Imperial College London

Data in Football

–

A machine learning-driven approach to predict game outcomes

(Word Count: 5,163)

Executive Summary

Over the recent years, an arms race to quantify the working mechanism of football games by making sense of the exploding amount of information and data available between opposing teams, analytics companies and bookmakers is visible. This development exposes the number-heavy sports betting market to newly arising opportunities and treats for bettors and bookmakers. The paper at hand develops a public data-based, machine learning-driven sports betting strategy for the English Premier League season 2018/19 by using team characteristics input features as well as time-series performance data in combination with dimensionality reduction methods and optimisation techniques. The final validated strategy achieved a return of investment of 38.38 % by betting on 8% of all possible games.

Table of Contents

List of Abbreviations

Bn	Billions
EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
ROI	Return of Investment

1 Introduction and Motivation

1.1 Research Subject

“The ball is round, the game lasts ninety minutes, and everything else is theory.”
- Sepp Herberger (1898-1977), Former Head Coach of the German National Football Team

Football was long seen as a chaotic, unstructured and complex game where a few decisive actions that are accompanied by a high amount of variance decide the outcomes of games. The common opinion was that the critical, controllable elements for success were mainly intangible factors such as talent, passion or team spirit. Room and application areas for structured information processing, analytics and forecasting were limited for a long time. However, with the technological progress for gathering and analysing data, the tighter margins between opposing teams and the increasing economic significance of the football business, data science departments within football clubs and football analytics companies such as StatsBomb or Opta Sports are ever since on the rise. (Biermann, 2019)

Football analytics companies collect up to 2,000 geospatial and action-based data points per game from an individual player-based perspective as well as from a team perspective (Burn-Murdoch, 2018). The collected data can then be used for the individual player and team performance assessment, scouting and transfers, training and game plan design or the development of sports betting strategies. This paper will focus on modeling and forecasting Premier League game outcomes for the season 2018/19 with publicly available data in order to design a profitable betting strategy. To put the economic relevance of the topic into perspective, the English Premier League is estimated to be a € 5.44 bn industry on which bets worth € 67 bn are placed every season (Deloitte, 2019; Statista, 2019a).

1.2 Literature Review

In the following, the goal is to provide an outline of the advancement of scientific approaches to modeling and forecasting football results. The review will start with traditional statistical approaches, followed by the upcoming use of modern machine learning techniques and end with today’s state-of-the-art use of individual player data.

The first statistical approaches to model football results were based on using results only as input variables. One of the earliest approaches of showing that football results are not occurring purely by chance goes back to Hill (1974) in 1974. The researcher found a strong correlation between a team's scored goals in the previous season and a team's scored goals in the current season. However, it was only in 1997 when researchers first came up with a complete statistical model to forecast English League game outcomes to systematically challenge the odds of bookmakers. Dixon & Coles (1997) used maximum likelihood to develop comprehensive parametric Poisson regression models to forecast goals in order to assign probabilities to potential scores. This model lied the foundation for a profitable betting strategy. Crowder et al.'s (2002) work builds upon Dixon & Coles's (1997) approach of applying Poisson regression models. The researchers improved the accuracy of Dixon & Coles model by enhancing the input data layer with a proxy variable for team strength based on past performances. In order to create this proxy, Markov Chains and Monte Carlo Simulations were used.

In the year 2005, Goddard (2005) made one of the first approaches of not only using previous results as input variables but to supplement the model with information such as the geographical distance between the two opposing teams or the teams' performances in other competitions. Moreover, the researcher further enhanced Dixon & Coles's approach with an ordered probit model. However, only slightly better forecasting power could be achieved. Some years later, Hucaljuk & Rakipović (2011) made a first approach of applying modern machine learning techniques such as KNN, random forest or neural networks while using input features such as the recent results, the teams' current rankings, the number of injured players, the average number of scored goals or the outcome of previous meetings between the two opposing teams. The researchers achieved a forecasting accuracy between 50-65% for predicting wins by the home team, draws or wins by the away team. However, the sample size was very small, and the variance of the accuracy was high depending on the statistical method used. Tax. et al. (2015) used a very similar approach to Hucaljuk & Rakipović (2011) by combining the previous work with dimensionality reduction techniques. The team achieved a stable accuracy of 49% using public data from the Dutch league.

In terms of input features, a recent popular trend is to model player and team strength. The basic concept goes back to Elo (1978) who modeled player strength to predict chess games. For football, Constantinou et al. (2013) and Lasek (2016) developed dynamic team rating systems based on features such as recent performances, previous meetings between the teams or the location of the match. Another hot trend is to use individual player data such as average number of shots, average distance of shots, average angle of shots, defender proximity for shots, or the number of passes to create significant goal-scoring opportunities, also called packing, in order to predict the number of expected goals by a team (Lucey, 2015; Impect, 2019). However, to design these input features, geospatial and detailed player individual data is necessary. This information is not publicly available.

1.3 Structure

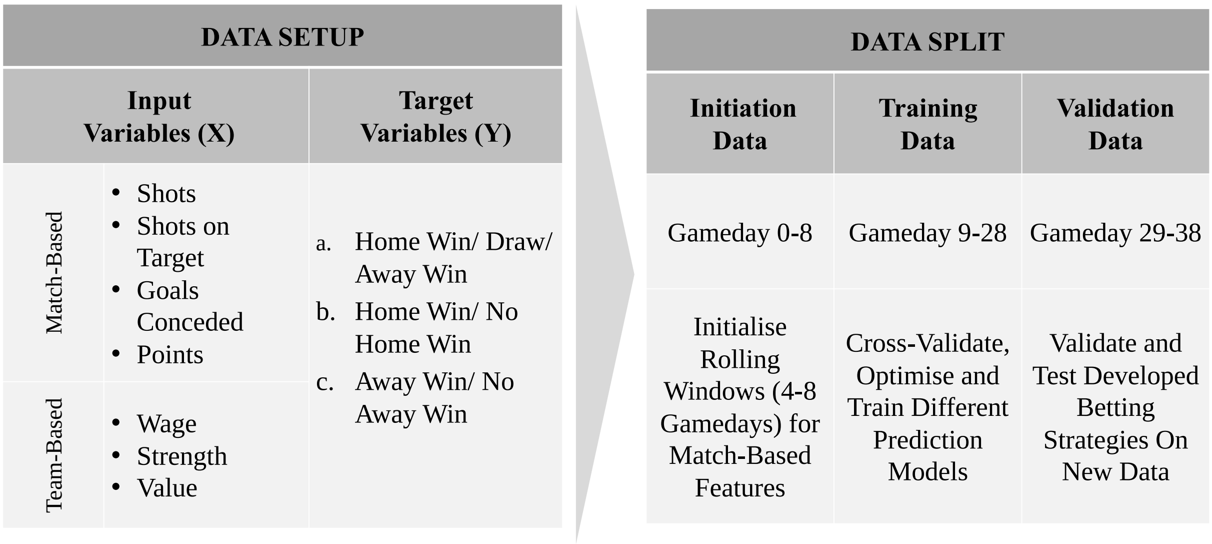
This paper is willing to combine different, in the previous chapter outlined modeling techniques and approaches from credit default forecasting methods to lay the foundation for the design of a profitable strategy in today's betting market. The paper's goal will be approached in four steps. First of all, based on the insights of the literature review, the methodology of the paper's modeling approach will be outlined. Subsequently, the sourcing of the data will be explained before an exploratory data analysis (EDA) will help to find relevant patterns in the data, check the validity of the study and guide the engineering of the most meaningful input features. Thirdly, different models will be assessed before the practical relevance of the results will be reflected. Finally, a conclusion chapter will summarise the results, outline limitations and highlight the further potential of the study.

2 Methodology

Based on the literature review, three findings for accurate football game outcome forecasting are apparent. First of all, the highest prediction accuracy was achieved by using state-of-the-art machine learning techniques in combination with dimensionality reduction techniques (Tax. et al., 2015). Secondly, team strength input features, as well as time-series performance data, are crucial to achieving high-scoring results (Constantinou et al., 2013; Lasek, 2016). Thirdly, player individual performance data enhances models further. As player individual performance data is not publicly available, the focus will lie on finding the best possible data sources, proxies and modeling techniques for findings one and two. Additionally, betting odds for the Premier League season 2018/19 will be sourced.

Graph 1 illustrates the planned usage of data. The input layer data consists of match-based features such as the number of collected points, scored goals and different kinds of shots taken for rolling windows of the last four to eight matches. Rolling windows are defined as averaging the number of actions for a certain metric over a defined period of past games. The modeling of team characteristic is considered from three different perspectives. Firstly, the team characteristics are modeled as team strength. The strength will be based on the ratings of the video game Fifa 19. Each year, a group of experts systematically evaluates the abilities of each player within a team to subsequently assign a strength score to each team. Secondly, the Premier League teams' squad net worth is considered. The third perspective of team characteristics is the average squad salary as it is assumed that there is, as in every competitive market, a strong correlation between performance and salary. Considering all the above-mentioned data points, a data input layer of over sixty independent variables will be provided. In terms of the target variable, there are three different target variables modeled. The first target variable will be *home win/ away win/ draw*. The second target variable will be *home win/ no home win* and the third target variable will be *away win/ no away win*. The prediction of draws will be disregarded for target variables two and three as draws are the least frequently occurring outcome. Therefore, draws are harder to predict and associated with more variance. As a next step, the data is split into three parts. The first part, the first 80 matches of the season (the first eight gamedays), are used to initialise the rolling window time-series data. The next 200 matches (gamedays 9-28) will constitute the training data set. Lastly, the last 100 games of the season (gamedays 29-38) will be used as a validation data set.

Graph 1: Planned Data Usage



DATA SETUP		
Input Variables (X)		Target Variables (Y)
Match-Based	• Shots • Shots on Target • Goals Conceded • Points	a. Home Win/ Draw/ Away Win b. Home Win/ No Home Win c. Away Win/ No Away Win
Team-Based	• Wage • Strength • Value	

DATA SPLIT		
Initiation Data	Training Data	Validation Data
Gameday 0-8	Gameday 9-28	Gameday 29-38
Initialise Rolling Windows (4-8 Gamedays) for Match-Based Features	Cross-Validate, Optimise and Train Different Prediction Models	Validate and Test Developed Betting Strategies On New Data

To start with the modeling of the training data, four machine learning techniques are used to predict the target variables. The reason for evaluating multiple machine learning techniques is based on the fact that certain models lead to better results given specific characteristics of the data. As a next step, each method is twice applied for each of the three target variables, once with the full set of input variables and once with a principal component-reduced (PCA) set of input variables. Dimensionality reduction techniques such as PCA allow reducing the number of input variables while conserving all critical information and therefore decreasing noise and creating sparse models (James et al., 2013). The number of principal components is determined by cross-validation. Cross-validation is a technique to further split the training data into smaller subsets in order to train and test a method on differently composed subsets. This allows to reduce variance and lower bias by addressing underfitting and overfitting tendencies as well as the low stability of results given the dependency on the composition of the particular training data set (James et al., 2013). Subsequently, the accuracy of each machine learning technique is evaluated by means of ten-fold cross-validation. Next, all the different techniques will be used to predict the probability for each possible match outcome.

To convert the predicted probabilities into a promising betting strategy, a recursion optimisation function will be developed to determine the optimal probability threshold (cutoff) of above which probabilistic certainty to invest into a predicted outcome in order to maximise the winnings. This idea comes from the area of credit default forecasting (Kürüm et al., 2012). In a first step, given the probabilities of the different outcomes for each match, the function places a bet of one on outcomes that exceed a given probability and does not bet on all other predicted match outcomes. As a second step, the placed bets are reconciled with the effective outcome. In case of a match between the bet and the outcome, the odds for the outcome are multiplied by the betting amount of one and noted as a win. In case of mismatch, the loss of the betting amount of one is noted. As a third step, the sum of wins and losses as well as the total betting amount are accumulated for each probability threshold between 0.5 and 1 in steps of 0.01. Finally, the probability threshold (cutoff value) with the highest return of investment (ROI), defined as the sum of wins and losses plus the betting amount divided by the betting amount, will be evaluated. Subsequently, based on the ROI-maximisation criteria and the stability of the achieved results, the most promising machine learning technique and its corresponding probability threshold will then be applied to evaluate the investment strategy’s out of sample performance on the validation data set. The stability of the results will be tested by the graphical inspection of the ROI development for all possible probability thresholds. The above-described function will be run for each of the three target variables and the details of the functions working mechanism can be found in Appendix 11.

3 Data

3.1 Sourcing

The match data for all 360 Premier League matches of the season 2018/19 was obtained from the website football-data.co.uk (2019). The dataset contains for each match 61 features. However, the majority of the features are betting odds from different betting companies. Therefore, a first reduced dataset with match prediction relevant data was created. The obtained data set includes features such as goals, shorts or shorts on target for both the home and the away team. For all features, time-series data as described in the methodology chapter was created. In terms of modeling team characteristics, the first feature squad net worth was based on the values assigned by the community of football enthusiasts and experts from transfermarkt.com (2019). The wage bills of each team for the season 2018/19 were obtained from the Statista (2019b). Lastly, as already mentioned, the strengths of the teams are approximated by the assigned values of the video game Fifa 19 and were scraped from fifaindex.com (2019).

The betting odds for home wins, away wins and draws for all 380 matches were considered from the betting company Bet365. These odds were included in the dataset from football-data.co.uk (2019) already. It is important to see that the betting odds of the major players in the British sports betting market only deviate marginally. Therefore, the odds of Bet365, one of the largest betting companies in Great Britain, constitute a representative market to develop a relevant betting strategy.

3.2 Exploratory Data Analysis

As a first step of the EDA, the goal is to evaluate whether there are significant differences between the characteristics of the different data sets. Table 1 shows the summary statistics of the match features for each of the three data sets.

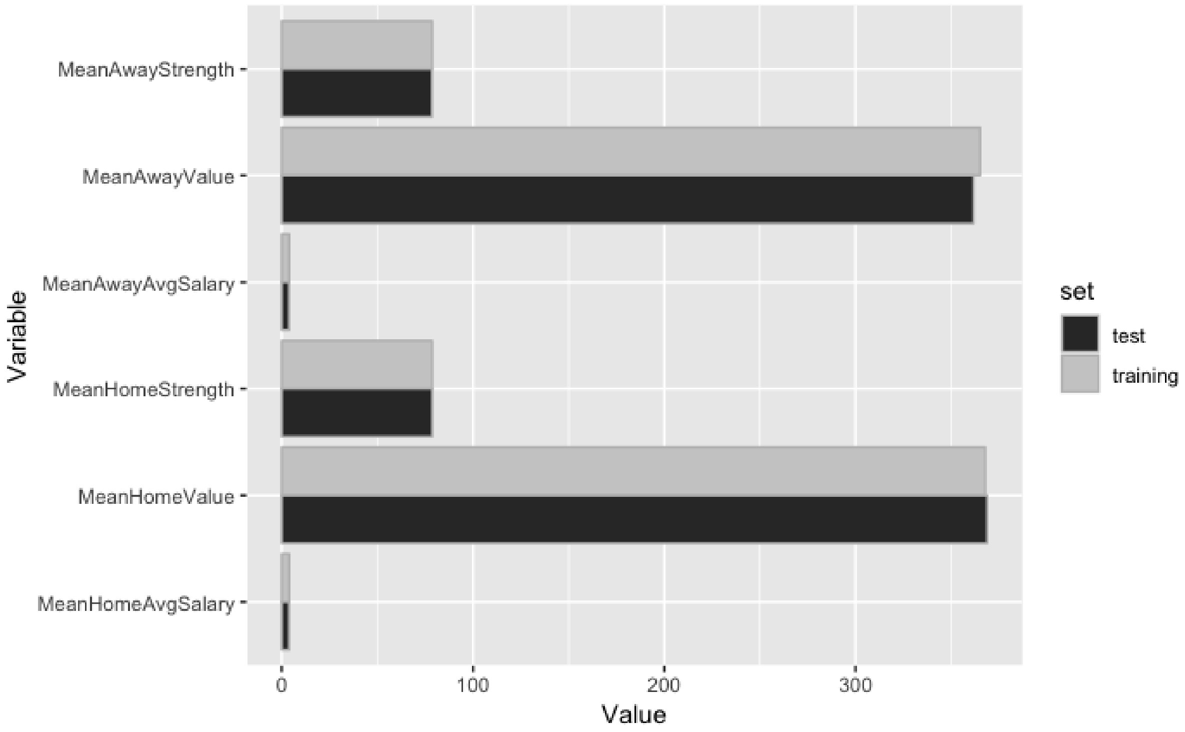
Table 1: Summary Statistics Match Features

	Initalisation	Training	Validation	Overall
HomeGoals (∅)	1.475	1.605	1.570	1.57
AwayGoals (∅)	1.338	1.230	1.230	1.25
HomeShots (∅)	13.863	14.270	14.080	14.13
AwayShots (∅)	11.438	10.710	11.780	11.15
HomeShotsTarget (∅)	4.725	4.875	4.630	4.78
AwayShotsTarget (∅)	4.525	3.705	3.900	3.93
HomeWins (%)	42.50	49.00	49.00	47.63
AwayWins (%)	37.50	32.00	34.00	33.68
Draws (%)	20.00	19.00	17.00	18.69
Count	80	200	100	380

The table provides two major insights. Firstly, it is important to notice that the outcomes of the target variables are not perfectly balanced. Home wins occur almost three times as likely as draws and approximately 50% more often than away wins. This is important to take into consideration when choosing appropriate machine learning prediction techniques at a later stage. Secondly, the characteristics of the training and validation data sets are highly similar while there is a slight discrepancy between the initialisation data set and the training and validation data sets. This is likely due to the small sample size and potentially also to the particular fixtures within each data set. For example, the higher amount of away goals and the lower amount of home goals in the initialisation data set could be caused by a higher proportion of weaker teams playing stronger opponents at home during that particular period of the season. Nevertheless, as no algorithm is trained, and no predictions are made with the initialisation data set, the differences have no impact on the results of the study. Therefore, in the following, the focus will lie on identifying patterns concerning the training and validation data set.

Graph 2 shows that also in terms of team-based features, the differences between the observations in the training data set and the observations in the validation data set are marginal. Therefore, the assumption of a homogenous test and validation data set holds. Consequently, the results of the study should be meaningful and representative.

Graph 2: Summary Statistics Team-Based Features



As a next step, a closer look at the correlation tables shows that all team-based features show relevant correlations with the *home win/ draw/ away win* target variable in the range of |0.21-0.26|. With regard to the time-series data, it appears that the rolling windows of the last six to eight games show the highest correlations with the target variables. The average shots on target, the average number of points and the average number of goals from both the away and the home team show the highest correlations with the *home win/ draw/ away win* target variable (range |0.14-0.19|) while other variables show lower correlations. Consequently, the results of the correlation tables indicate that dimensionality reduction techniques could be important tools to create models with a reduced amount of noise and therefore a higher predicting power. All the details for the correlation tables can be found in Appendix 1-10.

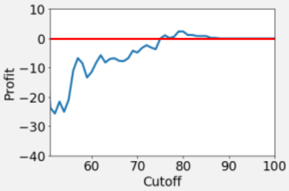
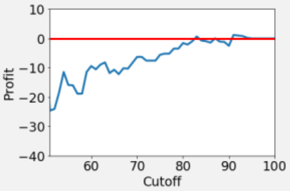
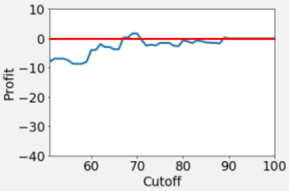
4 Modeling

4.1 Training Data

As outlined before, the analysis will be based on the application of four different machine learning models to different input variables and output variables. In the following, the working pattern of each of the four methods as well as their suitability for the particular problem at hand will be shortly introduced. Subsequently, the results for each of the modeling techniques for the training data set will be depicted.

The first considered machine learning algorithm to predict the Premier League games outcomes of the season 2018/19 is linear discriminate analysis (LDA). “LDA is a linear, multivariate function that models conditional distributions of the target variable in order to allocate observations to different groups or assigning a probability of belonging” (James et al., 2013; Studer, 2019). A summary of the results for the LDA method can be found in Table 2. LDA results in an accuracy of 46.35% for the full model and 58.00 % accuracy for the PCA-reduced set of input variables when predicting *Home Win/ Away Win/ Draw*. When predicting the dummy variable *Home Win/ No Home Win* the full model achieves an accuracy of 57.31% and an accuracy of 65.35% for the reduced model. For the dummy variable *Away Win/ No Away Win* the full model’s accuracy is 68.32% while the accuracy for the reduced model is again significantly higher with 76.48%. The ROI maximising cutoff value for the *Home Win/ Away Win/ Draw* prediction is 0.9 and leads to an ROI of 29.88% while for the *Home Win/ No Home Win* dummy the cutoff is 0.9 as well which leads to an ROI of 23.60%. For the *Away Win/ No Away Win* prediction the cutoff is 0.68 with an ROI of 5.70%. The betting volume is distinct higher for the *Away Win/ No Away Win* prediction compared to the prediction of the other two variables. However, considering the graphs, the results seem to be stable for the *Home Win/ Away Win/ Draw* and the *Home Win/ No Home Win* target variables while for *Away Win/ No Away Win* in some cases losses are realised even after the optimal cutoff value. This indicates less stable predictions and higher variance.

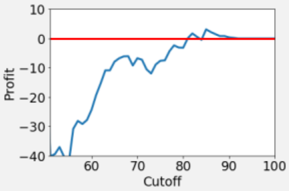
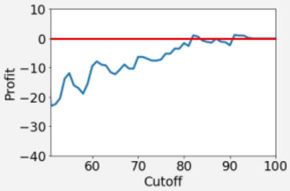
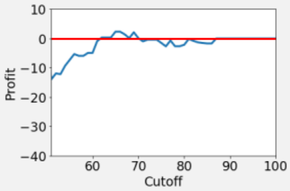
Table 2: Summary Results Training Data Set for LDA

	Home Win/ Away Win/ Draw	**Home Win/ No Home Win**	**Away Win/ No Away Win**
Accuracy Full Model (%)	46.35	57.31	68.32
Accuracy PCA-Reduced Model (%)	58.00	65.54	76.48
Graph Best Performing Model (Full or PCA-Reduced)			
Max Cutoff (%)	90	90	68
Max ROI (%)	29.88	23.60	5.7

Betting Volume (%)	4	2.5	15

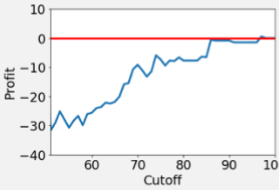
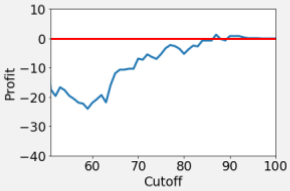
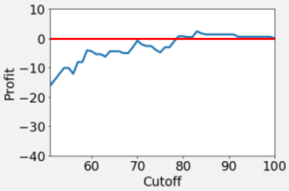
The second machine learning technique applied to the problem at hand is logistic regression. “Logistic regression is a linear modeling technique based on maximum likelihood specifically designed for classification and probability predictions” (James et al., 2013; Studer, 2019). This machine learning technique is particularly suitable as all classifications are based on the probability predictions of the method. The test data set results for logistic regression are displayed in Table 3. One can notice that also for the logistic regression method, higher accuracies can be achieved with the PCA-reduced input layer. Overall, the logistic regression performs similar but slightly worse than LDA except for the target variable *Home Win/ No Home Win* where the accuracy score could be slightly increased to 66.04%. Also in terms of betting volume, ROI and result stability for the optimal strategies, the obtained outcomes are highly similar to the results of LDA.

Table 3: Summary Results Training Data Set for Logistic Regression

	Home Win/ Away Win/ Draw	**Home Win/ No Home Win**	**Away Win/ No Away Win**
Accuracy Full Model (%)	52.44	61.42	73.95
Accuracy PCA-Reduced Model (%)	57.03	66.04	75.03
Graph Best Performing Model (Full or PCA-Reduced)			
Max Cutoff (%)	84	90	64
Max ROI (%)	31.10	23.60	7.48
Betting Volume (%)	5	2.5	15.5

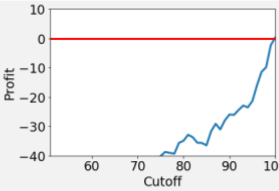
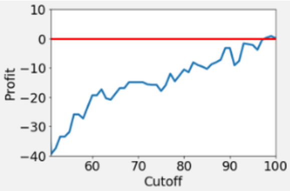
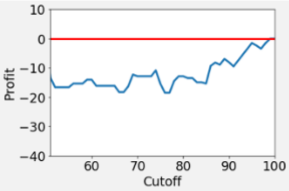
As a third machine learning prediction method, random forest will be tested. “Random forest is a machine learning technique suitable for probability prediction that reduces the overfitting tendency of decision trees by building a multitude of trees with different subsets of predictors” (Studer, 2019). Random forest is considered a promising approach for the task at hand due to the superior handling of unbalanced target variable distributions, insignificant input variables as well as lower sample sizes (James et al., 2013). However, the results in Table 4 show that random forest has a strictly lower accuracy compared to LDA and logistic regression. Nevertheless, when it comes to the optimal investment strategy, the model performs well and reliable for the variables *Home Win/ No Home Win* and *Away Win/ No Away Win*. However, the cutoff values for both strategies are high. The flip side of the high cutoff values and the high result stabilities are the low betting volumes.

Table 4: Summary Results Training Data Set for Random Forest

	Home Win/ Away Win/ Draw	Home Win/ No Home Win	Away Win/ No Away Win
Accuracy Full Model (%)	55.34	64.50	67.06
Accuracy PCA-Reduced Model (%)	57.55	58.96	71.99
Graph Best Performing Model			
Max Cutoff (%)	96	86	82
Max ROI (%)	56.99	16.25	49.99
Betting Volume (%)	0.5	8	2.5

Lastly, the gradient boosting algorithm will be applied. "Gradient boosting step-wise builds a decision tree to minimalise a defined loss function in order to correctly classify/assign probabilities to different classes. It can be particularly suitable for unbalanced data sets" (Studer, 2019; James et al., 2013). However, as one can see in Table 5, also the gradient boosting accuracies are significantly lower compared to the results obtained with LDA and logistic regression.

Table 5: Summary Results Training Data Set for Gradient Boosting

	Home Win/ Away Win/ Draw	Home Win/ No Home Win	Away Win/ No Away Win
Accuracy Full Model (%)	49.95	58.98	71.12
Accuracy PCA-Reduced Model (%)	53.15	56.51	73.12
Graph Best Performing Model (Full or PCA-Reduced)			
Max Cutoff (%)	99	98	-
Max ROI (%)	20.50	29.30	-

Betting Volume (%)	1	1.5	-

Overall, the results show that LDA and logistic regression lead to the highest accuracy scores. However, this does not mean these techniques are automatically the best foundations for the most rewarding betting strategy. The main criteria to define the best betting strategy are the total winnings, the betting amount multiplied with the ROI, of the profit optimised cutoff value. Moreover, it is important to check the graph to see whether the results above the given cutoff threshold are strictly positive and therefore stable. To get more concrete, it can be seen that for the prediction of *Home Win/ Away Win/ Draw,* logistic regression with a probability threshold equal or above 0.84 leads to the highest winnings. Additionally, the strategy is also stable as the winnings never turn negative above this threshold. Therefore, *Home Win/ Away Win/ Draw* will be predicted with logistic regression and the corresponding 0.84 cutoff value on the validation data set. For the target variable *Home Win/ No Home Win* random forest with a probability threshold of 0.86 and above promises the most profitable and stable strategy. Lastly, *Away Win/ No Away Win* will be predicted with random forest as well. There, the threshold of 0.82 and above will be used. As for all three target variables a profitable and stable betting strategy could be found, the final out of sample strategy will combine all three different approaches.

4.2 Validation Data

The validation results for the three different strategies as well as for the overall strategy are displayed in Table 6. There are three major findings. First of all, each sub-strategy results in a positive ROI. Secondly, by combining the three strategies, the achieved ROI of 38.38% is highly promising. Thirdly, the betting volume for the overall strategy is with 8%, out of all possible games, fairly high and therefore indicates stable and reliable results.

Table 6: Results Validation Data

	Machine Learning Technique	**Cutoff Threshold (%)**	**Betting Volume (%)**	**ROI (%)**
Home Win/ Away Win/ Draw	Logistic Regression	0.84	3	17.33
Home Win/ No Home Win	Random Forest	0.86	3	9.66
Away Win/ No Away Win	Random Forest	0.82	2	113.00
Overall Strategy	mixed	mixed	8	38.38

5 Discussion

As a first step, the obtained results should be put into perspective and hence be compared to other forecasting approaches. Overall, the obtained accuracy of 58% for *the home win/ draw/ away win* target variable is higher than or similar to the achieved accuracies by other researchers such as Hucaljuk & Rakipović (2011) or Tax et al. (2015). However, accuracy is highly dependent on the particular data set. Therefore, it makes sense to compare the results to benchmarks for the same data set. Thereby, the random benchmark of 49% was outperformed. With regard to betting companies, it has to be acknowledged that the bookmaker Bet365 has with an accuracy of 59.5% still a slightly higher predictive power for the *home win/ draw/ away win* target variable compared to the best performing model of this paper. However, this is not surprising. The bookmakers have more data available with regard to features per game, the time frame of past data as well as the experience and human capital knowledge to develop superior models. In addition to that, the betting companies offer odds which do not add up, the companies introduce a margin, in order to establish a further edge and guarantee their profitability. Having a look at the financial reports of Bet365 over recent years confirms the common belief that overall, the bookmaker always wins (Ahmed & Bounds, 2018).

However, there are two distinct advantages that allow bettors to create successful betting strategies despite the lower resources and the disfavorable odds. Contrary to the bookmakers, the bettor does not have to act in every match but can focus on events where the certainty of his model is high. This lowers the variance and allows focusing on events that the bettor's model understands best. The betting company does not have this advantage as it is expected to make the market and assign probabilities to every possible outcome. Secondly, the bettor can lower the complexity of his models and gain a different perspective by grouping outcomes as done in this paper with the target variables *home win/ no home win* and *away win/ no away win*. This again introduces flexibility the bookmakers do not have.

As indicated by the above results, these two advantages combined with modern statistical, optimisation and machine learning techniques as well as creative feature engineering still allows bettors to create a profitable, public data-based strategy in today's football betting market.

6 Conclusion

6.1 Results

The present paper modeled a machine learning-driven sports betting strategy for the English Premier League season 2018/19 by using team characteristics input features as well as time-series performance data in combination with dimensionality reduction and optimisation techniques. The final validated strategy achieved a return of investment of 38.38 % by betting on 8% of all possible games.

6.2 Limitations

As every statistical model, also this study does not come without any limitations. The major drawback of this paper's model is the small sample size of the different data sets. Even though it's comparable to other scientific research and a lot of attention was paid to reducing variance with statistical methods such as cross-validation and rolling windows, the betting volume of the final strategy is still exposed to the variance of a few events in the validation data set. Therefore, to prove the absolute validity and sustainability of the strategy over multiple seasons and leagues, the model would need to be applied to many other leagues and seasons. However, this would go beyond the goal and scope of this paper as it would be connected to a significant increase in data sourcing and data manipulation work.

With regard to the implementation of the betting strategy, it is important to notice that there are betting companies which simply exclude strong winning players from their service (Economist, 2017). Consequently, the opportunity to apply the developed strategy on such platforms is limited. Hence, alternatives such as implementing the strategy on betting exchange sites where the bet is matched with a counterposition of another bettor have to be considered. An example of such a platform would be betfair.com.

6.3 Outlook

There are three major suggestions on how the study can be enhanced to increase its value further. First of all, an optimisation function can not only be applied to evaluate the optimal probability cutoff in order to maximalise the wins but also to optimise the gap between betting odds and modeled probabilities. However, this would increase the model's and its computational complexity significantly. Moreover, it would also premise superior prediction accuracy in at least some identifiable cases, which directly leads to the second point.

As the publicly available data and its forecasting power are limited, there is a lot of room to create additional value by enhancing the developed modeling approach with data that is not publicly available. In order to create an even larger edge in the betting market, detailed team performance data, as well as player individual match data from companies such as OPTA, could be considered. With that, modeling will get much more complex but the potential rewards in the betting market will increase as well, as shown by the recent examples of Tony Bloom who bought with the winnings of his quant-based sports betting company the Premier League football club Brighton & Hove Albion (Biermann, 2019).

Lastly, moving away from the betting market, there is a lot of value for the football coaching staff by better understanding the relationship between detailed team-based features as well as player individual data and game outcomes. Examples of these input features are ball circulation speed, passing accuracy, room covering, individual player movement or defender proximity to opponents. Understanding the relationship of these features with game outcomes can help to develop better game plans, squad planning and transfer policies. For all these topics, this paper's model can be insightful by having a more detailed look at the feature importance of the random forest and the composition of the different principal components. However, once more, there would be a need for alternative, private data sources in order to get these insights. The Danish football champion of the season 2017/18 and former underdog FC Midtiylland already showed the power of analytics-based transfer strategies and game plan design (Biermann, 2019). In the future, with the ever-increasing amount of data, the most effective navigation through the flood of newly available information and its subsequent implantation onto the football pitch and into the football ball clubs board rooms', will likely be decisive tools for success in the increasingly competitive football world.

References

Ahmed, M. and Bounds, A. (2018). Bet365 stands out among rivals for more than pay. *FT* [online]. Available at: https://www.ft.com/content/f40e1b44-ee60-11e8-89c8-d36339d835c0 [Accessed 29 August 2019]

Barnard, M., Boor, S., Winn, C., Wood, C., and Wray, I. (2019). World in motion – Annual review of football finance. *Deloitte* [online]. Available at: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-annual-review-of-football-finance-2019.pdf [Accessed 19 August 2019]

Biermann, C. (2019). *Football hackers: The science and art of data revolution*. Blink: London.

Burn-Murdoch, J. (2018). How data analysis helps football clubs make better signings. *FT* [online]. Available at: ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8[Accessed 19 August 2019]

Conn, D. (2018). How data analysis helps football clubs make better signings. *FT* [online]. Available at: ft.com/content/84aa8b5e-c1a9-11e8-84cd-9e601db069b8 [Accessed 19 August 2019]

Constantinou, A.C. and Fenton, N.E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, *9*(1), pp.37-50. doi: https://doi.org/10.1515/jqas-2012-0036

Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society*, *46*(2), pp.265-280. doi: https://doi.org/10.1111/1467-9876.00065

Economist (2017). *How bookmakers deal with winning customers* [online]. Available at: https://www.economist.com/the-economist-explains/2017/10/04/how-bookmakers-deal-with-winning-customers [Accessed 29 August 2019]

Elo, A.E., 1978. *The rating of chessplayers, past and present*. Arco Pub.

Fifindex.com (2019). Available at: https://www.fifaindex.com/ [Accessed 10 August 2019]

Football-data.co.uk (2019). *Data files: England*. [online]. Available at: https://www.football-data.co.uk/englandm.php [Accessed 10 August 2019]

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, *21*(2), pp.331-340. doi: https://doi.org/10.1016/j.ijforecast.2004.08.002

Hill, I.D. (1974). Association football and statistical inference. *Journal of the Royal Statistical Society, 23*(2), pp.203-208. doi: https://doi.org/10.2307/2347001

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques [online]. In: *2011 Proceedings of the 34th International Convention MIPRO*. New Jersey*:* IEEE, pp. 1623-1627. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967321 [Accessed 19 August 2019]

Impect (2019). *Impect – Making success in soccer measurable* [online]. Available at: https://www.impect.com/en/ [Accessed 19 August 2019]

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Kürüm, E., Yildirak, K. and Weber, G.W. (2012). A classification problem of credit risk rating investigated and solved by optimisation of the ROC curve. *Central European Journal of Operations Research*, *20*(3), pp.529-557. doi: https://doi.org/10.1007/s10100-011-0224-5

Lasek, J. (2016). EURO 2016: Predictions using team rating systems. In: *MLSA* [online]. Dublin: PKDD/ECML. Available at: https://pdfs.semanticscholar.org/378d/f9520dec6fc2792adf377b47baa5009b495c.pdf [Accessed 19 August 2019]

Lucey, P., Bialkowski, A., Monfort, M., Carr, P. and Matthews, I. (2014). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In: *Proc. 8th annual MIT Sloan sports analytics conference* [online]. Cambridge: SSAC, pp. 1-9. Available at: http://www.sloansportsconference.com/wp-content/uploads/2015/02/SSAC15-RP-Finalist-Quality-vs-Quantity.pdf [Accessed 29 August 2019]

Statista (2019a). *Total collection from sport betting at an international level for season 2014/15, by league (in billion euros)* [online]. Available at: https://www.statista.com/statistics/620308/collection-from-betting-on-international-football-europe/ [Accessed 19 August 2019]

Statista (2019b). *Average annual player salary in the English Premier League in 2018/19, by team (in million U.S dollars)* [online]. Available at: https://www.statista.com/statistics/675303/average-epl-salary-by-team/ [Accessed 10 August 2019]

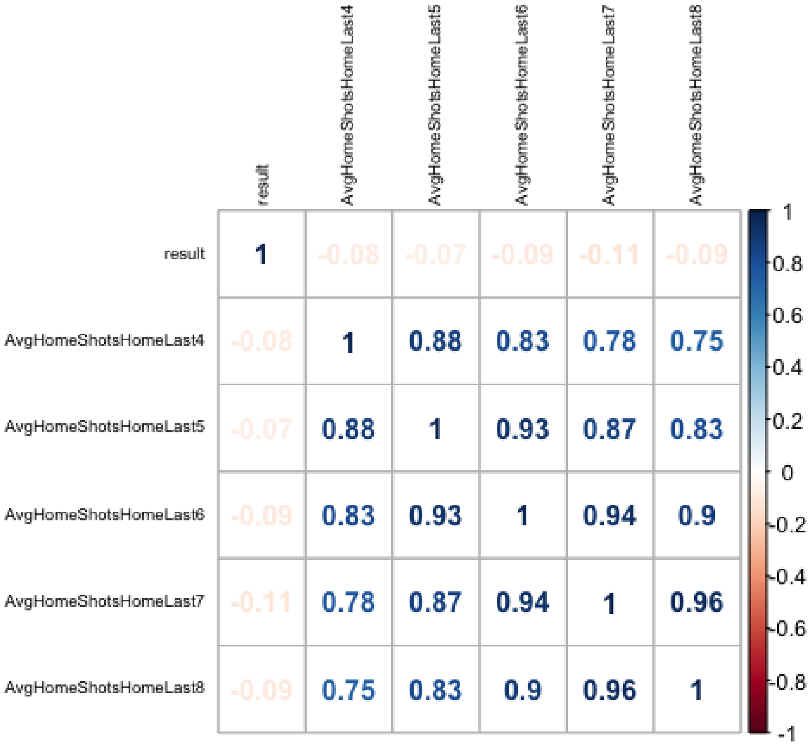
Studer, C. (2019). *The 10% annual ROI in the peer to peer lending market in 2018? – A machine learning-driven investment strategy*. MSc. Imperial College London.

Tax, N. and Joustra, Y., 2015. Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering, 10*(10), pp.1-13.

Transfermarkt.co.uk (2019). Clubs Premier League, 18/19 *[online]*. Available at: https://www.transfermarkt.co.uk/premier-league/startseite/wettbewerb/GB1/saison_id/2018/plus/[Accessed 10 August, 2019]

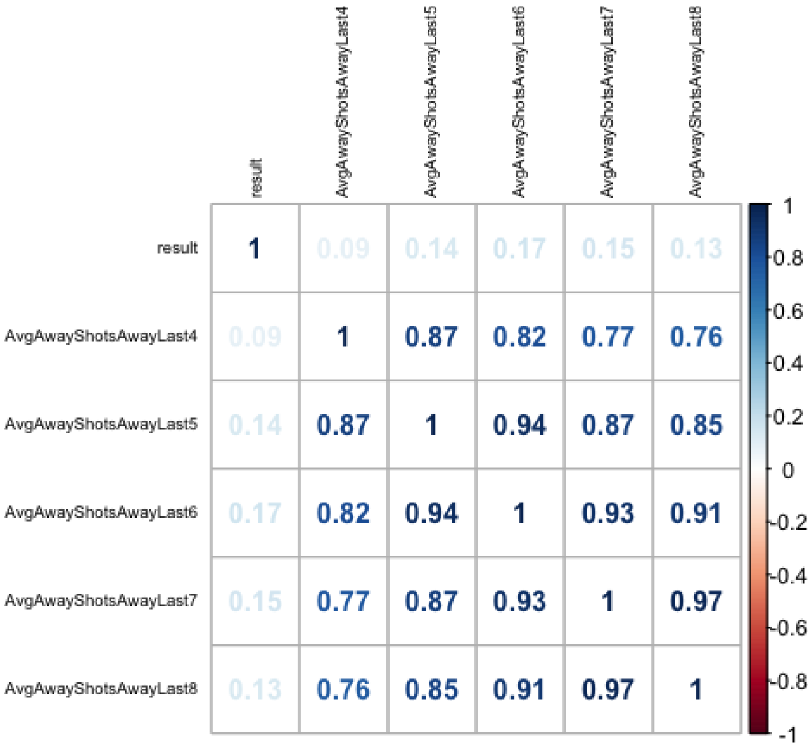
Appendix

Appendix 1: Correlation Table Result and Different Rolling Windows for Average Number of Shots Taken by the Home Team in Home Games



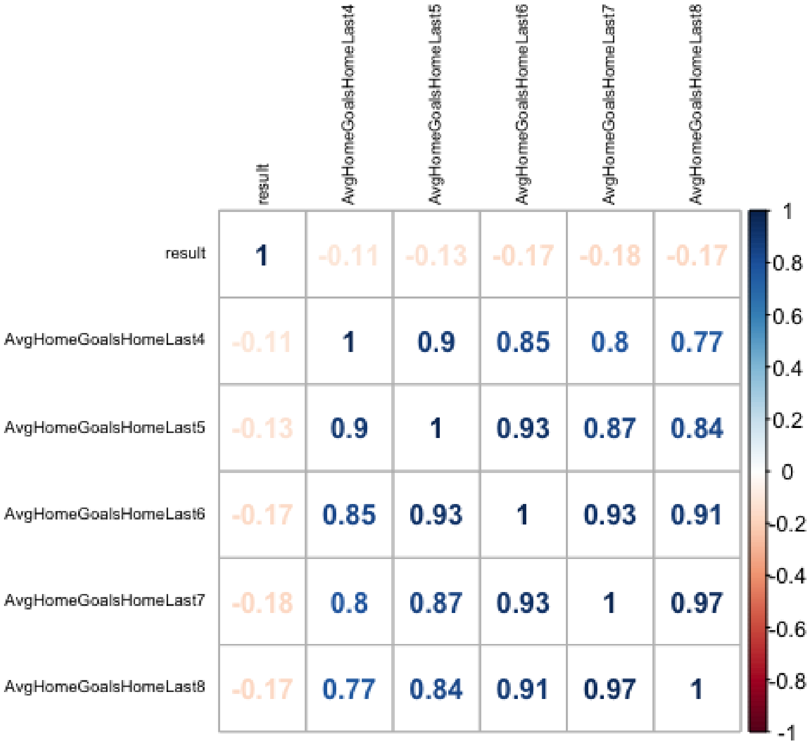
	result	AvgHomeShotsHomeLast4	AvgHomeShotsHomeLast5	AvgHomeShotsHomeLast6	AvgHomeShotsHomeLast7	AvgHomeShotsHomeLast8
result	1	-0.08	-0.07	-0.09	-0.11	-0.09
AvgHomeShotsHomeLast4	-0.08	1	0.88	0.83	0.78	0.75
AvgHomeShotsHomeLast5	-0.07	0.88	1	0.93	0.87	0.83
AvgHomeShotsHomeLast6	-0.09	0.83	0.93	1	0.94	0.9
AvgHomeShotsHomeLast7	-0.11	0.78	0.87	0.94	1	0.96
AvgHomeShotsHomeLast8	-0.09	0.75	0.83	0.9	0.96	1

Appendix 2: Correlation Table Result and Different Rolling Windows for Average Number of Shots Taken by the Away Team in Away Games



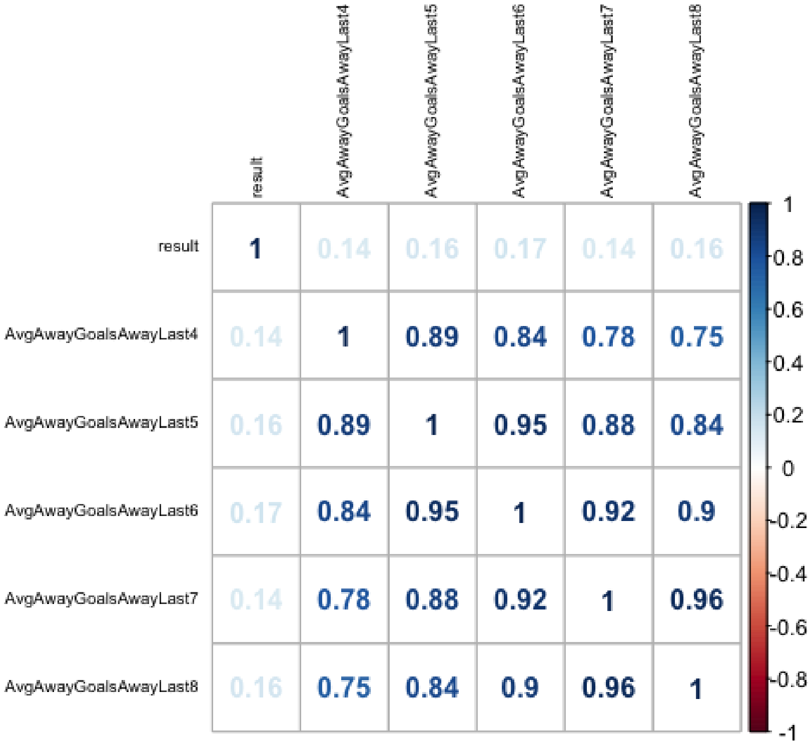
	result	AvgAwayShotsAwayLast4	AvgAwayShotsAwayLast5	AvgAwayShotsAwayLast6	AvgAwayShotsAwayLast7	AvgAwayShotsAwayLast8
result	1	0.09	0.14	0.17	0.15	0.13
AvgAwayShotsAwayLast4	0.09	1	0.87	0.82	0.77	0.76
AvgAwayShotsAwayLast5	0.14	0.87	1	0.94	0.87	0.85
AvgAwayShotsAwayLast6	0.17	0.82	0.94	1	0.93	0.91
AvgAwayShotsAwayLast7	0.15	0.77	0.87	0.93	1	0.97
AvgAwayShotsAwayLast8	0.13	0.76	0.85	0.91	0.97	1

Appendix 3: Correlation Table Result and Different Rolling Windows for Average Number of Goals Scored by the Home Team at Home



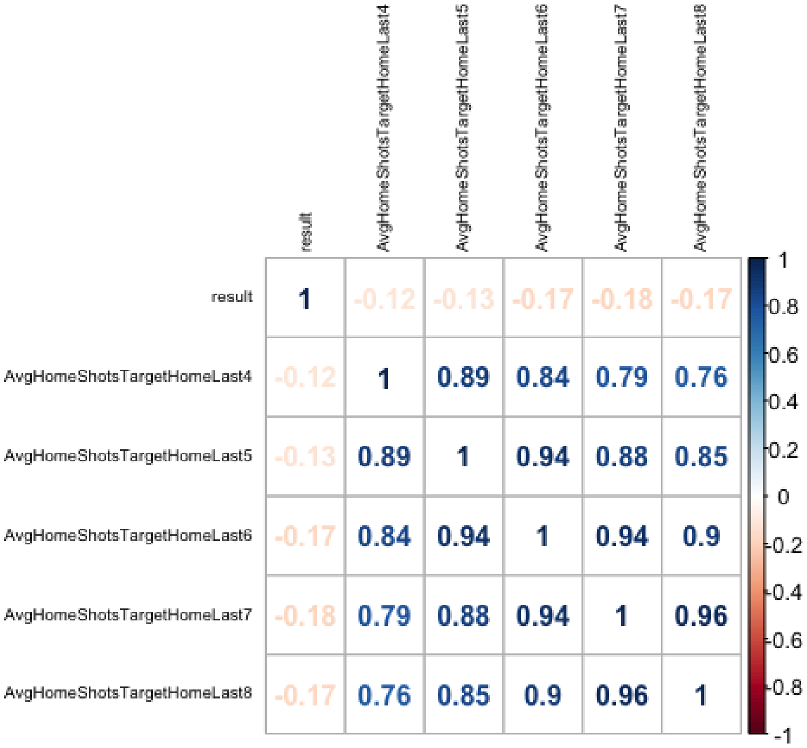
	result	AvgHomeGoalsHomeLast4	AvgHomeGoalsHomeLast5	AvgHomeGoalsHomeLast6	AvgHomeGoalsHomeLast7	AvgHomeGoalsHomeLast8
result	1	-0.11	-0.13	-0.17	-0.18	-0.17
AvgHomeGoalsHomeLast4	-0.11	1	0.9	0.85	0.8	0.77
AvgHomeGoalsHomeLast5	-0.13	0.9	1	0.93	0.87	0.84
AvgHomeGoalsHomeLast6	-0.17	0.85	0.93	1	0.93	0.91
AvgHomeGoalsHomeLast7	-0.18	0.8	0.87	0.93	1	0.97
AvgHomeGoalsHomeLast8	-0.17	0.77	0.84	0.91	0.97	1

Appendix 4: Correlation Table Result and Different Rolling Windows for Average Number of Goals Scored by the Away Team in Away Games



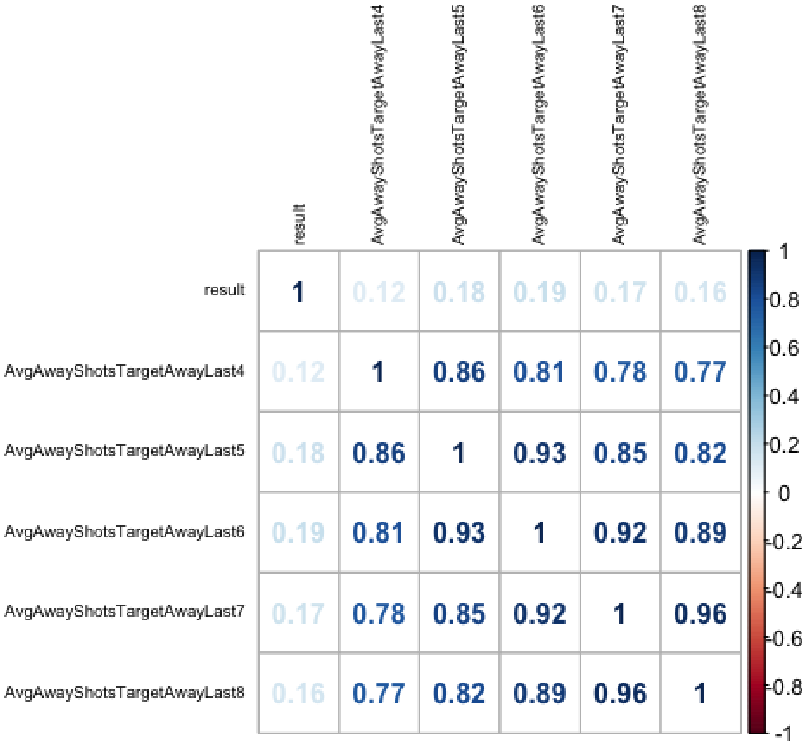
	result	AvgAwayGoalsAwayLast4	AvgAwayGoalsAwayLast5	AvgAwayGoalsAwayLast6	AvgAwayGoalsAwayLast7	AvgAwayGoalsAwayLast8
result	1	0.14	0.16	0.17	0.14	0.16
AvgAwayGoalsAwayLast4	0.14	1	0.89	0.84	0.78	0.75
AvgAwayGoalsAwayLast5	0.16	0.89	1	0.95	0.88	0.84
AvgAwayGoalsAwayLast6	0.17	0.84	0.95	1	0.92	0.9
AvgAwayGoalsAwayLast7	0.14	0.78	0.88	0.92	1	0.96
AvgAwayGoalsAwayLast8	0.16	0.75	0.84	0.9	0.96	1

Appendix 5: Correlation Table Result and Different Rolling Windows for Average Number of Shots on Target by the Home Team at Home



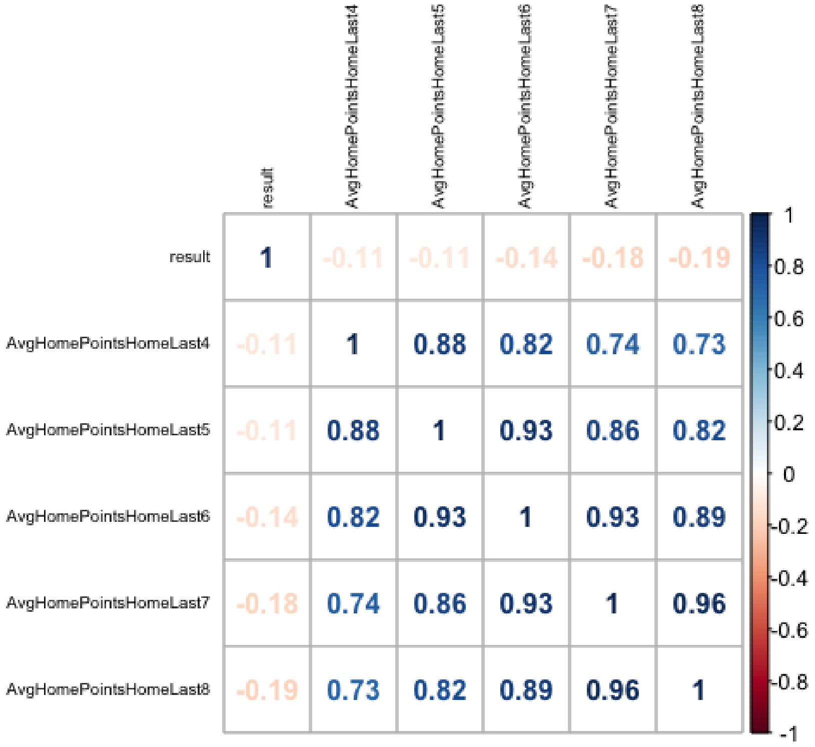
	result	AvgHomeShotsTargetHomeLast4	AvgHomeShotsTargetHomeLast5	AvgHomeShotsTargetHomeLast6	AvgHomeShotsTargetHomeLast7	AvgHomeShotsTargetHomeLast8
result	1	-0.12	-0.13	-0.17	-0.18	-0.17
AvgHomeShotsTargetHomeLast4	-0.12	1	0.89	0.84	0.79	0.76
AvgHomeShotsTargetHomeLast5	-0.13	0.89	1	0.94	0.88	0.85
AvgHomeShotsTargetHomeLast6	-0.17	0.84	0.94	1	0.94	0.9
AvgHomeShotsTargetHomeLast7	-0.18	0.79	0.88	0.94	1	0.96
AvgHomeShotsTargetHomeLast8	-0.17	0.76	0.85	0.9	0.96	1

Appendix 6: Correlation Table Result and Different Rolling Windows for Average Shots on Target by the Away Team in Away Games



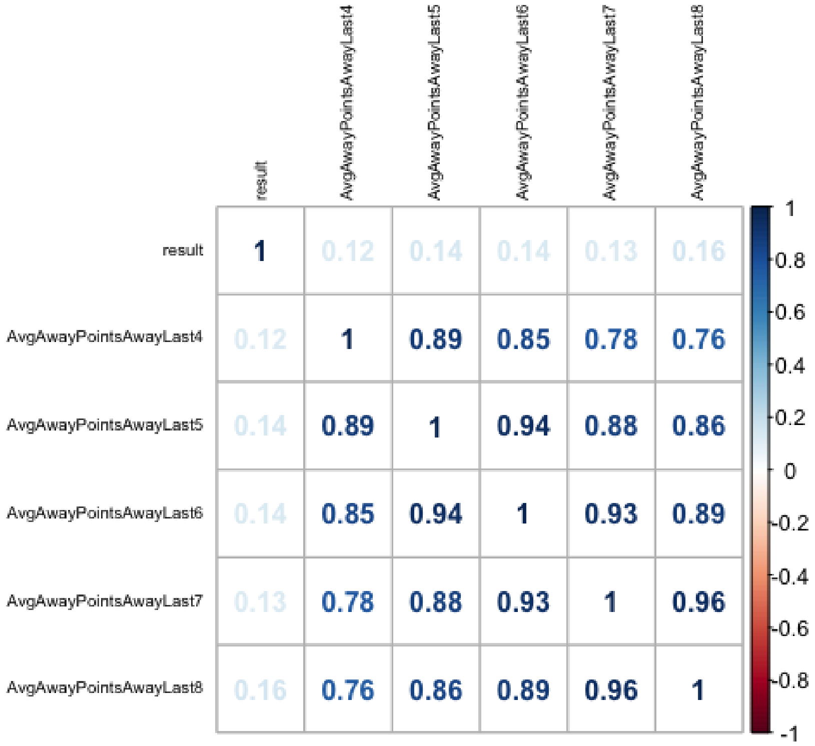
	result	AvgAwayShotsTargetAwayLast4	AvgAwayShotsTargetAwayLast5	AvgAwayShotsTargetAwayLast6	AvgAwayShotsTargetAwayLast7	AvgAwayShotsTargetAwayLast8
result	1	0.12	0.18	0.19	0.17	0.16
AvgAwayShotsTargetAwayLast4	0.12	1	0.86	0.81	0.78	0.77
AvgAwayShotsTargetAwayLast5	0.18	0.86	1	0.93	0.85	0.82
AvgAwayShotsTargetAwayLast6	0.19	0.81	0.93	1	0.92	0.89
AvgAwayShotsTargetAwayLast7	0.17	0.78	0.85	0.92	1	0.96
AvgAwayShotsTargetAwayLast8	0.16	0.77	0.82	0.89	0.96	1

Appendix 7: Correlation Table Result and Different Rolling Windows for Average Number of Points in by the Home Team at Home



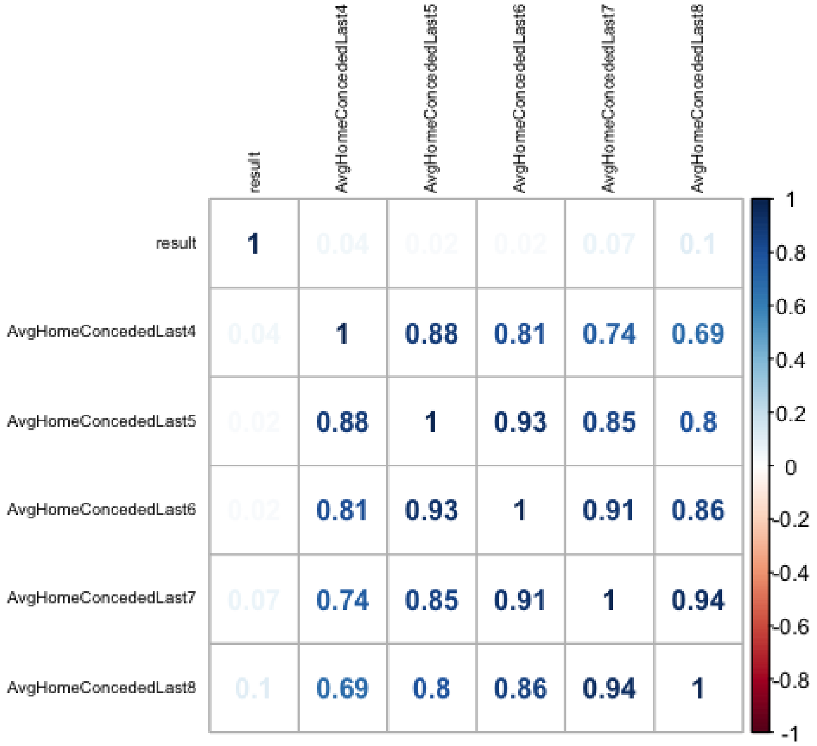
	result	AvgHomePointsHomeLast4	AvgHomePointsHomeLast5	AvgHomePointsHomeLast6	AvgHomePointsHomeLast7	AvgHomePointsHomeLast8
result	1	-0.11	-0.11	-0.14	-0.18	-0.19
AvgHomePointsHomeLast4	-0.11	1	0.88	0.82	0.74	0.73
AvgHomePointsHomeLast5	-0.11	0.88	1	0.93	0.86	0.82
AvgHomePointsHomeLast6	-0.14	0.82	0.93	1	0.93	0.89
AvgHomePointsHomeLast7	-0.18	0.74	0.86	0.93	1	0.96
AvgHomePointsHomeLast8	-0.19	0.73	0.82	0.89	0.96	1

Appendix 8: Correlation Table Result and Different Rolling Windows for Average Number of Away Points by the Away Team in Away Games



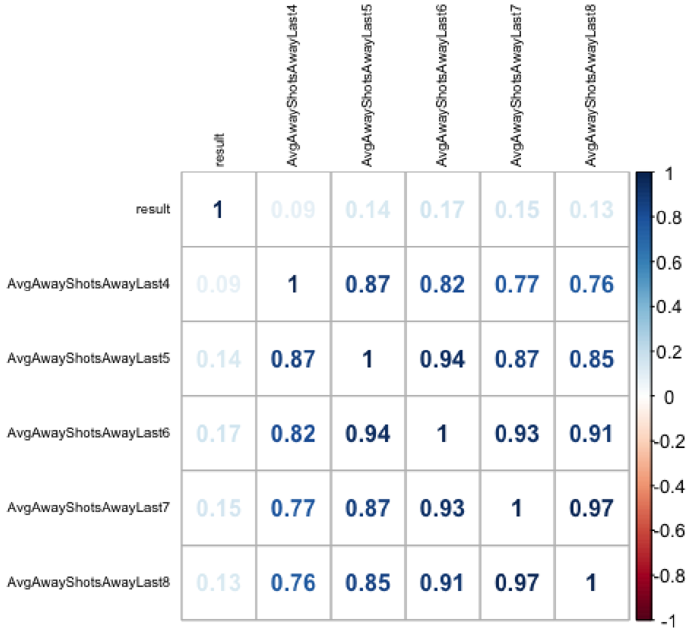
	result	AvgAwayPointsAwayLast4	AvgAwayPointsAwayLast5	AvgAwayPointsAwayLast6	AvgAwayPointsAwayLast7	AvgAwayPointsAwayLast8
result	1	0.12	0.14	0.14	0.13	0.16
AvgAwayPointsAwayLast4	0.12	1	0.89	0.85	0.78	0.76
AvgAwayPointsAwayLast5	0.14	0.89	1	0.94	0.88	0.86
AvgAwayPointsAwayLast6	0.14	0.85	0.94	1	0.93	0.89
AvgAwayPointsAwayLast7	0.13	0.78	0.88	0.93	1	0.96
AvgAwayPointsAwayLast8	0.16	0.76	0.86	0.89	0.96	1

Appendix 9: Correlation Table Result and Different Rolling Windows for Average Number of Conceded Goals by the Home Team



	result	AvgHomeConcededLast4	AvgHomeConcededLast5	AvgHomeConcededLast6	AvgHomeConcededLast7	AvgHomeConcededLast8
result	1	0.04	0.02	0.02	0.07	0.1
AvgHomeConcededLast4	0.04	1	0.88	0.81	0.74	0.69
AvgHomeConcededLast5	0.02	0.88	1	0.93	0.85	0.8
AvgHomeConcededLast6	0.02	0.81	0.93	1	0.91	0.86
AvgHomeConcededLast7	0.07	0.74	0.85	0.91	1	0.94
AvgHomeConcededLast8	0.1	0.69	0.8	0.86	0.94	1

Appendix 10: Correlation Table Result and Different Rolling Windows for Average Number of Conceded Goals by the Away Team



	result	AvgAwayShotsAwayLast4	AvgAwayShotsAwayLast5	AvgAwayShotsAwayLast6	AvgAwayShotsAwayLast7	AvgAwayShotsAwayLast8
result	1	0.09	0.14	0.17	0.15	0.13
AvgAwayShotsAwayLast4	0.09	1	0.87	0.82	0.77	0.76
AvgAwayShotsAwayLast5	0.14	0.87	1	0.94	0.87	0.85
AvgAwayShotsAwayLast6	0.17	0.82	0.94	1	0.93	0.91
AvgAwayShotsAwayLast7	0.15	0.77	0.87	0.93	1	0.97
AvgAwayShotsAwayLast8	0.13	0.76	0.85	0.91	0.97	1