Bitlis Eren University Journal of Science
BİTLİS EREN ÜNİVERSİTESİ FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 11 NO: 1 PAGE: 227-236 YEAR: 2022
DOI: 10.17798/bitlisfen.1015215

Factors Associated with Match Result and Number of Goals Scored and Conceded in the English Premier League

Günal BİLEK[1*], Betül AYGÜN[2]

[1]*Department of Business Administration, Izmir Democracy University, Izmir, Turkey*
[2]*Management Information Systems, İzmir Democracy University, Izmir, Turkey*
(ORCID: 0000-0001-6417-7129) (ORCID: 0000-0001-9610-9235)

Keywords: Support vector machine, Multinomial logistic regression, Poisson regression, Machine learning, Performance analysis, Football.

Abstract
The aim of this research is to identify the factors associated with the match result and the number of goals scored and conceded in the English Premier League. The data consist of 17 performance indicators and situational variables of the football matches in the English Premier League for the season of 2017-18. Poisson regression model was implemented to identify the significant factors in the number of goals scored and conceded, while multinomial logistic regression and support vector machine methods were used to determine the influential factors on the match result. It was found that scoring first, shots on target and goals conceded have significant influence on the number of goals scored, whereas scoring first, match location, quality of opponent, goals conceded, shots and clearances are influential on the number of goals conceded. On the other hand, scoring first, match location, shots, shot on target, clearances and quality of opponent significantly affect the probability of losing; while scoring first, match location, shots, shots on target and possession affect the probability of winning. In addition, among all the variables studied, scoring first is the only variable appearing important in all the analyses, making it the most significant factor for success in football.

1. Introduction

Football is the most popular sport in the world and its economy is worth billions of dollars. Also, rapid advances in technology make it easier to collect football data. As a result of these two, there has been a significant increase in the number of studies aiming to model the outcomes of the matches [1]. Modeling match results and effective parameter selections are not only popular in football, but also in other sport sciences. For instance, [2] combined three factors by adopting contribution parameters to simulate outcomes of the future games in Major League Baseball, [3] investigated the parameters that affect the National Basketball Association team values and [4] presented a new hybrid model, based on the definition of the Poisson distribution, to predict ice hockey match results. However, the sport field where the prediction of the match result is most frequently studied is football.

There are two different empirical literatures on modeling the results of matches in football. The first approach targets to model the match results (win, draw, and lose) directly, while the latter aims to model the number of scored and conceded goals [5]. This study intends to contribute to both of them. Therefore, this study is divided into two parts. The first part aims to investigate the performance indicators and situational variables affecting the number of scored and conceded goals with Poisson regression model. Although this paper is not the first one aiming to model scored and conceded goals, it is the only one (to our knowledge) aiming to investigate which variables affect the number of goals to what extent. The studies aiming to predict the number of goals [6]–[9] mainly focused on the statistical modelling rather than performance analysis. With this research, we aim to fill this gap by identifying the factors significantly associated with the number of goals conceded and scored and these variables' size of

*Corresponding author: gunalbilek@gmail.com
Received: 20.10.2021, Accepted: 10.02.2022

effect on goals. The second part of this study targets to detect the variables directly affecting the match results with multinomial logistic regression and support vector machine methods. The aim of using two different approaches for the same reason is to compare the performances of the models and investigate if the same variables have significant effect on the match result in the two models.

Variable selection is the first and the most important step of predicting the match result and the number of goals scored and conceded because there are many variables measured in a match and it is not possible or plausible to include all of them in the analyses. Therefore, it is wise to choose the ones that have the potential to be significant and this is done based on the previous studies. After reviewing the current literature, we ended up with the variables of scoring first [10]–[12], match location [10]–[18], quality of opponent [11], [12], [19], clearances [12], [20]–[22] corners, passes [21], [23], previous match result [24], goals scored per game [25], goals conceded per game [26], ball possession [27]–[31], shots, shots on target [14], [31]–[33], tackles [20], [22] and touches[34].

2. Material and Method

2.1. Data

In this study, the data consist of 17 variables, some of which performance indicators and some situational variables, of football teams in the English Premier League (EPL) in the season of 2017-18. Data of each team are analysed individually. Therefore, each match corresponds to two different observations, one for home team and one for away. Since there are 20 teams in the EPL, 380 games are played every season, leading to 760 observations. As some variables require information from the previous week and this is not possible for the first week, its observations are removed from the data set, which leaves a total of 740 observations to analyse. The variables used in this research and their definitions are as follows:

- Result ©: Result of the match – *win*, *draw* or *lose*.
- Goals scored (*GS*): Number of goals the team scored in the match.
- Goals conceded (*GC*): Number of goals the opposing team scored.
- Match location (*ML*): Where the team played the game – *home* or *away*
- Scoring first (*SF*): A dummy variable indicating whether or not the team scored first – *yes* (1) or *no* (0)
- Quality of opponent (*QO*): A metric showing the quality of the opposing team calculated by the difference between the rankings of the teams, that is, $R_1 - R_2$, where R_1 and R_2 are the rankings of the first-named team and its opponent in the EPL in the current week, respectively. The larger the QO, the stronger the opponent is.
- Goals for per game (*GFPG*): Average number of goals scored by the team per game.
- Goals against per game (*GAPG*): Average number of goals scored against the team per game.
- Shots (*S*): Total number of shots of the team in the match.
- Shots on target (*ST*): Total number of shots on target of the team in the match.
- Clearances ©: Total number of clearances of the team in the match.
- Corners (*Co*): Total number of corners of the team in the match.
- Passes (*P*): Total number of passes completed by the team in the match.
- Possession percentage (*PP*): Percentage of time in which the team possesses the ball in the match.
- Previous result (*PR*): Result of the team's previous match; *win*, *draw*, or *lose*.
- Tackles (*T*): Total number of tackles of the team in the match.
- Touches (*To*): Total number of touches of the team in the match.

The data were obtained from the official website of www.premierleague.com which retrieves data from OPTA whose data reliability range from 0.92 to 0.94 [35].

2.2. Statistical Analysis

Poisson regression (PR), multinomial logistic regression (MLR) and support vector machine (SVM) methods are applied for the feature extraction and predictive analysis. During this study, firstly the factors which significantly affect the match outcome and scored and conceded goal numbers are extracted and prediction models are applied to predict whether the match outcome is win, draw or lose. In this section, used methods for the analysis are detailed and, finally, the predictive classification performance metrics are described. For this analysis, Python *statsmodel* 0.12.0 library is used.

2.2.1. Poisson Regression (PR)

PR is a member of the generalised linear model family which can provide precise results for data sets with count, binary, ordinal and time-to-success dependent variables [36]. As this paper aims to model the number of scored and conceded goals, which are count variables, PR is used. The results are reported with coefficients, odds ratios and corresponding *p* values. This model is also used to predict the match results as draw, win or lose by considering the differences between the predicted number of goals scored and conceded by the teams. The prediction performance of the match results is statistically summarized and visualized with the heat map.

2.2.2 Support Vector Machine (SVM)

SVM is a type of a supervised machine learning algorithms for classification and feature extraction. Since SVM has advantages in dealing with high dimensional problems and solving small sample datasets as in this study, it is expected to yield better and meaningful results [37], [38]. There are three main parameters that must be optimized for the SVM algorithm: kernel function, regularization, and gamma values. Kernel function is used to transform input data into the required form. The study [39] claimed that, polynomial SVM kernel and tangential kernel performs poorer than radial and linear kernels for the groups coming from Poisson distribution [39]. Therefore, linear kernel is selected as kernel parameter. Besides, for the C and gamma parameters tuning, grid search view is applied; while for the implementation of SVC, Python Scikit Learn 0.23.2 library is used [40].

2.2.3. Multinomial Logistic Regression (MLR)

Since the number of the possible outcomes of the matches (win, lose and draw) is greater than two: the multinomial logistic regression approach is used to model the match results. The accuracy of the match results predicted is visualized with the heat map. Besides, regression coefficients, *p* values and odds ratios are evaluated to discuss the effects of the variables on the relationship with the dependent variable [41].

2.3. Predictive Performance Metrics of Classification

The match outcomes are predicted by using the defined methods. The dataset is divided into training and testing data sets with a ratio of 70:30. To evaluate the accuracy of the models, confusion matrix is used which summarizes the correctly and incorrectly classified outcomes. Accordingly, statistical metrics as precision, recall, and F1-score are evaluated to compare the prediction performances of the algorithms. High precision shows that more relevant results are returned than the irrelevant ones within predicted values by the algorithm and high recall means that of the relevant results are returned among actual results [42]. On the other hand, F1-score is evaluated to see the balance between the precision and recall. It provides more confidence result than accuracy for the dataset having unequal class distribution.

3. Results

First, which factors affect the number of goals scored and conceded and match result and how they affect them are discussed in the first two subsections. Second, the prediction performances of these algorithms are detailed.

3.1. Factors Associated with Goals Scored and Conceded

Table 1 presents Poisson regression results showing coefficients, odds ratios ($e^{coefficient}$) and p values for situational variables and performance indicators on the number of goals scored and conceded. Since the match result is determined by the number of goals conceded and

scored, it is not included in the model when analysing the number of goals scored and conceded. The regression coefficients indicate the change in the logarithm odds of the number of goals scored and conceded for a change in the explanatory variable. Additionally, if the sign of a coefficient is positive, this means that this variable positively affects the number of goals scored or conceded. Based on that, it is noted that GC, SF and ST have positive significant effects on the number of goals scored. To illustrate, if a team concedes a goal, that team is expected to increase the number of goals scored by 8.84% $((1.0884 - 1) \times 100)$. Similarly, each shot on target rises the number of goals scored by 18.05%. Additionally, scoring-first teams are likely to score 110.52% goals more. The remaining variables have no significant effect on the number of goals scored.

Table 1. Poisson regression results showing regression coefficients, odds ratios (ORs) and p values for situational variables and performance indicators on goals scored and conceded.

Goals scored				Goals conceded			
Variable	**Coef**	**OR**	**p**	**Variable**	**Coef**	**OR**	**p**
Intercept	-1.0294	0.3572	0.0035**	***Intercept***	2.1604	8.6750	<0.001***
GC	0.0847	1.0884	0.0154*	***GS***	0.0607	1.0626	0.024*
SF (yes)	0.7444	2.1052	<0.0001***	***SF (yes)***	-0.5154	0.5973	<0.001***
ML (home)	0.1290	1.1376	0.1060	***ML (home)***	-0.3156	0.7294	<0.001***
QO	-0.0105	0.9896	0.2990	***QO***	0.0162	1.0164	0.0069**
S	-0.0141	0.9860	0.4060	***S***	-0.0232	0.9770	0.0284*
ST	0.1659	1.1805	<0.0001***	***ST***	0.0029	1.0030	0.8874
C	0.0031	1.0031	0.3221	***C***	-0.0264	0.9740	<0.001***
Co	-0.0146	0.9855	0.5585	***Co***	0.0049	1.0049	0.7107
T	0.0062	1.0062	0.1051	***T***	0.0024	1.0024	0.7153
To	-0.0003	0.9996	0.4197	***To***	-0.0004	0.9996	0.6313
P	0.0006	1.0006	0.1572	***P***	-0.0004	0.9996	0.6892
PP	-0.0019	0.9981	0.5528	***PP***	-0.0041	0.9959	0.5505
GAPG	-0.0010	0.9901	0.9980	***GAPG***	-0.1032	0.9019	0.1806
GFPG	0.0219	1.0222	0.5229	***GFPG***	-0.0286	0.9718	0.6866
PR (loss)	-0.0068	0.9932	0.6764	***PR = loss***	0.0855	1.0893	0.2926
PR (win)	0.0405	1.0413	0.2507	***PR = win***	-0.0515	0.9498	0.5666

Coef: regression coefficient; OR: odds ratio; *: significant at the level 0.05; **: significant at the level 0.01; ***: significant at the level 0.001

Continuing with conceded goals, the variables which significantly adversely affect the number of goals conceded are SF, ML (home) and C. To clarify, if a team scores the first goal, that team is expected to concede 40.27% less goals. Furthermore, home teams are likely to concede 27.06% less goals compared to away teams and one unit increase in the number of clearances and shots leads to 2.6% and 2.3% less goals conceded, respectively. In contrast, GS and QO have positive effects on the number of conceded goals. Accordingly, each goal a team scores leads to 6.26% more goals conceded and one unit increase in QO leads to 1.64% more goals conceded. The other variables have no significant effect on the number of goals conceded.

3.2. Feature Extraction of Result, the Dependent Variable

Since the match result is nominal, MLR and SVM are applied to reveal the important features and to predict the match result. As the numbers of goals scored and conceded are direct indicators of winning and losing the match, these are excluded from the dataset while estimating the match result.

Figure 1. shows the contribution of each feature in the SVM classification model. The feature importance values are estimated by taking the square of the coefficient values of the SVM model [43]. According to Figure 1, SF, ML and ST are the most relevant features to the target value.

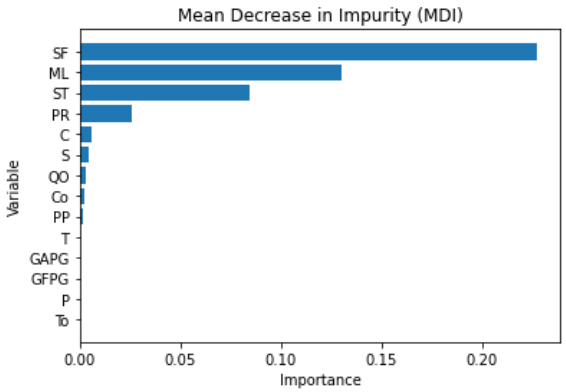


Figure 1: Feature importance of SVM algorithm

MLR results are given for match results win and lose (reference category = draw) in Table 2. Coefficient, odds ratio and *p* value for each feature of lose and win are detailed. First, *p* value is interpreted to measure the evidence whether or not that variable has a significant impact on the match outcome. It is seen that ML, S, ST, SF, QO and C are statistically significant on losing the match. Second, the coefficient value determines whether the event is more likely or less likely when there is a change in the variable. Further on, the sign of the coefficient shows the direction of the relationship between the predictor and the match result. An increase in the values of the variable having positive coefficient increases the probability of occurrence of the event; on the contrary, increase in the negative ones decreases the probability of occurrence of the event. From this point of view, if the match location is home or the team scores the first goal, the probability of losing the match decreases. Correspondingly, increasing in the number of shots, shots on target, and clearances also decreases the probability of losing the match. On the other hand, if the quality of opponent is higher (i.e. a stronger opponent), probability of losing the match raises.

The features that affect the result of win are slightly different from the lose result. There are five factors that significantly affect the likelihood of winning. Scoring first (OR=8.002), significant at level < 0.001, intensely increases the probability of winning the match when it is compared with other significant features. Next, comes ML (home) with an odds ratio of 2.386, which shows that home teams are twice as likely as to win. Additionally, a rise in PP significantly decreases the probability of winning. Contrarily, the coefficients of S and ST are positive, indicating a favourable effect on winning the match.

Table 2. Multinomial regression results showing the regression coefficients, odds ratios and p values for situational variables and performance indicators on result lose and win.

Result=*lose*				**Result=*win***			
Variable	**Coef**	**OR**	**p**	**Variable**	**Coef**	**OR**	**p**
Intercept	4.34	76.707	<0.001***	**Intercept**	-4.26	0.013	<0.001***
ML (home)	-0.861	0.423	<0.001***	***ML (home)***	0.870	2.386	<0.001***
PP	-0.030	0.970	0.161	***PP***	-0.050	0.952	0.038*
ST	-0.123	0.883	0.046*	***ST***	0.335	1.398	<0.001***
S	-0.086	0.917	0.006**	***S***	0.065	1.067	0.047*
To	0.001	1.001	0.778	***To***	-0.001	0.999	0.968
P	0.001	1.001	0.723	***P***	0.007	1.007	0.074
T	0.017	1.017	0.422	***T***	0.008	1.008	0.736
C	-0.085	0.918	<0.001***	***C***	0.024	1.024	0.055
Co	0.041	1.042	0.472	***Co***	-0.028	0.972	0.531
SF (yes)	-0.645	0.524	0.008**	***SF (yes)***	2.080	8.002	<0.001***
GAPG	0.008	1.008	0.973	***GAPG***	0.256	1.292	0.322
GFPG	-0.013	0.987	0.954	***GFPG***	0.242	1.273	0.308
PR (lose)	0.418	1.518	0.106	***PR (lose)***	0.018	1.018	0.952
PR (win)	0.019	1.539	0.944	***PR (win)***	0.431	1.539	0.134
QO	0.036	0.967	0.047*	***QO***	-0.033	0.967	0.108

Coef: regression coefficient; OR: odds ratio; *: significant at the level 0.05; **: significant at the level 0.01; ***: significant at the level 0.001

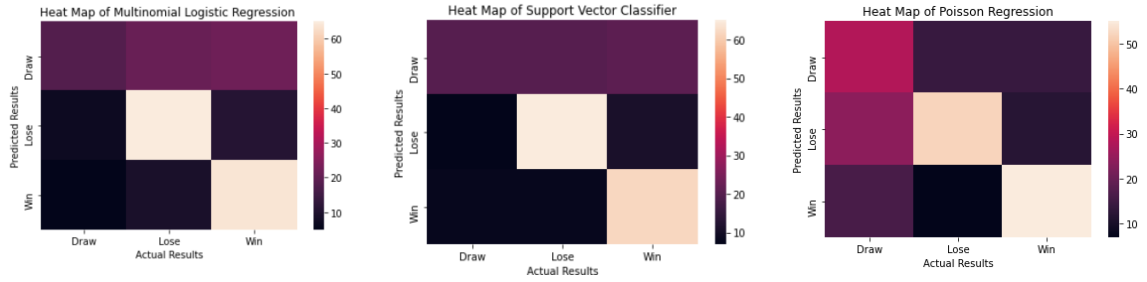
3.3. Predicting match result

Heretofore, the variables which affect the match result and the number of scored and conceded goals have been discussed by using PR, MLR and SVC. In this part of the study, the prediction performances of the algorithms are presented. The dataset is divided into two sets: training and testing datasets. The predictive models are trained on training data set and accuracy and

performance of the models are measured by using the testing data. The statistical measurements that presents the performance for each algorithm is represented in Table 3. It is noted that SVM outperforms MLR and PR.

Figure 2 parts (a), (b) and (c) represent the heat map graphics for the predicton results of MLR, SVM and PR, respectivel. Furthermore, (d), (e) and (f) show the confusio matrix of MLR, SVM and PR, respectiely. MLR and SVM algorithms yield superior results for the win and lose categories compared o PR. However, for the draw results, the accuracy performances of the MLR and SVM are not satisfactory. In contrast, PR yields convincing prediction results for draw.

(a) (b) (c)



(d)

Predicted Results / Actual Results	Draw	Lose	Win
Draw	18	21	22
Lose	7	65	11
Win	5	9	64

(e)

Predicted Results / Actual Results	Draw	Lose	Win
Draw	20	20	21
Lose	7	65	11
Win	8	8	62

(f)

Predicted Results / Actual Results	Draw	Win	Lose
Draw	28	14	14
Lose	24	52	12
Win	16	7	55

Figure 2. Heat maps of significantly related features based on match result.

The heat maps show the prediction results of the algorithms. The number of predictions in that intersection of actual and predicted category increases as the colour becomes lighter. The result of "win" was predicted more accurately in all three algorithms. Since, intersected area where actual and predicted value "win" are the most lighted part of the heat maps. In Figure 2 part (a) and (b), the colour of intersected area where actual and predicted value is "draw"close to darker which shows the prediction matches whose results is draw is not gratifying. On the contrary, even accuracy of the PR (0.61) method is lower than the accuracy of SVM (0.68) and higher than the accuracy of MLR (0.), it predicts the matches whose result is draw more accurately than other two algorithms. In Figure 2(d), low precision of the draw category of the outcome shows that the number of incorrectly classified as win or lose is considerably high. 43 matches among 61 matches resulted in draw are misclassified, causing low precision of 0.60. Besides, low recall means that the incorrectly classified win and lose category by draw category is high. 30 matches are categorized as draw, however, actual value of 12 of them are not draw. F1-score of win and lose matches are higher than draw matches proving that matches whose results are win and lose are categorized more competently.

When the heat map is investigated, the colour of the intersection of win and lose are lighter in both algorithms which indicates good performance. Although both algorithms provide consistent results with each other, accuracy value of SVM, (0.68) is larger than those of MLR (0.66) and PR (0.61). This shows that SVM yields better results than both regression algorithms.

Table 3. Prediction performance metrics including accuracy, precision, recall and F1 score

	Result value	Precision	Recall	F1 score
Multi Nominal	*draw*	0.60	0.30	0.40
Logistic	*lose*	0.68	0.78	0.73
Regression	*win*	0.66	0.82	0.73
	Accuracy			0.66
Support Vector	*draw*	0.56	0.34	0.42
Machine	*lose*	0.69	0.77	0.73
	win	0.72	0.84	0.77
	Accuracy			0.68
Poisson	*draw*	0,41	0,50	0,45
Regression	*lose*	0,71	0,59	0.64
	win	0,68	0,71	0.69
	Accuracy			0.61

4. Discussion and Conclusion

The aim of this research was to identify the performance indicators and situational variables which have significant effects on the number of goals scored and conceded and the match result. PR was implemented to model the number of goals scored and conceded, while SVM and MLR were applied to model the match results.

To start with the number of goals scored, our results showed that GC, SF and ST are the only significant variables influencing the number of goals scored and all of them have favourable effects on it. So, one thing a team can do to increase the number of goals scored is to score first. This finding is in line with a similar research [10] which reported that scoring-first teams scored 1.88 goals more than the opposing team. Additionally, increasing the number of shots on target leads to more goals scored. Furthermore, while similar studies [16], [23], [44] reported that home teams scored more goals; in our study, no significant association between match location and the number of scored goals was detected. Lastly, the remaining variables have no impact on the number of goals scored.

On the other hand; SF, ML, S and C have adverse impacts on the number of goals conceded. So, home teams or scoring-first teams concede less goals. Additionally, increasing the number of clearances or shots seem to work in decreasing the number of goals conceded. In contrast, QO has a positive impact on the number of goals scored, meaning stronger opponent score significantly more goals. Lastly, scored goals leads to more conceded goals and vice versa, indicating a significant positive association between the scored and conceded goals. The rest of the factors do not significantly affect the number of conceded goals.

It would have been very useful to compare all of these findings with those of similar studies. However, as mention in the introduction, in the current literature, there are few studies [10], [16], [23] investigating only few factors associated with the number of goals scored, but the number of goals conceded. Therefore, we are unable to make comparisons in terms of the number of scored and conceded goals.

To continue with the factors statistically significantly influencing the match result, MLR showed that SF is a significant indicator for the match outcome. It adversely influences the probability of losing and positively affects the chance of winning. These findings suggest that scoring first provides important advantage for teams to succeed. Other studies [10]–[12] also support this finding.

Likewise, S and ST have similar impacts on the match outcome. Having more shots and/or shots on target decreases the probability of losing, but increases the probability of winning, which implies that increasing the number of shots and/or shots on target increases the likelihood of success. These findings coincide with those obtained in similar studies [14], [32], [33].

ML (home) is another important influential factor on the match outcome and it has a negative influence on losing and positive on winning, indicating that playing at home has a great advantage in success. The advantage of playing at home in many leagues was pointed out by many studies [11]–[14], [17], [45].

Heretofore, the factors affecting both losing and winning have been discussed. However, there are also factors affecting either losing or winning. One of them is C which has a negative impact on losing, but no effect on winning. This indicates that increasing the number of clearances decrease the likelihood of losing but does not contribute to winning. This

finding was supported by a study [12] where the factors affecting the lose, draw and win were investigated separately. They found that C has a significant negative effect on lose only.

Another influential factor is QO. The findings pointed out that QO has a positive effect on losing the match but does not affect the winning probability. In other words, it is more likely to lose a match if the opponent is stronger. Other studies also indicated that, in football, favourite teams only win just over 50% of the matches, whereas in others sports such as basketball or handball, the favourite team wins more than 65% of the matches [11], [19].

The last influential factor is PP. Our results indicated that PP has an unfavourable impact on winning, meaning winning teams have less possession percentage. In support of this finding, other studies [27], [28] also found that winning teams have less possession percentage. Also, another research[29] stated that winning teams have lower possession percentage because they start to play with less risk, a well-structured defensive strategy, and place more players between the ball and its own goal. So that they can also prevent possible goal-scoring opportunities.

In conclusion, this research has identified the influential factors in the number of goals scored and conceded and the match result. In football, it is obvious that scoring goals increases the chance of winning, while conceding goals increases the probability of losing. In our study, the significant factors in scoring goals (SF and ST) found by PR are also among the significant factors in winning the match according to MLR. Similarly, the factors that have significant effects on conceding goals (SF, ML, S, C and QO) are among the factors that cause losing. In addition, SVM showed that SF, ML and ST are the most important factors affecting the match result. All these findings indicate that the results are consistent with each other, and SF is the only influential variable in all the analyses, making it the most important factor for success. Finally, further studies can be conducted to confirm the significant influential factors on the number of scored and conceded goals as their number in current literature is very limited.

Contributions of the authors

All authors contributed equally to the study.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics

References

[1] Y. Li, R. Ma, B. Gonçalves, B. Gong, Y. Cui, and Y. Shen, "Data-driven team ranking and match performance analysis in Chinese Football Super League," *Chaos, Solitons & Fractals*, vol. 141, p. 110330, 2020.

[2] T. Y. Yang and T. Swartz, "A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball," *J. Data Sci.*, vol. 2, no. 1, 2021.

[3] E. Ulas, "Examination of National Basketball Association (NBA) team values based on dynamic linear mixed models," *PLoS One*, vol. 16, no. 6, 2021, 2021.

[4] P. Marek, B. Šedivá, and T. ˛oupal, "Modeling and prediction of ice hockey match results," *J. Quant. Anal. Sport.*, vol. 10, no. 3, 2014.

[5] J. Goddard, "Regression models for forecasting goals and match results in association football," *Int. J. Forecast.*, vol. 21, no. 2, pp. 331–340, 2005.

[6] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 46, no. 2, pp. 265–280, 1997.

[7] D. Karlis and I. Ntzoufras, "Analysis of sports data by using bivariate Poisson models," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 52, no. 3, pp. 381–393, 2003.

[8] A. J. Lee, "Modeling scores in the Premier League: Is Manchester United really the best?," *CHANCE*, vol. 10, no. 1, pp. 15–19, 1997.

[9] M. J. Maher, "Modelling association football scores," *Stat. Neerl.*, vol. 36, no. 3, pp. 109–118, 1982.

[10] C. Lago-Peñas, M. Gómez-Ruano, D. Megías-Navarro, and R. Pollard, "Home advantage in

football: Examining the effect of scoring first on match outcome in the five major European leagues," *Int. J. Perform. Anal. Sport*, vol. 16, no. 2, pp. 411–421, 2016.
[11] J. García-Rubio, M. Á. Gómez, C. Lago-Peñas, and J. S. Ibáñez, "Effect of match venue, scoring first and quality of opposition on match outcome in the UEFA Champions League," *Int. J. Perform. Anal. Sport*, vol. 15, no. 2, pp. 527–539, 2015.
[12] G. Bilek and E. Ulas, "Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators," *Int. J. Perform. Anal. Sport*, vol. 19, no. 6, pp. 930–941, 2019.
[13] V. Armatas and R. Pollard, "Home advantage in Greek football," *Eur. J. Sport Sci.*, vol. 14, no. 2, pp. 116–122, 2014.
[14] C. Lago-Peñas, J. Lago-Ballesteros, A. Dellal, and M. Gómez, "Game-related statistics that discriminated winning, drawing and losing teams from the Spanish Soccer League.," *J. Sports Sci. Med.*, vol. 9, no. 2, pp. 288–93, 2010.
[15] R. Pollard, "Worldwide regional variations in home advantage in association football," *J. Sports Sci.*, vol. 24, no. 3, pp. 231–240, 2006.
[16] D. R. Poulter, "Home advantage and player nationality in international club football," *J. Sports Sci.*, vol. 27, no. 8, pp. 797–805, 2009.
[17] M. Saavedra García, O. Gutiérrez Aguilar, J. J. Fernández Romero, and P. Sa Marques, "Measuring home advantage in spanish football (1928-2011)," *Rev. Int. Med. y Ciencias la Act. Fis. y del Deport.*, vol. 15, no. 57, 2015.
[18] S. Thomas, C. Reeves, and S. Davies, "An analysis of home advantage in the English Football Premiership.," *Percept. Mot. Skills*, vol. 99, no. 3 Pt 2, pp. 1212–6, 2004.
[19] C. Anderson and D. Sally, *The numbers game: why everything you know about Football is wrong*. New York: Penguin Books, 2014.
[20] C. H. Almeida, A. P. Ferreira, and A. Volossovitch, "Effects of match location, match status and quality of opposition on regaining possession in UEFA champions league," *J. Hum. Kinet.*, vol. 41, no. 1, 2014.
[21] B. J. Taylor, D. S. Mellalieu, N. James, and P. Barter, "Situation variable effects and tactical performance in professional association football," *Int. J. Perform. Anal. Sport*, vol. 10, no. 3, 2010.
[22] H. Lepschy, A. Woll, and H. Wäsche, "Success Factors in the FIFA 2018 World Cup in Russia and FIFA 2014 World Cup in Brazil," *Front. Psychol.*, vol. 12, p. 525, 2021.
[23] C. Lago-Peñas and J. Lago-Ballesteros, "Game location and team quality effects on performance profiles in professional soccer.," *J. Sports Sci. Med.*, vol. 10, no. 3, pp. 465–71, 2011, Accessed: [11-Mar-2021]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24150619.
[24] L. M. Hvattum and H. Arntzen, "Using ELO ratings for match result prediction in association football," *Int. J. Forecast.*, vol. 26, no. 3, 2010.
[25] M. Crowder, M. Dixon, A. Ledford, and M. Robinson, "Dynamic modelling and prediction of English Football League matches for betting," *J. R. Stat. Soc. Ser. D Stat.*, vol. 51, no. 2, 2002.
[26] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, "'Quality vs quantity': Improved shot prediction in soccer using strategic features from spatiotemporal data," in *Proc. 8th Annu. MIT Sloan Sport. Anal. Conf.*, 2014.
[27] P. D. Jones, N. James, and S. D. Mellalieu, "Possession as a performance indicator in soccer.," *Int. J. Perform. Anal. Sport*, vol. 4, no. 1, pp. 98–102, 2004.
[28] C. Lago and R. Martín, "Determinants of possession of the ball in soccer," *J. Sports Sci.*, vol. 25, no. 9, pp. 969–974, 2007.
[29] C. Lago, "The influence of match location, quality of opposition, and match status on possession strategies in professional association football," *J. Sports Sci.*, vol. 27, no. 13, pp. 1463–1469, 2009.
[30] B. Mcguckin, J. Bradley, M. Hughes, P. O'donoghue, and D. Martin, "Determinants of successful possession in elite Gaelic football Determinants of successful possession in elite Gaelic football," *Int. J. Perform. Anal. Sport*, 2020.
[31] H. Liu, M. Á. Gomez, C. Lago-Peñas, and J. Sampaio, "Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup," *J. Sports Sci.*, vol. 33, no. 12, pp. 1205–1213, 2015.
[32] J. Castellano, D. Casamichana, and C. Lago, "The use of match statistics that discriminate

between successful and unsuccessful soccer teams," *J. Hum. Kinet.*, vol. 31, no. 1, 2012.
[33] F. A. Moura, L. E. B. Martins, and S. A. Cunha, "Analysis of football game-related statistics using multivariate techniques," *J. Sports Sci.*, vol. 32, no. 20, pp. 1881–1887, 2014.
[34] R. Ensum, R. Pollard, and S. Taylor, "Applications of logistic regression to shots at goal in association football," in *Science and Football V*, Routledge, 2005, pp. 211–218.
[35] H. Liu, W. Hopkins, M. A. Gómez, and J. S. Molinuevo, "Inter-operator reliability of live football match statistics from OPTA Sportsdata," *Int. J. Perform. Anal. Sport*, vol. 13, no. 3, 2013.
[36] S. Coxe, S. G. West, and L. S. Aiken, "The analysis of count data: A gentle introduction to Poisson regression and its alternatives," *J. Pers. Assess.*, vol. 91, no. 2, pp. 121–136, 2009.
[37] Y. Huo , L. Xin , C. Kang , M. W. Qin Ma and B. Yu, "SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso," *J. Theor. Biol.*, 2019.
[38] H. Pei, Q. Lin, L. Yang, and P. Zhong, "A novel semi-supervised support vector machine with asymmetric squared loss," *Adv. Data Anal. Classif.*, vol. 15, no. 1, pp. 159–191, 2021.
[39] D. A. Salazar, J. I. Vélez, and J. C. Salazar, "Comparison between SVM and logistic regression: Which one is better to discriminate?," *Rev. Colomb. Estadística*, vol. 35, no. SPE2, 2012.
[40] P. G. V. G. M. T. Fabian, "Scikit-learn: Machine learning in Python.," *J. Mach. Learn. Res.*, 2011.
[41] J. M. Bland and D. G. Altman, "Statistics notes. The odds ratio," *BMJ*, vol. 320, no. 7247, p. 1468, 2000.
[42] I. Soto-Valero, C., González-Castellanos,and M., Pérez-Morales, "A predictive model for analysing the starting pitchers' performance using time series classification methods.," *Int. J. Perform. Anal. Sport*, vol. 17, no. 4, pp. 492–509, 2017.
[43] V. Guyon, I., Weston, J., Barnhill, and S., Vapnik, "Gene selection for cancer classification using support vector machines.," *Mach. Learn.*, vol. 46, no. 1, pp. 389-422., 2002.
[44] T. Liu, A. García-De-Alcaraz, L. Zhang, and Y. Zhang, "Exploring home advantage and quality of opposition interactions in the Chinese Football Super League," *Int. J. Perform. Anal. Sport*, vol. 19, no. 3, pp. 289–301, 2019.
[45] T. Peeters and J. C. van Ours, "Seasonal Home Advantage in English Professional Football; 1974–2018," *Economist (Leiden).*, vol. 169, no. 1, pp. 107–126, 2021.