Human Decision Making and Artificial Intelligence - A Comparison in the Domain of Sports Prediction

Arnu Pretorius
Department of Statistics and Actuarial Science
Stellenbosch University
Stellenbosch, South Africa
arnu@ml.sun.ac.za

Douglas A. Parry
Department of Information Science
Stellenbosch University
Stellenbosch, South Africa
dougaparry@sun.ac.za

ABSTRACT

Artificial intelligence (AI) research has become prominent in both academia and industry. With this, an interest in AI's ability to make sound decisions when compared to human decision making has grown. Predicting the outcome of sporting events has traditionally been seen as a difficult task, due to the complex relationships between variables of interest. Attempts to make accurate predictions are fraught with biases owing to the bounded rationality within which human decision making functions. This study puts forward the position that an AI approach using machine learning will yield a comparable level of accuracy. A random forest classification algorithm was employed to predict match outcomes in the 2015 Rugby World Cup. The performance of this model was compared to aggregate results from *SuperBru* and *OddsPortal*. The machine learning based system achieved an accuracy of 89.58% with 95%-CI (77.83, 95.47) vs. 85.42% with 95%-CI (72.83, 92.75) for the platforms. These results indicate that for rugby, over the limited period of a specific tournament, the evidence was not strong enough to suggest that a human agent is superior in terms of accuracy when predicting match outcomes compared to a machine learning approach, at a significance level $\alpha = 0.05$. However, the model was better able to estimate probabilities as measured by monetary winnings from betting rounds compared to the two platforms.

CCS Concepts

•**Applied computing** → **Psychology;** •**Computing methodologies** → *Machine learning;* •**Information systems** → Web applications;

Keywords

Decision making; sports prediction; classification; random forests; rugby

SAICSIT '16, September 26-28, 2016, Johannesburg, South Africa
© 2016 ACM. ISBN 978-1-4503-4805-8/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2987491.2987493

1. INTRODUCTION

A large part of human decision making involves prediction, which is therefore an implicit requirement for any AI system comparable on a certain task. Furthermore, prediction and sport have gone together since the inception of the latter [23]. Indeed, in Ancient Rome the prediction and betting on the outcome of games was often of greater importance than the spectacle itself [9]. Today, sports prediction has grown into a sophisticated endeavour described by many as both an art and a science [30]. However, sports prediction is characterised by distinct differences between amateurs, experts, and sophisticated mathematical models.

1.1 Sport as Research Platform

There are a few reasons for choosing the domain of sport to serve as a platform for comparisons. First, sports prediction is a growing area of interest in which a large number of experts and laypeople attempt to make predictions regarding the outcome of an event. This has been mirrored by the rise of online sports prediction and betting services [36]. Combined, this allows for the representation of a large sample of predictions made by human agents for comparison. Furthermore, sports prediction has not only captured the interest of the public, it has become a growing area of academic interest, especially in the fields of AI and machine learning. Efforts within these fields have sought to create methods to enhance the accuracy of sports prediction by removing the human element and supplanting it with sufficient computational power and AI [19, 10, 42].

1.2 Research Goals

Sports prediction poses an interesting research problem, primarily due to the fact that the outcomes of sporting events are often determined by many interrelated and difficult to quantify factors. Potential factors influencing outcomes include the weather, previous results, players' experience, crowd support and even morale. However, it is clear from the continued popularity of sports betting and online-prediction leagues such as *SuperBru* [3] and betting websites such as *OddsPortal* [2] that factors such as these do not present a prohibitive burden to would-be pundits, punters or bookies engaging in sports prediction.

The primary issue explored in this study is a comparison between the sports prediction ability of the intuitive approach, and an algorithmic, rational approach. More specifically, this study compares the predictive ability of human agents to that of an artificially intelligent agent based on a machine-learning prediction model. This is to be done in the context of predicting the outcomes of rugby matches taking place throughout the 2015 Rugby World Cup (RWC). A comparison between the outcomes achieved by punters on *SuperBru* and *OddsPortal* (standing as proxies for the human prediction ability) and the predictions derived from this model is presented. To this end, the primary hypothesis is proposed:

- *H*1: *The performance of a human agent is superior to that of a machine agent (an AI system) in terms of accuracy in the domain of sport prediction.*

A test for the primary hypothesis is unfortunately not directly attainable, therefore using proxies for a human agent and restricting the scope of sport prediction to only include the 2015 RWC, the secondary hypothesis is proposed:

- *H*2: *The performance of online platforms SuperBru and OddsPortal, serving as proxies for the prediction abilities of a human agent when faced with the task of predicting the outcomes of matches at the 2015 RWC, is superior to that of a machine learning model.*

The paper is structured as follows. First, the theoretical basis for understanding the human prediction ability is established through a brief review of the literature. Next, the supervised learning approach to machine learning is discussed primarily within the classification setting. This section aims to provide the necessary background regarding key machine learning concepts relevant to the model created to generate predictions for the RWC. The next section details the development process for the machine learning model employed in this study. The penultimate section then presents an evaluation of the outcomes of the model in comparison with the outcomes achieved by other platforms representing human sports prediction capabilities. Finally, a discussion of findings that stand out, the shortcomings of the proposed system as well as the research design are presented.

2. BACKGROUND

There exist many different approaches to the task of sports prediction. This domain of research is saturated with studies exploring strategies for predicting the outcomes of particular sporting events [19, 10]. Two distinct classes emerge through which the study of predictions is approached. First, there is the analytic approach, attempting to improve the accuracy of forecasts through the aggregation of large datasets of information pertinent to the particular sporting domain. The second class of prediction research falls within the cognitive and behavioural sciences. This area seeks to understand how people make predictions. Stemming from this second class of research is a body of knowledge supporting the role of simplistic techniques for casting predictions. We begin our analysis by presenting a review the literature regarding these approaches to prediction.

2.1 Human Prediction

When making predictions people rely on their previous experiences, knowledge of past events as well as their judgement and intuition [28]. The ability to create predictions is based on our capacity to collect, filter and analyse vast quantities of information simultaneously, within a contextual framework. When broken down to its constituent parts, making predictions forms a decision making problem. The predictor is compelled to complete a series of decisions about the role particular factors might have on the outcome, ultimately making a decision in favour of one out of many possible futures. While there are clear examples of the power of the human decision making ability, a significant body of research exists underlying the bounded rationality that shapes human decision making and how this leads to cognitive biases and errors in forecasting and prediction [39]. Research into individual's abilities to make predictions is shaped around two dichotomous schools of thought. The first outlook advances the idea that decision making abilities are predisposed to cognitive biases, undermining the capacity to make accurate predictions [21]. Conversely, the second school of thought argues that these 'tricks' used to make decisions often lead to more accurate choices being made [4].

March [31] outlines deficiencies in peoples' abilities to interpret and extrapolate from evidence of past events, explaining that these limitations undermine the accuracy of predictions stemming from observations of previous events. These limitations include: inaccurate recall, remembering history in ways dependent on current beliefs; superstitious learning, assigning causal significance to correlated but inconsequential actions; and compounding current wishes and hopes with their expectations of reality [31]. While these limitations impair judgement and the ability to create accurate predictions, other research has shown that the simpler methods commonly employed by experts, fans and punters alike can in fact produce predictions with comparable accuracy to more complex, statistical methods [21].

2.1.1 Rationality

People make decisions and predictions under the constraints of limited time, knowledge and cognitive-processing capability [17]. The concept that rationality in decision making is constrained by these forces was coined 'bounded rationality' by Simon [39] in 1972. Building upon the idea of bounded rationality, decision theorists have shown that instead of considering all alternatives, people typically consider only a few potential options [31]. Moreover, relevant information potentially aiding the analysis of future outcomes is not sought [31]. Information constraints manifest as limitations in attention, memory, comprehension and communication abilities [31]. Kahneman and Tversky note that three classes of information are required when making predictions: background information, specific contextual information and information about the expected accuracy of the prediction [27]. Bounded rationality hampers the gathering of this information, inhibiting accurate prediction making.

2.1.2 Heuristics and Biases

When making predictions or decisions people recognise patterns in the situations they face and apply rules of appropriate behavior to those situations [31]. These rules are known as fast and frugal heuristics. Fast, because they avoid estimation and frugal, because information is ignored [21]. Kahneman and Tversky establish the cognitive basis underlying common errors in decision making, prediction and probability judgement; arising from various heuristics and biases. In their 1973 study [27] exploring the psychology of prediction Kahneman and Tversky investigate the rules that determine intuitive judgements of confidence, contrasting these rules with the normative principles of statistical prediction. This study empirically confirms that people rely on a limited number of heuristics to arrive at predictions, ignoring prior probabilities and evidence potentially useful in improving the accuracy of their predictions [27]. Extending this idea March [31] explains that people spurn the calculation of the probability of future events through the use of complex decision trees, in favour of using the output of memory to inform how frequently similar events have occurred in the past. March summarised this idea by stating that people "use the results of memory as a proxy for the projection of future probability"[31].

Kahneman and Tversky [26] posit that when making predictions, individuals place a disproportionate amount of emphasis on novel information. However, this idea has been countered by experimental evidence suggesting that in fact people anchor on previous results and undervalue new information when making predictions [42]. A further bias that emerges relating to the interpretation of information has been termed the notion of a "hot hand". This bias is characterised by the belief that a streak of positive results is always followed by more positive results [18] — a misunderstanding of the statistical property of event independence. Paul and Weinbach [33] show that bettors frequently over-bet on teams experiencing winning streaks. A further element introducing biases decision making is sentiment for particular teams. Research in this area has shown that sentiment plays a significant role in the prediction and placing of bets in the context of sporting events [44].

While many biases in decision making emerge from the proclivity to use heuristics over more complicated statistical procedures, Goldstein and Gigerenzer [21] present an argument that heuristics used intuitively can in fact be as capable or even more accurate than statistical procedures for arriving at predictions. Goldstein and Gigerenzer analyse a study exploring the performance of a recognition heuristic employed by amateurs to that of more established rankings and prediction benchmarks in tennis. The recognition heuristic works as follows: "If only one of the teams is recognized, predict that the recognized one will win"[20]. In two studies [38, 37] use of this heuristic led to accurate predictions as often or better than the official benchmarks. These results lend credence to the idea that heuristic based prediction can outperform more complex prediction methods.

2.1.3 Extracting knowledge from experience

Making accurate predictions requires the assimilation of knowledge from previous experiences with specific information relevant to the context [27]. March describes three biases pertaining to the quality of information extracted from previous experiences, inhibiting peoples' ability to accurately forecast future events [31]. In order for accurate predictions to be made an understanding of the causal structure for the phenomena needs to be arrived at. This understanding is reduced by the tendency to interpret previous experiences on the basis of insufficient information [31]. Moreover, the comprehension of causal relationships is further undermined by the ambiguity of information received. March explains that people face great difficulty in separating causal effects from random, extraneous forces. Finally, past successes in predictions cloud future judgements, March showed that successful past actions tend to be repeated [31] — this tendency leads to the redundancy of experience and the inability of people to adequately extract knowledge useful in making predictions.

2.1.4 The role of expertise

Research into sports prediction has typically focused on the forecasts of experts in the domain [11, 41, 4]. The accuracy of experts' predictions is compared to those of amateurs, prediction markets or other sources of predictions. Comparing the predictions of experts, amateurs and prediction markets with algorithmic, machine learning based predictions has not emerged in the research as a prominent focus area. In a study comparing the accuracy of prediction markets and sports experts, Spann and Skiera [41] note that the empirical evidence for the accuracy of experts in making sports predictions is limited. Interestingly, this study indicated that prediction markets and betting odds are far more accurate in predicting results. Similarly, in a study exploring the differences between experts and non-experts in predicting outcomes of the 2002 Football World Cup, Andersson *et al.* [4] show that experts in this domain fail to achieve more accurate predictions than individuals with limited domain knowledge. Goldstein and Gigerenzer [21] comment on this study describing the positive performance of the recognition heuristic as a result in favour of simplistic prediction techniques. However, while this may be the case, these results also confirm the cognitive biases described by Kahneman and Tversky [27] and the flawed judgement utilised in creating forecasts and predictions by both experts and non-experts alike. While both Spann and Skiera [41] and Andersson *et al.* [4] underscore the deficiencies in expert-based predictions, they do not make any comparison to statistics based prediction techniques, implying that this area of research requires further investigation.

2.2 Machine Learning

Machine learning is the approach taken in this study to create an artificially intelligent agent capable of challenging the sport prediction capabilities of humans. Specifically, machine learning for prediction usually takes the form of using a set of inputs $X_1, ..., X_p$ to predict an unknown value Y. An approach often considered when faced with such a setting is *supervised learning*. The basic idea of supervised learning is to monitor the system of interest over a period of time and collect data of both the inputs and the corresponding outputs. Once the data has been collected both the recorded input and output values can be used to extract rules by which knowledge of the input values can accurately produce the corresponding output values. The rule extraction is typically performed by a predefined computer algorithm known as a *learning* algorithm which is "trained" to approximate the unknown mechanism governing the system using the data. It is for this reason that the recorded data consisting of input-output pairs is aptly named the *training* data. The architecture of a learning algorithm often depends on the nature of the output value of interest. The main distinction is between quantitative outputs Y which lead to the development of *regression* algorithms and qualitative outputs C which lead to the development of *classification* algorithms. The focus of this paper is classification (whether a team won or lost a match), however estimating the probabilities of winning or losing can be seen as modeling a quantitative output.

2.2.1 *Supervised Learning for Classification*

In classification the output C takes on values $k \in \{1, ..., K\}$ representing different groups or classes to which inputs fed to the system can belong. In this paper we will focus on the (common) binary case in which $K = 2$ (a match is won or lost), meaning that the outcome of a tied match was not considered in the modeling process ($K = 3$). For the time being, without loss of generality let $C \in \{0, 1\}$. A classification algorithm then aims to be able to take an observed input $\mathbf{x}$ without knowledge of the corresponding output and assign it to the correct class. However, it is rare for any system that a given input can be uniquely determined as belonging to a certain class. As a consequence classification often boils down to estimating class probabilities. Suppose $\mathbf{x} \in \mathbb{R}^p$, a point in a p-dimensional input space, then $\mathbf{x}$ can belong to class 1 only with a certain probability $P(C = 1|\mathbf{x}) = 1 - P(C = 0|\mathbf{x})$ based on its location.

2.2.2 *Expected Loss and the Bayes Classifier*

Naturally, misclassification of a point $\mathbf{x}$ by a classifier $g(\mathbf{x})$ will have associated with it some form of cost or loss. Suppose ℓ_0 is the loss incurred for misclassifying $\mathbf{x}$ as belonging to class 0 instead of class 1 and ℓ_1 vice versa. Let $I_0 = I(g(\mathbf{x}) = 0)$, then the expected loss is

$$E\big[L(C, g(\mathbf{x}))\big] = \ell_0 I_0 P(C = 1|\mathbf{x}) + \ell_1 I_1 P(C = 0|\mathbf{x}), \quad (1)$$

where $I(\cdot)$ is the indicator function equal to 1 if its argument is true, 0 otherwise and $L(\cdot, \cdot)$ is a loss function. If $\ell_0 = \ell_1 = 1$, the loss function becomes $L_{0\text{-}1}(C, g(\mathbf{x})) = I(g(\mathbf{x}) \neq C)$ known as the *0-1 loss* with an expected value

$$E\big[L_{0\text{-}1}(C, g(\mathbf{x}))\big] = I_1 P(C = 0|\mathbf{x}) + I_0 P(C = 1|\mathbf{x}). \quad (2)$$

Note that (2) implies that if $P(C = 1|\mathbf{x}) > 0.5$ the expected loss incurred for misclassifying $\mathbf{x}$ as belonging to class 1 is $P(C = 0|\mathbf{x}) = 1 - P(C = 1|\mathbf{x}) < P(C = 1|\mathbf{x})$, the loss for the opposite mistake. The optimal classifier which will minimise the expected loss in this situation will therefore be a rule classifying to the most probable class, i.e.

$$g_B(\mathbf{x}) = I\left(P(C = 1|\mathbf{x}) \geq \frac{\ell_1}{\ell_0 + \ell_1} = \frac{1}{2}\right), \quad (3)$$

called the *Bayes* classifier [13]. It is important to realise that even the optimal (Bayes) classifier will rarely achieve perfect classification due to the intrinsic probabilistic nature underlying observable systems. In addition, it is often the case that $\ell_0 = \ell_1$, but also not uncommon to be faced with the setting where $\ell_0 \neq \ell_1$. Asymmetry related to different types of misclassification departs from the 0-1 loss framework and classification by way of most probable class. However the Bayes classifier is still well defined, since the threshold is simply adjusted (away from 0.5). A key assumption in the approach taken in this paper is that $\ell_0 = \ell_1$. In other words, the loss associated with misclassifying a match as a win for a particular team when in fact they lost is equal to misclassifying a loss when in reality a win was observed.

2.2.3 *Generalisation Error*

A related quantity to that of expected loss is the *generalisation* or *test* error of a classifier which is dependent on a particular training set. Consider the following set of input-output pairs forming the training set (of size N) $\Omega_{tr} = \{(\mathbf{x}_i, c_i), i = 1, ..., N\}$ then the generalisation error for $g(\mathbf{x})$ is given by the following expectation

$$Err^* = E\big[L(C, g(\mathbf{x}))\big|\Omega_{tr}\big]. \quad (4)$$

There is a greater interest in obtaining an estimate of (4) since it is an error measure that is a closer reflection of the real world. The expectation in (2) is taken over all possible training data sets sampled from the joint distribution (X, C), which is clearly a quantity further from reality. So in summary, supervised learning for classification seeks to find a model (function) $g(\mathbf{x})$, by way of a learning algorithm trained using a particular training set Ω_{tr}, capable of minimising the generalisation error Err^*.

2.2.4 *Trees for Classification*

Concretely, consider a situation where there are two input variables X_1 and X_2 believed to be related to a qualitative binary output $C \in \{b, o\}$. The two-dimensional input space depicting the locations of training data points is presented in Figure 1.

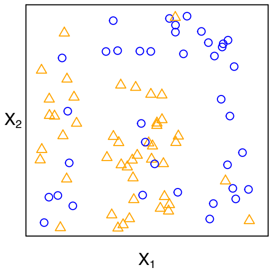

Figure 1: Two-dimensional input space.

The first step is to use these points to learn a rule capable of separating as well as possible the observations illustrated by the blue circles in Figure 1, from the orange triangles.

Consider the following strategy: take the input space and split it into two rectangular regions achieving a reasonable degree of separation. Next, treat each new region as the first and split *it* into two smaller rectangular regions. Continue in this way until each region is fairly homogeneous with respect to class representation. The steps are illustrated in the left panel of Figure 2 where the input space is split into regions $R_1, R_2, ..., R_5$.

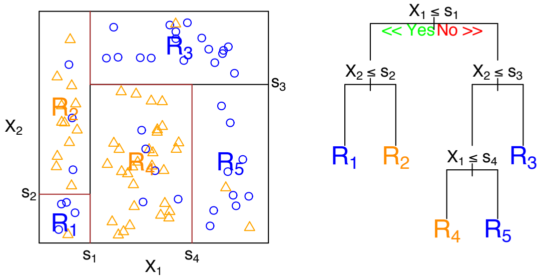

Figure 2: Binary classification tree.

The right side of Figure 2 is an isomorphic representation of each recursive binary partitioning. The picture is resemblant to that of an upside down tree and therefore the appropriate name given to this type of classifier is a *classification tree*. Each rule (i.e. $X_1 \leq s_1$) is called a *node* in the tree and corresponds to a partition of the input space. When a node is split, each resulting node is referred to as a child node of the original. The *CART* [8] algorithm for classification trees finds the best split at each node by searching over all available input variables and split-points and selecting the optimal variable and split-point pair minimising the sum of child node impurities [1]. For other approaches towards tree induction, see [22, 34, 35, 25].

At prediction time a new unseen observation $\mathbf{x}_0$ can be "dropped" from the *root* node at the top and based on which rules are satisfied follow a specific path down to a *terminal* node at the bottom of the tree. The quantity $P(C = b|\mathbf{x}_0) = 1 - P(C = o|\mathbf{x}_0)$ can be estimated by computing the respective proportions of each class inside the terminal node to which $\mathbf{x}_0$ belongs. Symmetric loss classification is then performed by way of the most probable class.

Tree classifiers are intrinsically attractive since they naturally handle quantitative and qualitative data types as well as missing values, are robust to outliers and possess a form of implicit variable selection able to deal with many irrelevant input variables [24].

2.2.5 Training Error and the Bias-Variance Trade-off

The *training error* of a classifier $\bar{Err} = \frac{1}{N}\sum_{i=1}^{N} L(c_i, g(\mathbf{x}_i))$ measures the average loss over the training set. For trees the training error $\bar{Err}$ can be made arbitrarily small by simply continuing the splitting procedure until each region contains only points belonging to a single class. However, this by no means guarantees that the tree classifier will generalise well to data presented to it in the future.

The drawback with this approach is that the measure of variability dependent on the locations of the sampled training points, the *variance* of the classifier, is high. In other words, it can be imagined that if a new round of data recording took place from the same system under study that the layout of the partitioned space might change considerably from the one presented in Figure 2. This is in contrast with a simple linear separating boundary that is less susceptible to changes in the data. Less complex approaches such as linear regression are "rigid" in this sense and is said to have low variance. However, they rely heavily on the rather strict assumption that the separating boundary appropriate for the data is a $(p-1)$-dimensional hyperplane. This high *bias* might cause linear classifiers to suffer if in truth these assumptions are incorrect. Trees on the other hand are more complex making very few assumptions regarding the shape of the decision boundary and therefore have low bias. This trade-off based on model complexity is referred to as the *bias-variance trade-off.*

2.2.6 Beyond a Single Tree

Much of the recent success in the development of machine learning algorithms have gone the way of using multiple classification trees as *base learners*, combined in clever ways to produce a better performing *ensemble classifier* [5, 12, 7, 14, 15]. To loosely motivate this approach, from a variance reduction perspective, consider the following general argument [24]. Let $X_1, ..., X_B$ be identically distributed variables, not necessarily independent with $Var(X_i) = \sigma^2$ and $Cov(X_i, X_j) = \rho\sigma^2, \forall i, j, i \neq j$. This gives

$$Var(\bar{X}) = Var\left(\frac{1}{B}\sum_{i=1}^{B} X_i\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \qquad (5)$$

Taking the average over a large number of random variables reduces the variance and ensures that the second term in (5) can be made arbitrarily small by increasing B. This motivates aggregation. However, the first term remains unaffected by the size of B and only interacts with the magnitude of the correlation between the variables. Going one step further, many proposals have been made that in addition to aggregation, some degree of artificial randomness is injected into the algorithm [7, 16, 32, 46]. This can actually have an effect of reducing the correlation ρ and thus further reduce the variance. An algorithm using this approach is called a *random forest* [7].

3. METHODOLOGY

In order to generate predictions for the outcomes of the matches, a two stage approach was adopted. The first stage was to collect past data of international rugby matches. Once the data had been collected, cleaned and processed, different random forests were trained and compared with each other. The best model was selected based on metrics such as training time and test error. This provided a machine agent ready at the start of the tournament. The second stage involved creating an automated cloud-based system able to collect the most recent match data after every match. The model was then automatically retrained incorporating the newest available data and updated estimates were produced.

A summarised version of the methodology discussed thus far is provided in Figure 3.

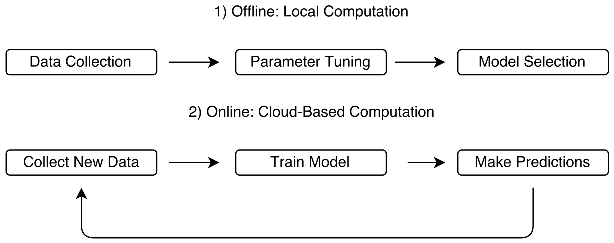

Figure 3: Project pipeline.

Comparisons were conducted with the aggregate results achieved by human agents on the sports prediction platform *SuperBru* as well as an analysis of potential monetary winnings through sports betting website *OddsPortal*.

There exist several reasons underlying the decision to employ these platforms as proxies for the predictive ability of human agents. Primarily it is the view of the researchers that both of these platforms espouse several of the characteristics of prediction markets. Prediction markets are commonly defined as markets facilitating the exchange of trades on the outcome of events — the market price is an indica-

[1] The impurity of a node is a measure of the heterogeneousness inside the defined region with respect to class representation.

tion of the crowd's perception of the probability associated with a particular outcome [40]. *OddsPortal* meets all of the traditional requirements of a prediction market.

The reasons behind the conceptualisation of *SuperBru* as possessing several of the qualities of prediction markets requires further clarification. While it is true that the platform does not facilitate the trading of predictions for sporting outcomes, there exist other reasons why this platform can be portrayed as sharing the characteristics of prediction markets. Users compete in pools, earning points based on the accuracy of their predictions. In addition to the leaderboard-facilitated competition incentivising accuracy, in many cases there exist monetary and material prizes for performance within these pools. For these reasons it is argued that the predictions derived from *SuperBru* share the characteristic of prediction markets, representing the collective predictions of the crowd. Snowberg, Wolfers and Zitzewitz [40] explain that because superior performance produces monetary rewards, there exists a financial incentive for users to provide the most accurate predictions. This is useful because it implies that the predictions made are accurate representations of the individuals' true feelings about the outcomes — a factor missing from many other forms of opinion gathering used in sports forecasting [40].

3.1 Data Collection

The data was collected from public websites including:

- *http://www.rugbydata.com*: for each of the 20 teams in the RWC we collected some general historical team statistics and past match data stretching back to the beginning of 2013 [2].
- *http://wrr.live555.com*: we collected rankings of each team as well as their recent change in rank.
- *http://en.wikipedia.org/wiki/World_Rugby_Rankings*: we collected ranking points.

Once the data was cleaned and structured into a tidy data set where each observation represented a match between two teams together with team related statistics for each team, it was split into a training and test set. The test set contained every second match from the start of 2015 until the most recent match prior to the tournament, which included 21 matches in total. All the remaining matches were used for training (379 matches in total). The output variable was binary and coded to reflect whether the home team won the match or not.

Figure 4 shows a heatmap of correlation between all the variables used in the training data set.

A dark green square in Figure 4 indicates a strong positive correlation between two variables and a light (white) square a strong negative correlation. Many of the variables were found to be correlated with each other (such as the average points scored against a team at home and the number of matches the team lost at home). Nevertheless, the first column of the heat map provides the degree of correlation between all the variables (indexed by the rows) with the outcome of the match. As expected the rank of the home and away team seemed to be correlated with the match outcome when compared to other input variables.

3.2 Model Selection

Arguably the most popular random forest algorithm for classification is Breiman's *Forest-RI* [7]. The strategy is to build an ensemble of randomised trees by sampling with replacement from the training data and at each node only selecting a random subset of the input variables as candidates for splitting.

Forest-RI splits the variable space using orthogonal (perpendicular to the variable axis) splits, whereas *oblique random forests* [32] use some linear combination of the variables for splitting. It has been suggested that oblique trees are better suited for splitting a space consisting of many correlated variables, therefore in addition to Forest-RI, an oblique random forest using partial least squares for node splits was also considered [32, 47].

3.2.1 Parameter Tuning

Both approaches have the same two tuning parameters, namely the size of the ensemble (number of trees) *ntree*, and the size of the subset of randomly selected variables at each node split, *mtry*. The *mtry* parameter was tuned using five-fold cross-validation [43] over a grid of values. Table 1 shows the grid search values where those in bold were the optimal selected by cross-validation.

Table 1: Tuning parameter grid

mtry: *Forest-RI*	mtry: *Oblique-RF*
1	**1**
7	7
14	14
27	27
40	40
53	53

The *ntree* parameter was fixed at 200 after observing at which point the out-of-bag error curve of the random forests "flattened out" [6].

3.2.2 Training Time and Test Error

To make the final decision regarding the best model, each model's training time and test error were compared [3]. The results are given in Table 2.

Table 2: Model performance

Algorithm	**Training Time**	**Test Error**
Forest-RI	14.32 secs	19.05%
Oblique-RF-PLS	6.39 mins	23.81%

Forest-RI outperformed the oblique random forest both in terms of training time and test error. Therefore, the model selected was Forest-RI with an ensemble size 200 and an *mtry* set equal to one.

[2] More data were available but 2013 was selected as the cut-off, since as matches stretch further back in time they become less relevant.

[3] Training time was deemed important since it was required to retrain the model between every match.

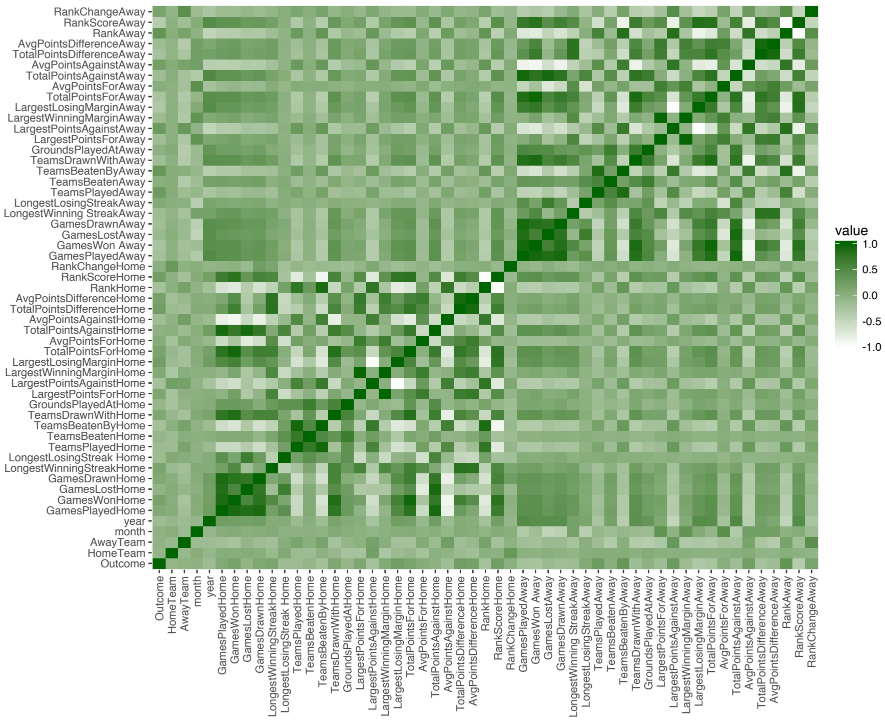

Figure 4: Heatmap of correlations between input variables.

3.3 Automated Cloud-Based Predictions

The analytics platform *Domino*[4] [1] was used to create an automated cloud-based prediction system accessible through a RESTful API. A schedule was created on the platform to recollect data after every match based on match dates and times. Furthermore, once the new data had been collected, the model was retrained and the prediction API republished using Domino's API endpoints service.

Due to the way in which the output variable was coded (indicator for a home team win), the model was able to pick up on home team advantage. This meant that different estimates were given for the same match-up based on which team was playing at home. Since the tournament took place in England, it was decided to remove this advantage for all the other teams. To this end, when the API was called with a match-up (not involving England), two estimates were produced with each team in turn presented as the home team and the average estimate of the two scenarios was then sent back. England was treated as always playing at home.

4. RESULTS

In Table 3 the accuracy of each approach for predicting matches in the 2015 RWC is provided. The second column is the number of matches correctly predicted out of a total of the 48 matches that took place throughout the tournament, with the corresponding accuracy percentage and 95% confidence interval [5] given in the third column. Both OddsPortal and SuperBru correctly predicted 41 out of 48 matches (85.42%), but were outperformed by the random forest model with 43 out of 48 correct (89.58%). The difference in performance is not statistically significant at a significance level $\alpha = 0.05$. However, not rejecting the null hypothesis of lesser or equal performance is not the same as accepting it. Therefore, the results can only be interpreted as indicating that there is not enough evidence to suggest that human agents are significantly superior in terms of prediction performance when compared to a machine learning approach.

Table 3: Approach Prediction Accuracy

Approach	**#Correct**	**Accuracy (95%-CI)**
Forest-RI Model	43/48	89.58% (77.83, 95.47)
OddsPortal	41/48	85.42% (72.83, 92.75)
SuperBru	41/48	85.42% (72.83, 92.75)

The estimated probabilities for each match is given in Figure 5. The way the tournament played out is presented from left to right (with matches on the x-axis) and the probability presented on the y-axis is the estimated probability for the team labelled at the top of Figure 5 to win the match. As previously mentioned, the assumption of symmetric loss was made regarding misclassification and therefore following the Bayes classifier given in (3) the threshold for prediction was chosen to be 0.5.

The model is far more conservative than the other approaches (almost always being closer to the 0.5 threshold). The red crosses and green stars in Figure 5 indicate the matches in which differences in probability estimates be-

[4] Domino is an enterprise-grade platform enabling researchers and industry practitioners to run, scale, share and deploy analytical models.

[5] Confidence intervals estimated according to Wilson's interval.

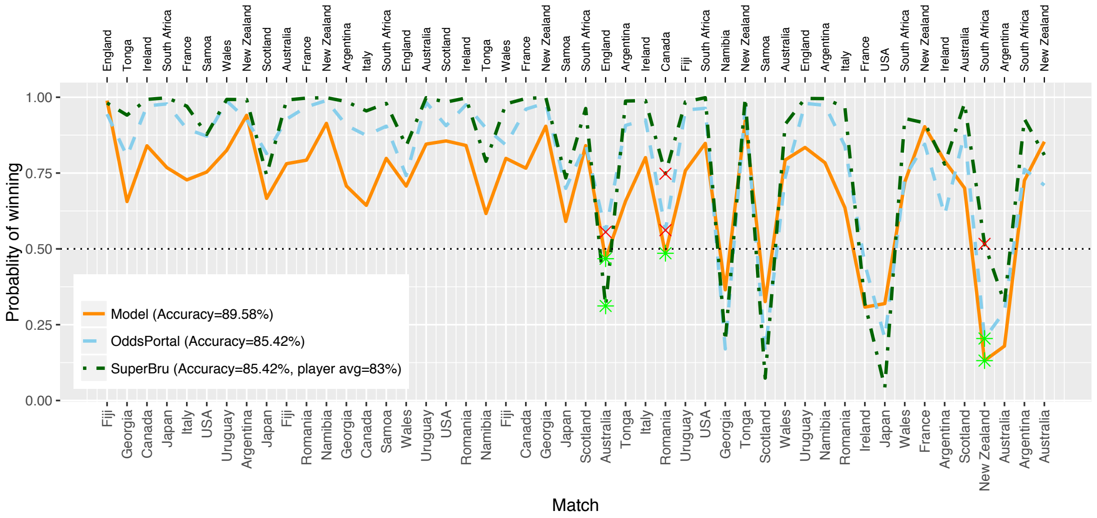

Figure 5: Comparing human versus AI prediction accuracy over the duration of the 2015 RWC.

tween the approaches resulted in disagreement regarding the winning team. An interesting case is the match between New Zealand and South Africa. Here the model was more confident than both the other approaches, however SuperBru is seen to differ somewhat, being in favour of South Africa winning the match.

The quality of the probability estimates themselves were assessed using the Kelly criterion [29] defined as

$$f^* = \frac{bp - q}{b}, \tag{6}$$

where b is the payout given a one unit bet, p is the estimated probability of winning and $q = 1 - p$ [45]. The criterion is used to calculate the optimal percentage of a given starting capital to bet at each match using the knowledge of the aforementioned terms. Therefore, probability estimates can be compared by specifying an equal starting amount of capital for each approach and making bets based on (6). The best estimator is then the approach with the most capital at the end of the tournament. Figure 6 shows the evolution of winnings given a starting capital of $R100$.

Two big losses early on in the tournament (Tonga vs. Georgia and South Africa vs. Japan) cost all the approaches a considerable sum of their initial capital. But, due to the conservative nature of the estimates produced by the model a smaller sum was lost and a subsequent recovery was seen until Ireland faced Argentina. On the other hand, the confidence of the other two approaches cost them, making recovery much less feasible given the little capital left.

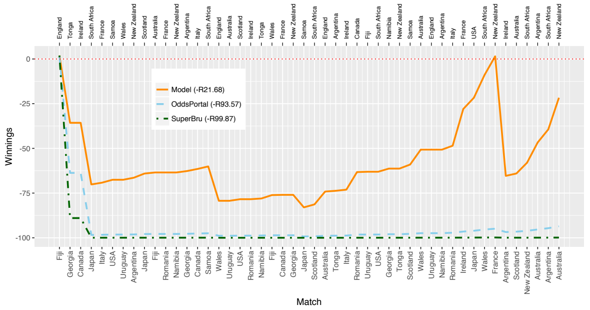

Figure 6: Comparing human versus AI monetary winnings given bets were made throughout the tournament based on the Kelly criterion with a starting capital of R100.

5. DISCUSSION

It was shown that their is a lack of evidence to support H2 stating that a human agent is superior to a machine learning model in terms of prediction accuracy. More specifically, the results indicate that both proxies of human agents achieved an accuracy statistically no greater than a *Forest-RI* model. However, in terms of a comparison in monetary winnings calculated via the Kelly criterion the model seemed to show superior performance.

The study findings are of interest because a large pool of amateurs, experts and stakeholders alike have not been able to clearly outperform an artificially intelligent agent using a machine learning algorithm, fed limited data, ignorant of many factors perceived to have an impact on predicting match outcomes. This research underscores the potential of machine learning based prediction models taking limited data inputs and creating accurate predictions. Furthermore, it highlights the progress of artificially intelligent systems in the realm of prediction and decision sciences.

Several of the more important limitations require attention. Of central concern is the merit of *SuperBru* and *OddsPortal* as proxies for the prediction ability of human agents. These platforms were selected due to qualities they posses which are inherent in prediction markets. However, in the case of *OddsPortal*, probability estimates were derived from aggregated odds collected from many bookmakers. This means that the validity of the assumption that these bookmakers all arrived at their odds through human knowledge and expertise is by no means guaranteed. *SuperBru* was selected due to its power as an aggregator of predictions. Over 220 000 people placed close to 10 million predictions on the platform for the RWC [3]. Nonetheless, there do exist biases in the predictions derived from both of these platforms. For instance, a significant proportion of the participants on *SuperBru* are South African (180 000), potentially skewing the predictions due to sentiment and other cognitive biases such as representativeness.

A further limitation to the generalisability of the findings to the sport of rugby as a whole exist due to the limited sample size of 48 matches. In a typical year there are over 100 tier 1 and tier 2 matches [3]. Moreover, because the 48 matches considered were RWC matches, they cannot be considered typical of all matches in general. There exists a far greater level of pressure, training and focus put into these matches than any other. While these results certainly hold true for RWC matches, and could possibly be generalised to other non-world cup matches, the extent to which the results shown could be generalised to other sports is not clear. Future research in this area should focus on creating a model capable of predicting results for all forms of rugby matches, as well as determining whether such a general model could be extended to other sports.

6. CONCLUSION

This study set out to conduct a comparison between human prediction ability and an artificially intelligent prediction system. This comparison was conducted in the domain of sports prediction. More specifically, the intention was to determine whether the degree of accuracy of two proxies for human based predictions were significantly greater than an artificially intelligent system using machine learning when predicting the outcome of matches at the 2015 RWC. This process was accomplished through the design and development of an automated cloud based prediction system using a random forest classification algorithm (*Forest-RI*). For the case of the 2015 RWC there was not enough evidence to suggest that the two proxies for human prediction capabilities outperformed the machine learning approach. Moreover, the random forest model produced superior probability estimates when used for betting compared to the other two platforms.

7. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their insightful comments which improved the overall quality of this paper. The work in this paper was partially supported by the MIH Media Lab at Stellenbosch University and the National Research Foundation (NRF) under grant number: UID 94615. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

8. REFERENCES

[1] Domino data lab: Enterprise data science platform. https://www.dominodatalab.com.

[2] Odds portal - betting odds monitoring service. http://www.oddsportal.com.

[3] Superbru - social sports prediction and fantasy game. https://www.superbru.com.

[4] P. Andersson, J. Edman, and M. Ekman. Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting*, 21(3):565–576, 2005.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] L. Breiman. Out-of-bag estimation. Technical report, Statistics Department, University of California Berkeley, 1996.

[7] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[8] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees.* CRC press, 1984.

[9] L. Casson. *Everyday life in ancient Rome.* JHU Press, 1998.

[10] M. J. Dixon and P. F. Pope. The value of statistical forecasts in the uk association football betting market. *International Journal of Forecasting*, 20(4):697–711, 2004.

[11] D. Forrest and R. Simmons. Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting*, 16(3):317–331, 2000.

[12] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[13] J. H. Friedman. On Bias , Variance , 0 / 1 - Loss , and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[14] J. H. Friedman and B. E. Popescu. Importance Sampled Learning Ensembles. *Machine Learning*, 94305(2):1–32, 2003.

[15] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3):916–954, 2008.

[16] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[17] G. Gigerenzer and R. Selten. *Bounded rationality: The adaptive toolbox.* Mit Press, 2002.

[18] T. Gilovich, R. Vallone, and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314, 1985.

[19] J. Goddard. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, 21(2):331–340, 2005.

[20] D. G. Goldstein and G. Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

[21] D. G. Goldstein and G. Gigerenzer. Fast and frugal forecasting. *International Journal of Forecasting*, 25(4):760–772, 2009.

[22] G.V.Kass. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society*, 29(2):119–127, 1980.

[23] P. Halpern. *The pursuit of destiny: A history of prediction.* Perseus, 2000.

[24] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2009.

[25] D. Heath. Induction of Oblique Decision Trees.

Proceedings of the 13th International Joint Conference on Artiffcial Intelligence, 21218(410):1002–1007, 1993.

[26] D. Kahneman, P. Slovic, and A. Tversky. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[27] D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological review*, 80(4):237, 1973.

[28] D. Kahneman and A. Tversky. Intuitive prediction: Biases and corrective procedures. Technical report, DTIC Document, 1977.

[29] J. L. Kelly Jr. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.

[30] K. Mansfield. Top advice from the worlds top 4 sports gamblers.

[31] J. G. March. *Primer on decision making: How decisions happen*. Simon and Schuster, 1994.

[32] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht. On oblique random forests. In *Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer, 2011.

[33] R. J. Paul and A. P. Weinbach. Bettor misperceptions in the nba: The overbetting of large favorites and the "hot hand". *Journal of Sports Economics*, 6(4):390–400, 2005.

[34] J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.

[35] J. R. Quinlan. *C4.5: Program for Machine Learning*. 1993.

[36] S. Salaga and S. Tainsky. Betting lines and college football television ratings. *Economics Letters*, 132:112–116, 2015.

[37] B. Scheibehenne and A. Bröder. Can lay people be as accurate as experts in predicting the results of wimbledon 2005. *International Journal of Forecasting*, 23:415–426, 2007.

[38] S. Serwe and C. Frings. Who will win wimbledon? the recognition heuristic in predicting sports events. *Journal of Behavioral Decision Making*, 19(4):321–332, 2006.

[39] H. A. Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

[40] E. Snowberg, J. Wolfers, and E. Zitzewitz. Prediction markets for economic forecasting. Technical report, National Bureau of Economic Research, 2012.

[41] M. Spann and B. Skiera. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55–72, 2009.

[42] H. O. Stekler, D. Sendor, and R. Verlander. Issues in sports forecasting. *International Journal of Forecasting*, 26(3):606–621, 2010.

[43] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147, 1974.

[44] K. S. Strumpf. *Illegal sports bookmakers*. s.n, 2003.

[45] E. O. Thorp. The kelly criterion in blackjack, sports betting, and the stock market. *Finding the Edge: Mathematical Analysis of Casino Games*, 1(6), 1998.

[46] E. E. Tripoliti, D. I. Fotiadis, and G. Manis. Modifications of the construction and voting mechanisms of the Random Forests Algorithm. *Data and Knowledge Engineering*, 87:41–65, 2013.

[47] H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.