Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 463–470

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

Prediction of football match results with Machine Learning

Fátima Rodrigues[a], Ângelo Pinto[b]

[a]*Interdisciplinary Studies Research Center (ISRC), Polytechnic of Porto - School of Engineering, Porto, Portugal*
[b]*Polytechnic of Porto - School of Engineering, Rua Dr. António Bernardino de Almeida, Porto, Portugal*

Abstract

Football is one of the most popular sports in the world, so the perception of the game and the prediction of results is of general interest to fans, coaches, media and gamblers. Although predicting football results is a very complex task, the football betting business has grown over time. The unpredictability of football results and the growing betting business justify the development of prediction models to support gamblers. In this article, we develop machine learning methods that take multiple statistics of previous matches and attributes of players from both teams as inputs to predict the outcome of football matches. Several prediction models were tested, with the experimental results showing encouraging performance in terms of the profit margin of football bets.

© 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Industry Sciences and Computer Sciences Innovation

Keywords: data mining, sports betting, feature selection, classification, football;

1. Introduction

Football is one of the most impactful sports in the world. Over the last decades this sport has evolved significantly, and so has the revenue of the associated betting business. Profiting from betting on sports events is very difficult, especially when it comes to football. Anticipating results is a complex task due to the large number of factors that can influence games. Due to the characteristics of the game itself, it is possible that a team loses against a clearly inferior side, which makes any type of bet even more difficult. The unpredictability of the game makes it difficult to bet without doing any kind of analysis on information of the games.

Currently there are sites that offer free football betting tips. The predictions of these sites are based on team strength, goals scored [1] and mathematical calculations [2]. Although these sites provide a variety of information, they are not entirely reliable, and betting based on these predictions can lead to great losses. A concrete example is the website forebet [2]. Taking the 2017/18 English Premier League season as an example, this website could only correctly predict, on average, 2 out of the 10 matches of each round, which demonstrates the risk of relying on these predictions.

1877-0509 © 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Industry Sciences and Computer Sciences Innovation
10.1016/j.procs.2022.08.057

There is therefore an opportunity to develop better forecasting tools that indicate the most likely outcome of a football match and the respective confidence in the result, thus leading to more informed bets.

Football match results can be predicted by analysing historical data from previous seasons. The availability of data related to matches in the various football leagues is increasingly detailed, which enables the collection of data with distinct features. In this article, the prediction of results of football matches using machine learning (ML) algorithms will be carried out with multiple features that incorporate match statistics and attributes of all players from both teams.

The rest of the article is organized as follows. Section 2 gives a summary of the associated work. After analysing and processing the dataset in Section 3, several models are developed with different machine learning algorithms and tested in Section 4. Lastly, Section 5 concludes the manuscript and provides an outlook on future work.

2. Context and State of the Art

Sports betting consists of risking a certain amount of money by attempting to predict the outcome of a sports event, yielding a profit if a correct prediction is made and to the loss of the full amount otherwise. Although there are several types of sports bets, in this article we shall focus on the most common type encountered in football: 1X2 bets. These have three possible outcomes: 1 – home team win, X – draw, and 2 – away team win.

The profit that can be obtained from a given bet is calculated by multiplying the amount of money gambled by the odd of the bet. An odd is a value greater than 1 that reflects the probability of the associated event happening. The higher the probability, the lower the odd (i.e., closer to 1). Since a football match can have three different outcomes, the probability of hitting the right result by chance is 1/3, so odds are set by betting companies in such a way that random betting leads to losses, on average.

The use of ML in sports has increased considerably in recent years, having helped in the decision-making process in this area. For example, ML was successfully applied in the 2014 World Cup by the German national team scouting staff to study opposing teams and monitor their players to support the head coach in decision-making processes [3]. Gomes et al. [4] used statistics from more than 4900 games spanning 13 seasons (from 2000/01 to 2012/13) to develop prediction models. Such models were tested in 7 rounds of the 2013/14 season, making a total of 70 games and achieving an accuracy rate of 54.29%, with a profit margin of 20%. An important phase in this study was the creation of variables that could exist before the start of a game, such as the average goals of a team.

Although the unpredictability of sport is widely known, the football community is occasionally taken aback by some very surprising results, such as Leicester City's English Premier League title in the 2015/16 season. An in-depth investigation [5] was carried out to try to understand what led to this surprising victory and to try to understand how future predictions can be made. It was found that the achievement was attained due to the excellent performance of the Leicester goalkeeper and the fact that they were very effective at scoring on counterattacks. Another important factor was that several Leicester players made a large number of interceptions of passes that had a probability of more than 80% of being successful. In this case study, a model was also created to predict the number of shots and goals that a team would score during a game. It was found that a model that included information on the types of shots made (e.g., shots made in counter-attack pieces, shots from crosses into the penalty area), obtained better prediction results.

There is also a case study where the prediction of football matches was made using a multi-agent system [6]. The learning method used was the Multilayer Perceptron and was tested with data from the 2015/2016 season of the Spanish Premier Division. The success rate obtained was 61%. In another case study [7], logistic regression was used to predict the games of the 2015/16 season of the English Premier League. The prediction model only predicted the victory or defeat of the home team, excluding the possibility of a draw. The hit rate was around 69.5% and it was concluded that the variables that most influenced the prediction were home and away team defense records.

Another study [8] used data from the 2010/11 season of the Italian Serie A using 300 games for training and 80 for testing. One of the conclusions drawn in this study was that a team that makes many plays with aerial shots is more likely to draw or lose the game. Finally, the work [9] explored machine learning to forecast the outcome of football games based on match and player attributes. A simulation study that included all matches of the five top European football leagues and the corresponding second divisions between 2006 and 2018 revealed that an ensemble strategy achieves statistically significant returns of 1.58% per match.

The analysis performed showed that the studies with lower models' performance are related to the lack of variables that best characterize the players and the games itself. In addition, it is essential that the models are trained with games from different seasons as there are great variations in teams from season to season.

3. Analysis and Processing of Data

3.1. Data Description

The data that will be used in the present study correspond to a total of 1900 football matches spanning 5 seasons, from 2013/2014 to 2018/2019. The games are related to the top division of football in England, the official name of which is Premier League. In the 1900 matches collected, the home team won in 861 (45.3%), there were 470 draws (24.7%) and the away team won in 569 (29.9%).

Besides free statistical data on football matches, information about the performance of both teams was also collected, including goals scored, shots made, number of corners, number of faults committed, number of yellow and red cards, odds for each match, final result of match and match referee. In addition to these data, data extracted from the website sofifa.com, containing descriptive statistics of individual features and skills of all football players (e.g., pass accuracy, agility, reaction, aggression) as well as statistical information on the quality of football teams, were also used. Most of those variables are rated on a scale from 0 to 100. Other variables related to the performance of teams, such as overall rating, and rating of a team's attack, midfield and defence, are taken to be constant for each team over an entire season.

After data collection the data were inspected to resolve problems that might exist. However, no missing values and no duplicate records were found, so the number of records after data cleaning remained at 1900 records.

3.2. Data Processing

New variables were created to better predict the final result of matches. As in [4], variables related to the number of home wins of the team playing at home and the number of away wins of the visiting team were created. In the case studies analysed, only data relating to the goals scored by the teams were used and never the goals conceded. For this reason, we decided to calculate the goals conceded by the teams since these statistics could improve the results of the predictions. The variables created were average of goals conceded at home by the home team and average of goals conceded in away matches by the visiting team.

In order to make predictions of football matches before they take place, it is necessary to provide the prediction models with data that is available before the start of each match. Since the data extracted were related to the end of each match, such as the number of goals and shots of each team, this data could not be used directly to train the prediction models. For this reason, it was necessary to transform these data. The solution adopted in this case was to consider the averages of the available data such as the average of goals or the average of shots of a team before a given game. For each game and for each team, the averages of the following attributes were calculated: number of goals scored, shots, shots on target, number of corners, number of fouls committed, number of yellow cards, and number of red cards.

Thus, for a given match, the average goals for the home team are calculated based on the previous home matches played by that team during that season. For away teams, averages are also calculated based on previous games played as visitors during the season in which the game is played. For the remaining variables the averages are calculated in the same way.

3.3. Data Exploration

Then, in the total of 31 variables available, it was verified which ones were most related to each other, which variables could more clearly predict the goal attribute and which could be excluded. To verify the relationship between variables a correlation matrix was made between all numerical variables shown in Fig. 1.

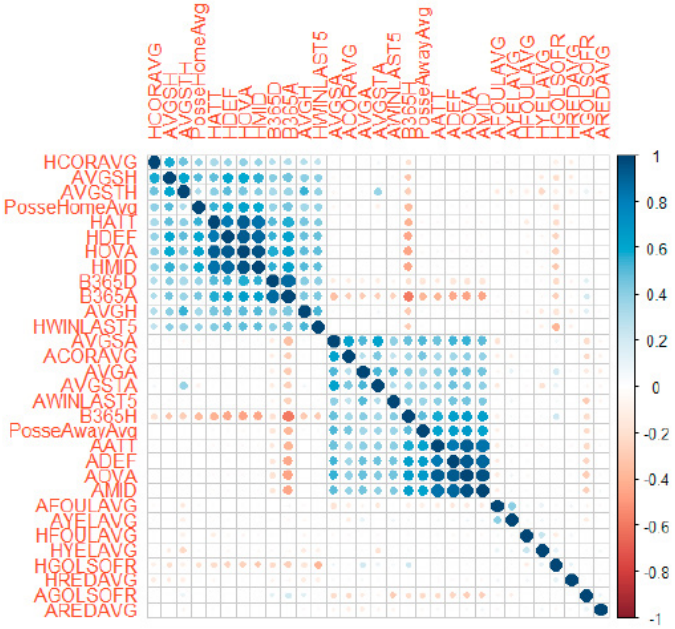


Fig. 1 Matrix correlation between variables.

The variables that have the highest positive correlation with each other are the variables that assess the quality of a football team (HATT, HDEF, HOVA and HMID and AATT, ADEF, AOVA, and AMID). There is also a strong correlation between B365D and B365A which represent the odds for the occurrence of a draw and a win for the visiting team. In addition, it can be seen that the variables related to the home team have a greater correlation with each other, as the variables related to the visiting team also have a greater positive correlation with each other. The variables related to the number of fouls, yellow and red cards and the number of goals conceded do not have a positive relationship with other variables. However, it appears that HGOLSOFR has a negative correlation with HWINLAST5, which would be expected since the teams that concede more goals have more difficulties to win games. It can also be concluded that the variable B365H is negatively related to variables such as B365A, HATT or HMID. This is normal since the higher the odds for the home team, the lower the odds for the visiting team. Likewise, the better the home team, that is, the higher the HATT and HMID values, the lower the odds for the home team to win.

The correlation matrix was also important to identify variables to be removed from the initial data set. In a classification process, variables with high correlation with other variables must be removed [10]. This must be done so that the importance of these variables is not overestimated, harming the prediction of results. In case there are two identical variables, one of them becomes redundant, not adding relevant information to the training model. For this reason, the variables HATT and AATT, corresponding to the attack strength of the home and away team, with correlation greater than 0.9 were removed.

An important step before the forecasting phase is to try to identify the variables that help to more easily predict the target attribute/variable, in this case the FTR variable. Thus, several graphics were made to verify the relationship between all variables and the FTR values: win, draw and loss. The graphics allowed us to verify that there are four variables that best help to predict the final result of a game. These variables are B365H, B365A, AOVA, HWINLAST5 shown in Fig 2. In addition to verifying the variables that could offer a better prediction, those that would have the worst ability to predict the objective attribute were also analysed. It was found that the variables HREDAVG, AREDAVG, HCORAVG and ACORAVG (Fig. 2) would be of little relevance to train the forecast model. For this reason, these 4 variables were not used in the training of the forecast models.

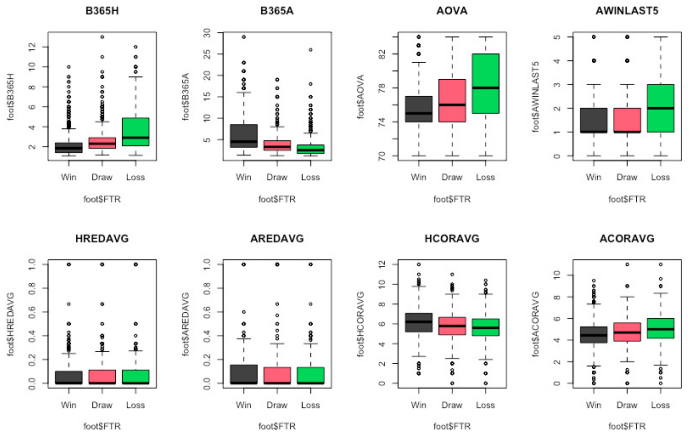


Fig. 2. Boxplot of 4 variables best and worst related with goal attribute.

4. Data Mining

Next, the data were separated into training and test sets. In this work, it was considered that the test set should include all the games of a season, because throughout the season the team performance varies. At the beginning of the season the teams may not be at their normal level and at the end of the season they may be worn out, or have already reached their goals, leading some teams to have a performance below normal. These factors can lead to unexpected results, so for a classification model to be considered credible, it must be tested over an entire season.

Concerning the training set, if a model is trained with few examples, there may be a risk of overfitting, that is, the model fits too much to the training data and has poor predictive ability. For this reason, 4 seasons were used for training and 1 season for testing. The 4 training seasons (2013/2015 to 2016/2017) correspond to a total of 1520 games and the test season (2018/2019) to 380 games.

To find the best classification model, several algorithms with different characteristics were tested in order to verify which one best fits the data. Next, the algorithms used and the R software libraries:

- Naive Bayes (NB) – e1071 package;
- K-nearest neighbors (KNN) – kknn package;
- Random Forest (RF) – randomForest package;
- Support Vector Machines (SVM) – svm method from e1071 package;
- C5.0 (decision trees) –C50 package;
- Xgboost –xgboost package;
- Multinomial Logistic Regression (MLR) – multinom method from nnet package;
- Artificial Neural Networks (ANN) – nnet method of the nnet package

Before testing the different algorithms, the data was normalized using the “z-score” method to eliminate the effect of large variations in values [10]. After normalizing the data, the most important variables for predicting the results

were identified. This process was done so that the forecast models use only the most relevant variables and guarantee better forecasts. To identify the best variables, the Boruta algorithm [11] was used. Boruta is a heuristic variable selection algorithm based on the random forests algorithm that aims to find the most relevant variables in a dataset. The results of running this algorithm show the relevant and non-relevant variables. This algorithm was used because it does not look for a suboptimal solution, instead it tries to find all variables with relevant information, thus allowing to eliminate variables that would negatively affect the forecast models. The Boruta algorithm eliminated 7 non-relevant variables, thus leaving 18 variables for the construction of classification models.

4.1. Prediction Results – Phase 1

In order to correctly verify the differences between the classification models, different measures were used, such as the accuracy of the model and the percentage of games correctly predicted for the draws and for the victories of the home and away team. It was also considered the profit that would be obtained if each bet was correct or incorrect. Since the forecast model will be included in a betting support decision system, calculating the profit made is essential to verify that the model is successful. The profit was calculated considering a value of 2 euros per bet. Bearing in mind that there are 380 test games, the total amount wagered would be 760 euros, which would be wagered over 9 months. If a bet was missed, the profit would decrease by 2 euros. If correct, the profit is calculated according to equation: $Profit = bet_{amount} \times bet_{odd} - bet_{amount}$. For example, in a game with a bet on a tie and the draw odd is 1.5 the profit would be 1 euro. As the value bet is always 2 euros the profit would be equal to 2x1.5-2 = 1. Table 2 presents the results of predictions made with the 8 models developed with the selected algorithms.

Table 2. Forecast results with 18 variables.

Algorithm	Accuracy	Profit	% Victories Home Team	Draws	% Victories Away Team
Bayes	53,42%	17,40€	51,87%	30,95%	73,79%
KNN	57,63%	78,02€	78,07%	15,48%	55,05%
RF	59,21%	85,20€	75,40%	21,43%	60,55%
SVM	61,32%	95,06€	88,77%	3,57%	58,72%
C5.0	55,26%	42,52€	72,73%	23,81%	49,54%
Xgboost	59,47%	72,80€	77,54%	10,71%	66,06%
RLM	57,63%	32,56€	78,07%	5,95%	62,34%
RNA	50,00%	18,28€	58,29%	30,95%	50,46%

The forecast results were satisfactory. The best algorithm was SVM achieving a percentage of success above 61.32%. The profit of 95.06 euros, although not high, corresponds to a reasonable profit margin of 12.51%. It should also be noted that all algorithms achieve a profit. However, the best hit rate obtained did not exceed the best hit rates found in the referenced case studies. The best model was correct in only 3.57% of the ties, which is a low value. Since the results were still not as expected, it was decided to proceed with the development of new forecast models.

4.2. Prediction Results – Phase 2

In this second phase, it was decided to test all possible combinations with the 18 pre-selected variables, in order to achieve the best possible hit rate. However, with 18 variables, the total number of combinations to be tested would be over 260,000. This number is too high, so this approach would not be reasonable, so it was decided to identify among the 18 pre-selected variables, the most important variables using the Backwards Feature Selection “rfe” method of the “caret” package of the software R. This algorithm starts by verifying the importance of the variables using all the variables provided, then makes successive iterations where it removes some variables, leaving only the most important in each iteration. In the end, the most important variables are those used in the test that obtained the best result. The attributes identified by the algorithm as the most important were: B365H, B365D, B365A, AVGH, AGVA, HOVA and AGOLSOFR. Thus, from the 18 initial attributes, 11 attributes remain, so the total number of combinations needed

in this case is 2048, and this approach is feasible. In this new approach, all combinations tested used the 7 attributes, varying only the remaining 11 attributes, and 8 models were always generated with each combination of attributes. Table 3 presents the best results obtained with each of the combinations of variables.

Table 3. Better predictions with the various combinations of variables.

N. of variables combined	Algorithm	Accuracy	Profit	% Victories Home Team	Draws	% Victories Away Team
1	SVM	62,62%	134,50€	51,87%	30,95%	73,79%
2	SVM	63,68%	157,22€	90,91%	4,76%	62,39%
3	SVM	63,95%	160,12€	91,44%	3,57%	63,30%
4	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
5	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
6	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
7	RF	63,95%	191,58€	80,21%	28,57%	63,30%
8	RF	65,26%	203,24€	81,28%	29,76%	65,14%
9	RNA	62,89%	181,72€	82,35%	27,38%	56,88%
10	SVM	62,36%	123,36€	88,77%	8,33%	58,72%
11	SVM	62,11%	115,06€	88,77%	7,14%	58,72%

4.3. Results Analysis

The approach taken to test different combinations of variables yielded good results. In all cases, the best models obtained higher success rates than the initial model, the success rate of which was 61.32%. The algorithms that allowed reaching the best models in the different cases were SVM, RF, Xgboost and RNA. The best model was obtained by testing combinations of 8 variables with the 7 variables identified as the most important, thus having a total of 15 variables. The best model used the RF algorithm, which obtained a hit rate of 65.26% and a profit margin of 26.74%. Comparing with the initial model, the success rate increased by almost 4%. The percentage of correct bets on home team wins decreased by 7% but increased by 26% on draws and close to 7% on away team wins. The profit margin rose by 14%. This rise is justified by the increase in the number of correctly predicted draws. Bets on draws tend to have higher odds than bets on wins, so the profit obtained was higher.

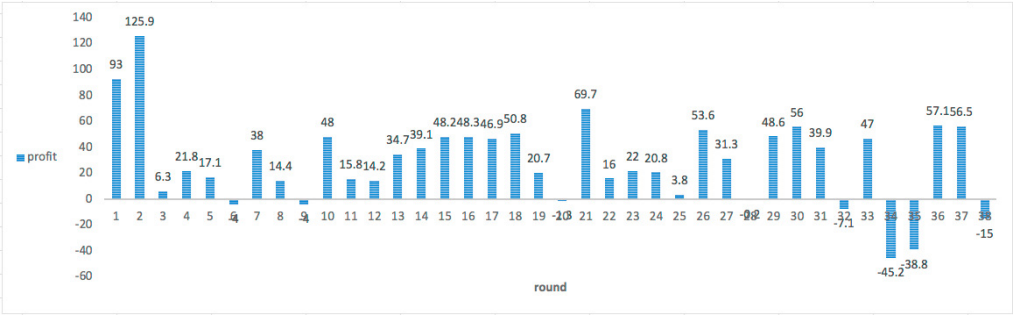
In addition to obtaining better results, this forecasting model has a more balanced success rate for different classes than the initial one. This will, therefore, be the model chosen to be included in the support decision system. Table 4 details the measures used to evaluate the forecast model. These measures were the accuracy, the macro-average of precision and recall, and the profit margin. The macro average is calculated by averaging the precision and recall for each class − in this case, home team win, draw and away team win. Analysing the results (Table 4), it is possible to verify that the macro-average of precision and recall are close to the value of the accuracy, so the forecast model is balanced.

Table 4. Performance measures of the best model.

Algorithm	Random Forest	
Accuracy = 65,26%	Profit Margin = 26,78%	
Victory Home Team	Precision=68,47%	
	Recall = 81,29%	
Draw	Precision=50,0%	Macro media Precision = 61,40%
	Recall = 29,76%	
Victory Away Team	Precision=65,74%	Macro media Recall= 58,73%
	Recall = 65,14%	

In addition to the global analysis of the forecast model throughout the whole season, an assessment on each round was also carried out. This analysis is crucial to verify that the prediction model has a constant performance. A model of this type should not achieve too low success rates in individual rounds, as this could lead to incurring in great losses. Fig. 3 shows the profit of the forecast model at each of the 38 rounds.

Fig. 3. Profit of best model by round.



The model performed well throughout the season. In the 38 rounds of the season, a profit was achieved in 30 and only 8 led to a loss. This yields an average profit margin of 26.78%, which competes with the works analysed.

5. Conclusions

In this article the process of developing models for predicting the results of football matches to support sports betting was described. Data from two different sources were used, one to obtain statistical data about previous games and the other to collect data related to the teams. The analysis and processing of the data made it possible to draw important conclusions about the variables to be used in the models. The study compared several algorithms in order to create the best prediction model. The algorithms were trained with data from 4 seasons and tested with all the games of the 2016/2017 season of the English Premier League, which allowed a detailed assessment of the behaviour of the model over the various rounds of the season, namely the match success rate and profit margin that would be obtained in each betting week. The percentage of games correctly predicted by the model was 65.26%, which competes with the best works analysed in the area. The profit margin obtained was also higher than that of the referenced case studies.

As future work, the forecast model will be integrated in a decision support system that will assess the risk of bets based on the probability of occurrence of the forecast model results. This will allow the gambler to know the risk associated with the bet, thus having greater support in obtaining profit from sports betting.

References

[1] SoccerVista - football results, predictions and betting picks, SoccerVista, 2018. http://www.soccervista.com [Last access: 10-Jan-2022]

[2] Forebet, Mathematical football predictions, Tips, Statistics, Previews, Forebet, 2018. https://www.forebet.com [Last access: 10-Jan-2022]

[3] Bojanova, I. (2014). It enhances football at world cup 2014. IT Professional, 16(4), 12-17.

[4] Gomes, J., Portela, P. and Santos, M. F., (2015). Decision Support System for predicting Football Game result. 348-353.

[5] Ruiz, H., Power, P., Wei, X., & Lucey, P. (2017). " The Leicester City Fairytale?" Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1991-2000.

[6] Cañizares, P. C., Merayo, M. G., Núñez, M. and Suárez-Paniagua, V., (2017) 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), pp. 572-576.

[7] Prasetio, D. and Harlili, D., (2016) International Conference On Advanced Informatics: Concepts, Theory And Application, pp. 1-5.

[8] Zuccolotto, P., Carpita, M., Sandri, M. and Simonetto, A., (2014) 'Football Mining with R', in Data Mining Applications with R.

[9] Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: Prediction of match results based on player characteristics. Applied Sciences, 10(1), 46.

[10] Dasgupta, N., (2018) Practical Big Data Analytics. Packt Publishing Ltd.

[11] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), pp. 1-13.