level factors such as venue and pitch conditions, it is assumed that efficiency-based metrics, such as percentage boundaries, percentage dots and team strike rate, inherently contain this information. For example, if the bowling team is considered strong or the pitch conditions are *not* batter friendly, this information is captured in batting efficiency metrics. Although the expected runs model does not utilise venue and bowling strength, it is assumed that batting efficiency metrics incorporate such information.

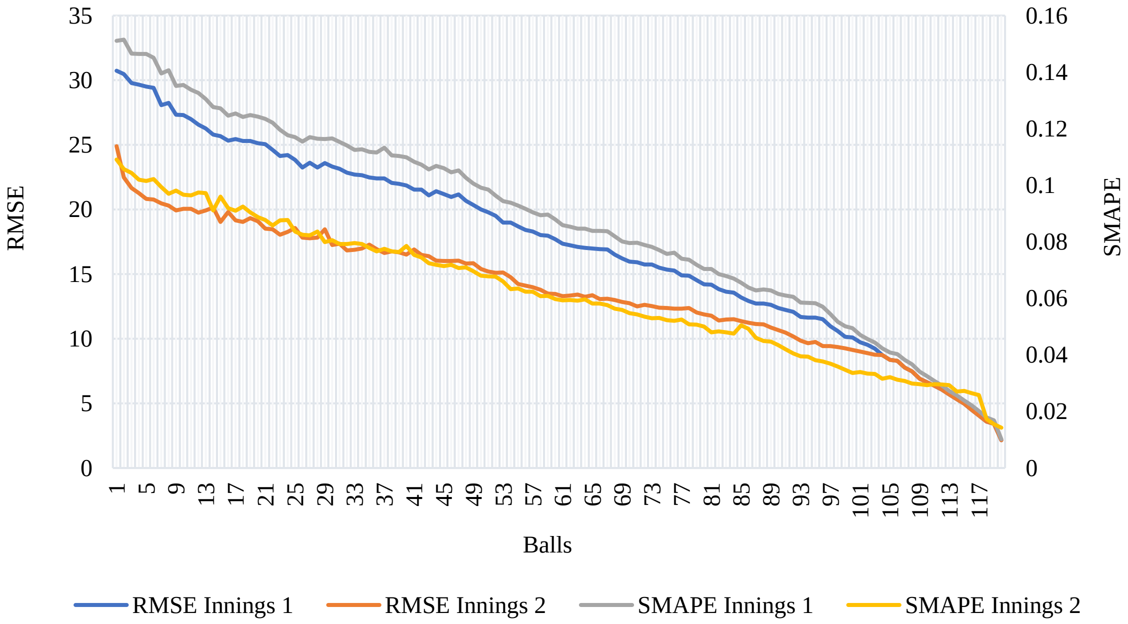


Figure 14: First and second innings RMSE and SMAPE expected runs

Applying a gbm technique with a Poisson distribution and 25,000 iterations, two individual models (first and second innings) predicting ball-by-ball expected runs were developed. The metrics included: projected total, team strike rate, run rate, current runs, wickets, percentage dots, percentage boundaries, resources remaining, balls, required run rate and runs remaining. Projected total (*current total/ resources available*) was incredibly important on first innings expected runs model (r^2 = 0.58). However, there is a significant drop in its importance in the second innings. The importance of projected total is equivalent to team strike rate. This result is expected as the runs scored in the second innings is dictated by first innings total and the runs remaining to achieve the target score, given resources available, which is heavily dependent on the scoring rate. Figure 14 outlines the ball-by-ball predictive power using root mean square error, of the two models predicted runs against observed runs.

It is shown that the predictive power improves as the innings progresses as an inning matures the metrics become more indicative of actual total and incorporates more information surrounding the end-of-innings total. The results reveal that model 2 outperforms model 1, because in the 2nd innings the batting team's target score is known, and the model utilises runs remaining and run rate required. Therefore, model 2 has greater 'knowledge' about match state regarding the team's 'optimal' scoring pattern and distance between expected total and target total, relative to model 1.

5.5.2.1 Expected runs-based features

The expected runs metric is used to engineer a volume-based metric measuring a player's ball-by-ball contribution and the effect each ball outcome has on a team's expected total. Given the outcome of $ball_i$, a player's contribution is evaluated through the change in expected runs between $ball_i$ and $ball_{i-1}$:

$$batter\ runs\ contribution_i = \ expected\ runs_i - expected\ runs_{i-1} \tag{8}$$

$$bowler\ runs\ saved_i = \ expected\ runs_{i-1} - expected\ runs_i$$

For example, if $expected\ runs_{i-1}$ = 129, and batter A hits a boundary four off $ball_i$, delivered by bowler B, which increases the $expected\ runs_i$ to 135, then batters A's contribution for $ball_i$ = 6 (135 -129), while bowler B's contribution for $ball_i$ = -6. This context-based metric measures player contribution at a team level. A player's *total batting contribution* is calculated by summing their ball-by-ball run contributions, while their *total bowling contribution* is calculated by summing their ball-by-ball runs saved.

A player's contribution is interpreted in terms of how many runs they contributed to the final total and shows how many runs a batter contributes to the innings total. For example, if batter *k* scored 35 runs but their total contribution is 43, this is interpreted as the batter's true *value* to the end-of-innings total was 43 runs.

5.5.3 Pressure Model

It is assumed that team-level pressure is a function of the number of runs a team is expected to score at each stage of an innings, given the proportion of resources remaining. For example, if the first innings batting team has a slow scoring rate and are trending towards a sub-par total, the batting team is under greater pressure than if they had a stronger scoring rate and trending towards an above par winning total. Based on this assumption the expected runs metric was used to engineer a pressure feature.

Bhattacharjee & Lemmer (2016) stated the pressure the batting team experiences is determined by a combination of batting performance and retaining wickets, while the pressure the bowling team experiences is determined by its ability to take wickets and to restrict the number of runs scored by the opponent.

On average a total of 175 runs are required for the first innings batting team to win. Using this value, a raw pressure metric for any given ball *i* is established:

$$raw\ pressure_i = \begin{cases} if\ innings = 1;\ 175 - expected\ total_i \\ if\ innings = 2;\ target\ total - expected\ total_i \end{cases}$$

Given expected runs accounts for the batting teams scoring rate and resources remaining, the pressure metric adopts an assumption like that outlined in Bhattacharjee & Lemmer (2016). Bhattacharjee & Lemmer (2016) suggested that second innings pressure is a function of the remaining runs and required run rate to achieve the target total. This has been incorporated in the pressure metric, as runs remaining and required are embedded in the expected runs model. “It is customary to assess the scoring process by calculating the required run rate time to time to see whether the scoring rate is satisfactory (Bhattacharjee & Lemmer, p. 684, 2016). The Bhattacharjee & Lemmer (2016) pressure index decreases when the batting progress is ‘good’ and if resources are kept until the target has been reached.

At any stage of the first innings, the greater the difference between a team’s expected total and 175 the greater the pressure, as the team is tracking towards a sub-par total. The first innings pressure metric can take negative and positive values as a team’s expected total can be above or below the par total of 175. The second innings pressure only takes positive values, as the expected total can be below or equal to the target, leading to positive pressure or no pressure, respectively. Figure 15 shows the distribution of raw pressure across both innings. The first innings [raw] pressure follows an approximate normal distribution, while second innings [raw] pressure follows an approximately gamma distribution.

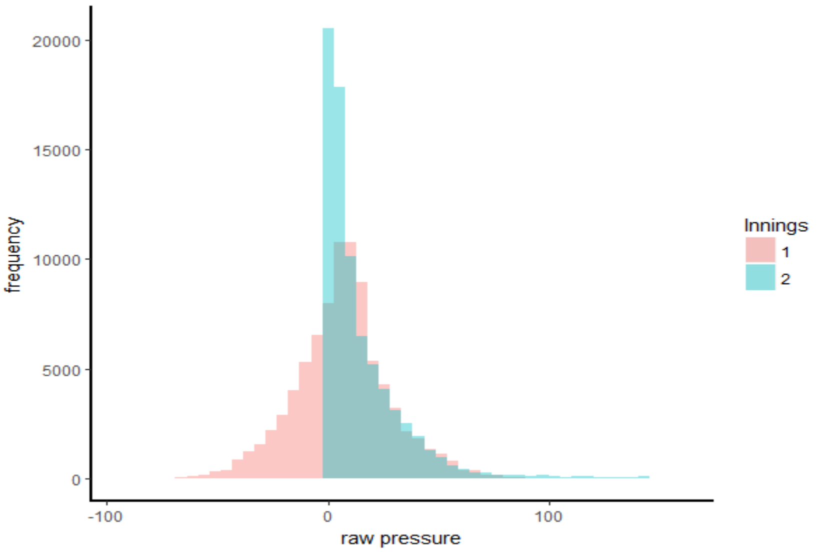


Figure 15: [raw] pressure distribution across the 1st and 2nd innings

Here, pressure can be interpreted in terms of ‘runs’, however the standard deviation across the first and second innings is 21 and 19.5, respectively. Consequently, there are instances where exaggerated estimates are produced. This issue was addressed by ‘forcing’ pressure into a logit-normal distribution, which was achieved by extracting the empirical cumulative density functions (Figure 15) associated with [raw] pressure across the two innings. Pressure was fit to a logit-normal distribution with shape parameter = 0.6, on a [0,1] support.

Applying the empirical CDF transforms the [raw] pressure values such that they are uniformly distributed. These uniformly distributed values are malleable and can be compelled into any distribution.

Applying a logit-normal distribution produces a distribution on [0,1]. This is a desired property as it produces an interpretable measure of pressure; for example, pressure values close to 0 represent low pressure, while values close to 1 represent high pressure. It is assumed that pressure follows a bell curve distribution because changes in pressure are assumed to be normal, hence the shape parameter 0.6.

Given pressure is a function of expected runs, the metric measures the distance between the winning and expected total, overall measuring the amount of pressure felt by the batting team at ball i, while the $bowling\ team\ pressure_i = 1 - batting\ team\ pressure_i$, at ball i.

Figure 16(a) shows that the empirical cumulative density function of first innings pressure follows a logistic function, and Figure 16(b) shows that the ecdf of second innings pressure follows an exponential function. Given [raw] pressure follows an approximate normal and gamma distribution across the first and second innings, respectively, the empirical cumulative density functions are as expected.

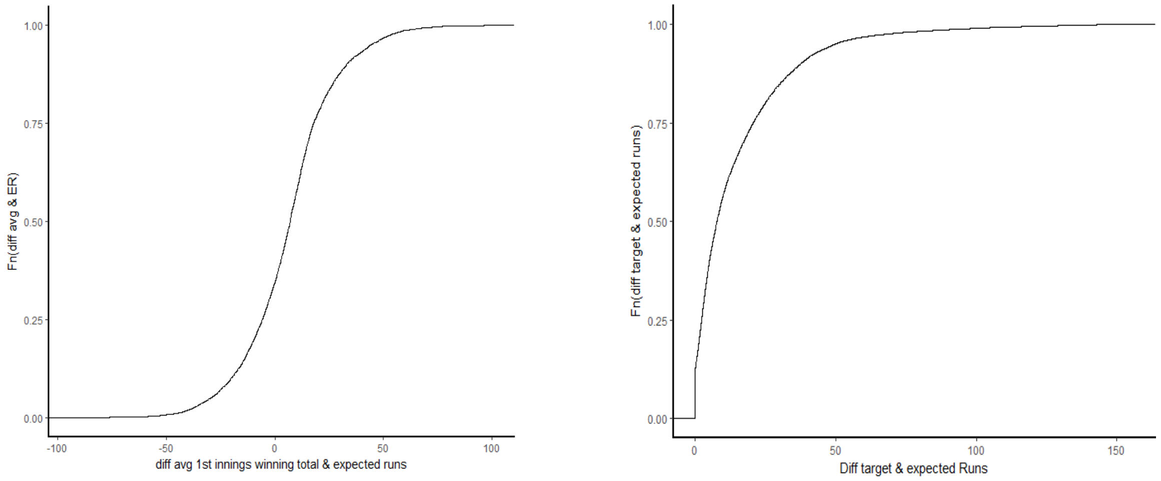


Figure 16(a) and 16(b): Empirical cumulative distribution of raw pressure across the 1st and 2nd innings

Figure 17(a) and 17(b) shows how pressure evolves across both innings for the batting and bowling team depending on the match results. Figure 17(a) shows first inning pressure for losing teams steadily increasing as the innings progresses, however the losing batting team experiences a slight decrease after the 8th over, while the losing bowling team experience an increase after the 18th over. Figure 17(a) shows first innings pressure steadily decreasing as the innings

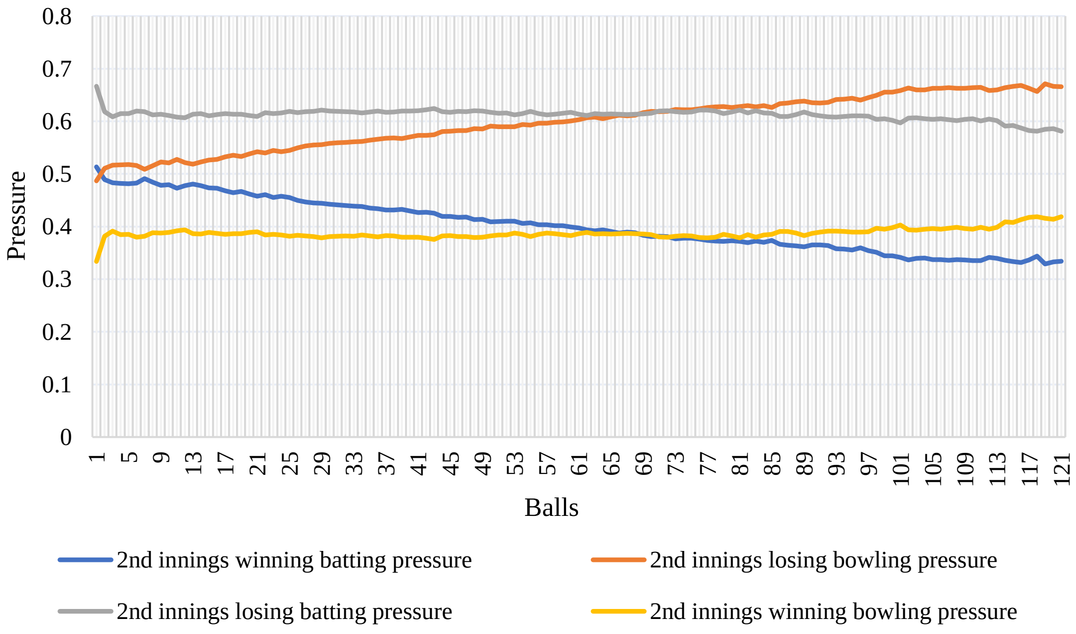


Figure 17(a): First innings pressure by batters and bowlers progresses, however the winning bowling team experience an increase after the 18th over. This maybe because in the final overs of the first innings, regardless of match situation, the batting team begins to take high risk shots or attempting to achieve greater reward by hitting boundaries. This increases the team's run rate, leading to an increase in expected runs. This claim was reinforced from the data which showed the 18th -20th overs are the most expense when the first innings bowling team goes on to win the match – on average 25 runs are conceded during these overs.

The most significant difference between batting and bowling pressure occurs in the second innings (Figure 17(b)) when the batting side loses. Moreover, the pressure remains relatively constant during the innings. This shows that, in general, when the team batting second loses, they are under pressure from the outset of the innings, most likely due to a large target total and required run rate. Finally, Figure 17(b) shows second innings pressure for the losing bowling team steadily increases as the inning progresses. This maybe because as the batting team draws closer to the target total or are tracking towards achieving an expected total equal to the target, the pressure for the bowling team increases. Further, the deviation between the first innings pressure among winning and losing teams is not as pronounced as it is in the second innings, this shows that there is implicit pressure in the first innings (setting a total) while there is explicit pressure in the second innings (chasing a known total).

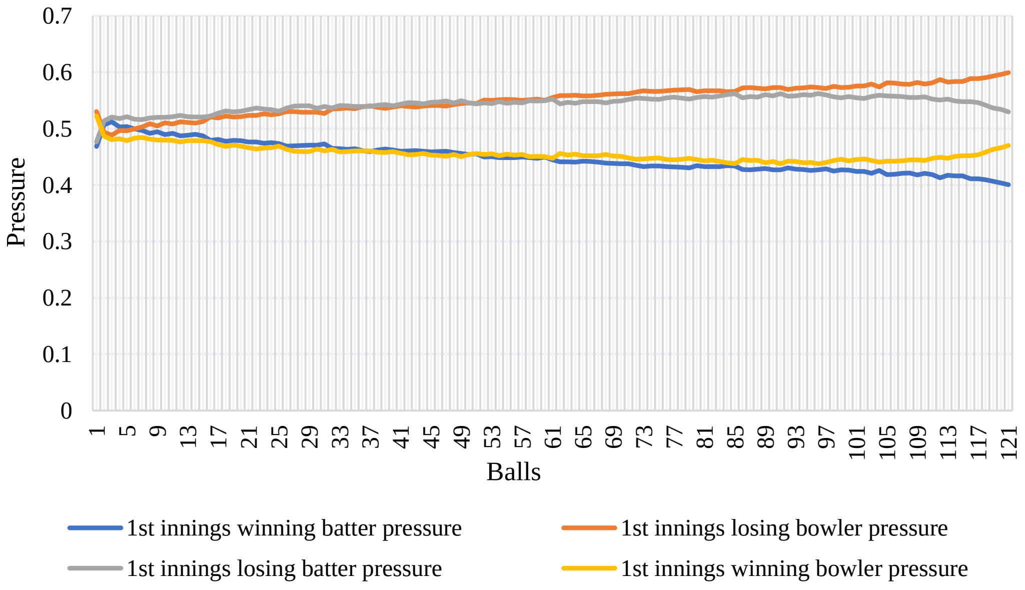


Figure 17(b): First innings pressure by batters and bowlers

Sample		K-S	P-value
One	**Two**		
1st inn winning batter pressure	1st inn losing bowler pressure	0.934	< 0.0001
1st inn losing batter pressure	1st inn winning bowler pressure	0.983	< 0.0001
2nd inn winning batting pressure	2nd inn losing bowling pressure	0.983	< 0.0001
2nd inn losing batting pressure	2nd inn winning bowling pressure	1	< 2.2e-16

Table 10: K-S statistics across groups

A Kolmogorov-Smirnov (K-S) test was conducted on the first and second innings pressure values across winning and losing, batting, and bowling, teams. Table 10 outlines the *p-value* for the K-S statistic across groups. The results show that the distance between the empirical distribution functions of the two samples is statistically significant, illustrating that the four groups are statistically significantly different from each other.

5.5.3.1 Pressure-based features

The pressure metric was used to engineer a *"contribution under pressure"* metric which assesses a team's run scoring and run restricting ability per unit of pressure, at ball *i*:

$$runs\ per\ unit\ pressure_i = \frac{current\ total_i}{pressure_i}$$

This metric was used to assess the amount of contribution player *k* has made under varying levels of pressure. For example, a batter's contribution under pressure at $ball_i$ is $\frac{total\ runs\ scored_i}{batting\ pressure_i}$, and a bowler's contribution under pressure at $ball_i$ is $\frac{total\ runs\ conceded_i}{bowling\ pressure_i}$.

5.5.4 Probability of Win Model

The derivation of match-level metrics, resources, expected runs and pressure, allows the evaluation of the probability of winning. These features relate to different dimensions within the data and explain different aspects of a cricket match. Combining the derived features with traditional statistics a highly predictive probability of winning model is built for both innings.

A logistic regression model is adopted for estimating the probability of the batting team winning the match. Two models are developed, one for each innings of T20 cricket. The reason two separate models were developed is twofold: 1) the batting team (reference team) in each innings plays with a different strategy. The first innings batting team aims to score as many runs as possible to maximise their chances of winning, while the second innings batting team aims to achieve the target before all wickets have been lost or the pre-allotted overs have been played.

To obtain the 'best' inning-dependent logistic regression models the *bestglm()* function in R was used adopting a cross-validation delete d-method. The list of candidate model covariates for inclusion: 1) pressure, 2) expected total, 3) total runs, 4) strike rate, 5) percentage dots, 6) percentage boundaries, 7) resources remaining, 8) scoring efficiency and 9) scoring pressure. In addition, runs remaining and run rate required are included in the second innings model. Table 11 outlines the practically and statistically significant covariates across the two models. The results reveal that the probability of winning in the first innings is dependent on scoring volume and scoring efficiency, while the probability of winning in the second innings is dependent on scoring efficiency under pressure. This result validates the adage “scoreboard pressure”, showing that in the second innings pressure has a greater impact on probability of winning relative to the first innings. However, the Hosmer-Lemeshow test statistic, with 10 groups, for the first and second innings models was 98.6 and 369.77 with *p-values* < *0.0001* and < *0.0001*, respectively, indicating evidence of poor fit. This poor fit is because the independence assumption of observation is violated.

	Innings			
	First	**Coeff.**	**Second**	**Coeff.**
Metrics	intercept	-7.71	intercept	3.77
	expected total	0.04	bat team pressure	-6.70
	strike rate	0.87	resources remaining	2.90
	% dots	-1.55	innings runs	0.02
	% boundaries	4.17	required run rate	-0.36
	scoring efficiency	0.04	scoring efficiency	0.03

Table 11: Probability of winning model coefficients

The statistical and practical significance of the expected runs, pressure, and resources remaining metrics within the probability of win models and their predictive power (Table 12, Figure 18 and 19) validates the relevance of the engineered features. For example, probability of winning increases by 3% in the first innings for every unit increase in percentage boundaries; and the probability of winning decreases by 16.4% in the second innings for every unit increase in batting team pressure.

A downfall to the method is that the response variable (match outcome) with respect to the ball-by-ball data in a specific innings of a match remains unchanged after each ball, meaning the independence assumption of observations is violated. An alternative approach would be a series of *k* = 120 independent models (i.e. fitting a different model to each ball). Fitting a series of *k* independent models is appealing, in that the sample of matches over which the regression coefficients that are estimated are played independently (McHale & Asif (2016)).

A first step in assessing model validity is to compare the predicted probabilities with the actual outcome, for different categories of the predicted probability of winning from the model. Table 12 shows the observed proportion of matches that finish in a victory. In general, the model-predicted probabilities and the corresponding empirical probabilities are well-aligned, in that, there is a monotonic increase in the observed proportion of win for each increase in predicted probability band.

Predicted probability	**Overs remaining**							
	First innings				**Second innings**			
	20	**15**	**10**	**5**	**20**	**15**	**10**	**5**
0-0.1	0.135	0.128	0.108	0.067	0.105	0.093	0.082	0.065
0.1-0.2	0.268	0.167	0.131	0.120	0.233	0.188	0.199	0.197
0.2-0.3	0.314	0.270	0.248	0.211	0.312	0.273	0.273	0.268
0.3-0.4	0.409	0.370	0.375	0.377	0.388	0.352	0.339	0.356
0.4-0.5	0.467	0.458	0.456	0.465	0.488	0.446	0.450	0.420
0.5-0.6	0.475	0.555	0.573	0.554	0.584	0.512	0.524	0.519
0.6-0.7	0.555	0.616	0.615	0.631	0.628	0.602	0.586	0.612
0.7-0.8	0.684	0.735	0.745	0.745	0.713	0.711	0.684	0.737
0.8-0.9	0.695	0.784	0.776	0.815	0.816	0.818	0.844	0.862
0.9-1.0	0.818	0.857	0.857	0.889	0.918	0.955	0.971	0.986

Table 12: Model predicted win probabilities and the proportion of matches resulting in a win. The figure in parentheses show the number of matches in each category. For example, there were 63 matches with a predicted probability of victory between 0 and 0.1, when there were 15 overs remaining in the first innings. Of the 63 matches, 0.128*63 = 8 were won.

A further two model validation exercises are conducted to examine predictive power. First, a log-loss evaluation is carried out on ball-by-ball predictive for winning and losing results, respectively across both innings. Second, a leave-one-out cross validation (LOOCV) to examine the proportion of match results that were predicted correctly by the models as the match progresses. The proportion of correct out-of-sample predictions made by the two logistic regression models are examined using LOOCV. Figure 18 shows the proportion of correct predictions for each ball across both innings.

5.5.4.1 Logarithmic loss

Minimising the log-loss is equivalent to maximising accuracy of the classifier, therefore a lower log-loss value means better predictions. Log-loss closer to 0 indicates high accuracy, whereas if the log-loss is away from 0 indicates lower accuracy. Log-loss works by heavily penalises classifiers that are confident about an incorrect classification.

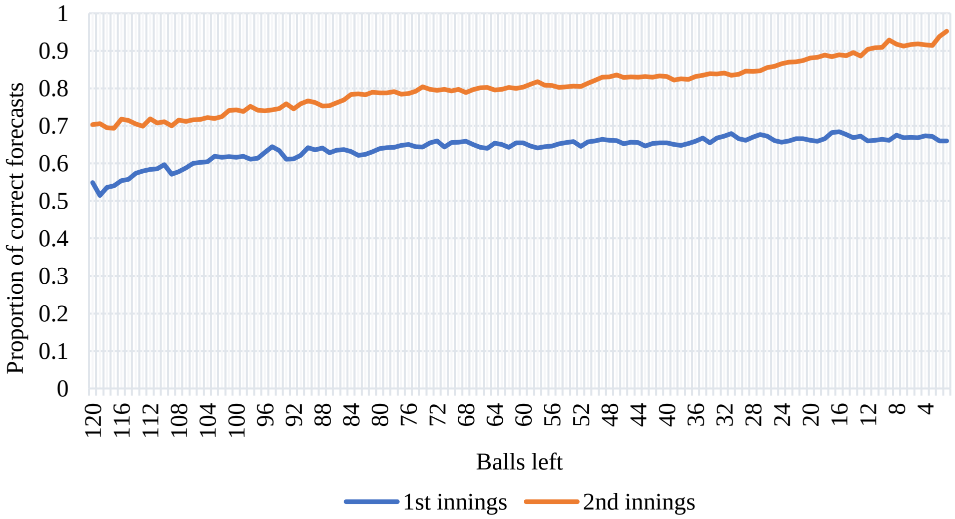


Figure 18: Leave-one-out cross validation: proportion of 'correct' forecasts made for first and second innings

Figure 19 shows log loss decreasing as the inning matures and the models' high predictive power from the beginning of the models. This result is expected because as the inning nears completion the metrics store more information regarding match outcome, increasing the model's "informativeness". Moreover, the second innings model is superior in terms of predictive power, relative to the first innings model, as it incorporates additional information such as required run rate and runs remaining, and therefore is more informative about state-of-play. Surprisingly, the models perform better when predicting losing outcomes, across both innings, however, overall, the second innings models have greater predictive power than the first innings model. Moreover, the second innings model, for winning outcomes, has greater predictive power between balls 6 and 72 (over 1-12), however this change between balls 73-120 where second inning losing outcome the model has greater predictive power. However, after the 12th overs the log-loss for the losing model experiences a rapid decline, while the log-loss for the winning team experiences a slight increase, followed by a steady declined until inning completion. This maybe because teams that win in the second innings generally finish well before 12 balls have been delivered. Although, if the team batting second winnings with only a few resources remaining then it is assumed the match is "close", therefore more difficult to predict match outcome, leading to an increase in log-loss. Moreover, if the team batting second needs more time (i.e. resources) to reach the target total than it is more than likely that they will lose the match.

This could possibly explain the decrease in log-loss for [second innings] losing results after 13 overs.

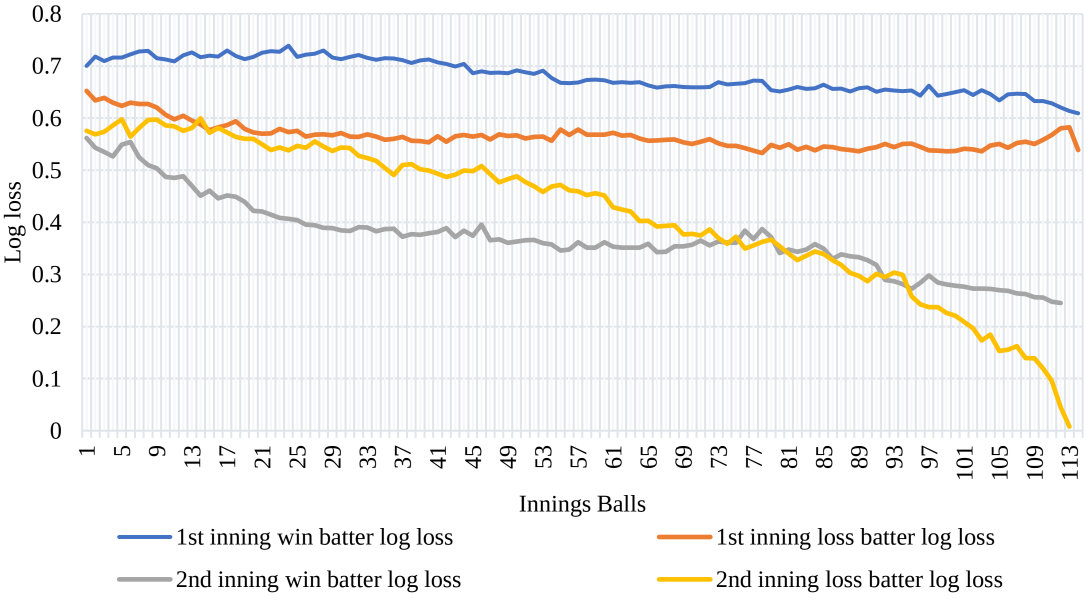


Figure 19: Ball-by-ball log-loss of actual outcome and predicted outcome

5.5.4.2 The Distance and Magnitude Spherical Performance metric

Maximising the DMS performance metric is equivalent to maximising the classifiers predictive accuracy, therefore a higher DMS score means better predictive power.

Figure 20 shows that the DMS metric outperforms the log-loss during the middle and latter stages of the first innings, given the meaningful change in the DMS score relative to the match context. Between overs 1-4 a power-play phase is conducted and during this time of the match, a lot of uncertainty is present due to scoring rate and possible resource depletion which affects match volatility. Figure 18 and Figure 19 clearly show both metrics indicating improvement in model performance as more information is obtained as the match advances.

Figure 20 shows the DMS performance metric increasing as the first and second innings mature, showing the model's predictive power increasing as the match progresses[12]. The DMS score reveals similar model insights to that of the log-loss (Figure 19): 1) The predictive power of the second inning model's is significantly better than the first innings.

[12] The update procedure applied to the weights, produced by the AHP, to calculate the DMS is the same as the linear process outlined in Chapter Five (section 3.11).

2) The models perform better when predicting losing results relative to winning results, across both innings. 3) The second inning losing model has the greatest predictive power between balls 72-120.

Compared to the log-loss metric, the DMS metric output better scores when using the first inning model to predict match results. Figure 19 shows that the DMS score is fairly constant around 0.5 throughout the first innings, for both winning and losing outcomes. This shows that the DMS metric is unable to define a winner or loser based on first inning results. Although, a score of 0.5 does not indicate 'good' predictive power it does indicate slight predictivity, however, it reveals the metrics ability to pick up match context, in that, on average, match results are not clear cut, at the end of the first innings. Therefore, a DMS score of 0 indicates no predictivity, 1 indicates 'extremely good' predictivity, and 0.5 indicates an undefined or 50:50 result.

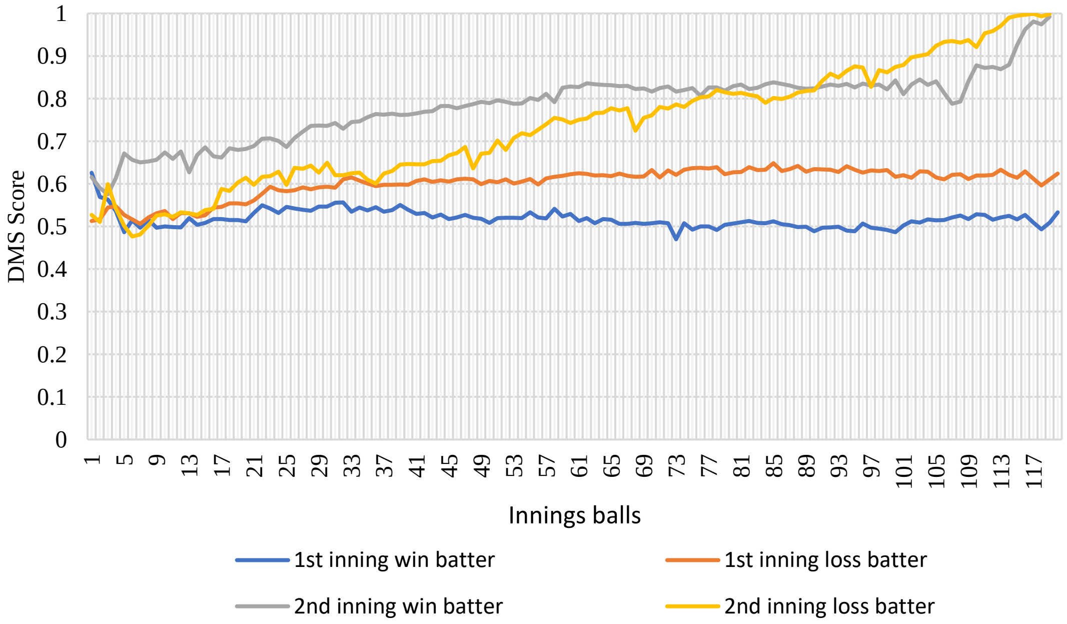


Figure 20: Ball-by-ball DMS score of actual outcomes and predicted outcome

Figure 19 shows the log-loss steadily during the first innings. Compared to the DMS metric, the log-loss produces poorer results, especially for the first innings win model, although the log-loss declines throughout the first innings, it never goes below 0.60 and 0.53 for the winning and losing models, respectively.

Figure 20 shows a relatively rapid movement in the first five overs of the first innings with the DMS score going from 0.52 to 0.65 between ball 1 and 30. Conversely, the log-loss score does not experience significant movement until after the sixth over, with a gradual decline from 0.45 to 0.68 after 56 balls (end of eight overs). This indicates slight improvements in score from an interpretation of the log-loss compared to the early and meaningful escalation of the DMS metric which provides greater insight into model performance relative to match context.

During the second innings, for both winning and losing models, the DMS score is approximately monotonically increasing between ball 1 and 72, going from 0.65 to 0.82. This demonstrates the DMS metric rewards the probability of win models based on its ability to predict accurately from long-range, indicating the metric is providing reliable, intuitive, robust, and transparent outputs.

Surprisingly, for the second innings, across winning results, the log-loss is approximately exponentially decreasing between ball 1 and ball 72. Although, between ball 72 and ball 78, it experiences a slight increase, and after experiences a slow monotonic decrease to 0.25. Given that in the second inning win model the DMS score is monotonically increasing until the 18th over (ball 108), after which it experiences an exponential increase, it shows that the DMS metric provides greater insight into model performance relative to match context.

Finally, for the second innings win model, both DMS and log-loss produce similar performance up until the 12th over, overall indicating that the second innings win model outperformed the second innings loss model. Thereafter, the second innings loss model outperforms its winning counterpart. Moreover, after the 12th over, the log-loss reflects this decrease in predictive power of the second innings loss model with a slow rate of decline. The DMS ignores this poor performance and continues to produce measures indicative of match outcome at a faster rate.

In the second innings the DMS and log-loss metric have similar performance in terms of final predictions, however, the DMS has better performance during the early stages of the second innings as it better accounts for match context. Surprisingly, for the second innings loss model, the log-loss steadily declines until the 72nd ball (12th over) and thereafter experiencing a rapid decline. This is because after the 12th, the DMS score does not experience a rapid increase it still converges to the actual outcome at a similar period of the inning to that of the log-loss.

This is represented in Figure 20, as the DMS trend converges faster to what actually happened in the second innings, while the first innings trend remains relatively flat. Specifically, compared to the log-loss across both innings the DMS metric converges at a faster rate to the actual outcome. This shows that from the outset of the second innings (and the first innings) the DMS better accounts for match context, and that appropriate weight adjustments have been applied.

In both instances the log-loss and DMS metric scores are best in the 2nd innings, which is expected as more reliable or informed data is available because the first innings has passed. Figure 20 reveals that the DMS metric is better at utilising new information and converges faster to actuality relative to the log-loss metric, this is due to the assigned weights representing the forecasting scenario.

These shows results reveal that the DMS performance metric is an appropriate metric to assess the effectiveness of meaningful sport-based rating systems and its ability to outperform well-known performance metric of the log-loss.

5.5.4.1.1 Win-based features

The results show the two winning models produces predictive results and confirm the relevance and predictive power of the engineered features. Given the models statistical and practical significance a player metric evaluating how a player's action has affected match outcome can be derived. The equation below shows how ball-by-ball changes in the probability of winning can be used to evaluate a batter's and bowler's contribution to match outcome, at ball$_i$.

$$Batters\ win\ contribution_i = prob(batting\ team\ win)_i - prob(batting\ team\ win)_{i-1}$$

$$Bowlers\ win\ contribution_i = prob(batting\ team\ win)_{i-1} - prob(batting\ team\ win)_i$$

For example, if batter A hits a boundary four off $ball\ _i$ delivered by bowler B, and the $prob(batting\ win)_i$ increases to 0.64 from 0.62 between $ball_i$ and $ball_{i-1}$, the batters win contribution for $ball_i$ is 0.02 (0.64-0.62), while bowler B's win contribution for $ball_i$ is -0.02. Moreover, a player's *total* contribution is calculated using the following equations:

$$batter\ total\ win\ contribution = \sum_i batter\ win\ contribtuion_i\ ;\ \text{for batter}\ k$$

$$bowler\ total\ win\ contribution = \sum_i bowler\ win\ contribtuion_i\ ;\ \text{for bowler}\ k$$

Given the probability of win is dependent on pressure, resources remaining, and scoring efficiency (including percentage dots and percentage boundaries), a players' total winning contribution is classified as a volume, efficiency, and contributions under pressure feature.

5.5.5 Batting Survival Model

Given the match level metrics provided adequate insight surrounding match outcomes and the individual metrics have been statistically engineered using these match-level metrics, the next step is to engineer a batter specific metric. A batter survival metric is engineered representing the probability of dismissal on ball *i*.

Extending the methodology outlined in Brown, Patel, and Bracewell (2017) to non-opening batter (i.e. top, middle, lower, and tail) across both innings, a batter's ball-by-ball survival probability is calculated. Brown *et al.* (2017) established three model development criteria: 1) model coefficients must be practically and statistically significant, 2) a decrease in resources leads to a decrease in the probability of surviving the next ball, and 3) probability of survival decreases on a ball-by-ball basis as resources are monotonically decreasing. Models that met these three criteria were included in the candidate set.

Applying a Cox proportional hazard technique, the total number of balls faced, by any given batsmen, represented the time till failure (i.e. batsmen dismissal). The Cox proportional hazard model has the following form:

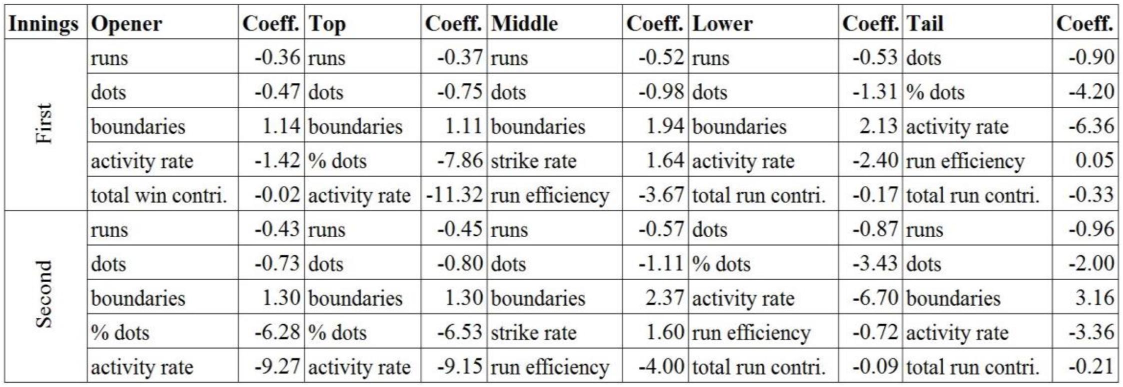
$$h(t,\boldsymbol{X}) = h_0(t,\boldsymbol{\alpha})\boldsymbol{e}^{(\boldsymbol{\beta}'\boldsymbol{X})}$$

Here, $h_0(t,\boldsymbol{\alpha})$ represents the hazard function at baseline levels of covariates, and varies over time, and α is a vector of parameters influencing the baseline hazard function. The Cox model has the following survival function:

$$S(t,\boldsymbol{X}) = S_0(t,\boldsymbol{X},\boldsymbol{\beta})\mathbf{e}^{(\boldsymbol{\beta}'\mathbf{X})},$$

Here, $S_0(t,\boldsymbol{X},\boldsymbol{\beta})$ represents the survival function at baseline levels of covariates. A right censoring methodology was adopted as a batsman may not be dismissed during an innings.

The validity of the Cox model relies on two assumptions: 1) the effect of each covariate is linear in the log hazard function, and 2) The ratio of the hazard function for two individuals with different sets of covariates do not depend on time. Both assumptions were met across all models.



Innings	Opener	Coeff.	Top	Coeff.	Middle	Coeff.	Lower	Coeff.	Tail	Coeff.
First	runs	-0.36	runs	-0.37	runs	-0.52	runs	-0.53	dots	-0.90
	dots	-0.47	dots	-0.75	dots	-0.98	dots	-1.31	% dots	-4.20
	boundaries	1.14	boundaries	1.11	boundaries	1.94	boundaries	2.13	activity rate	-6.36
	activity rate	-1.42	% dots	-7.86	strike rate	1.64	activity rate	-2.40	run efficiency	0.05
	total win contri.	-0.02	activity rate	-11.32	run efficiency	-3.67	total run contri.	-0.17	total run contri.	-0.33
Second	runs	-0.43	runs	-0.45	runs	-0.57	dots	-0.87	runs	-0.96
	dots	-0.73	dots	-0.80	dots	-1.11	% dots	-3.43	dots	-2.00
	boundaries	1.30	boundaries	1.30	boundaries	2.37	activity rate	-6.70	boundaries	3.16
	% dots	-6.28	% dots	-6.53	strike rate	1.60	run efficiency	-0.72	activity rate	-3.36
	activity rate	-9.27	activity rate	-9.15	run efficiency	-4.00	total run contri.	-0.09	total run contri.	-0.21

Table 13: Practically and statistically significant metrics across both innings

To identify the optimal models the study implemented the *glmulti* R package. The models implemented an exhaustive genetic algorithm to explore the candidate set in conjunction with an *AIC* criterion to dictate model selection. Model interactions were not considered and a model with a minimum set of three and maximum set of five metrics were required. All five optimal models met the Cox assumptions, declared convergence, and met model development criteria. Table 13 outlines the metrics that met the requirements and the coefficients, for all batting positions, across both innings.

Figure 21 illustrates the exponentially decreasing nature of a batter's survival probability across both innings for winning and losing teams. A batter's survival, irrespective of position, is dependent on volume, efficiency, and pressure-based contributions. Moreover, these results prove the validity and statistical significance of the engineered features when measuring player ability. Finally, Figure 21 illustrates the decreasing survival probability across both innings for winning and losing teams, for all five bating positions. It also reveals that the survival probability of first innings openers and top order batter is greater than their second innings counterpart. Moreover, the survival probability of the lower and tail-ender batters, for winning teams, is greater than the lower and tail ender batters, for losing teams. This is because the lower-order and tail-end batters, for winning teams, occupy the crease for longer which leads to more runs and less wickets.

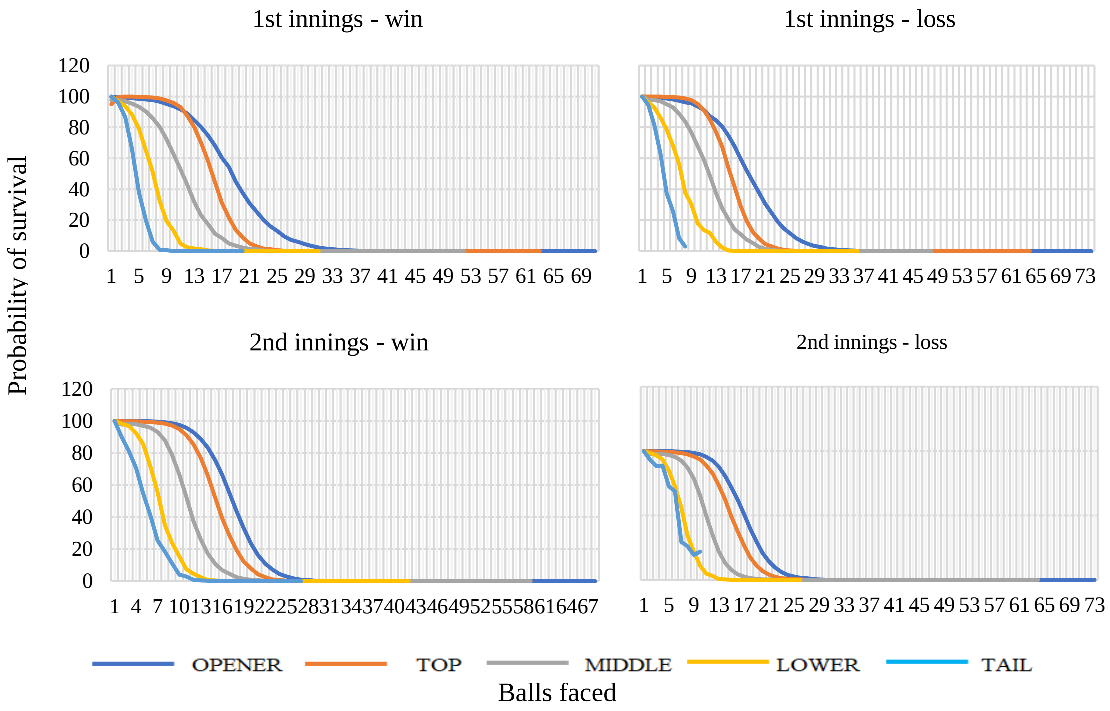


Figure 21: Survival probabilities by innings result

5.5.6 Player Rating Model

The derivation of match-level and individual player metrics allows the evaluation of player ratings which measure the amount of influence a player exerts on a T20 match of cricket, at each stage of an innings. ‘Ensembling’ the derived features with traditional statistics makes it possible to build an accurate player ratings model for both innings.

Four separate logistic regression models are developed to estimate a player’s rating, i.e. batter and bowler models, across each innings. The reason four separate models were developed is twofold: 1) The batting and bowling team in each innings plays with a different strategy. The first innings batting team aims to score as many runs as possible to maximise their chances of winning, while the second innings batting team aims to achieve the target before either all wickets have been lost or the pre-allotted overs have been played. 2) The first innings bowling team aims to restrict the number of runs conceded, while the second innings bowling teams aims to exhaust the batting team resources before they reach the target total.

Again, to identify the ‘best’ dependent logistic regression models the *bestglm()* function in R was applied adopting a cross-validation delete *d*-method. The list of candidate batting metrics for inclusion in the models are runs scored, balls faced, contribution, total runs contributed, strike rate, dots faced, total boundaries, percentage dots, percentage boundaries, pressure contribution, activity rate and survival probability. The list of candidate bowling metrics for inclusion in the model are balls bowled, runs conceded, percentage dots, percentage boundaries, economy rate, pressure contribution, dots bowled, boundaries bowled, maidens, total runs saved, wickets. Table 14 outlines the practically and statistically significant covariates across the two models and reveals the validity of the engineered features.

	INNINGS			
	FIRST		**SECOND**	
BATTING	Intercept	-0.98	Intercept	-0.73
	runs scored	0.22	runs scored	0.04
	dots faced	-0.03	boundaries	0.23
	survival prob.	0.05	survival prob.	0.01
	pressure contribution	2.01	percentage dots	-0.21
	activity rate	0.38	pressure contribution	9.77
BOWLING	Intercept	0.09	Intercept	-0.50
	runs conceded	-0.13	runs conceded	-0.34
	percentage dots	0.13	percentage dots	0.18
	economy rate	-0.17	economy rate	-1.01
	pressure contribution	6.38	pressure contribution	18.10
	runs saved	0.18	runs saved	0.24

Table 14: Inning-based player influence model coefficients

Table 14 reveals that a batter's and bowler's ability to influence match outcome, across both innings, is dictated through volume, efficiency and pressure-based metrics, validating the hypothesis that to accurately measure player influence it is necessary to implement 3 key metrics-types: 1) volume of contribution, 2) efficiency of contribution and 3) contributions under pressure.

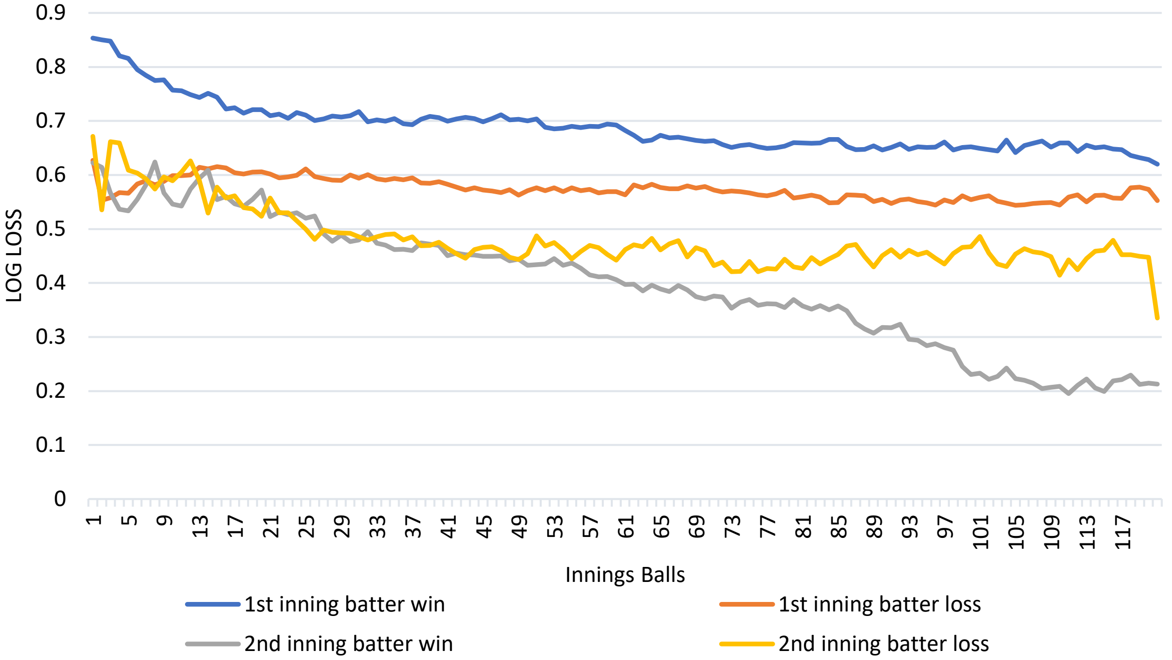


Figure 22: Ball-by-ball log-loss actual match outcome vs. player influence rating

Unsurprisingly, pressure and efficiency metrics have a greater impact on player influence in the second innings compared to the first innings. This is an expected result because in the second innings batters experience "scoreboard pressure" and need to efficiently score runs and score runs under pressure, while bowlers need to efficiently restrict runs and restrict runs under pressure.

It is evident from Figure 22 the second innings player ratings models have significantly better predictive power than the first innings models, with the second innings model producing better predictions for winning results than losing results. Figure 22 shows the log-loss decreasing as the inning matures, illustrating that a player's influence becomes more indicative of match outcome as an innings mature. This result is expected because as the inning nears completion a player's performance metrics store more information regarding match outcome, increasing model "informativeness". Moreover, the Hosmer-Lemeshow test statistics, with 10 groups, for the first and second innings models was 30,457 and 33,910, with *p-values* <0.0001 and <0.0001, respectively, indicating evidence of poor fit. These results indicate that a dynamic player rating framework has successfully allowed the construction of a ratings framework that measures the amount of influence they exert on a T20 match of cricket.

These player ratings are intuitive and transparent, as the results can be mapped to real-world observable outcomes and the context to which the system is being applied, and are interpretable and easily communicated, respectively. Moreover, the ratings are reliable and robust, as they yield good performance during different stages of the first and second innings, across winning and losing performance, and are well-calibrated and sharp, respectively.

Surprisingly, Figure 23 shows the first innings player models to have better predictive power than the second inning models, until the 12th over. After, the 12th over the second innings model produce better player ratings across both winning and losing results.

Figure 23 reveals that the DMS[13] score better accounts for first innings match context relative to the log-loss metric (Figure 22) as it never produces scores below 0.50. Although, the DMS score shows that the first inning player-rating models are unable to account for variation in player influence at the same rate as the second innings. The DMS metric and the log-loss show that the second innings model produce significantly better player ratings as the match near completion, due to the amount of information that is available. Surprisingly both the DMS metric and log-loss report similar predictive power for the second inning rating models, however the predictive power as reported by the log-loss, for the corresponding losing model, slows after the 12th over.

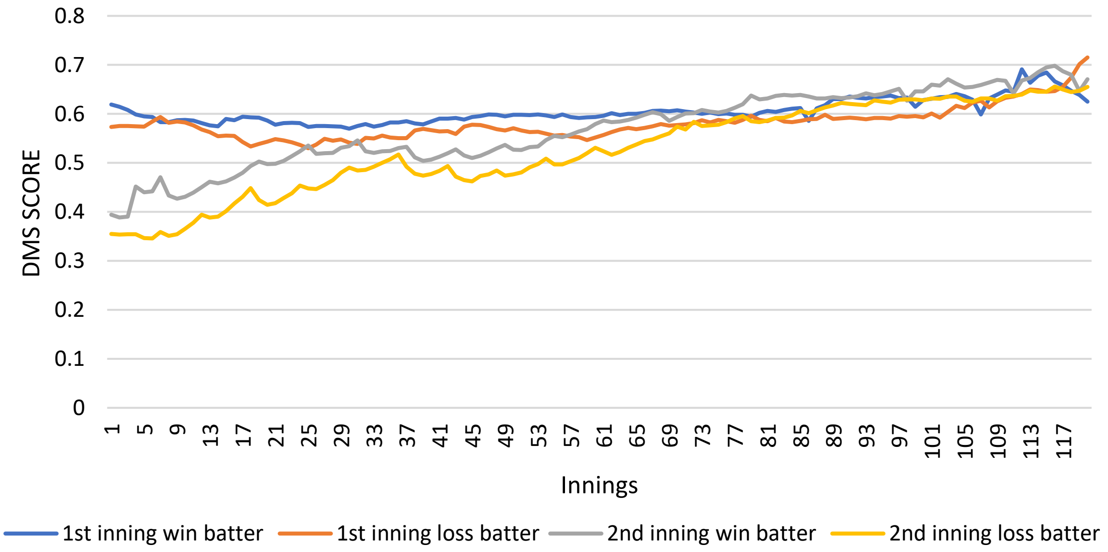


Figure 23: Ball-by-ball DMS score of actual outcomes vs. player influence rating

[13] The update procedure applied to the weights, produced by the AHP, to calculate the DMS is the same as the linear process outlined in Chapter Five (section 3.11).

Figure 22 and Figure 23 reveal both DMS and log-loss producing fairly consistent measures for first innings player models. For both the DMS and log-loss measures produce monotonically increases and decreases, respectively. Although, for the second innings, the monotonically decreasing aspect of the log-loss slows after the 9th over (ball 54), after which the log-loss declines to 0.42. This trend phenomenon is not experienced by the DMS and continues it monotonically increasing trend until match competition. Again, these results reveal that the DMS performance metric is an appropriate metric to assess the effectiveness of meaningful sport-based rating systems and its ability to outperform well-known metric of the log-loss.

The results from the player rating models are intuitive and transparent, as they can be mapped to real-world observable outcomes and the context to which the system is being applied, and are interpretable and easily communicated, respectively. Moreover, the ratings are reliable and robust, as they yield good performance during different stages of an innings, and are well-calibrated and sharp, respectively.

5.6 DISCUSSION AND CONCLUSION

To accurately measure a player's influence on a T20 cricket match three key dimensions of a player's game must be considered: 1) volume of contribution, 2) efficiency of contribution and 3) contributions made under pressure. To derive volume, efficiency and pressure-based metrics, this chapter applies the ratings framework to build models capable of calculating ball-by-ball performance metrics associated with these three dimensions, leading to a novel perspective of dynamically assessing player impact on match outcome. Through the application of the inning-based models, it was established that inning-based models possess the capability to dynamically evaluate a player's ability to influence match outcome. The in-play player ratings represent the level of influence a player is exerting on a match and quantifies the amount of influence (or impact) an action has on match outcome. The results highlight the 3 key implications: 1) the model's strong predictive power to evaluate match outcome during any stage of an innings for batters and bowlers, 2) the 'inverse' relationship between the predictive power of corresponding models (i.e. model 1 & 2, model 3 & 4), and the ability to dynamically account for match outcome variation based on significant changes in performance metrics, and 3) the model's ability to dynamically evaluate shifts in batting and bowling metrics, and the ability to evaluate the changes effect of match outcome.

The models developed can be used by various stakeholders such as commentators, to establish live odds and deliver more insightful commentary on match state, create player specific training regimes to optimise in-game situations where players thrive and reduce match situations where the players performance deteriorates. The model results validate the assumption that a player's influence had three key dimensions: volume of contribution, efficiency of contribution and contributions made under pressure. Moreover, it is revealed that to effectively measure a player’s in-play influence many conventional cricket metrics are inapplicable due to two primary reasons: *undefinable* and *event dependent metrics.*

Given the magnitude of the batting metrics that affect match influence vary depending on batting position and playing-role, it is recommended that future research build a batter influence model based on batting position (i.e. opener, top, middle, lower and tail).

Using the ratings framework constructed in Chapter Three, this chapter developed a novel player-based rating system. The rating system adopted a multi-objective modelling strategy whereby each objective corresponded to a trait significantly affecting performance, and applied dimension reduction, feature selection and feature engineering techniques to ensure these traits were sufficiently and appropriately quantified. This chapter shows that rating systems must implement these key communalities to produce meaningful ratings of a players sporting performance. Moreover, it was shown that different feature-types across varying levels of complexity can be applied to build meaningful trait-based ratings; and that performance ratings are only as ‘good’ as the individual trait used to produce the final ratings.

It has been shown that the framework can be applied within the sporting context to produce intuitive, robust, reliable, and transparent player ratings. It is concluded that the framework constructs sport-based rating systems which output meaningful ratings of performances within the sporting context.

The predictive power of the probability of win and the player rating models were assessed using the DMS performance metric and the log-loss scoring rule. The DMS metric was shown to outperform the log-loss during the middle and latter stages of a second innings, however it performed equivalently during the first innings. This is due to the weightings assigned to the rate of change magnitude and angular metrics. The DMS was shown to better account for match-context for latter stages of an innings, this is due to the weight adjustments that reflect match situation and better incorporates match information.

This is represented in the latter stages of the DMS as the trend converges faster to what happened in the second innings, while the first innings trend remains relatively flat. Further, compared to the log-loss predictions across both innings the DMS metric converges at a faster rate to the actual outcome, revealing a better use of information.

In both instances (Figure 22 and Figure 23) the log-loss and DMS metric predictions are better in the 2nd innings, which is expected as more reliable or informed data is available because the first innings has passed. Overall, it is revealed that the DMS metric is better at utilising new information and converges faster to actuality relative to the log-loss metric, this is due to the assigned weights representing the forecasting scenario.

These results reveal that the DMS metric is an appropriate metric to assesses the effectiveness of rating systems and outperforms well-known performance metric of the log-loss.

Chapter Six

DISCUSSION, CONCLUSION AND FUTURE RESEARCH

> *“Now it is time for the next chapter. I have new dreams and aspirations, and I want new challenges”.*
>
> Derek Jeter (2014)
>
> On his retirement from professional Baseball.

6.1 DISCUSSION AND CONCLUSION

Formally, this thesis has three research objectives: 1) develop a quantitative ratings framework to construct sport-based rating systems that output meaningful ratings. 2) Develop a novel evaluation metric to quantify the effectiveness of meaningful sport-based ratings. 3) Demonstrate the applicability of the developed ratings framework and novel performance metric within the sporting context. This chapter outlines the key outcomes and findings from this research, briefly describes the research limitation, future areas of research and possible ways to extend and improve the work presented throughout this thesis.

Through the literature review, it became abundantly clear that the growing application of big data and machine learning within the commercial environment has significantly increased the need for data-driven performance-based evaluation systems, referred to as rating systems or scoring models.

Specifically, rating systems have recently experienced a major growth in three major industry verticals, specifically 1) credit-risk, 2) sport and 3) the computer developer environments. Specifically, this growth is most prevalent when evaluating an applicants' creditworthiness and repayment behaviour (credit-risk), evaluating team and player performance (sports) evaluating a developers' coding ability (developer). Consequently, DOT Loves Data was approached by Umano (a software company), Penny (a peer-to-peer lending service) and New Zealand Cricket to develop rating systems. Specifically, Umano wanted to develop a real-time computer programming and developer rating system which monitors individual, project, and team performance. Penny a wanted to develop a dynamic credit-risk scorecard which evaluates an applicant's credit worthiness (ability to make timely repayments). New Zealand cricket wanted to develop a team and player optimisation tool allowing managers and coaches to select players and build strategy based on data-driven player ratings.

As a result of these three diverse, yet similar, projects, DOT funded this research to develop a ratings framework for constructing rating systems that could be applied and commercially deployed across multiple domains. Although this research was funded to develop a ratings framework to construct ratings systems across multiple domains, this thesis purely focusses on the development of a novel framework to construct rating systems within the sporting context, referred to as sport-based rating system. Specifically, the aim of this research was to develop an approach for constructing systems that produce reliable, robust, intuitive, and transparent ratings, or more simply, meaningful ratings, within the sporting context.

This rating systems disclosed within this thesis were restricted to the sports domain due to the commercial sensitivity of credit-risk and developer data, and intellectual property and non-disclosure agreements. Moreover, sporting data is easily accessible, a large amount which is publicly available, and does not suffer from commercial sensitive, such as credit-risk and developer data.

This thesis began by evaluating the current state of the ratings literature, specifically within the credit risk and sporting environments, and use the identified distinctions, limitations, and communalities to develop a ratings framework for constructing meaningful sport-based rating systems.

The research extends Bracewell's (2003) definition of ratings; Bracewell (2003) who stated that ratings are an elegant form of dimension reduction and enable the simplification of massive amounts of data into a single quantity. Specifically, ratings are an elegant and excessive form of dimension reduction whereby a numerical value provides a meaningful quantitative interpretation of performance. Meaningful ratings must have the following characteristics: 1) Robust – ratings must yield good performance where data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes. 2) Reliable – ratings must be accurate and highly informative predictions that are well-calibrated and sharp. 3) Transparent – ratings must be interpretable and easy to communicate. 4) Intuitive – ratings should relate to real-world observable outcomes and the context to which the system is being applied.

The criteria of intuitive and transparency were necessary as sport-based ratings systems required commercially deployment, and therefore the outputs need to relate to observable real-world outcomes and be easy to communicate to decision-makers.

Based on the literature review, exploratory research was conducted and several rating systems, using commonly applied methodologies within the sporting context, were developed. Consequently, key limitations and communalities within the rating methodologies were identified. The limitations were the 1) lack of a ratings framework, 2) lack of meaningful ratings and 3) lack of an evaluation metric which quantifies the effectiveness of sport-based rating systems and accounts for different sporting context. The communalities were the application of 1) dimension reduction and feature selection techniques, 2) feature engineering tasks, 3) a multi-objective framework, 4) time-based variables and 5) an ensembling procedure to produce an overall rating. Therefore, a ratings framework was developed to apply methodologies to address these limitations and incorporate methodologies to implement these communalities.

Using these findings, a ratings framework was developed which implemented a dynamic multi-objective ensembling forecasting strategy. The framework applies a methodology for constructing rating systems within the sporting environments to produce meaningful ratings of performance. Rating systems built using the framework utilise action, context, and time-based metrics at varying levels of complexity from traditional and environmental-based metrics to complex metrics. Effectively, the framework ensembles ratings corresponding to traits that significantly affect behaviour, expressed as performance on context specific traits. These trait-based ratings are derived by identifying the significant attribute-types (action, context and time), at varying levels of complexity, that affect the individual traits, and combining these trait-based ratings through a modelling function to output meaningful trait-based ratings. Therefore, to produce meaningful ratings, the trait-based ratings that are ensembled must also be meaningful, in that, they are robust, reliable, transparent, and intuitive. Moreover, to truly capture the definition of ratings as an elegant and excessive form of dimension reduction, the framework adopts dimension reduction techniques.

An ensemble approach was adopted because it is assumed that performance is a function of the individual traits significantly affecting performance. Therefore, performance is defined as $performance = f(trait_1, \ldots, trait_n)$. Moreover, the framework is a form of model stacking where information from multiple models is combined to generate a more informative model.

The framework was used to construct a novel team and player-based rating system (Chapter Four and Chapter Five, respectively), specifically within the cricketing context. These two rating systems were shown to output meaningful ratings, confirming that a multi-objective ensembling strategy is an appropriate approach to construct meaningful rating systems within the sporting context. It was concluded that ensembling trait-based ratings, derived by combining the action, context, and time-based feature-types that significantly affect each trait is an appropriate strategy. Moreover, it was confirmed that to construct meaningful sport-based rating systems the following elements must be implemented 1) dimension reduction and feature selection techniques to identify the traits significantly affecting performance and identify the feature-types (action, context and time) of varying complexity that significantly affect each trait, respectively, 2) feature engineering to extract the latent traits affecting performance, 3) multi-objective framework to derive trait-based ratings, 4) time-based variables to dynamically evaluate ratings and 5) applying an ensembling procedure to combine trait-based ratings and produce outputs that have better predictive performance compared to single predictions and are more stable. Given its validity, the ratings approach was used to develop the underlying models which are currently deployed within the Umano and Penny environments. Furthermore, the ratings framework was used to develop a player and team optimisation tool for New Zealand Cricket to select the optimal team for the T20 2019 cricket world cup. Although, these results and models were not disclosed in this thesis due to commercial agreements (intellectual property and non-disclosure agreements).

During the development of exploratory and framework-based rating systems, numerous regression and classification-based evaluation metrics were used to evaluate model performance. Throughout this process the limitations of commonly applied performance metrics such as RMSE, MAPE, accuracy etc. was realised. It was realised that commonly used performance metrics were not completely suitable to evaluate the effectiveness of sport-based ratings, in that various performance metrics are required to evaluate their effectiveness but no single evaluation index can be applied across all systems and none is universally regarded as the ‘gold-standard’ metric to evaluate rating performance. To address this issue, a novel performance metric was developed to evaluate the effectiveness of ratings. Before constructing such a metric, the shortcomings of commonly used performance metrics needed to be understood and therefore a comprehensive review of evaluation metric was conducted.

A set of criteria were identified to construct a performance metric to quantify the effectiveness of meaningful sport-based rating systems. These criteria were: 1) sensitivity to distance, 2) sensitivity to time-dependence, 3) evaluates the ratings on the entire probability of distribution, 4) provides an incentive for calibration and sharp ratings and 5) adjusts incentives based on forecasting difficulty.

The literature review revealed that ensemble forecasts are generally assessed through two key statistics: reliability and resolution (i.e. calibration and sharpness, respectively). The reliability, or calibration, of a forecast indicates how confident the assessor is in their predictions and can be evaluated by comparing the standard deviation of the error in the ensemble mean with the forecast spread (Gneiting, Balabdaoui & Raftery, 2007). The resolution, or sharpness, of a forecast indicates how much the forecasts deviates from the climatological event frequency, given that the ensemble is reliable, increasing this deviation will increase the usefulness of the forecast. Therefore, given the five ideal criteria and the need for calibration and sharpness to assess ensembled ratings, a proper scoring rule methodology, specifically a spherical scoring rule, was identified as the most suitable approach to construct an evaluation metric to quantify the effectiveness of meaningful sport-based rating systems. Distance and magnitude-based measures derived from a non-local spherical scoring rule were used to develop this novel evaluation metric. This distance and magnitude-based spherical (DMS) metric implements an analytical hierarchy process (AHP), which enables the incorporation of prior knowledge surrounding sporting scenario and difficulty, and accounts for the time-element of sports.

Applying the DMS performance metric to team and player-based rating systems, specifically within the cricketing context, it was found to output prediction measures more aligned with actuality compared to traditional evaluation metrics such as the log-loss. This is because the DMS performance metric incorporates time-specific and scenario specific adjustments based on domain knowledge.

In conclusion this thesis has successfully identified the communalities and limitations of rating systems and developed a quantitative ratings framework to construct sport-based rating systems that produce meaningful ratings. The value of this framework was proved by demonstrating its applicability within the sporting context. Finally, the thesis developed a novel evaluation metric (DMS) to quantify the effectiveness of meaningful sport-based rating systems. which outperformed traditional metrics in certain forecasting scenarios. Successfully addressing the research objectives proves this thesis’s academic contribution.

6.2 FUTURE RESEARCH

This thesis assumes that performance is an ensemble of individual traits, without considering the probabilistic nature of each trait and measuring how each trait could change from state-to-state over time. Pentland & Liu (1999) described human behaviour as a set of dynamic models sequenced together by a Markov chain. Pentland & Liu (1999) considers the human as a device with many internal mental states, each with its own particular control behaviour and interstate transition probabilities (Pentland & Liu, 1999). Moreover, the state of each model was hierarchically organised to described both short-term and long-term behaviours. Future research could consider evaluating performance as an amalgamation of the work presented in this thesis and the work conducted by Pentland & Liu (1999). The framework could be extended by identifying long and short-term traits that affect performance. For example, when rating a cricket player there are short-term traits such as consecutive dots and consecutive boundaries, and long-term traits such as runs scored, percentage dots and percentage boundaries, that affect performance. In a dynamic Markov model, a probabilistic transition matrix could be assigned to measure the probability of moving from one state to another and anticipating the change for each player trait. Such an approach could possibly improve the framework as it would anticipate change in traits and where the trait transitions to, and therefore would adjust or recalculate performance based on probabilistic trait predictions. Such strategies of updating Markov chain models using ensembling techniques have been heavily investigated by Oliver, Chen & Naevdal (2010); Emerick & Reynolds (2011); Goodman & Weare (2010); Iba (2001); Moradkhani, DeChant & Sorooshian (2012); Posselt & Bishop (2012); Vrugt, Diks & Clark (2008).

Another area of future research is the application of the ratings framework to domains outside of sports. Although the ratings framework was only applied within the sporting-context for the purpose of this thesis, it is hypothesised that when applied to areas outside of sports, meaningful results can also be produced. Further research must be conducted to confirm this hypothesis. For example, there is a commercial need for such systems within digital marketing to assign ratings to online campaigns and evaluate their effectiveness and identify the key attributes and touchpoints that lead to sales and conversions.

Moreover, it is stated that the application of the DMS performance metric is not restricted to the evaluation of cricket-based rating systems and is applicable in any sporting code where an evaluation between the actual and predicted outcome is required. An area of future research is the application of the DMS performance metric outside the cricketing context. Therefore, future research into the applicability of the DMS metric across other sports, other than cricket is also recommended. It is hypothesised that when applied to areas outside of sports and cricket, the DMS metrics can still outperform traditional model evaluation metrics. Further research must be conducted to confirm this hypothesis. Further research should also be conducted into the type of weighting procedures that could be applied to update the rate of change attributes.

It is hypothesised that the DMS metric is applicable in any field where an evaluation between the actual and predicted outcome is required. To test this hypothesis, the metric should be applied in field such as digital marketing to evaluate the effectiveness of ratings assigned to online campaigns and key attributes and touchpoints that lead to sales and conversions.

REFERENCES

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. *Accounting and Management Information Systems*, *14*(1), 79.

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems with Applications, 73, 1-10.

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases. *Foundations of data organization and algorithms*, 69-84.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215-222). Springer, New York, NY.

Akhtar, S., & Scarf, P. (2012). Forecasting test cricket match outcomes in play. International Journal of Forecasting, 28(3), 632-643.

Akhtar, S., Scarf, P., & Rasool, Z. (2015). Rating players in test match cricket. *Journal of the Operational Research Society, 66(4)*, 684-695.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity, and predictive values. *Acta paediatrica*, *96*(3), 338-341.

Alabi, M. A., Issa, S., & Afolayan, R. B. (2013). An application of artificial intelligent neural network and discriminant analyses on credit scoring. *Journal of Modern Mathematics and Statistics*, *7(4)*, 47-54.

Alamar, B., & Mehrotra, V. (2011). Beyond 'Moneyball': The rapidly evolving world of sports analytics, Part I. *Analytics Magazine*.

Ala'raj, M., & Abbod, M. (2015, September). A systematic credit scoring model based on heterogeneous classifier ensembles. In *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). IEEE.

Alaraj, M., Abbod, M., & Hunaiti, Z. (2014, January). Evaluating Consumer Loans Using Neural Networks Ensembles. In *International Conference on Machine Learning, Electrical and Mechanical Engineering*.

Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (Eds.). (2017). Handbook of Statistical Methods and Analyses in Sports. CRC Press.

Aldous, D. (2017). Elo ratings and the sports model: A neglected topic in applied probability? *Statistical science*, *32*(4), 616-629.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, *13*(3), 469-475.

Allison, K., Crocker, G., Tran, H., & Carrieres, T. (2014). An ensemble forecast model of iceberg drift. *Cold Regions Science and Technology*, *108*, 1-9.

Almeida, L. G., Backović, M., Cliche, M., Lee, S. J., & Perelstein, M. (2015). Playing tag with ANN: boosted top identification with pattern recognition. *Journal of High Energy Physics*, *2015*(7), 86.

Al Maliki, A., Owen, G., & Bruce, D. (2012). *Combining AHP and TOPSIS approaches to support site selection for a lead pollution study* (Doctoral dissertation, IACSIT Press).

Alpaydin, E. (2012). *Introduction to Machine Learning* (2nd ed.). Massachusetts, USA: Massachusetts Institute of Technology.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus*, *43*(3), 332-357.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: receiver operating characteristic plots. *BMJ: British Medical Journal*, *309*(6948), 188.

Amini, M., Rezaeenour, J., & Hadavandi, E. (2015). A cluster-based data balancing ensemble classifier for response modeling in Bank Direct Marketing. *International Journal of Computational Intelligence and Applications*, *14*(04), 1550022.

Anagnostopoulos, Y., & Abedi, M. (2016). Risk pricing in emerging economies: credit scoring and private banking in Iran. *International Journal of Finance & Banking Studies*, 5(1), 51-72.

Analytics, M. (2016). The age of analytics: competing in a data-driven world.

Annis, D. H., & Craig, B. A. (2005). Hybrid paired comparison analysis, with applications to the ranking of college football teams. *Journal of Quantitative Analysis in Sports*, 1(1).

Argyle, M., & Little, B. R. (1972). Do personality traits apply to social behaviour? *Journal for the Theory of Social Behaviour*, *2*(1), 1-33.

Armstrong, J. S. (2001). Evaluating forecasting methods. In *Principles of forecasting* (pp. 443-472). Springer, Boston, MA.

Armstrong, J. S. (2001). Combining forecasts. In Principles of forecasting (pp. 417-439). Springer, Boston, MA.

Arsovski, S., Markoski, B., Pecev, P., Ratgeber, L., & Petrov, N. (2014, November). An ontology driven credit risk scoring model. In *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)* (pp. 301-305). IEEE.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, *12*(4), 929-935.

Avci, E., Ketter, W., & van Heck, E. (2018). Managing electricity price modeling risk via ensemble forecasting: The case of Turkey. Energy policy, 123, 390-403.

Babič, A. I. (2017). *A rating model for individual player qualities based on team results, applied in football* (Master's thesis, University of Twente).

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741-755.

Baez-Revueltas, F. B. (2009). *Residual logistic regression* (Doctoral dissertation, The Graduate School, Stony Brook University: Stony Brook, NY.).

Bahrammirzaee, A., Ghatari, A. R., Ahmadi, P., & Madani, K. (2011). Hybrid credit ranking intelligent system using expert system and artificial neural networks. *Applied Intelligence*, *34(1)*, 28-46.

Bailey, M., & Clarke, S. R. (2006). Predicting the match outcome in one day international cricket matches, while the game is in progress. *Journal of sports science & medicine*, *5*(4), 480.

Bakoben, M., Bellotti, T., & Adams, N. (2019). Identification of credit risk based on cluster analysis of account behaviours. *Journal of the Operational Research Society*, 1-9.

Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, *54(8)*, 822-832.

Banasik, J., & Crook, J. (2005). Credit scoring, augmentation, and lean models. *Journal of the Operational Research Society*, *56(9)*, 1072-1081.

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301-315.

Bartlett, J. (2014). The Hosmer-Lemeshow goodness of fit test for logistic regression. *Retrieved from The Stats Geek: http://thestatsgeekcom/2014/02/16/the-hosmer-lemeshow-goodness-offit-test-for-logistic= regression.*

Basel Committee. (2010). Basel III: A global regulatory framework for more resilient banks and banking systems. *Basel Committee on Banking Supervision, Basel*.

Bastos, J. (2007). Credit scoring with boosted decision trees.

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, *41*(1), 68-75.

Baxt, W. G. (1991). Use of an artificial neural network for the diagnosis of myocardial infarction. *Annals of internal medicine, 115*(11), 843-848.

Becker, S. (1991). Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, *2*(01n02), 17-33.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., & Frieder, O. (2008, June). Analysis of varying approaches to topical web query classification. In *Proceedings of the 3rd international conference on Scalable information systems* (p. 15). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Belanger, D., & McCallum, A. (2016, June). Structured prediction energy networks. In *International Conference on Machine Learning* (pp. 983-992).

Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical association*, *72*(357), 46-53.

Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting run production and run prevention in baseball: the impact of Sabermetrics. *Int J Bus Humanit Technol*, *2*(4), 67-75.

Benjamini, Y., & Hechtlinger, Y. (2013). Discussion: an estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. *Biostatistics*, *15*(1), 13-16.

Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, *134*, 213-227.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70-83.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bhattacharya, R., Gill, P. S., & Swartz, T. B. (2011). Duckworth–Lewis and twenty20 cricket. *Journal of the Operational Research Society*, *62*(11), 1951-1957.

Bhattacharjee, D., & Lemmer, H. H. (2016). Quantifying the pressure on the teams batting or bowling in the second innings of limited overs cricket matches. *International Journal of Sports Science & Coaching*, *11*(5), 683-692.

Bielecki, T. R., Cousin, A., Crépey, S., & Herbertsson, A. (2014). Dynamic hedging of portfolio credit risk in a Markov copula model. *Journal of Optimization Theory and Applications*, *161*(1), 90-102.

Bird & Bird LLP. (2018). *PASPA Repeal: What does it mean for the American sports betting market?* Retrieved from https://www.lexology.com/library/detail.aspx?g=b5b38add-e604-4424-ad0a-98001e155d4b.

Blume, M. (2017). *Sports Betting and R: How R is changing the sports betting world*. Retrieved from https://channel9.msdn.com/Events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/Sports-Betting-and-R-How-R-is-changing-the-sports-betting-world.

Bolger, F., & Rowe, G. (2015). The aggregation of expert judgment: do good things come to those who weight? *Risk Analysis, 35*(1), 5-11.

Bolton, C. (2010). *Logistic regression and its application in credit scoring* (Doctoral dissertation, University of Pretoria).

Bolton-Smith, C., Woodward, M., Tunstall-Pedoe, H., & Morrison, C. (2000). Accuracy of the estimated prevalence of obesity from self-reported height and weight in an adult Scottish population. *Journal of Epidemiology & Community Health*, *54*(2), 143-148.

Bouaguel, W., & Limam, M. (2015). An Ensemble Wrapper Feature Selection for Credit Scoring. In *Proceedings of Fourth International Conference on Soft Computing for Problem Solving* (pp. 619-631). Springer, New Delhi.

Bracewell, P. (2003). Monitoring meaningful rugby ratings. *Journal of Sports Sciences*, *21*(8), 611-620.

Bracewell, P., Blackie. E., Blain, P., & Boys, C. (2016, July 12). Understanding the impact of demand for talent on the observable performance of individuals. *Paper published in The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports*. (pp. 40-45). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Bracewell, P. J., Farhadieh, F., Jowett, C. A., Forbes, D. G., & Meyer, D. H. (2009). Was Bradman Denied His Prime? *Journal of Quantitative Analysis in Sports*, *5*(4).

Bracewell, P. J., Forbes, D. G., Jowett, C. A., & Kitson, H. I. (2009). Determining the evenness of domestic sporting competition using a generic rating engine. *Journal of Quantitative Analysis in Sports*, *5(1).*

Bracewell, P., Downs, M., and Sewell, J. The development of a performance-based rating system for limited overs cricket. *In MATHSPORT 2014* (2014).

Bracewell, P. J. (2015). *N.Z. Patent No. 076682*. Wellington, New Zealand: New Zealand Intellectual Property Office.

Bracewell, P., Blackie. E., Blain, P., & Boys, C. (2016, July 12). Understanding the impact of demand for talent on the observable performance of individuals. *Paper presented at The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports*. (pp. 40-45). Melbourne, Victoria, Australia.

Bracewell, P. J., Patel, A. K., Blackie, E. J., & Boys, C. (2017). Using a Predictive Rating System for Computer Programmers to Optimise Recruitment. *Journal of Cases on Information Technology (JCIT)*, *19*(3), 1-14.

Bracewell, P. J., Coomes, M., Nash, J., N., Rooney, S. J., Patel, A. K., & Meyer, D. H. (2017). Application of Reject Inference to T20 cricket bowlers: Calculating the probability of taking a wicket using a behavioural credit risk scorecard framework. *Australian & New Zealand Journal of Statistics.*

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324-345.

Brill, J. (1998). The importance of credit scoring models in improving cash flow and collections. *Business Credit*, *100(1)*, 16-17.

Broadie, M. (2012). Assessing golfer performance on the PGA TOUR. *Interfaces*, *42*(2), 146-165.

Broadie, M., & Rendleman, R. J. (2013). Are the official world golf rankings biased? *Journal of Quantitative Analysis in Sports*, *9*(2), 127-140.

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and forecasting*, *22*(3), 651-661.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The binormal assumption on precision-recall curves. In *2010 20th International Conference on Pattern Recognition* (pp. 4263-4266). IEEE.

Brooker, S., & Hogan, S. (2011). A Method for Inferring Batting Conditions in ODI Cricket from Historical Data.

Brown, P., Patel, A. K. & Bracewell, P. J. (2017, June 23). Optimising a Batting Order in Limited Overs Cricket using Survival Analysis. *Paper published in The Proceedings of the 17th MathSport International 2017 Conference Proceedings*. (pp. 71-80). Padua, Italy. ISBN: 978-88-6938-058-7.

Brown, P., Patel, A. K. & Bracewell, P. J. (2017). A Survival Analysis Approach to Optimising a Batting Order in One-Day International Cricket. *Journal of Quantitative Analysis in Sport.*

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*, 3446-3453.

Brown, P., Patel, A. K., & Bracewell, P. J. (2016, July 12). Real Time Prediction of Opening Batsmen Dismissal in Limited Overs Cricket. *Paper published in The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports*. (pp. 80-85). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Brownlee, J. (2014). Classification accuracy is not enough: More performance measures you can use. *Machine Learning Mastery, 21*.

Bücker, M., Kampen, M., Krämer, W. (2013). Reject inference in consumer credit scoring with nonignorable missing data. *Journal of Banking & Finance, 37*, 1040-1045.

Buizza, R. (2018). Ensemble forecasting and the need for calibration. In Statistical Postprocessing of Ensemble Forecasts (pp. 15-48). *Elsevier*.

Buja, A., Stuetzle, W., & Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*, *November 3*.

Bukiet, B., Harold, E. R., & Palacios, J. L. (1997). A Markov chain approach to baseball. *Operations Research, 45*(1), 14-23.

Bukiet, B., & Ovens, M. (2006). A mathematical modelling approach to one-day cricket batting orders. *Journal of sports science & medicine*, *5*(4), 495.

Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I. (2007, July). Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 63-70). ACM.

Bureau of Labor Statistics. (December 2017). *Occupational Outlook Handbook: Software Developers.* Retrieved from https://www.bls.gov/ooh/computer-and-information-technology/software-developers.html.

Cai, W., Yu, D., Wu, Z., Du, X., & Zhou, T. (2019). A hybrid ensemble learning framework for basketball outcomes prediction. *Physica A: Statistical Mechanics and its Applications*, 528, 121461.

Campbell, E & Patel. A. K. (2017, December 10th -14th). Optimising New Zealand Junior Rugby Weights Limits. *Paper Presented at The Joint Meeting of the 10th Asian Regional Section (ARS) of the International Association for Statistical Computing (IASC) and the NZ Statistical Association (NZSA).*

Campbell, E. C., Patel A. K., & Bracewell, P. J. (2018). Optimizing junior rugby weight limits in New Zealand. *Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport.

Campbell, E. C., Bracewell, P. J., Blackie, E., & Patel, A. K. (2018). The impact of Auckland junior rugby weight limits on player retention. *Journal of sport and health research*, 10(2), 317-326.

Cano, J. R., Herrera, F., & Lozano, M. (2007). Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability. *Data & Knowledge Engineering*, *60*(1), 90-108.

Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J. T., Cummings, E., & Yang, Q. (2009, July). Context-aware query classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-10). ACM.

Carlberger, J., Dalianis, H., Duneld, M., & Knutsson, O. (2006, May). Improving precision in information retrieval for Swedish using stemming. In *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*.

Carter, M., & Guthrie, G. (2004). Cricket interruptus: fairness and incentive in limited overs cricket matches. *Journal of the Operational Research Society*, *55*(8), 822-829.

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis*, *13*(4), 223-242.

CGI Group Inc. (2013). *Predictive Analytics: The rise and value of predictive analytics in enterprise decision making* [White paper]. Retrieved December 10, 2016, from CGI: https://www.cgi.com.

Chain, N. (2018). *Sports Betting Is Evolving More Rapidly Than Expected. The Medium.* Retrieved from https://medium.com/@nexuschain/sports-betting-is-evolving-more-rapidly-than-expected-72c010064024.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.

Charlton, G. (2013). *UK's online gambling sector worth £2bn in 2012: stats*. *Econsultancy*. Retrieved from http://econsultancy.com/nz/blog/62407-uk-s-online-gambling-sector-worth-2bn-in-2012-stats.

Chatfield, C. (1988). Apples, oranges and mean square error. *International Journal of Forecasting*, *4*(4), 515-518.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 875-886). Springer, Boston, MA.

Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, *24*(4), 433-441.

Chen, Y., Rennie, D., Cormier, Y., & Dosman, J. (2005). Sex specificity of asthma associated with objectively measured body mass index and waist circumference: the Humboldt study. *Chest*, *128*(4), 3048-3054.

Chen, X., Jiang, Y., Yu, K., Liao, Y., Xie, J., & Wu, Q. (2017). Combined time-varying forecast based on the proper scoring approach for wind power generation. *The Journal of Engineering*, *2017*(14), 2655-2659.

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical science*, 2-30.

Chen, X. (2007). Banking deregulation and credit risk: Evidence from the EU. *Journal of Financial Stability*, *2*(4), 356-390.

Chen, C., Twycross, J., & Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PloS one*, *12*(3), e0174202.

Cheng, C. L., & Garg, G. (2014). Coefficient of determination for multiple measurement error models. *Journal of Multivariate Analysis*, *126*, 137-152.

Cheng, C., Zhang, X. Y., Shao, X. H., & Zhou, X. D. (2016, October). Handwritten Chinese character recognition by joint classification and similarity ranking. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 507-511). IEEE.

Cheung, K. K. (2001). A review of ensemble forecasting techniques with a focus on tropical cyclone forecasting. *Meteorological Applications*, *8*(3), 315-332.

Choi, T. M., Hui, C. L., & Yu, Y. (Eds.). (2013). *Intelligent fashion forecasting systems: models and applications*. Springer Science & Business Media.

Chuang, C. L., & Huang, S. T. (2011). A hybrid neural network approach for credit scoring. *Expert Systems, 28(2),* 185-196.

Cision – PR Newswire. (2018). *The global gambling market is expected to reach revenues of over $525 billion by 2023*. Retrieved from https://www.prnewswire.com/news-releases/the-global-gambling-market-is-expected-to-reach-revenues-of-over-525-billion-by-2023-300714934.html.

Clarke, S. R. (1988). Dynamic programming in one-day cricket-optimal scoring rates. *Journal of the Operational Research Society,* 331-337.

Clarke, S. R. (2011). Rating non-elite tennis players using team doubles competition results. *Journal of the Operational Research Society, 62(7),* 1385-1390.

Clark, J. (2011). Changing the game-outlook for the global sports market to 2015. *PricewaterhouseCoopers. Luettavissa https://www. pwc. com/en_GX/gx/hospitality-leisure/pdf/changing-the-game-outlookfor-the-global-sports-market-to-2015. pdf (Luettu 17.2. 2014).*

Cochocki, A., & Unbehauen, R. (1993). *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc.

Colin, C., Lanoir, D., Touzet, S., Meyaud-Kraemer, L., Bailly, F., Trepo, C., & HEPATIS Group. (2001). Sensitivity and specificity of third-generation hepatitis C virus antibody detection assays: an analysis of the literature. *Journal of viral hepatitis*, *8*(2), 87-95.

Collins, G. S., & Altman, D. G. (2012). Predicting the 10-year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *Bmj*, *344*, e4181.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(Aug), 2493-2537.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*(7), 928-935.

Coomes, M. Comparison of reject inference methods on complete data with gradient

boosting machine variable selection. 2014.

Cormack, G. V., & Lynam, T. R. (2006, August). Statistical precision of information retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 533-540). ACM.

Craparo, E., Karatas, M., & Singham, D. I. (2017). A robust optimization approach to hybrid microgrid operation using ensemble weather forecasts. *Applied energy*, *201*, 135-147.

Crawley, M. J. (2012). *The R books*. John Wiley & Sons.

Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). *A comparison of discriminators under alternative definitions of credit default.* Paper presented at the IMA conference on credit scoring and credit control.

Cserepy, N., Ostrow, R., & Weems, B. (2015) Predicting the Final Score of Major League Baseball Games. (available at: http://cs229.stanford.edu/proj2015/113_report.pdf. Accessed 1 May 2018).

Cummings, R., Pennock, D. M., & Wortman Vaughan, J. (2016, July). The possibilities and limitations of private prediction markets. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 143-160). ACM.

Cummins, R. (2016). *Computer Science Tripos Part II: Information Retrieval- Linkage algorithms and web search, week 1, session 2 notes* [PowerPoint slides]. Retrieved from https://www.cl.cam.ac.uk/teaching/1516/InfoRtrv/lecture8-link-analysis-2x2.pdf.

Dahiya, S., Handa, S. S., & Singh, N. P. (2015). Credit scoring using ensemble of various classifiers on reduced feature set. *Industrija*, *43*(4), 163-174.

Dahl, M., Brun, A., Kirsebom, O., & Andresen, G. (2018). Improving short-term heat load forecasts with calendar and holiday data. *Energies*, *11*(7), 1678.

David, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2), 169-183.

Davis, J., Perera, H., & Swartz, T. B. (2015). A simulator for Twenty20 cricket. *Australian & New Zealand Journal of Statistics*, *57*(1), 55-71.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). ACM.

De Finetti, B. (1962). Does it make sense to speak of ‘good probability appraisers. The scientist speculates: An anthology of partly baked ideas, 257-364.

De Fontnouvelle, P., Jesus-Rueff, D., Jordan, J. S., & Rosengren, E. S. (2003). Using loss data to quantify operational risk. *Available at SSRN 395083*.

Delaney, L., Harmon, C., & Ryan, M. (2013). The role of noncognitive traits in undergraduate study behaviours. *Economics of Education Review*, *32*, 181-195.

DeLong, C., Terveen, L., & Srivastava, J. (2013, August). Teamskill and the NBA: applying lessons from virtual worlds to the real-world. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 156-161). IEEE.

DesJardins, S. L. (2002). An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher education*, *43*(5), 531-553.

Developer are in high demand. (2016). Retrieved from https://learningfuze.com/why-coding.

Dhamija, P. (2012). E-recruitment: A roadmap towards e-human resource management. *Researchers World, 3(3),* 33.

Ding, J., Chen, B., Liu, H., & Huang, M. (2016). Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and remote sensing letters*, *13*(3), 364-368.

Dodge, Y. (2008). Kolmogorov–Smirnov Test. *The concise encyclopedia of statistics*, 283-287.

Duckworth, F. C., & Lewis, A. J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society*, *49*(3), 220-227.

Duda, R.O., Hart, P.E., Stork, D.G. (2001). *Pattern classification* (2nd ed.). New York, USA: Wiley.

Dyte, D., & Clarke, S. R. (2000). A rating-based Poisson model for World Cup soccer simulation. *Journal of the Operational Research society, 51(8),* 993-998.

Dziuda, D.M. (2010). *Data mining for genomics and proteomics: Analysis of gene and protein expression data*. New Jersey, USA: John Wiley & Sons, Inc.

Ellickson, P. L., Bird, C. E., Orlando, M., Klein, D. J., & McCaffrey, D. F. (2003). Social context and adolescent health behavior: do school-level smoking prevalence affect students' subsequent smoking behavior? *Journal of Health and Social Behavior*, *44*(4), 525.

El Maarouf, I., Bradbury, J., Baisa, V., & Hanks, P. (2014, May). Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing. In *LREC* (pp. 1001-1006).

Elo, A. (1978). The Rating of Chess Players Past and Present. 1978. *London: Batsford*.

Emerick, A. A., & Reynolds, A. C. (2011, January). Combining the ensemble Kalman filter with Markov chain Monte Carlo for improved history matching and uncertainty characterization. In *SPE Reservoir Simulation Symposium*. Society of Petroleum Engineers.

Engsted, T. (2009). Statistical vs. Economic Significance in Economics and Econometrics: Further Comments on Mccloskey & Ziliak. *Journal of Economic Methodology, 16*, 393-408.

Ericsson, J., Dansingani, A., O’Hair, J., Jackson, K., & Edin, P. (2018). *KPMG: Data-driven growth*. Retrieved from https://advisory.kpmg.us/content/dam/advisory/en/pdfs/data-driven-growth-2018.pdf.

ESPNCricinfo. (2017). Trent Woodhill’s brave new, data-driven world. Retrieved from https://www.espncricinfo.com/story/_/id/18316732/jarrod-kimber-trent-woodhill-data-driven-methods-team-selection.

Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V., & Modha, D. S. (2015). Backpropagation for energy-efficient neuromorphic computing. In *Advances in Neural Information Processing Systems* (pp. 1117-1125).

Fadaei Noghani, F., & Moattar, M. (2017). Ensemble classification and extended feature selection for credit card fraud detection. *Journal of AI and Data Mining*, *5*(2), 235-243.

Fausett, L. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, Inc.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Feldmann, K., & Thorarinsdottir, T. (2012). Statistical postprocessing of ensemble forecasts for temperature: The importance of spatial modeling. *Ruprecht-Karls-Universitat Heidelberg*.

Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing*, *65*, 139-151.

Ferro, C. A., Richardson, D. S., & Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *15*(1), 19-24.

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1917-1923.

Figini, S., & Maggi, M. (2014). *Performance of credit risk prediction models via proper loss functions* (No. 064). University of Pavia, Department of Economics and Management.

Fogel, D. B., Hays, T. J., Hahn, S. L., & Quon, J. (2004). A self-learning evolutionary chess program. *Proceedings of the IEEE*, *92*(12), 1947-1954.

Foley-Train, J. (2014). Sports betting: Commercial and integrity issues. *Report prepared for the Association of British Bookmakers, European Gaming and Betting Association, European Sport Security Association and Remote Gambling Association. Retrieved January 21*, 2015.

Franses, P. H. (2016). A note on the mean absolute scaled error. *International Journal of Forecasting*, *32*(1), 20-22.

Freeze, R. A. (1974). An analysis of baseball batting order by Monte Carlo simulation. *Operations Research*, *22*(4), 728-735.

F. W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and climate extremes*, *18*, 65-74.

García, V., Mollineda, R. A., & Sánchez, J. S. (2009, June). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis* (pp. 441-448). Springer, Berlin, Heidelberg.

Gary, A. (2019). *The Size and Increase of the Global Sports Betting Market.* Retrieved from https://www.sportsbettingdime.com/guides/finance/global-sports-betting-market/.

Gjoreski, H., Kaluža, B., Gams, M., Milić, R., & Luštrek, M. (2015). Context-based ensemble method for human energy expenditure estimation. *Applied Soft Computing*, 37, 960-970.

Glickman, M. E., Hennessy, J., & Bent, A. (2016). A comparison of rating systems for competitive women's beach volleyball.

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of clinical epidemiology*, *67*(8), 850-857.

Gneiting, T., Raftery, A. E., Westveld III, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098-1118.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration, and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243-268.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, *17*(2), 211.

González, S., Mens, K., Colacioiu, M., & Cazzola, W. (2013, March). Context traits: dynamic behaviour adaptation through run-time trait recomposition. In *Proceedings of the 12th annual international conference on Aspect-oriented software development* (pp. 209-220). ACM.

Good. I.J. (1971). Comments on "Measuring information and uncertainty" (by R.J. Buehler), In Godambe, V.P. and Sprott, D.A. (eds), *Foundations of Statistical Inference,* Holt, Rinehart and Winston, Toronto, Canada, pp. 337-339.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. Journal of Forecasting, 23(1), 51-66.

Goodarzi, L., Banihabib, M. E., & Roozbahani, A. (2019). A decision-making model for flood warning system based on ensemble forecasts. *Journal of hydrology*, 573, 207-219.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, *5*(1), 65-80.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International journal of forecasting*, *15*(4), 405-408.

Gorman, R. P., & Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, *1*(1), 75-89.

Grady, N. W., Schryver, J. C., & Leuze, M. R. (1999). Mining for personal profiles. In *AFCEA Second Federal Data Mining Conference*.

Grant, M. J., Button, C. M., & Snook, B. (2017). An evaluation of interrater reliability measures on binary tasks using d prime. *Applied psychological measurement*, *41*(4), 264-276.

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media.

Greene, H. J., & Milne, G. R. (2010). Assessing model performance: The Gini statistic and its standard error. *Journal of Database Marketing & Customer Strategy Management*, *17*(1), 36-48.

Greer, S.J., Patel, A. K., Trowland, H.E., & Bracewell, P. J. (2018). The Impact of Injury on the Future Performance Ratings of Domestic T20 Cricketers. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hadi, M. N. (2003). Neural networks applications in concrete structures. *Computers & structures*, *81*(6), 373-381.

Halligan, S., Altman, D. G., & Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, *25*(4), 932-939.

Halder, S., Roy, A., & Chakraborty, P. K. (2017). The influence of personality traits on information seeking behaviour of students. *Malaysian Journal of Library & Information Science*, *15*(1), 41-53.

Hand, D. J. (1981). Discrimination and classification. Wiley Series in Probability and *Mathematical Statistics, Chichester: Wiley, 1981.*

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3),* 523-541.

Hanson, R. (2012). Logarithmic markets coring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, *1*(1), 3-15.

Hastie, T., Tibshirani, R., Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Haykin, S., & Network, N. (2004). A comprehensive foundation. *Neural networks*, *2*(2004), 41.

Hebb, D.O. (1949). *The Organisation of Behaviour*. New York: John Wiley & Sons. Introduction and Chapter 4 reprinted in Anderson & Rosenfeld (1988), pp. 45-56.

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, *4*.

Henley, W. E. (1995). *Statistical aspects of credit scoring* (Doctoral dissertation, The Open University).

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559-570.

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill™: a Bayesian skill rating system. In *Advances in neural information processing systems* (pp. 569-576).

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105-117.

Heinström, J. (2003). Five personality dimensions and their influence on information behaviour. *Information research*, *9*(1), 9-1.

Hill, D. (2019). *Sports betting bettors' sharps kicked out spanky William Hill New Jersey*. Retrieved from https://www.theringer.com/2019/6/5/18644504/sports-betting-bettors-sharps-kicked-out-spanky-william-hill-new-jersey.

Hollard, G., Massoni, S., & Vergnaud, J. C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, *80*(3), 363-387.

Holmes, D., & McCabe, M. C. (2002, April). Improving precision and recall for soundex retrieval. In *Proceedings. International Conference on Information Technology: Coding and Computing* (pp. 22-26). IEEE.

Hosmer, D. W., Taber, S., & Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: a case study. *American journal of public health*, *81*(12), 1630-1635.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hsieh, N. C. (2004). An integrated data mining and behavioural scoring model for analysing bank customers. *Expert systems with applications, 27(4),* 623-633.

Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert systems with Applications*, *37*(1), 534-545.

Hudson, D. (2017). Ensemble Verification Metrics. Australian Government, Bureau of Meteorology, ECMWF Annual Seminar, https://www.ecmwf.int/sites/default/files/elibrary/2017/17626-ensemble-verification-metrics.pdf.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting*, *26*(3), 460-470.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

Iba, Y. (2001). Extended ensemble monte carlo. *International Journal of Modern Physics C*, *12*(05), 623-656.

Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, *10(3),* 233-240.

Ingram, M. (2019). Gaussian Process Priors for Dynamic Paired Comparison Modelling. *arXiv preprint arXiv:1902.07378*.

Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1-60.

Jackson, K. (2016). *Assessing Player Performance in Australian Football Using Spatial Data* (Doctoral dissertation, PhD Thesis, Swinburne University of Technology).

Jackson, K. (2016). Measuring the similarity between players in Australian football. In *Thirteenth Australasian Conference on Mathematics and Computers in Sport, Melbourne*.

Jaffery, T., & Liu, S. X. (2009). Measuring campaign performance by using cumulative gain and lift chart. In *SAS Global Forum* (p. 196).

Jager, L. R., & Leek, J. T. (2013). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*(1), 1-12.

Jayadevan, V. (2002). A new method for the computation of target scores in interrupted, limited-over cricket matches. *Current Science, 83*(5), 577-586.

Jayadevan, V. (2004). An improved system for the computation of target scores in interrupted limited over cricket matches adding variations in scoring range as another parameter. *Current Science*, *86*(4), 515-517.

Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial finance, 18(6)*, 15-26.

Jensen, M. B. (2006). Characteristics of B2B adoption and planning of online marketing communications. *Journal of targeting, measurement, and analysis for marketing*, *14*(4), 357-368.

Jensen, M. B., & Jepsen, A. L. (2008). SMS/MMS: A rising star in online marketing communications? *Journal of Website Promotion*, *2*(3-4), 31-41.

Jewson, S. (2004). The problem with the Brier score. *arXiv preprint physics/0401046*.

Jhanwar, M. G., & Pudi, V. (2016, September). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016)*.

Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q, Chen, J. Tsai1, C.J., Zhang, S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics 2004, 5,* 81.

Jiang, H. (2009, May). Study on the performance measure of information retrieval models. In *2009 International Symposium on Intelligent Ubiquitous Computing and Education* (pp. 436-439). IEEE.

Jiao, J., Courtade, T. A., Venkat, K., & Weissman, T. (2015). Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, *61*(10), 5357-5365.

Johnson, R. A., & Wichern, D. W. (2014). *Applied multivariate statistical analysis* (Vol. 4). New Jersey: Prentice-Hall.

Jordan, A., Krüger, F., & Lerch, S. (2017). Evaluating probabilistic forecasts with scoringRules. *arXiv preprint arXiv:1709.04743*.

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision,* 66(3), 263-281.

Joyce, J. M. (2011). Kullback-Leibler divergence. *International encyclopedia of statistical science*, 720-722.

Kaluarachchi, A., & Aparna, S. V. (2010, December). CricAI: A classification-based tool to predict the outcome in ODI cricket. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on* (pp. 250-255). IEEE.

Kampe, T., Edman, G., Bader, G., Tagdae, T., & Karlsson, S. (1997). Personality traits in a group of subjects with long-standing bruxing behaviour. *Journal of oral rehabilitation*, *24*(8), 588-593.

Katz, R. W., & Murphy, A. H. (Eds.). (2005). *Economic value of weather and climate forecasts*. Cambridge University Press.

Kauppi, T., Kamarainen, J. K., Lensu, L., Kalesnykiene, V., Sorri, I., Kälviäinen, H., ... & Pietilä, J. (2009, June). Fusion of multiple expert annotations and overall score selection for medical image diagnosis. In *Scandinavian Conference on Image Analysis* (pp. 760-769). Springer, Berlin, Heidelberg.

Kelz, R. R., Gimotty, P. A., Polsky, D., Norman, S., Fraker, D., & DeMichele, A. (2004). Morbidity and mortality of colorectal carcinoma surgery differs by insurance status. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *101*(10), 2187-2194.

Kennedy, K., Mac Name, B., Delany, S. J., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications, 40(4),* 1372-1380.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34(11),* 2767-2787.

Kharrat, T., Pena, J. L., & McHale, I. (2017). Plus-minus player ratings for soccer. *arXiv preprint arXiv:1706.04943*.

Kim, Y. (2009). Boosting and measuring the performance of ensembles for a successful database marketing. *Expert Systems with Applications*, *36*(2), 2161-2176.

Kim, I. (2016). Directors' and officers' insurance and opportunism in accounting choice. *Accounting & Taxation, 7*(1), 51-65.

Kim, E., Lee, J., Shin, H., Yang, H., Cho, S., Nam, S. K., ... & Kim, J. I. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications*, *128*, 214-224.

Király, F. J., & Qian, Z. (2017). Modelling Competitive Sports: Bradley-Terry-\'{E} l\H {o} Models for Supervised and On-Line Learning of Paired Competition Outcomes. *arXiv preprint arXiv:1701.08055*.

Kitts, B., Freed, D., & Vrieze, M. (2000, August). Cross-sell: a fast promotion-tuneable customer-item recommendation method based on conditionally independent probabilities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 437-446). ACM.

Klehe, U. C., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology*, *92*(4), 978.

Klotz, L., Vesprini, D., Sethukavalan, P., Jethava, V., Zhang, L., Jain, S., ... & Loblaw, A. (2014). Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. *Journal of Clinical Oncology*, *33*(3), 272-277.

Klugman, S. A., & Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance: mathematics and economics*, *24*(1-2), 139-148.

Koehrsen, W. (2018). Beyond accuracy: Precision and recall. *Towards Data Science*.

Kunst, R. M., & Jumah, A. (2004). *Toward a theory of evaluating predictive accuracy* (No. 162). Reihe Ökonomie/Economics Series, Institut für Höhere Studien (IHS).

Lachmi, K. (2018). *The Evolution of Ratings.* Retrieved from https://www.forbes.com/sites/forbestechcouncil/2018/05/23/the-evolution-of-ratings/#2fcd09e15b72.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (pp. 6402-6413).

Landgrebe, D. (2000, July). On the relationship between class definition precision and classification accuracy in hyperspectral analysis. In *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)* (Vol. 1, pp. 147-149). IEEE.

Langley, P. (1997, September). Machine learning for adaptive user interfaces. In *Annual Conference on Artificial Intelligence* (pp. 53-62). Springer, Berlin, Heidelberg.

Lanz, C., Marti, U., & Thormann, W. (2003). Capillary zone electrophoresis with a dynamic double coating for analysis of carbohydrate-deficient transferrin in human serum: precision performance and pattern recognition. *Journal of Chromatography A*, *1013*(1-2), 131-147.

Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, *1*(1), 27-46.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Lee, T. S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28(4)*, 743-752.

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, *26(3)*, 471-481.

Lemmer, H. H. (2011). The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of sports science & medicine*, *10*(4), 630.

Lessmann, S., Sung, M. C., Johnson, J. E., & Ma, T. (2012). A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. *European Journal of Operational Research*, 218(1), 163-174.

Leung, K., Cheong, F., & Cheong, C. (2007, September). Consumer credit scoring using an artificial immune system algorithm. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on* (pp. 3377-3384). IEEE.

Lewis, A. J. (2005). Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society*, *56*(7), 804-815.

Lichtendahl Jr, K. C., & Winkler, R. L. (2019). Why do some combinations perform better than others? *International Journal of Forecasting*.

Li, D. L., Shen, F., Yin, Y., Peng, J. X., & Chen, P. Y. (2013). Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chinese medical journal*, *126*(6), 1150-1154.

Li, L., Zhong, L., Xu, G., & Kitsuregawa, M. (2012). A feature-free search query classification approach using semantic distance. *Expert Systems with Applications*, *39*(12), 10739-10748.

Lin, A. Z. (2013). Variable reduction in SAS by using weight of evidence and information value. In *SAS Global Forum* (pp. 095-213).

Linder, J. A., Rigotti, N. A., Brawarsky, P., Kontos, E. Z., Park, E. R., Klinger, E. V., ... & Haas, J. S. (2013). Peer Reviewed: Use of Practice-Based Research Network Data to Measure Neighborhood Smoking Prevalence. *Preventing chronic disease*, *10*.

Liu, H., Xu, Y., & Chen, C. (2019). Improved pollution forecasting hybrid algorithms based on the ensemble method. *Applied Mathematical Modelling*, 73, 473-486.

Liu, L., Zhan, M., & Bai, Y. (2019). A recursive ensemble model for forecasting the power output of photovoltaic systems. *Solar Energy*, 189, 291-298.

Loh, W. Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, *3*(4), 1710-1737.

Lopes, R. H. (2011). Kolmogorov-smirnov test. *International encyclopedia of statistical science*, 718-720.

Machete, R. L. (2013). Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, *143*(10), 1781-1790.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions, and implications. *International journal of forecasting*, *16*(4), 451-476.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines. *Methods of information in medicine, 51*(01), 74-81.

Manly, B. F., & Alberto, J. A. N. (2016). *Multivariate statistical methods: a primer*. Chapman and Hall/CRC.

Mansell, Z., Patel, A. K., McIvor, J. T, & Bracewell, P. J. (2018). Managing run rate in T20 cricket to maximize the probability of victory when setting a total. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Marikkannu, P., & Shanmugapriya, K. (2011, April). Classification of customer credit data for intelligent credit scoring system using fuzzy set and MC2—Domain driven approach. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 3, pp. 410-414). IEEE.

Marin, J., Robert, C.P. (2014). *Bayesian Essentials with R* (2nd ed.). New York, USA: Springer.

Maskey, S. (2004). *Modelling uncertainty in flood forecasting systems*. CRC Press.

Masood, S. Z., Ellis, C., Nagaraja, A., Tappen, M. F., LaViola, J. J., & Sukthankar, R. (2011, November). Measuring and reducing observational latency when recognizing actions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 422-429). IEEE.

Massey, K. (1997). Statistical models applied to the rating of sports teams. *Bluefield College*.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68-78.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087-1096.

McHale, I. G., & Asif, M. (2013). A modified Duckworth–Lewis method for adjusting targets in interrupted limited overs cricket. *European Journal of Operational Research*, *225*(2), 353-362.

McIvor, J. T, Patel, A. K., Hilder, T.A., & Bracewell, P. J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Mease, D. (2003). A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins. *The American Statistician, 57(4),* 241-248.

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating forecasting methods by considering different accuracy measures. *Procedia Computer Science*, *95*, 264-271.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., ... & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, *21*(1), 1.

Mendoza, N. S., Rose, R. A., Geiger, J. M., & Cash, S. J. (2016). Risk assessment with actuarial and clinical methods: Measurement and evidence-based practice. *Child abuse & neglect*, *61*, 1-12.

Minka, T., Cleven, R., & Zaykov, Y. (2018). Trueskill 2: An improved bayesian skill rating system. *Tech. Rep.*

Misztal, M. (2014). On the Selected Methods for Evaluating Classification Models. *Acta Universitatis Lodziensis. Folia Oeconomica*, *3*(302).

Mohammed, A. A., Naugler, C., & Far, B. H. (2015). Emerging business intelligence framework for a clinical laboratory through big data analytics. *Emerging trends in computational biology, bioinformatics, and systems biology: algorithms and software tools. New York: Elsevier/Morgan Kaufmann*, 577-602.

Moore, W. E., Rooney, S. J., Bracewell, P. J., & Ray. S. (2018). Systematic optimization of the Elo rating system. *Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Moore, W. E., McIvor, J. T., & Bracewell, P. J. (2018). Deriving result-driven rugby ratings. *Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Moradkhani, H., DeChant, C. M., & Sorooshian, S. (2012). Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. *Water Resources Research*, *48*(12).

Morgan, J.N., Sonquist, J.A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association, 58*, 415-434.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, *12*(1), 215-223.

Murphy, A. H., & Winkler, R. L. (1982). Subjective probabilistic tornado forecasts: Some experimental results. *Monthly Weather Review*, 110(9), 1288-1297.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79(387), 489-500.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, *7*(4), 435-455.

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015, February). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691-692.

Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632). ACM.

Nock, R., Ali, W. B. H., D'Ambrosio, R., Nielsen, F., & Barlaud, M. (2014). Gentle nearest neighbors boosting over proper scoring rules. *IEEE transactions on pattern analysis and machine intelligence*, *37*(1), 80-93.

Nomura, S., Oyama, S., Hayamizu, T., & Ishida, T. (2004). Analysis and improvement of hits algorithm for detecting web communities. *Systems and Computers in Japan*, *35*(13), 32-42.

Nonhoff, C., Rottiers, S., & Struelens, M. J. (2005). Evaluation of the Vitek 2 system for identification and antimicrobial susceptibility testing of Staphylococcus spp. *Clinical microbiology and infection*, *11*(2), 150-153.

Noordzij, M., Dekker, F. W., Zoccali, C., & Jager, K. J. (2010). Measures of disease frequency: prevalence and incidence. *Nephron Clinical Practice*, *115*(1), c17-c20.

Offerman, T., Sonnemans, J., Van de Kuilen, G., & Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, *76*(4), 1461-1489.

Ohkusa, Y. (2001). An empirical examination of the quit behavior of professional baseball players in Japan. *Journal of Sports Economics, 2(1),* 80-88.

Oliver, D. S., Chen, Y., & Nævdal, G. (2011). Updating Markov chain models using the ensemble Kalman filter. *Computational Geosciences*, *15*(2), 325-344.

O'Neill, D. (2015). Measuring obesity in the absence of a gold standard. *Economics & Human Biology*, *17*, 116-128.

O’Riley, B. J., & Ovens, M. (2006). Impress Your Friends and Predict the Final Score: An analysis of the psychic ability of four target resetting methods used in One-Day International Cricket. *Journal of sports science & medicine*, *5*(4), 488.

Pacelli, V., & Azzollini, M. (2011). An artificial neural network approach for credit risk management. *Journal of Intelligent Learning Systems and Applications, 3(02),* 103.

Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, *61*, 27-40.

Pan, F., Converse, T., Ahn, D., Salvetti, F., Donato, G. (2009). Feature Selection for Ranking using Boosted Trees. *Paper presented at Conference on Information and Knowledge Management*, Hong Kong, China.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118, 33-45.

Papakonstantinou, A., & Pinson, P. (2016). Information uncertainty in electricity markets: Introducing probabilistic offers. *IEEE Transactions on Power Systems*, *31*(6), 5202-5203.

Pardowitz, T., Osinski, R., Kruschke, T., & Ulbrich, U. (2016). An analysis of uncertainties and skill in forecasts of winter storm losses. *Natural Hazards and Earth System Sciences*, *16*, 2391-2402.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity, and predictive values. *Indian journal of ophthalmology*, *56*(1), 45.

Parry, M. (2016). Extensive scoring rules. *Electronic Journal of Statistics*, 10(1), 1098-1108.

Patel, A. (2016). Roster-Based Optimisation for Limited Overs Cricket. *Unpublished Master’s Thesis*. Victoria University of Wellington.

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2016, July 12). Team Rating Optimisation for T20 Cricket. *Paper published in The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports*. (pp. 91-96). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2017). An Individual-Based Team Rating Method for T20 Cricket. *Journal of Sports and Human Performance 5(1): 1-17.*

Patel, A. K., Bracewell, P. J., Gazley, A. J., Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. *Journal of Sports Science and Coaching 12(3): 328-338.*

Patel, A. K., Bracewell, P. J., & Wells, J. D. (2017, June 23). Real-time measurement of individual influence in T20 cricket. *Paper published in The Proceedings of the 17th MathSport International 2017 Conference Proceedings*. (pp. 61-70). Padua, Italy. ISBN: 978-88-6938-058-7.

Patel, A. K., & Bracewell, P. J. (2018). A framework for quantifying the effectiveness of human-based rating systems. Paper presented at *The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel. A. K., Bracewell. P.J., & Bracewell, M.G. (2018). Estimating expected total in the first innings of T20 cricket using gradient boosted learning. Paper presented at *The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Bracewell, P. J., Wells. J. D., & Brown, P. (2018). Prediction football crowd attendance with public data. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Rooney. S. J., Bracewell, P. J., & Wells. J. D. (2018). Constructing a predictive PGA performance rating using hierarchical variable clustering. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Cook, M. K. A., Bracewell, P. J., & West, M. B. (2018). A framework to quantify the impact of social engagement on data driven creative. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., & Bracewell, P. J. (2019). Dynamic evaluation of player performance in T20 cricket. *Journal of Quantitative Analysis in Sport.*

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box data-driven explanation of black box decision systems. *arXiv preprint arXiv:1806.09936*.

Peimankar, A., Weddell, S. J., Jalal, T., & Lapthorn, A. C. (2018). Multi-objective ensemble forecasting with an application to power transformers. *Applied Soft Computing, 68, 233-248.*

Pentland, A., & Liu, A. (1999). Modeling and prediction of human behavior. *Neural computation*, *11*(1), 229-242.

Perera, H. P., & Swartz, T. B. (2012). Resource estimation in T20 cricket. *IMA Journal of Management Mathematics*, *24*(3), 337-347.

Pettersson, N. (2005, May). Measuring precision for static and dynamic design pattern recognition as a function of coverage. In *ACM SIGSOFT Software Engineering Notes* (Vol. 30, No. 4, pp. 1-7). ACM.

Peussa, A. (2016). Credit risk scorecard estimation by logistic regression.

Plomin, R., Owen, M. J., & McGuffin, P. (1994). The genetic basis of complex human behaviors. *Science*, *264*(5166), 1733-1739.

Posselt, D. J., & Bishop, C. H. (2012). Nonlinear parameter estimation: Comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. *Monthly Weather Review*, *140*(6), 1957-1974.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, Informedness, markedness and correlation.

Prasad, K., Dash, S. K., & Mohanty, U. C. (2010). A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts. *International Journal of Climatology*, *30*(10), 1577-1588.

Pratt, M. K., & White, S. K. (2018). What is business analyst? A key role for business–IT efficiency. *CIO FROM IDG*.

Preston, I., & Thomas, J. (2000). Batting strategy in limited overs cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(1), 95-106.

Pretto, L. (2002, September). A theoretical analysis of Google's PageRank. In *International Symposium on String Processing and Information Retrieval* (pp. 131-144). Springer, Berlin, Heidelberg.

Qin, Q., Xie, K., He, H., Li, L., Chu, X., Wei, Y. M., & Wu, T. (2019). An effective and robust decomposition-ensemble energy price forecasting paradigm with local linear prediction. *Energy Economics*, *83*, 402-414.

Quinlan, J. R. (1993). *C4.5 programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Raboin, B. (2013). Accepting a Double-Fault: How ADR Might Save Men's Professional Tennis. *Miss. Sports L. Rev.*, *3*, 211.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885-3900.

Rello, J., Luján, M., Gallego, M., Vallés, J., Belmonte, Y., Fontanals, D., ... & PROCORNEU Study Group. (2010). Why mortality is increased in health-care-associated pneumonia: lessons from pneumococcal bacteremic pneumonia. *Chest*, *137*(5), 1138-1144.

Robu, V., Chalkiadakis, G., Kota, R., Rogers, A., & Jennings, N. R. (2016). Rewarding cooperative virtual power plant formation using scoring rules. *Energy*, *117*, 19-28.

Roby, T. B. (1965). Belief states and the uses of evidence. *Behavioral Science*, 10(3), 255-270.

Rocco, B., de Cobelli, O., Leon, M. E., Ferruti, M., Mastropasqua, M. G., Matei, D. V., ... & Djavan, B. (2006). Sensitivity and detection rate of a 12-core trans-perineal prostate biopsy: preliminary report. *European urology*, *49*(5), 827-833.

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Roughgarden, T., & Schrijvers, O. (2017). Online prediction with selfish experts. *In Advances in Neural Information Processing Systems* (pp. 1300-1310).

Rosset, S., Neumann, E., Eick, U., Vatnik, N., & Idan, I. (2001, August). Evaluation of prediction models for marketing campaigns. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 456-461). ACM.

Rouam S. (2013) False Discovery Rate (FDR). In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) *Encyclopedia of Systems Biology*. Springer, New York, NY

Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of clinical epidemiology*, *63*(8), 938-939.

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 419-430.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*, 40.

Saaty, T. L. (1988). What is the analytic hierarchy process? In *Mathematical models for decision support* (pp. 109-121). Springer, Berlin, Heidelberg.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Sano, S. M., Quarracino, M. C., Aguas, S. C., González, E. J., Harada, L., Krupitzki, H., & Mordoh, A. (2008). Sensitivity of direct immunofluorescence in oral diseases. Study of 125 cases. *Medicina Oral Patologia Oral y Cirugia Bucal*, *13*(5), 287.

Saunders, L. L., Krause, J. S., & Acuna, J. (2012). Association of race, socioeconomic status, and health care access with pressure ulcers after spinal cord injury. *Archives of physical medicine and rehabilitation*, *93*(6), 972-977.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783-801.

Scarf, P., & Shi, X. (2005). Modelling match outcomes and decision support for setting a final innings target in test cricket. *IMA Journal of Management Mathematics*, 16(2), 161-178.

Schapire, R.E. (1999), A brief introduction to boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.*

Schärli, N., Ducasse, S., Nierstrasz, O., & Black, A. P. (2003, July). Traits: Composable units of behaviour. In *European Conference on Object-Oriented Programming* (pp. 248-274). Springer, Berlin, Heidelberg.

Scheuerer, M., & Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, *143*(4), 1321-1334.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*, *18*(3), 457-490.

Scully, G. W. (1994). Managerial efficiency and survivability in professional team sports. *Managerial and Decision Economics, 15(5),* 403-411.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

Seddon, P. B., Constantinidis, D., Tamm, T., & Dod, H. (2017). How does business analytics contribute to business value? *Information Systems Journal*, *27*(3), 237-269.

Seliya, N., Khoshgoftaar, T. M., & Van Hulse, J. (2009, November). A study on the relationships of classifier performance metrics. In *2009 21st IEEE international conference on tools with artificial intelligence* (pp. 59-66). IEEE.

Setiono, R., Thong, J. Y., & Yap, C. S. (1998). Symbolic rule extraction from neural networks: An application to identifying organizations adopting IT. *Information & management*, *34(2),* 91-101.

Shad, M. Y., Rehman, M. K. (2012). Credit Risk Modelling/ Scorecard. Retrieved from https://pdfs.semanticscholar.org/presentation/e026/4e75803235de2e152668a8637c78983d7cdc.pdf.

Shah, A., Jha, D., & Vyas, J. (2016). Winning and Score Predictor (WASP) Tool. *International Journal of Innovative Research in Science and Engineering.*

Shah, N., Zhou, D., & Peres, Y. (2015, June). Approval voting and incentives in crowdsourcing. In *International Conference on Machine Learning* (pp. 10-19).

Sharma, D., Overstreet, G., Beling, P. (2009). Not if affordability data adds value but how to add real value by leveraging affordability data: Enhancing predictive capability of credit scoring using. *Affordability Data. CAS (Casualty Actuarial Society) Working Paper*.

Sharma, D. (2012). Improving the art, craft and science of economic credit risk scorecards using random forests: why credit scorers and economists should use random forests. *Academy of Banking Studies Journal, 11*, 93-115.

Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, *526*, 121073.

Shung, K. P. (2018). Accuracy, Precision, Recall or F1? *Towards Data Science*.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring* (Vol. 3). John Wiley & Sons.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., ... & Zwiers, F. W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and climate extremes*, 18, 65-74.

Silver, N., & Fischer-Baum, R. (2015). CARMELO NBA player projections.

Silverman, S. (2019). Legalized Sports Gambling Passes $10 Billion, Likely just Tip of the Iceberg. Retrieved from https://www.forbes.com/sites/stevesilverman/2019/08/29/legalized-sports-gambling-passes-10-billion-likely-just-tip-of-the-iceberg/#185ade1cc223.

Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual differences*, 44(4), 1012-1021.

Simmonds, P., Patel, A. K., & Bracewell, P. J. (2018). Using network analysis to determine optimal batting partnership in T20 cricket. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Singh, T., Singla, V., & Bhatia, P. (2015, October). Score and winning prediction in cricket through data mining. In *Soft Computing Techniques and Implementations (ICSCTI), 2015 Int.Conference* (pp. 60-66). IEEE.

Singh, S. (2017). E-Recruitment: a new dimension of human resource management in India. *International Journal*, *5*(3).

Sing'oei, L., & Wang, J. (2013). Data mining framework for direct marketing: A case study of bank marketing. *International Journal of Computer Science Issues (IJCSI)*, *10*(2 Part 2), 198.

Sinuany-Stern, Z. (1988). Ranking of sports teams via the AHP. *Journal of the Operational Research Society*, *39*(7), 661-667.

Smith, B. C. (2011). Stability in consumer credit scores: Level and direction of FICO score drift as a precursor to mortgage default and prepayment. *Journal of Housing Economics*, *20*(4), 285-298.

Smits, N. (2010). A note on Youden's J and its cost ratio. *BMC medical research methodology*, *10*(1), 89.

Soper, T. (2014). Analysis: the exploding demand for computer science education, and why America needs to keep up. *Geekwire*.

Sorensen, S. P. (2000). An overview of some methods for ranking sports teams. *University of Tennessee. Knoxville.*

Soureshjani, M. H., & Kimiagari, A. M. (2013). Calculating the best cut off point using logistic regression and neural network on credit scoring problem-A case study of a commercial bank. *African Journal of Business Management*, *7*(16), 1414-1421.

Sun, H., & Guo, M. (2015, December). Credit risk assessment model of small and medium-sized enterprise based on logistic regression. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1714-1717). IEEE.

Sun, S., Wang, S., & Wei, Y. (2019). A new multiscale decomposition ensemble approach for forecasting exchange rates. *Economic Modelling*, *81*, 49-58.

Surma, J., & Furmanek, A. (2010, August). Improving marketing response by data mining in social network. In *2010 International Conference on Advances in Social Networks Analysis and Mining* (pp. 446-451). IEEE.

Su, W., Yuan, Y., & Zhu, M. (2015, September). A relationship between the average precision and the area under the ROC curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (pp. 349-352). ACM.

Standard, M. J. (2018). The Statistical Definition of Entropy. *Chemistry 360: Fall 2018*.

Stanski, H. R., Wilson, L. J., & Burrows, W. R. (1989). Survey of common verification methods in meteorology.

Stefani, R. T. (1997). Survey of the major world sports rating systems. *Journal of Applied Statistics*, *24(6)*, 635-646.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4).

Steinberg, L. (2015). Changing the game: the rise of sports analytics. *Forbes. Retrieved March 14*, 2017.

Steiner, S. H. (1999). EWMA control charts with time-varying control limits and fast initial response. *Journal of Quality Technology*, *31*(1), 75-86.

Stephenson, D. B., Coelho, C. A., & Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, *23*(4), 752-757.

Stern, S. E. (2016). The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society*, 67(12), 1469-1480.

Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and remote sensing*, *52*(3), 397-399.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. *BMC Bioinformatics*, *8*, 25.

Swartz, T. (2003). The best batsmen and bowlers in one-day cricket: general. *South African Statistical Journal*, *37*(2), 203-222.

Swartz, T. B., Gill, P. S., & Muthukumarana, S. (2009). Modelling and simulation for one-day cricket. *Canadian Journal of Statistics*, *37*(2), 143-160.

Tang, L., Wu, Y., & Yu, L. (2018). A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting. *Applied Soft Computing*, *70*, 1097-1108.

Thibodeau, P. (2012). IT jobs will grow 22% through 2020, says *US Computerworld*.

Thomas, L. C. (2000). A survey of credit and behavioural scoring forecasting financial risk of lending to consumers. *International journal of forecasting*, *16(2)*, 149-172.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2004). Readings in Credit Scoring: recent developments, advances, and aims.

Thomson, M. E., Pollock, A. C., Önkal, D., & Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, *35*(2), 474-484.

TotalSportTrek (April 2015). 25 World's Most Popular (Ranked by 13 factors). Retrieved from http://www.totalsportek.com/most-popular-sports/.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health*, *5*, 307.

Tsaih, R., Liu, Y. J., Liu, W., & Lien, Y. L. (2004). Credit scoring system for small business loans. *Decision Support Systems*, *38*(1), 91-99.

van Strien, T. (1986). *Eating behaviour, personality traits and body mass* (Doctoral dissertation, Van Strien).

Vaziri, B., Dabadghao, S., Yih, Y., & Morin, T. L. (2018). Properties of sports ranking methods. *Journal of the Operational Research Society*, *69*(5), 776-787.

Veček, N., Mernik, M., & Črepinšek, M. (2014). A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms. *Information Sciences*, *277*, 656-679.

Veček, N., Mernik, M., Filipič, B., & Črepinšek, M. (2016). Parameter tuning with Chess Rating System (CRS-Tuning) for meta-heuristic algorithms. *Information Sciences*, *372*, 446-469.

Vrooman, J. (2012). The economic structure of the NFL. In *the Economics of the National Football League* (pp. 7-31). Springer New York.

Vrugt, J. A., Diks, C. G., & Clark, M. P. (2008). Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environmental fluid mechanics*, *8*(5-6), 579-595.

Vuk, M., & Curk, T. (2006). ROC curve lift chart and calibration plot. *Metodoloski zvezki*, *3*(1), 89.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, *38*(1), 223-230.

Wang, G., Ma, J., Huang, L., Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61-68.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, *87*, 12-20.

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27*, 74-82.

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The American statistician*, *50*(4), 284-293.

Weiten, W. (2007). *Psychology: Themes and variations: Themes and variations*. Cengage Learning.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27(11*), 1131-1152.

West, B. T., & Lamsal, M. (2008). A new application of linear modeling in the prediction of college football bowl outcomes and the development of team ratings. *Journal of Quantitative Analysis in Sports, 4(3)*.

Wilks, D. S. (2010). Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 2109-2118.

Wilks, D. S. (2018). Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, *144*(710), 76-84.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of applied Meteorology, 7*(5), 751-758.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, *64*(327), 1073-1078.

Winkler, R. L., & Murphy, A. H. (1970). Nonlinear utility and the probability score. *Journal of Applied Meteorology*, *9*(1), 143-148.

Winkler, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science*, *40*(11), 1395-1405.

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., ... & Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1-60.

Yang, Y., Hong, W., & Li, S. (2019). Deep ensemble learning based probabilistic load forecasting in smart grids. *Energy*, 116324.

Yao, K., Zweig, G., Hwang, M. Y., Shi, Y., & Yu, D. (2013, August). Recurrent neural networks for language understanding. In *Interspeech* (pp. 2524-2528).

Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics, 11(2),* 111-125.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Young, R. M. B. (2010). Decomposition of the Brier score for weighted forecast-verification pairs. *Quarterly Journal of the Royal Meteorological Society*, *136*(650), 1364-1370.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert systems with applications*, *34*(2), 1434-1444.

Yu, W., Nakakita, E., & Jung, K. (2016). Flood Forecast and Early Warning with High-Resolution Ensemble Rainfall from Numerical Weather Prediction Model. *Procedia Engineering*, *154*, 498-503.

Zandi, M. (1998). Incorporating economic information into credit risk underwriting. *Credit Risk Modeling: Design and Application (Dearborn Publishers, Chicago/ London)*.

Zhang, A. (2009). Statistical Methods in Credit Risk Modeling.

Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician*, *71*(4), 310-316.

Zhang, H., He, H., & Zhang, W. (2018). Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. *Neurocomputing*, *316*, 210-221.

Zhou, L., Lai, K. K., & Yen, J. (2009). Credit scoring models with AUC maximization based on weighted SVM. *International journal of information technology & decision making*, *8*(04), 677-696.

Zhuang, Y., Liu, X., & Pan, Y. (1999, December). Apply semantic template to support content-based image retrieval. In *Storage and Retrieval for Media Databases 2000* (Vol. 3972, pp. 442-449). International Society for Optics and Photonics.

Zhao, T., Wang, Q. J., Schepen, A., & Griffiths, M. (2019). Ensemble forecasting of monthly and seasonal reference crop evapotranspiration based on global climate model outputs. *Agricultural and forest meteorology*, *264*, 114-124.

Zhu, Y., Zhou, L., Xie, C., Wang, G. J., & Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, *211*, 22-33.

Zięba, M., & Świątek, J. (2012, February). Ensemble classifier for solving credit scoring problems. In *Doctoral Conference on Computing, Electrical and Industrial Systems* (pp. 59-66). Springer, Berlin, Heidelberg.

Ziliani, M. G., Ghostine, R., Ait-El-Fquih, B., McCabe, M. F., & Hoteit, I. (2019). Enhanced flood forecasting through ensemble data assimilation and joint state-parameter estimation. *Journal of Hydrology*, *577*, 123924.

Zion Market Research. (2019). *Global Sports Betting Market Size & Share will reach USD 155.49 Billion by 2024.* Retrieved from https://www.globenewswire.com/news-release/2019/08/29/1908388/0/en/Global-Sports-Betting-Market-Size-Share-Will-Reach-USD-155-49-Billion-By-2024-Zion-Market-Research.html.

APPENDICES

Appendix A

List of published papers and conference proceedings

Published (as part of Ph.D. thesis)

Bracewell, P. J., **Patel, A. K.,** Blackie, E. J., & Boys, C. (2017). Using a Predictive Rating System for Computer Programmers to Optimise Recruitments. *Journal of Cases on Information Technology 19(3): 1-14.*

Brown, P., **Patel, A. K.** & Bracewell, P. J. (2017, June 23). Optimising a Batting Order in Limited Overs Cricket using Survival Analysis. *Paper published in The Proceedings of the 17th MathSport International 2017 Conference Proceedings*. (pp. 71-80). Padua, Italy. ISBN: 978-88-6938-058-7.

Campbell, E. C., **Patel A. K.,** & Bracewell, P. J. (2018). Optimizing junior rugby weight limits in New Zealand. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Campbell, E. C., Bracewell, P. J., Blackie, E., & **Patel, A. K.** (2018). The impact of Auckland junior rugby weight limits on player retention. *Journal of sport and health research,* 10(2), 317-326.

Greer S., **Patel, A.K.,** Trowland, H., & Bracewell P.J. (2018). The impact of injury on the future performance ratings of domestic T20 cricketers. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Mansell, Z., **Patel, A. K.,** McIvor, J. T, & Bracewell, P. J. (2018). Managing run rate in T20 cricket to maximize the probability of victory when setting a total. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

McIvor, J. T, **Patel, A. K.,** Hilder, T.A., & Bracewell, P. J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2017). An Individual-Based Team Rating Method for T20 Cricket. *Journal of Sports and Human Performance 5(1): 1-17.*

Patel, A. K., Bracewell, P. J., Gazley, A. J., Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. *Journal of Sports Science and Coaching 12(3): 328-338.*

Patel, A. K., Bracewell, P. J., & Wells, J. D. (2017, June 23). Real-time measurement of individual influence in T20 cricket. *Paper published in The Proceedings of the 17th MathSport International 2017 Conference Proceedings*. (pp. 61-70). Padua, Italy. ISBN: 978-88-6938-058-7.

Patel, A. K., Rooney. S. J., Bracewell, P. J., & Wells. J. D. (2018). Constructing a predictive PGA performance rating using hierarchical variable clustering. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., & Bracewell, P. J. (2018). A framework for quantifying the effectiveness of human-based rating systems. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Cook, M. K. A., Bracewell, P. J., & West, M. B. (2018). A framework to quantify the impact of social engagement on data driven creative. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel. A. K., Bracewell. P.J., & Bracewell, M.G. (2018). Estimating expected total in the first innings of T20 cricket using gradient boosted learning. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel. A. K., Bracewell. P.J., Wells, J.D., & Brown, P. (2018). Predicting football crowd attendance with public data. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel A. K., & Bracewell, P. J. (2019). Quantifying the evolution of first-class rugby in New Zealand. *Paper published in The Proceedings of the 18th MathSport International 2019 Conference Proceedings*. (pp. 51-60). Athens, Greece. ISBN: 978-618-5036-53-9.

Simmonds, P., **Patel, A. K.,** & Bracewell, P. J. (2018). Using network analysis to determine optimal batting partnership in T20 cricket. *Paper published in The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. Sunshine Coast, Queensland, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Contributing Publications (pre-Ph.D.)

Brown, P., **Patel, A. K.,** & Bracewell, P. J. (2016, July 12). Real Time Prediction of Opening Batsmen Dismissal in Limited Overs Cricket. *Paper published in The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports*. (pp. 80-85). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2016, July 12). Team Rating Optimisation for T20 Cricket. *Paper published in The Proceedings of the 13th Australian Conference* *on Mathematics and Computers in Sports*. (pp. 91-96). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.

Under Review

Patel, A. K., & Bracewell, P. J. (2019). Dynamic evaluation of player performance in T20 cricket. *Journal of Quantitative Analysis in Sport;* (under review).

Patel, A. K., Bracewell, P. J., & Coomes, M. (2019). Inferring bowling strike rate in limited overs cricket. *Journal of Sports Analytics;* (under review).

In-Preparation

Patel A. K., & Bracewell, P. J. (2020). Demonstrating the evolution of first-class rugby in New Zealand using a modified Elo ratings system. In preparation to submit to the *Journal of Quantitative Analysis in Sport.*

Patel A. K., & Bracewell, P. J. (2020). A novel performance metric to measure the predictive accuracy of probability of win model in T20 cricket. In preparation to submit to *The Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports*. Wellington, New Zealand: ANZIAM MathSport.

Appendix B

Analytical Hierarchy Process

Analytical Hierarchy Process

The analytical hierarchy process (AHP) is a multi-criteria decision-making tool developed by Thomas Saaty (Saaty, 1988). Given a user defined pairwise comparison matrix, the AHP translate the matrix into a vector of relative weights for each criterion element using a mathematical model. The pairwise comparison matrix provided a numerical comparison of each attributes with respect to the other attributes being evaluated. These matrix entries are determined using the fundamental AHP scale (Table 15) and are based on prior experience or expert knowledge. Applying the AHP to the pairwise comparison matrix translates the subjective weights into objectives weights, representing the importance of the attribute relative to the other attributes. Moreover, the method implements a consistency measure for each attribute to ensure that the 'user' defined weights are consistent and reduces bias in the decision-making process. "The aim is to provide the decision maker a precise reference to make adequate decisions and reduce the risk of making biased decisions by decomposing the problem into a hierarchy of more easily comprehended sub-problems" (Sinuany-Stern, 1988, p. 74). According to (Maliki, Owen & Bruce, 2006, p. 4) the following steps are computed applied to conduct the AHP:

Compute the value of criteria weights

The user defines an $n \times n$ pairwise comparison matrix, A, where n represents the number of evaluation criteria. Each a_{ij} entry evaluates the importance of attribute i with respect to j. The entries a_{ij} and a_{ji} must satisfy: $a_{ij} \times a_{ji} = 1$, while criteria with the same level of importance must satisfy: $a_{ij} = a_{ji} = 1$. The importance of criteria i relative to j can be established via the fundamental scale of the AHP:

Value of a_{ij}	Interpretation
1	i and j are equally important
3	i is slightly more important than j
5	i is more important than j
7	i is strongly more important than j
9	i is absolutely more important than j

Table 15: Fundamental scale of AHP

Synthesis judgement

Derive the normalised pairwise comparison matrix, A_{norm}, by the equating the sum of column entries to 1. The entries in matrix A_{norm} are computed as:

$$\bar{a}_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}$$

Create a criteria weight vector

$$w_i = \frac{\sum_{j=1}^{n} \bar{a}_{ij}}{n}$$

A relationship exists between the pairwise comparison matrix A and the weights vector, $w,$ such that $Aw = \lambda_{max}w$. The maximum eigenvector λ_{max} can be found by computing a consistency check:

$$CV_i = \frac{\sum_{i=1}^{n} a_{ij} \times w_j}{w_j},$$

And dividing the summation of consistency check values by, $n,$ the number of criteria:

$$\lambda_{max} = \frac{\sum_{i=1}^{n} CV_i}{n}$$

Consistency check of pairwise comparison matrix

The λ_{max} parameter enables the deviation of a consistency ratio (CR) which validates the consistency of the estimates vector:

$$CI = \frac{\lambda_{max} - n}{n - 1}$$

$$CR = \frac{CI}{RI}$$

n	2	3	4	5	6	7	8	9	10
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.51

The random index (RI) is dependent on n, if $CR \leq 0.1$ then the values of subjective judgement (i.e. pairwise comparison matrix) and the weights generated in step 3 are regarded as acceptable.

Appendix C

Metric types and definitions

Performance metric	Definition	Attribute type
Resources remaining	Proportion of balls and wicket left in an innings	Context/ Time
Over number	Number of over bowled	Action/ Time
Wickets	Number of wickets	Action/ Time
Innings balls	Number of balls bowled	Action/ Time
Innings runs	Ball-by-ball runs scored	Action/ Time
Ball run rate	Current total / balls bowled	Action/ Time
Projected total	Current total / resources remaining	Time
Team percentage dots	Total dots / balls bowled	Context/ Time
Team percentage boundaries	Total boundaries / balls bowled	Context/ Time
Runs remaining	Number of runs needed to reach the target total	Context/ Time
Required run rate	Run rate needed to reach the target total	Context/ Time
Batting team pressure	The amount of pressure experienced by the batting team	Context/ Time
Bowling team pressure	The amount of pressure experienced by the bowling team	Context/ Time
Batting team run efficiency	Inning runs / resources consumed	Context/ Time
Balls faced	Number of balls delivered to a batter	Action/ Time
Batter total runs	Total number of runs a batter scored	Action/ Time
Batter runs contributed	Batter runs / innings runs	Action/ Time
Batter strike rate	Batter runs / ball faced	Context/ Time
Batter total dots	Balls faced by a batter in which no runs are scored	Context/ Time
Batter total boundaries	Number of boundaries hit by the batter	Action/ Time
Batter percentage dots	Batter total dots/ balls faced	Context/ Time
Batter percentage boundaries	Batter total boundaries/ balls faced	Context/ Time
Batter activity rate	Percentage of balls a batter score runs off	Context/ Time
Batter total run contribution	Sum of batter runs contributed	Action/ Time
Balls bowled	Number of balls bowled by a bowler	Action/ Time
Bowler runs saved	Number of runs saved by a bowler	Action/ Time
Bowler total runs contributed	Sum of bowler runs saved	Action/ Time
Batter run efficiency	Batter runs scored / resources consumed by the batter	Context/ Time
Batters pressure contribution	Batter’s contribution made under pressure	Context/ Time
Bowlers pressure contribution	Bowler’s contribution made under pressure	Context/ Time
Batters win contribution	$prob(batting\ team\ win)_i - prob(batting\ team\ win)_{i-1}$	Context/ Time
Bowlers win contribution	$prob(batting\ team\ win)_{i-1} - prob(batting\ team\ win)_i$	Context/ Time
Batters total win contribution	Sum of a batters win contribution	Context/ Time
Bowlers total win contribution	Sum of a bowlers win contribution	Context/ Time
Batter’s survival probability	Probability of a batter being dismissed	Context/ Time

Table 16: Metric definitions and attribute-type for player-based ratings framework (Chapter Five)