Rochester Institute of Technology

Soccer Result Prediction and Analysis

Submitted as a Capstone Project Report in Partial fulfillment of a Master of Science Degree in Professional Studies at the Rochester Institute of Technology

Karan Chauhan

December 2, 2016

CONTENTS

Soccer Result Prediction and Analysis

Karan Chauhan

ABSTRACT

With more than 200 countries playing, Soccer is the most popular sport of our times. A large part of the world follows soccer on a day-to-day basis. With advent of technology in sports, more advancements have been made over a period of time to make traditional soccer not just a physical sport played on the field but also a mental game played off the field involving statistics that analyzes the opposition's strengths and weaknesses, players' form, factors such as recent form of the team and so on. Team Managers decide their on-field strategy only after detailed analysis considering these factors in depth. Hence, it is highly imperative that some sort of efficient pre-match analysis is carried out for higher success rates. In this project, a model for soccer result and score prediction has been proposed that will help soccer statisticians to get and in-depth forecast of various match factors such as future match outcomes, future scorelines and predicted cup winner. Predicted betting odds will not help people to bet on appropriate teams but also give them a proper insight as to which team is most likely to dominate. Using data from official soccer website for premier league, multinomial regression, random forest and Poisson regression models have been implemented to predict various future match outcomes. With decent accuracies achieved using all the models, multinomial regression proves to be the most accurate model in predicting the results. Results prove that the proposed system would prove to be a handy tool in aiding statisticians, betters and soccer fans to have an in-depth prediction of whats going to happen in the world of soccer.

Keywords - Multinomial Regression, Random Forest, Betting Odds, Correlation, Imputation, Prediction, Confusion Matrix, Linear Regression

1. INTRODUCTION

With most countries from all continents officially playing soccer, it is the most popular sport on the planet today without any doubt. The recent FIFA World Cup was watched by over 3.2 billion viewers and Champions League is watched by over 1 billion people every year. The magnanimity of this sport has reached new bounds with fans considering it as a religion. With increasing popularity comes tremendous evolution in terms of the way the game is played and what external factors can be involved in deciding the outcomes of the game. In modern soccer, we see every organization equipped with a team of statisticians and analyzers trying their best provide accurate analysis ranging from player details, previous match statistics and probability to come up with suggestions so as to improve team performance.

Hence, it is the necessity of the hour to combine data mining and machine learning techniques with soccer analysis to help soccer team organizations to dominate the world of soccer in their respective leagues. The use of past match information including information on match statistics such as goals scored, corners won, freekicks won, manager of the team, home advantage and recent form can be used with machine learning as the basis to predict future outcomes along with betting odds prediction to give an overall prediction and analysis. With this as its underlying principle, the proposed system in this project attempts to predict future scores, future winners and predicted outcome along with betting odds and winning probability percentages for English Premier League season. For this, multinomial regression, random forest and Poisson model proved to be most suitable for determining future predictions.

Section 2 reasons for the motivation behind this project. It explains why such a model was implemented taking into consideration the need and scope of the model. Section 3 enlists previous related works in this domain and how this project differs from the those. Section 5 walks us through the dataset used and various data cleaning and preparation methods used. The next section explains the feature selection techniques used. Section 7 illustrates the methodology followed for implementing this project. Section 8 demonstrates the results with tables and diagrams. Section 9 shows the Graphical User Interface used for displaying predictions. The last section concludes this analysis and talks about the work that can be improvised in the future.

2. MOTIVATION

The need to blend data mining and machine learning with the world of sports in the most efficient manner was the major motivation factor behind attempting to implement such a project. Making use of past soccer statistical data along with betting odds to predict future outcomes has never been done before. Another motivating factor is the need to address factors such as team recent form, manager reputation and home advantage and examine the extent to which they affect the outcome of the game. Soccer being a very unpredictable game, it is difficult to accurately estimate what the future results would be given the factors that constantly act upon the game. Usually stronger teams dominate but past results have shown chances of weak team causing upsets

and retaining a sense of unpredictability in the game. Implementing accurate models that try to to come up with the solution to answer this unpredictability is a major task. This project would provide an answer to the question - whether data analysis would be able to handle this job of accurately predicting soccer match outcomes or not.

3. RELATED WORK

Previous works attempt to make use of data mining to predict results but with lesser accuracy and using alternative approaches. Haghighat[7], made use of various techniques such as support vector machines, Bayesian networks and decision tree to predict soccer outcomes. Comparing accuracies of different models was the main aim of this project. The work conducted by Zimmerman[23] makes use of two types of data - one which is real and other that is virtual - based on video game data. The authors focused on improving the strategy of weaker teams using both supervised and unsupervised machine learning methods. Lots of research and data mining was done in this project.

Huang and Chang[9] proposed a neural network-based prediction of 2006 World Cup matches using a small subset of features aggregating the statistics for the entire team. The attribute list includes hand-selected features such as shots on target, corners, goals scored and possession. The authors were able to achieve accuracy of around 62 percent but the method bookmakers betting data was not utilized properly. In 2003, Buchdahl[17] researched the effect of betting odds in predicting results. He concluded that with larger feature set, it should be possible to predict football matches more accurately. Research in this field is still going on but it is a major challenge to predict the outcomes more accurately and by considering all possible factors that influence the game.

4. FLOW OF THE PROJECT

Figure 1 shows the flow diagram of the process that is followed in this project. First, the raw dataset containing 15 datasets were merged to 10 datasets using common attributes. Then, on each of the 10 files, attribute removal and data imputation was performed. Once the new features were added and final merged dataset was obtained, feature selection takes place followed by dividing the dataset into training and testing sets. Finally, different models were implemented to get the desired results.

5. DATA CLEANING AND PREPARATION

5.1 Dataset Description

The dataset was collected from www.football-data.co.uk which is the official football dataset website. The data files are in csv format. Each csv file corresponds to an official seasonal year in English Premier League. The detailed description of the attributes of the dataset can be seen in Figure 2 and Figure 3.

As seen from the Figure 2 and Figure 3, the feature set in the raw data contains attributes including statistical information as well as betting odds data from popular betting websites[21].The number of attributes vary from year to year.

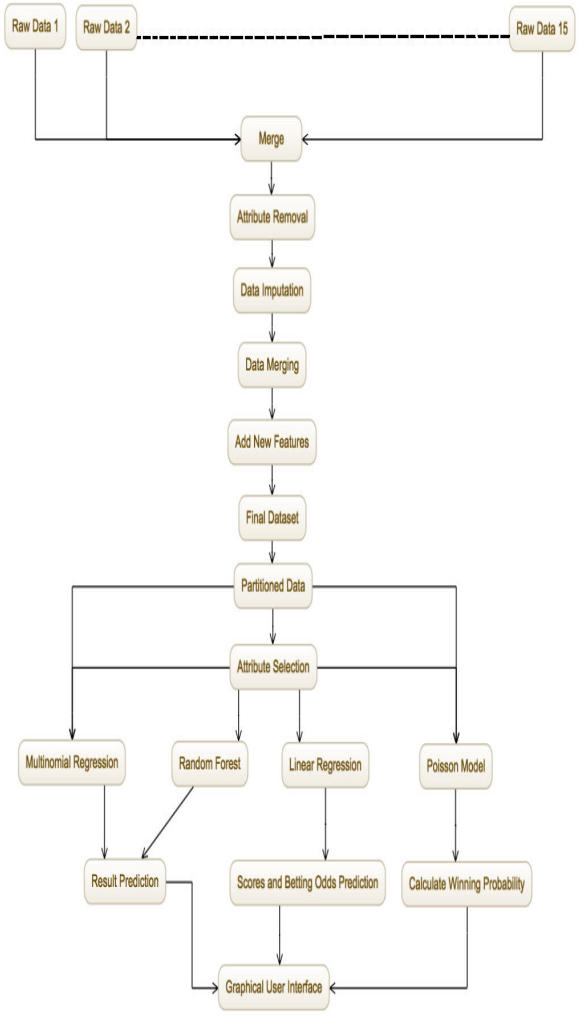

Figure 1: The Flow Diagram of the Project

5.2 Data Preparation

To perform analysis and predict the outcomes using various models, it was necessary to have one data file in the csv format as a whole. Since every file corresponds to a year, merging of 15 different files into one was necessary. But since all had different number of attributes both common and uncommon, data merging was the most important step.But before data merging, data cleaning had to be performed to ensure that there are no null values, redundant values and attributes or any unwanted columns.

First, all the attributes that would remain in the final dataset were decided. For this, attributes that were common in most individual datasets were retained. Based on these attributes, 10 different files were created that were later merged into one final dataset.

Among the attributes retained, attributes that had majority of null values were eliminated. Similarly, attributes that had some missing values were taken care of using data imputation as discussed in the next subsection.

```
Div = League Division
Date = Match Date (dd/mm/yy)
HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Match Statistics (where available)
Attendance = Crowd Attendance
Referee = Match Referee
HS = Home Team Shots
AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HHW = Home Team Hit Woodwork
AHW = Away Team Hit Woodwork
HC = Home Team Corners
AC = Away Team Corners
HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HO = Home Team Offsides
AO = Away Team Offsides
HY = Home Team Yellow Cards
AY = Away Team Yellow Cards
HR = Home Team Red Cards
AR = Away Team Red Cards
HBP = Home Team Bookings Points (10 = yellow, 25 = red)
ABP = Away Team Bookings Points (10 = yellow, 25 = red)
```

Figure 2: Description of Attributes-1

```
B365H = Bet365 home win odds
B365D = Bet365 draw odds
B365A = Bet365 away win odds
BSH = Blue Square home win odds
BSD = Blue Square draw odds
BSA = Blue Square away win odds
BWH = Bet&Win home win odds
BWD = Bet&Win draw odds
BWA = Bet&Win away win odds
GBH = Gamebookers home win odds
GBD = Gamebookers draw odds
GBA = Gamebookers away win odds
IWH = Interwetten home win odds
IWD = Interwetten draw odds
IWA = Interwetten away win odds
LBH = Ladbrokes home win odds
LBD = Ladbrokes draw odds
LBA = Ladbrokes away win odds
PSH = Pinnacle home win odds
PSD = Pinnacle draw odds
PSA = Pinnacle away win odds
SOH = Sporting Odds home win odds
SOD = Sporting Odds draw odds
SOA = Sporting Odds away win odds
SBH = Sportingbet home win odds
SBD = Sportingbet draw odds
SBA = Sportingbet away win odds
SJH = Stan James home win odds
SJD = Stan James draw odds
SJA = Stan James away win odds
SYH = Stanleybet home win odds
SYD = Stanleybet draw odds
SYA = Stanleybet away win odds
VCH = VC Bet home win odds
VCD = VC Bet draw odds
VCA = VC Bet away win odds
WHH = William Hill home win odds
WHD = William Hill draw odds
WHA = William Hill away win odds
```

Figure 3: Description of Attributes-2

5.2.1 *Data Imputation*

Data imputation is the process of accounting for the missing values in the dataset[12]. Various approaches can be used for data imputation, but the one used for this project was averaging out the data column wise. The attributes that were imputed were those with betting odds information. The list of attributes that were imputed are shown in Figure 4.

List of Attributes Imputed
IWD, IWH, IWA
GBD, GBH, GBA
WHH, WHA, WHD
LBH, LBD, LBA

Figure 4: List of Imputed Attributes

5.2.2 *Data Merging*

Once attributes with null values were dropped and missing data was handled using imputation, different data files were merged into one. Initially, data files that had common attributes were merged together. As a result, we had 10 different data files.

After data cleaning, all these 10 files were merged into one single final dataset. The final dataset had 56 final features and 6080 instances. As a final step, 6 more attributes were added. The newly added 6 attributes form play a vital role in the analysis. HomeRF and AwayRF represent home team recent form and away team recent form that determine how good the teams have been performing in the recent matches. These values were calculated based on the team's performance in the last few matches both home and away. HMRL and AMRL correspond to the Manager's reputation level at each time for the team. All these attributes are categorical on a scale of 5 and collected from various informative soccer sources[1].

6. FEATURE SELECTION

Before heading on to the predictive analysis, the most important part is to select features that would efficiently serve as appropriate predictors for the model. Various approaches can be followed to select features that are important to the analysis. In this project, attribute selection using correlation plots, Boruta package[11] and methods such as Chi-Squared and InfoGain Attribute Ranking of Weka have been used.

6.1 Correlation Matrix

To test for important attributes, correlation matrix is produced. Correlation matrix gives the relation between two attributes and it is very important where many features are involved. Below are some correlation plots that help to select the important attributes.

Figure 5 shows the correlation plot of FTAG against B365 Betting Odds. It shows the effect of betting points data in predicting away team goals. As we can see for the correlation matrix, away goals increase when B365A is less and vice versa. This proves that more that betting points data can effectively determine the results that we are trying to predict. Similar conclusion can be drawn with respect to home goals and B365H after examining the correlation plot from Figure 6.

Figure 7 examines the relation between full time result and newly added attributes home and away recent forms.

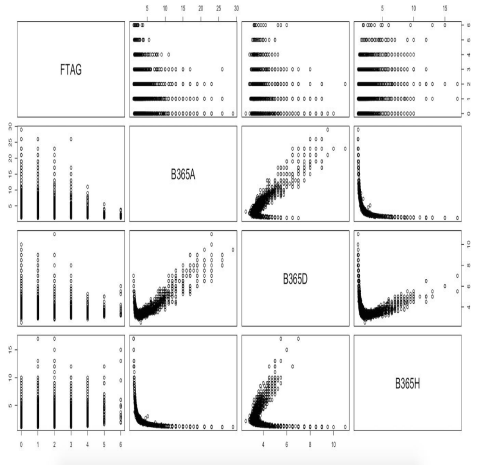

Figure 5: Correlation Matrix for FTAG against B365

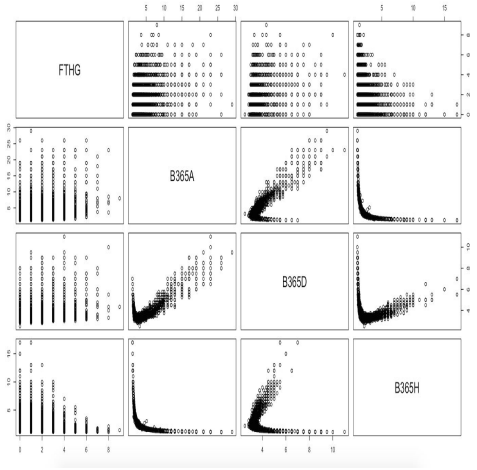

Figure 6: Correlation Matrix for FTHG against B365

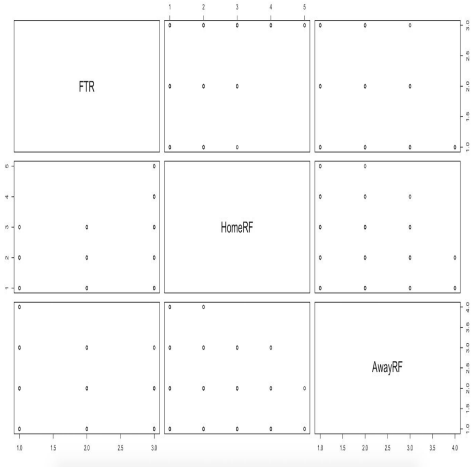

Figure 7: Correlation Matrix for FTR against HomeRF and AwayRF

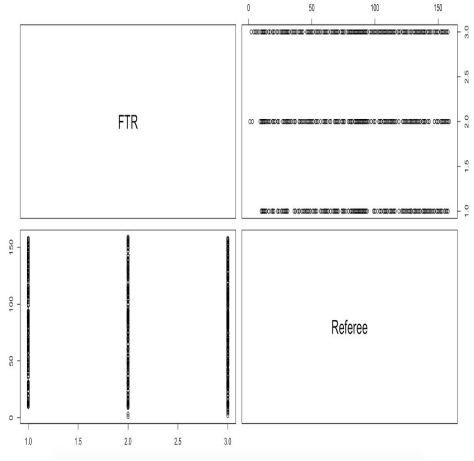

Figure 8: Correlation Matrix for FTR against Referee

For the Correlation plot, we can see that when HomeRF is high, then the result turns out to be in favor of the home team. Similarly, when AwayRF is high, then result is in the favor of the away team. For a draw, the values are intermediate. This strongly suggests that HomeRF and AwayRF are strong features in predicting the results.

Figure 8 shows the correlation plot for FTR against Referee. From the nature of the plot, we can conclude that Referee attribute has no effect on predicting the Full Time results and hence can be discarded.

6.2 The Boruta Package

The reason for using the Boruta package is that it adds randomness to the dataset as it creates shuffled copies of all features[4]. It then trains a random forest classifier on the dataset and applies a feature importance measure to evaluate the importance of each feature. At each iteration, it performs a check whether a feature has a higher importance than the rest of the features and goes on removing features which are relatively less important. Finally, the algorithm stops when all features are confirmed and rejected. In this case, 99 iterations were run using the Boruta package. The priority of the features importance-wise can be seen from figure 9.

As seen, FTAG and FTHG are most important but we eliminate them as they tend to bias the results because they represent scores and results can be easily figured out from the scores. The next most important features are HomeRF and AwayRF that correspond to forms of the teams.The next set of important features turn out to be half-time scores and results. Following are the other major statistics and betting odds such as B365H, SBH and many more. We can see that the least important features include Date,Referee, AC, AY, HY , AC ,Bb0u and others. These are discarded from the analysis. Of all the three methods used, feature selection using the Boruta package is most reliable as it conducts

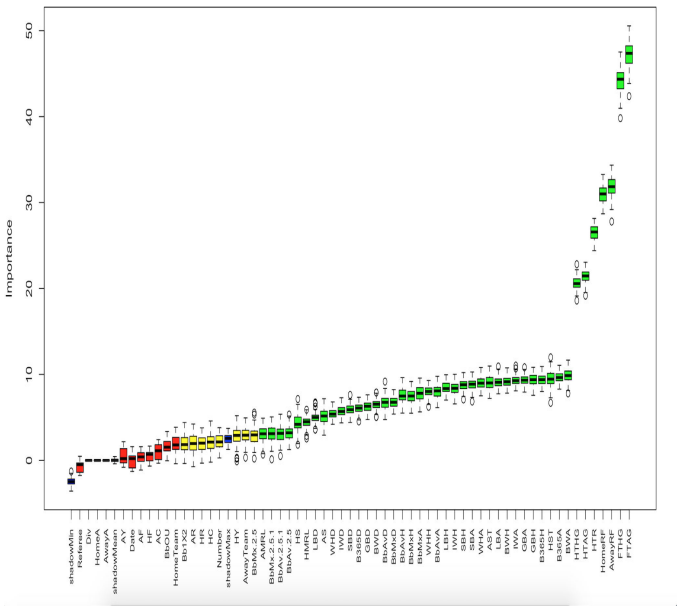

Figure 9: Attribute Importance Using Boruta Package

multiple runs on the attributes.

6.3 Attribute Ranking in Weka

To cross verify the features selected from the above two methods are important, the attribute ranking techniques Chi-Squared[18] and InfoGain[19] were used in Weka. The important attributes obtained using these methods are shown in Figures 10 and 11 respectively.

```
Attribute Evaluator (supervised, Class (nominal): 5 FTR):
        Chi-squared Ranking Filter

Ranked attributes:
2509.52322    8 HTR
2351.93724    1 HomeRF
2188.57279    2 AwayRF
1324.65713    6 HTHG
1140.71655    7 HTAG
1084.50239   21 B365H
1066.44567   51 SBH
1049.12396   24 BWH
1047.02708   50 GBA
1042.40057   29 IWA
1041.21197   27 IWH
1041.00552   26 BWA
1038.49204   30 LBH
1038.26415   48 GBH
1019.02009   53 SBA
1009.27712   23 B365A
 994.82814   33 WHH
 991.19457   32 LBA
 961.4752    35 WHA
 655.78037   38 BbAvH
 653.53842   37 BbMxH
 630.93276   41 BbMxA
 615.77286   42 BbAvA
 537.24954   11 HST
```

Figure 10: Chi-Squared Attribute Ranking

As seen from Figures 10 and 11, feature selection using the attribute ranking techniques from Weka matches closely to the feature selection performed using the Boruta package.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.312 +- 0.004      1.1 +- 0.3      1 HomeRF
 0.305 +- 0.003      1.9 +- 0.3      8 HTR
 0.263 +- 0.004      3   +- 0        2 AwayRF
 0.17  +- 0.003      4   +- 0        6 HTHG
 0.133 +- 0.002      5.3 +- 0.9      7 HTAG
 0.129 +- 0.004      7.2 +- 2.23    21 B365H
 0.129 +- 0.003      7.8 +- 2.14    51 SBH
 0.127 +- 0.004      9.1 +- 2.3     50 GBA
 0.127 +- 0.004      9.5 +- 2.73    29 IWA
 0.126 +- 0.003     10.6 +- 2.37    30 LBH
 0.126 +- 0.004     10.8 +- 2.23    26 BWA
 0.126 +- 0.003     11.6 +- 2.06    24 BWH
 0.124 +- 0.004     13   +- 2.65    27 IWH
 0.124 +- 0.004     13.2 +- 2.75    53 SBA
 0.124 +- 0.003     13.9 +- 1.7     23 B365A
 0.123 +- 0.003     14.5 +- 1.5     48 GBH
 0.121 +- 0.003     16.7 +- 1.27    32 LBA
 0.118 +- 0.003     18.1 +- 0.3     33 WHH
 0.116 +- 0.004     18.7 +- 0.64    35 WHA
 0.078 +- 0.003     20.5 +- 0.67    37 BbMxH
 0.077 +- 0.002     20.9 +- 0.94    38 BbAvH
 0.075 +- 0.002     22.2 +- 0.6     41 BbMxA
 0.074 +- 0.003     22.4 +- 0.8     42 BbAvA
 0.067 +- 0.001     24.3 +- 0.46    11 HST
 0.066 +- 0.001     24.7 +- 0.46    12 AST
```

Figure 11: InfoGain Attribute Ranking

Hence, considering the attributes obtained using feature selection techniques mentioned above, we proceed towards implementing the models for predicting results and scores.

7. APPROACH

7.1 Data partitioning

Data partitioning is the process of dividing the dataset into training and testing sets. The model is applied on the training dataset and is tested for accuracy on the testing dataset. For this project, training data was selected using the condition applied on the Date attribute. All the matches that took place before the Premier League 2013-14 season are considered as the training dataset whereas remaining part is the testing dataset. 4560 instances belonged to training dataset whereas 1520 instances belonged to the testing dataset. This data partioning has been used in all models throughout this project.

7.2 Result Prediction

Based on the training dataset, the aim is to predict the results for the testing dataset. The results would be in terms of A,D and H corresponding to Away Win, Draw and Home Win respectively. A multinomial logistic Regression[2] model has been used for this purpose. Since there are more than two classes to be predicted, the multinomial logistic regression model is the most suitable model. Multinomial logistic regression is a classification method for multi-class problems and is used to predict the probabilities of different possible outcomes of a dependent variable that is categorically based on a set of independent variables. To predict the outcome, we use as many important features as possible that are obtained using feature selection[6]. We apply multinomial regression using 2 different set of features. The results are written to a separate file. Figure 12 shows the different set of features that are used to perform Multinomial Logistic Regression.

Other approach used to predict results is the random forest classifier[13]. Random Forest produces multiple decision trees in order to classify the outcome in two or more classes. When the training set for the current tree is drawn by sam-

Feature Set 1	Feature Set 2
HMRL	HomeRF
AMRL	AwayRF
HomeRF	HTR
AwayRF	HTAG
HTR	HTHG
HTAG	HST
HTHG	AST
HST	SBH
HC	B365H, B365A
HY	GBA, GBH
AS	BWA, BWH
HS	SBA, SBD
SBH	BWH
B365H, B365D	IWA, IWH
GBH	WHA, WHH
BWA, BWD	BWD
SBA, SBD	B365D
BWH	AS
B365A	GBD
IWA, IWD	BbAvA
LBH, LBA	BbAv.2.5
WHD	BbAv.2.5.1
IWH	BbMx.2.5
BbAvA, BbMxA	BbMx.2.5.1
BbAvH, BbMxH	IWD
BbAvD, BbMxD	WHD

Figure 12: Different Features Set

pling, about one-third of the cases are left out known as the out-of-bag data. It is used to estimates of the classification error. After each tree is built, all of the data are run down the tree and nearness is calculated for each pair. We can control the number of trees using the *ntree* parameter which is set to 500 in this model[15]. Also, *mtry* value is set to 20. The features for random forest classification are same as those used in multinomial regression.

	left daughter	right daughter	split var	split point	status	prediction
1	2	3	20	3.150000	1	0
2	4	5	16	4.756250	1	0
3	6	7	5	6.500000	1	0
4	8	9	3	1.500000	1	0
5	10	11	3	1.500000	1	0
6	12	13	3	1.500000	1	0
7	14	15	3	1.500000	1	0
8	16	17	38	32.500000	1	0
9	18	19	4	1.500000	1	0
10	20	21	4	1.500000	1	0
11	22	23	4	1.500000	1	0
12	24	25	4	1.500000	1	0
13	26	27	4	1.500000	1	0
14	28	29	4	1.500000	1	0
15	30	31	25	7.500000	1	0
16	32	33	4	1.500000	1	0
17	34	35	36	1.585000	1	0
18	0	0	0	0.000000	-1	3
19	36	37	3	2.500000	1	0
20	38	39	29	3.850000	1	0
21	0	0	0	0.000000	-1	1
22	0	0	0	0.000000	-1	3
23	40	41	6	2.000000	1	0
24	42	43	37	1.855000	1	0
25	0	0	0	0.000000	-1	1
26	0	0	0	0.000000	-1	3
27	44	45	1	3.500000	1	0
28	46	47	23	4.637500	1	0
29	0	0	0	0.000000	-1	1
30	48	49	4	1.500000	1	0
31	0	0	0	0.000000	-1	1
32	50	51	1	2.500000	1	0

Figure 13: Tree Formation process

Figure 13 shows the results of using tree formation process of the first tree. The first line describes the root split[3]. The root split is based on variable 20 which will then decide to which node to go to - the left child or the right child. If the status of a line is -1, then a leaf node is reached and a prediction will be made.

7.3 Score Prediction

Score prediction makes use of the linear model function[20] (linear regression model) to fit the data. As far as score prediction is concerned, continuous values need to be predicted for which linear model function is the most appropriate choice. Linear regression was applied two times using two different target variables, once using FTHG for predicting home team goals and other using FTAG for predicting away team goals. For former, predictors used were those pertaining to home team properties such as HTAG, HS, HomeRF, HMRL, SBH, IWH and so on. In case of latter, predictors corresponding to away team were used including AwayRF,AS, SBA and so on. Home and Away predicted scores are stored in separate data files.

7.4 Betting Odds Prediction

We have the betting odds from various websites that provide odds based on certain factors. For example, if a team has odds 1.33 and other team has 4.33, then the winning odds are in favor of the first team with odds 1.33[16]. Similarly, these websites provide draw odds too. This system provides its own betting odds using previous betting odds and other features. For this purpose, linear regression has been used to predict continuous output in the form of home betting odds, draw betting odds and away betting odds. Linear regression has been used three times likewise.

7.5 Win Probabilities

Win probabilities are calculated in terms of win percentages. Home Team, Away Team and Draw probability percentages are predicted. For this, Poisson model has been used. The Poisson distribution is a probability distribution that gives the probability of a given number of events occurring with a certain rate[22]. For this model, munging of data is done where an additional variable known as home advantage is inserted. Data munging restructures the data in a suitable format as needed. Dataset after performing Data Munging[10] is shown in figure 14. The model is then run which gives the probability which is then converted into percentages. The *dpois* function is used to get the probabilities. Home Team and Away Team are to be entered by the user.

	team	opponent	goals	home
HomeTeam	Bournemouth	Aston Villa	0	1
AwayTeam	Aston Villa	Bournemouth	1	0
HomeTeam1	Chelsea	Swansea	2	1
AwayTeam1	Swansea	Chelsea	2	0
HomeTeam2	Everton	Watford	2	1
AwayTeam2	Watford	Everton	2	0

Figure 14: Data Munging for Poisson Model

8. RESULTS AND COMPARATIVE ANALYSIS

8.1 Multinomial Logistic Regression

Using multinomial regression, a prediction is made that categorizes the outcome in three classes - A, D and H. Figure 15 and Figure 16 show the confusion matrix for the two sets of features. Accuracy is 79 percent meaning that 79 percentage of attributes in the test data were correctly classified. The diagonals represent the true positives and for good prediction, the values in diagonals must be as high as possible. From the confusion matrices we can see that both the feature sets have nearly the same classification rate.

	A	D	H
A	334	125	1
D	27	333	26
H	5	138	531

Figure 15: Confusion Matrix for Logistic Regression using Feature set-1

	A	D	H
A	320	125	5
D	12	328	46
H	2	120	552

Figure 16: Confusion Matrix for Logistic Regression using Feature set-1

Figure 17 shows the distribution of classified instances of test data after performing Multinomial Logistic Regression using Feature Set 1.

Figure 18, Figure 19 and Figure 20 show the ROC curves that depict the results of Multinomial Logistic Regression.True positive are represented on the Y-axis and False positives are represented on the X-axis[5].Using Feature Set 1, the ROC curves show that the model is highly accurate as the curve inclines more towards the Y-axis which is true positives.This is true for ROC curves for Away, Draw and Home instances. The AUC (Area Under the curve) can also be considered an important measure. It is denoted by the area that falls

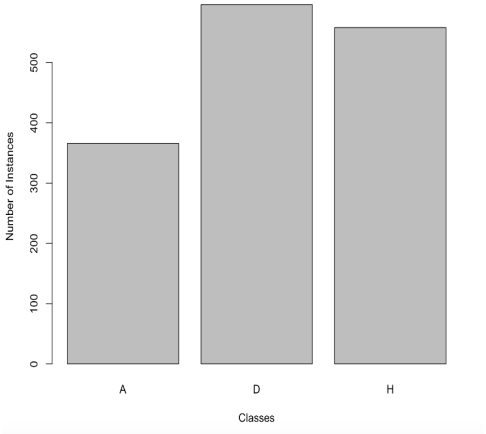

Figure 17: Distribution of Predicted Classes

within the scope of the curve and looking at the curves, we can deduce that AUC is pretty good.

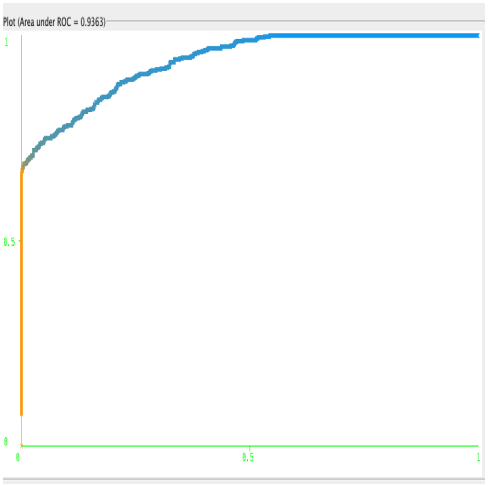

Figure 18: ROC curve for away instances

8.2 Random Forest

Random Forest produces decision trees to predict the correct class. A sample portion of one of the decision trees has been shown in Figure 17. We can see that the root node is HomeRF and based on the condition on the edges, the tree is traversed in the form of left child and right child which are denoted by other features. This process goes on until the leaf node is reached which denotes that a classification has occurred.

The confusion matrix for random forest classification is shown in Figure 22. Diagonal represents the correctly classified instances. The classification accuracy obtained using random forest is 75 percent which is slightly less as compared to multinomial regression.

Figure 23 shows the comparison between the accuracy of Both the approaches used. Using Multinomial Regression,

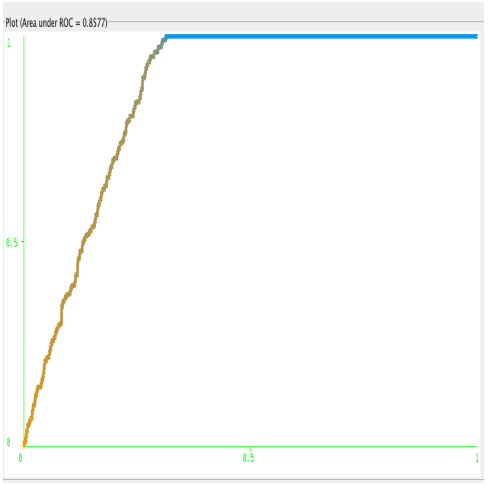

Figure 19: ROC curve for draw instances

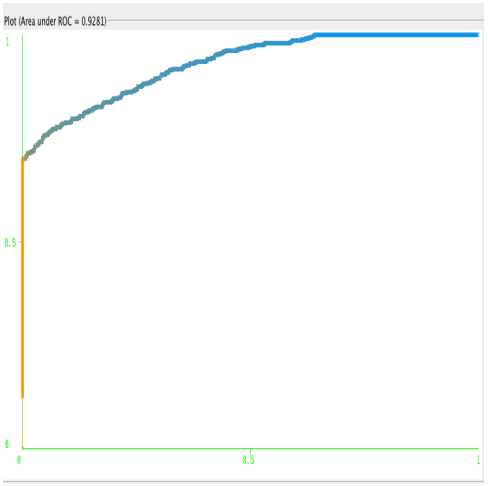

Figure 20: ROC curve for home instances

we can see that the classification rate is almost 79% using both set of features. The accuracy is slightly less in case of Random Forest Classifier which is 75%. Hence, we conclude that Multinomial Regression is slightly better in predicting the results.

8.3 Linear Regression

The linear regression model has been used to predict the home scores, away scores, home betting odds, draw betting odds and way betting odds. The results of the model are depicted in Figure 24 and Figure 25. From the summary in Figure 24, we can see that the model is quite efficient as the standard error values are very less. The p-value must be as small as possible for a good prediction[14], typically less than 0.05(alpha value) and it is significantly less in case of some attributes. The significance codes show that attributes like HomeRF, HTHG and HST prove to be important. Similarly in case of Figure 25, summary results depict low standard errors, less p-values and significance codes show that most important features are HTAG, AwayRF, AST and AC.

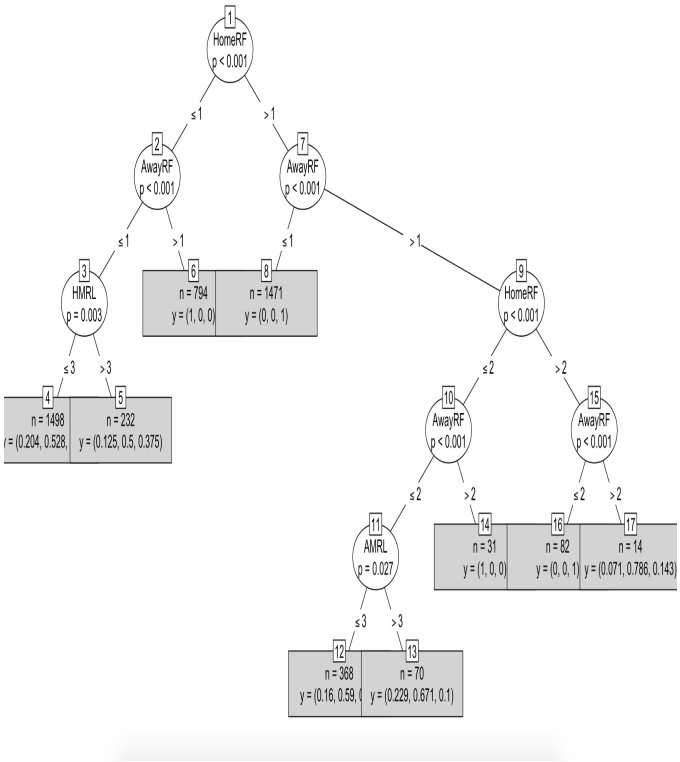

Figure 21: A Sample Decision Tree in RandomForest

	A	D	H
A	329	114	17
D	34	307	45
H	14	151	509

Figure 22: Confusion Matrix Random

Model	Accuracy	
Multinomial Logistic Regression	Feature Set 1	Feature Set 2
	78.81 %	78.94 %
Random Forest	75.32 %	

Figure 23: Comparative Analysis for both models

Figure 26 and Figure 27 show the relationship between the

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.84136 -0.42144 -0.04001  0.39818  1.81465

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0322785  0.0399620 -25.832  < 2e-16 ***
HomeRF       1.5510393  0.0135827 114.193  < 2e-16 ***
HMRL         0.0054021  0.0061990   0.871   0.3836
HTHG         0.2531362  0.0104184  24.297  < 2e-16 ***
HS          -0.0053191  0.0024843  -2.141   0.0323 *
HC          -0.0057409  0.0024342  -2.358   0.0184 *
HST          0.0286300  0.0035767   8.005 1.51e-15 ***
HR          -0.0456142  0.0251010  -1.817   0.0692 .
HY           0.0008066  0.0058064   0.139   0.8895
B365H        0.0262460  0.0326529   0.804   0.4216
WHH          0.0013452  0.0198583   0.068   0.9460
LBH          0.0186422  0.0268060   0.695   0.4868
IWH         -0.0774635  0.0343228  -2.257   0.0241 *
BWH          0.0010633  0.0448207   0.024   0.9811
SBH          0.0040118  0.0345765   0.116   0.9076
GBH         -0.0162146  0.0287315  -0.564   0.5725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4456 on 4544 degrees of freedom
Multiple R-squared:  0.8815,    Adjusted R-squared:  0.8811
F-statistic:  2253 on 15 and 4544 DF,  p-value: < 2.2e-16
```

Figure 24: Summary of Linear Regression to predict Home Goals

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.00258 -0.39001 -0.07916  0.39840  1.42662

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.084331   0.036018 -30.105  < 2e-16 ***
AwayRF       1.497836   0.015581  96.132  < 2e-16 ***
AMRL         0.005251   0.006132   0.856   0.3918
HTAG         0.306944   0.011721  26.188  < 2e-16 ***
AS          -0.006336   0.002826  -2.242   0.0250 *
AC          -0.010901   0.002687  -4.056 5.07e-05 ***
AST          0.039636   0.004164   9.519  < 2e-16 ***
AR          -0.017159   0.021123  -0.812   0.4166
AY           0.003266   0.005187   0.630   0.5290
B365A       -0.018342   0.010428  -1.759   0.0786 .
WHA          0.002503   0.007420   0.337   0.7359
LBA          0.003056   0.011171   0.274   0.7844
IWA         -0.032483   0.013384  -2.427   0.0153 *
SBA          0.004224   0.014628   0.289   0.7728
BWA          0.018779   0.015503   1.211   0.2258
GBA          0.006154   0.014835   0.415   0.6783
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4402 on 4544 degrees of freedom
Multiple R-squared:  0.8415,    Adjusted R-squared:  0.8409
F-statistic:  1608 on 15 and 4544 DF,  p-value: < 2.2e-16
```

Figure 25: Summary of Linear Regression to predict Away Goals

goals scored by home and away teams respectively against the predicted home and away betting odds. From Figure 26, we can see that as the predicted home odds increase, the number of goals scored by home team decreases. Same occurs in case of Figure 27 where the number of goals scored is inversely proportional to the predicted betting odds. These results prove that the model behaves in an efficient way.

Linear Regression proves good for continuous predictions. Features such as B365H, SBH, WHH, SBA, LBH and many more have been used to this model. The that actual predicted results may not be in tandem with the predicted betting odds as betting odds are just an estimation before the match as determined by the bookmakers.

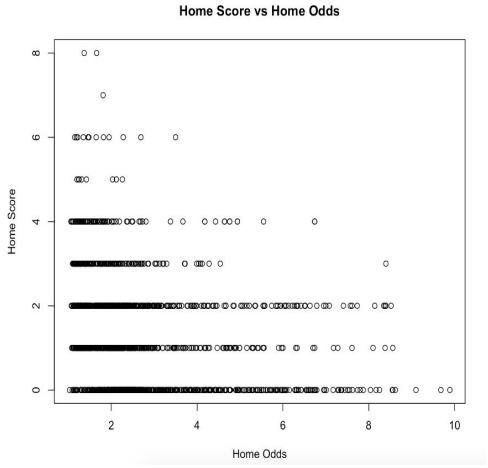

Figure 26: Home Score against predicted Home Odds

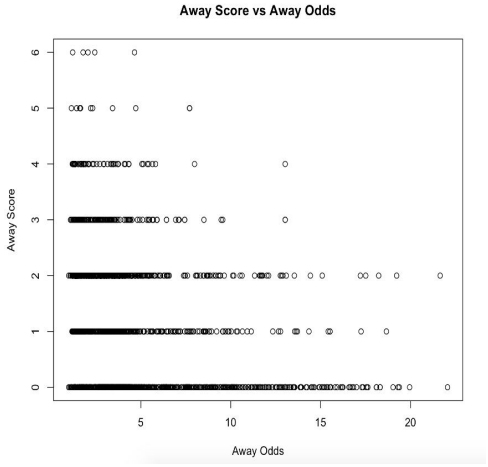

Figure 27: Away Score against predicted Away Odds

8.4 Poisson Model

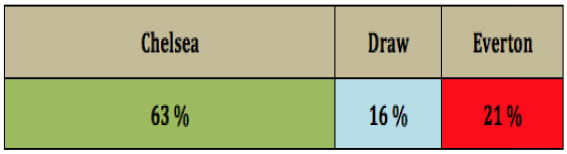

Figure 28: Winning Probabilities for two teams

The Poisson model calculates the percentage probabilities[8].After munging the data and, Poisson regression model

using the *dpois* function is used to get the predicted win percentages. A Sample run of the Poisson model for two teams is shown in figure 28 which predicts the winning probabilities when home team and away teams are entered.

8.5 Top 3 Teams

Based on the analysis and predictions that have been performed, an estimate has been made predicting the team that will win the League this current season along with the Runner-up and the third team with points. The prediction results are shown in figure 29 which show Man City as league Winners, Chelsea as runner-ups and Arsenal at third position. Apart from from this, their respective points have been calculated based on their number of wins, draw and losses as predicted in the analysis.

Position	Team	Points
1	Manchester City	99
2	Chelsea	84
3	Arsenal	79

Figure 29: Predictions show that Man City will win the league with Chelsea as Runner-up and Arsenal as second Runner-up

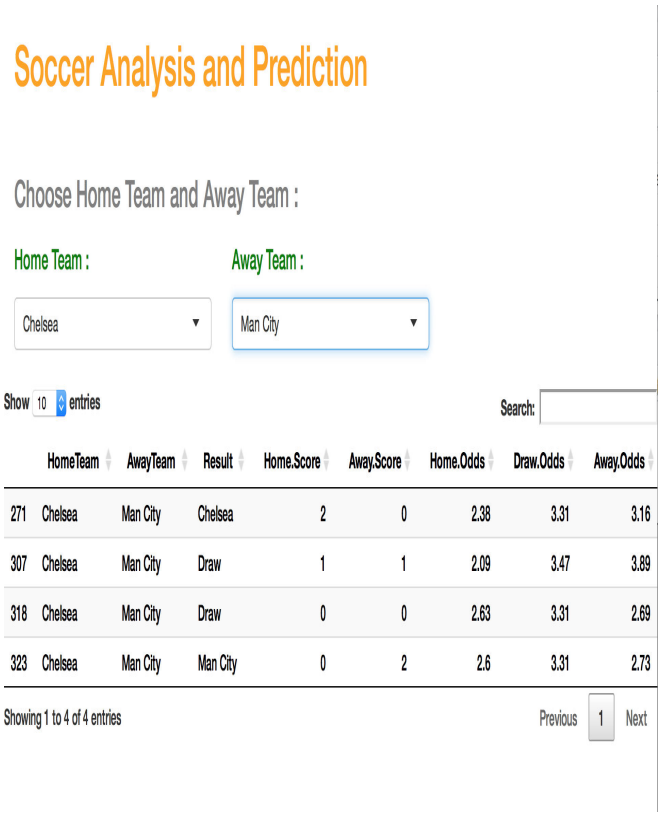

Figure 30: Displaying Predictions using Home Team as Chelsea and Away Team as Man City

9. GRAPHICAL USER INTERFACE

To display the results of the prediction, an interactive graphical user interface (GUI) was implemented using the Shiny apps for R. Figure 30 shows the GUI where home team and away team are to be selected and based on the teams selected, appropriate prediction will be displayed in terms of the predicted winner, predicted scoreline and predicted home, draw and away betting odds for the next few fixtures between the selected teams in future.

10. CONCLUSION & FUTURE WORK

To conclude, a model for soccer prediction and analysis has been implemented using data mining and machine learning techniques to predict match outcomes,match scores, betting odds and winning probabilities. By performing data cleaning and preparation tasks such as attribute removal, data imputation and attribute addition on raw dataset followed by feature selection, an appropriate dataset was generated that would be imperative to perform analysis pertaining to this project. As seen, multinomial logistic regression, random forest classification and linear regression models have been incorporated to perform the task of future result prediction. The Poisson Model has been used that makes use of probability to predict win percentages of various fixtures. Results show that we get high accuracy in all cases but multinomial regression proves to be the best model for prediction. This project also proves that more the number of features used for prediction, more efficient the model is. By performing comparative analysis, we come to the conclusion that with proper application of machine learning algorithms and proper handling of dataset, accurate predictions can be made even against unpredictability.

This project will ultimately aid soccer statisticians in improving their team's overall strategy and approach against specific oppositions by looking at the predicted results and factors that influence them. It will give a fairer chance to weak teams to rise up against relatively stronger teams and come out on top. This project will also serve as a basis for fans who are involved in betting and allow them to choose their teams to bet on based on the betting odds prediction.

As far as the future work is concerned, some additional work can be done to improve predictability. As of the dataset, some more attributes can be added. Factors such as player transfers, weather conditions and health status of individual players can be considered that might play a pivotal role in predicting the outcomes. For winning probability calculation, Dixon and Coles[1] adjustment can be applied to Poisson's Model to give the precision that is needed. Classifiers other than the ones used in this project can be tested for performance. Furthermore, this project can be extended not only to English Premier League but to the entire soccer world as well as to other form of sports as well. With the involvement of data mining techniques and analysis like this, the future of world soccer might just not remain specific to gameplay.

11. REFERENCES

[1] C. P. Barros and S. Leach. Analyzing the performance of the english fa premier league with an econometric frontier model. *Journal of Sports Economics*, 7(4):391–407, 2006.

[2] A. Bayaga. Multinomial logistic regression: usage and application in risk analysis.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Vsurf: An r package for variable selection using random forests.

[5] M. Gönen et al. Receiver operating characteristic (roc) curves.

[6] W. H. Greene. Econometric analysis.

[7] M. Haghighat, H. Rastegari, and N. Nourafza. A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, 2(5):7–12, 2013.

[8] H. Hu. Poisson distribution and application.

[9] K.-Y. Huang and W.-L. Chang. A neural network method for prediction of 2006 world cup football game. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[10] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[11] M. B. Kursa, W. R. Rudnicki, et al. Feature selection with the boruta package.

[12] C. Lerman, R. A. Schnoll, L. W. Hawk, P. Cinciripini, T. P. George, E. P. Wileyto, G. E. Swan, N. l Benowitz, D. F. Heitjan, R. F. Tyndale, et al. Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial. *The lancet Respiratory medicine*, 3(2):131–138, 2015.

[13] A. Liaw and M. Wiener. Classification and regression by randomforest.

[14] M. Lin, H. C. Lucas Jr, and G. Shmueli. Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, 2013.

[15] G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[16] I. Milliner, P. White, and D. J. Webber. A statistical development of fixed odds betting rules in soccer. *Journal of Gambling, Business and Economics*, 3(1):89–99, 2009.

[17] F. E. Moya. Statistical methodology for profitable sports gambling. 2012.

[18] M. Ramaswami and R. Bhaskaran. A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*, 2009.

[19] D. Roobaert, G. Karakoulas, and N. V. Chawla. Information gain, correlation and support vector machines. In *Feature Extraction*, pages 463–470. Springer, 2006.

[20] A. Schneider, G. Hommel, and M. Blettner. Linear regression analysis.

[21] J. A. L. Snyder. What actually wins soccer matches: Prediction of the 2011-2012 premier league for fun and profit. 2013.

[22] E. Štrumbelj. On determining probability forecasts from betting odds. *International journal of forecasting*, 30(4):934–943, 2014.

[23] A. Zimmermann. Wages of wins: could an amateur make money from match outcome predictions?