Academic Journal of Mathematical Sciences
ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

Research on the winning factors of football matches based on machine learning

Yuqing Yang

Shanghai World Foreign Language Academy, Shanghai, 200233, China

Abstract: *In order to explore the influence of various factors on the outcome of football matches, a total of 1520 matches of English Premier League, French Ligue 1, Spanish Liga, and Italian Serie A in the 2021-2022 season were selected as the research object, and Analytic hierarchy process (AHP) and Binary classification machine learning algorithm were used for modeling. The paper strength and onfield factors of the team are taken as variables, and the final winning or losing results of each participating team is taken as the result for statistical modeling. After obtaining the results of the model, it can be applied to the team's future result to lay out strategies in advance, thus providing a more intuitive understanding of the team's tactical style and scoring result. In general, the paper strength of attacking, passing, defense, goalkeeping and so on reflects the strength of the team to a large extent, and also determines the result and goal difference to a certain extent. In addition, the influence of the host and visiting venues on crowd support, player adaptation, and travel time has a further impact on the outcome of football matches, especially for teams with a small difference in paper strength. Even for some of the weaker teams or teams with similar paper strength, home field advantage will compensate for the difference in strength; for different leagues, the league with stronger average strength will be less affected by home and away, while the league with weaker average strength will be affected with greater variation. In summary, the accuracy of the overall prediction results is relatively significant.*

Keywords: *Analytic hierarchy process, Football match, Binary classification, Machine learning*

1. Introduction

In recent years, with the scientific training and diverse tactics of sports competitions, unexpected results have emerged frequently, and the importance of team comprehensive strength and on-site performance has continuously increased. The influence of star athletes on the direction of competitions has also been gradually weakened. Therefore, how to use scientific methods to analyze competitive competitions has become a hot topic. A large number of participating clubs in strong competitive events have started hiring professional data analysts to develop corresponding strategies and tactics by analyzing the various indicators of both sides of the competition, in order to help the team win the competition to a greater extent. Football is a representative global sport with a long history, demonstrating the highest intensity of team competition. Therefore, exploring the winning factors of football matches through data science methods will provide objective guidance for the strategic and tactical development of participating teams.

The development of data science and big data technology is becoming increasingly mature, and the application of artificial intelligence has emerged in various fields. Previous football league data and detailed indicator data of each team's players are publicly available to varying degrees on popular sports websites. Utilizing a large amount of historical data to train various machine learning models to achieve team data goals is one of the efficient data science guidance methods nowadays. [1] The five major leagues are top world-class football club leagues, and their participating teams and game results largely reflect the current development situation of football matches. This article will use mathematical modeling and data science methods to explore the main factors that affect the success or failure of football matches using data from the five major leagues, providing guidance for each team to scientifically formulate strategies and tactics.

ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

2. Methodology

2.1 Analytic Hierarchy Process

In order to study the influence of football match on winning rate, the team's own strength is a very important factor. All technical data in this study are from Football Reference data website (https://fbref.com/en/). By comparing the data of different companies, it is verified that the data collected by the company has a good consistency with the strength rating of players in football matches and the actual situation.

By the nature of the game, the most important 12 factors is selected from the following aspects to evaluate and analyze the overall paper strength of a team: squad standard stats, squad goalkeeping, squad shooting, squad passing, squad pass types, squad goal and shot creation, squad defensive actions, squad possession, squad miscellaneous stat. The 12 indicators selected for the analytic hierarchy are Goals Scored per 90 Minutes, Shots Total, Shots on Target, Pass Completion, Save Percentage, Tackles Win, Percentage of Dribblers Tackled, Blocks, Possession, Successful Take-On, Penalty Kicks Won and Market Value (As shown in Table 1).

Table 1: 12 Indicators and corresponding categories

Variable Categories	Indicators
Shoot	Goals Scored per 90 Minutes, Shots Total, Shots on Target
Pass	Pass Completion
Defense	Save Percentage, Tackles Win, Percentage of Dribblers Tackled, Blocks
Possession	Possession, Successful Take-On
Penalty	Penalty Kicks Won
Market Value	Market Value

From data collection website, after obtaining 12 index data of each team in four European leagues with overall eighty teams, the Analytic Hierarchy Process model[2] is constructed comparing pairwise. A Matrix judgment scale (1-9 scale method) was used to digitize the importance of the above indexes. When setting the weight for each indicator, the factors that have a greater impact and more significant effect on the football field are taken into account. For example, the market value, the more powerful the player's price is higher, the greater the value brought to the team, therefore, the market value reflects the strength of the player to a large extent, the highest weight in the 12 indicators. Among other indicators, goals and shots that can play a decisive role in victory also account for more than 0.1 percent, and the rest of the weight of ball control, passing, defense, etc., is similar. The judgement matrix is listed in table 2.

Table 2: Analytic hierarchy process of 13 factors affecting the paper strength

	Goal	Save	Sh	ShoT	Cmp	TKLW	Vs	Block	Poss	Succ	MV	PK
Goal	1	3/2	5/4	1	7/3	5/2	5/3	5/3	2	3	1/3	6
Save	2/3	1	2/5	1/3	1	1/2	1/3	2/3	1	1	3/5	1
Sh	4/5	5/2	1	2/3	5/4	7/3	6/5	2	3/2	5/2	2/3	1
ShoT	1	3	3/2	1	3/2	2	4/3	5/3	4/3	4/3	1/3	2
Cmp	3/7	1	4/5	2/3	1	1	1	1/2	1/3	2	3/5	2/5
TKLW	2/5	2	3/7	1/2	1	1	2	1	3/7	3/4	1/3	2
Vs	3/5	3	5/6	3/4	1	1/2	1	1/2	3/5	3/4	1/4	2/3
Block	3/5	3/2	1/2	3/5	2	1	2	1	2/3	1	2/3	1/2
Poss	1/2	1	2/3	3/4	3	7/3	5/3	3/2	1	3	1	3/2
Succ	1/3	1	2/5	3/4	1/2	4/3	4/3	1	1/3	1	2/5	1/2
MV	3	5/3	3/2	3	5/3	3	4	3/2	1	5/2	1	1
PK	1/6	1	1	1/2	5/2	1/2	3/2	2	2/3	2	1	1

After passing the consistency test, the Analytic Hierarchy Process, which sets an index based on the importance of each weight, can obtain the respective proportion of each w. As shown in Table 3, here are the values from w_1 to w_{12}:

ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

Table 3: Criteria weight

w	value
w_1	0.13511979
w_2	0.05155768
w_3	0.09626862
w_4	0.10118025
w_5	0.05441326
w_6	0.06314015
w_7	0.05729609
w_8	0.06654176
w_9	0.09717374
w_{10}	0.04917969
w_{11}	0.15205842
w_{12}	0.07607055

The weight vector is used to calculate the paper strength of the sample, and 12 paper strength variables are respectively a_1, a_2…… a_{12}. The final score of the team's paper strength of j is w_{ij} , then:

$$x_{1j} = \sum_{i=1}^{12} w_i a_i \tag{1}$$

The sample of the paper strength table is shown in table 4.

Table 4: Sample of paper strength

	Goal	save	Sh	SoT	Cmp	TklW	Tkl	Block	Poss	Succ	mv	pk	Paper Strength
A	1.97	76.5	583	200	85.1	348	45.3	408	61.8	58.3	20.78	8	149.559747
B	1.03	71	403	128	79	338	42.6	360	46.9	60.6	3.46	3	115.8750312
C	1.47	73.2	517	166	83	316	44.6	392	51.5	57.6	14.94	5	134.1371131
D	1.47	70.2	470	150	78.1	392	46.8	381	49.5	54.6	3.64	6	129.7809733

2.2 Binary prediction model

Under normal circumstances, the outcome of the game is determined by the strength of the team and the factors at hand. The contingent factors such as injury and player status cannot be estimated in advance, so they may be considered as natural and unavoidable errors. Weather, stadium comfort and other on site factors have the same impact on both sides, and are not included in the winning factors; Home and away factors have different degrees of influence on players' mental states, scene atmosphere and so on, which is an important variable affecting the outcome of the game. Therefore, assuming that the outcome of a match is determined by the paper strength of both sides and home-and-away factors, league's sample data set is constructed and a sample is given as shown in the table 5:

Table 5: Sample data

	Former team strength	Home/Away	Latter team strength	Goal Difference
Comp163	152.6171601	1	139.2180699	7
Comp272	149.559747	1	116.7073227	7
Comp162	160.5252508	1	139.2180699	6
Comp1630	149.559747	0	131.2769958	6
Comp11	152.6171601	1	141.6837392	5

In each match, there are two equivalent samples, home and away factors indicate whether team A is at home, and labels indicate whether team A wins. Away is 0, home is 1, wins are 1, losses are 0; To eliminate the effect of ties on sample uniformity, all tied games were removed.

The binary classification model is established to predict the result of the contest, and the classical machine learning method is considered. The paper strength of team A, whether team A is home and away, and the paper strength of team B are x_1, x_2, x_3 respectively, and whether team A wins is y_0, then:

$$y_0 = f_k(x_1, x_2, x_3), \tag{2}$$

ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

where $f_l$represents any binary model.

The method of control variables is used to explore the influence of different winning factors on the outcome of the match. The home and away factor is removed and only the remaining two paper strength variables are used to predict the match, then:

$$y_0 = f_l(x_1, x_3), \tag{3}$$

where $f_l$represents any binary model.

Using f_k and f_l to predict can help to observe the potential of this study in predicting the outcome of football league. On the other hand, it can compare and analyze the influence of paper strength factors and home and away factors on the results of matches. In order to get a better classification effect, this paper selects different representative binary classification algorithms to train the model from four aspects, such as statistics, and analyzes the prediction performances of different models of the competition.

3. Results

3.1 Experiment Setup

In order to take into account the different styles and tactical differences of different leagues, the collected data of English Premier League, Ligue 1, Serie A and La Liga are modeled respectively, and a total of 5 data are obtained, so as to explore the system differences of different leagues and analyze the result of winning rate as well.

English Premier League, Serie A, La Liga and French Ligue 1 are taken as League 1, League 2, League 3 and League 4 respectively. Data sets are established for the four major leagues from season 2021 to 2022. Sample size of each data set is shown in the table 6:

Table 6: Sample size of each data set

	league 1	league2	league3	league4	total
sample size	585	565	539	557	2256

In this paper, four binary classification algorithms including Logistic Regression, random forest, naive Bayes classifier and Back Propagation neural network were selected for model training and prediction, and parameter settings of each algorithm were given:

a. LR regression[3]: Use the lbfgs algorithm to optimize the loss function.

b. Random forest[4]: Using gini coefficients to divide molecular trees, using 100 decision tree classifiers.

c. Naive Bayes[5]: A prior selection of a Gaussian distribution.

d. BP neural network[6]: Adam optimizer[7], learning rate of 0.1, 15 hide layers, iteration 1000 times.

10-fold cross-validation method was used to test the performance of the model. Each model test of each data set was conducted 100 times, and the average value of each result index was taken.

3.2 Experiment Results and Analysis

The experiment was carried out in python3.8 enviroment, and the classification accuracy results were obtained as shown in Table 7.

Table 7: Accuracy of binary classification algorithm

	All	League 1	League 2	League 3	League 4
Logistic Regression	0.72	0.75	0.73	0.71	0.7
Logistic Regression(home/away)	0.75	0.8	0.78	0.76	0.71
Random Forest	0.64	0.73	0.65	0.64	0.58
Random Forest(home/away)	0.71	0.78	0.67	0.7	0.71
Naive Bayes	0.7	0.68	0.7	0.72	0.68
Naive Bayes(home/away)	0.72	0.72	0.77	0.76	0.71
BP neural network	0.71	0.76	0.72	0.67	0.71
BP neural network(home/away)	0.74	0.78	0.77	0.74	0.72

ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

According to the results, although the difference between the classification results after the removal of home and away variables is small, the average accuracy is lower than that before the removal. It can be concluded that paper strength is the most important factor to determine the outcome of the game, and home field advantage can increase the team's winning probability to a certain extent, but the impact is small. In addition, the conclusion accords with objective common sense, the strength of the team is the first standard to measure the level of the team, and the on-site factors play a small role.

Secondly, it can be seen from the table that among the four binary classification algorithms, the performance of random forest on the league data set is obviously inferior to the other three algorithms. Therefore, league data implies a large amount of information that can be fitted, and the classification algorithm performs poorly on this data set, while the algorithms of statistical perspective and fitting perspective perform better.

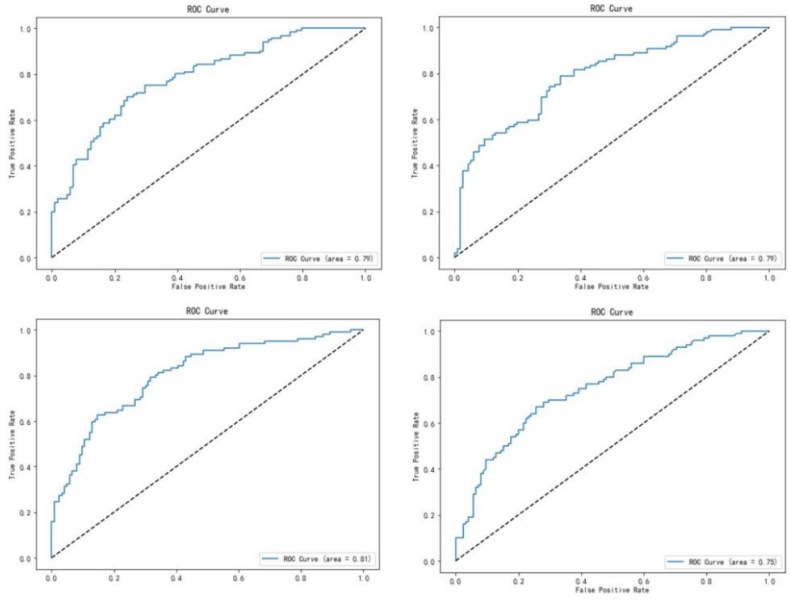
The differences among major leagues were analyzed through the test results. According to the paper strength evaluation results, league 1 has the strongest strength, league3 has the weakest strength, league2 and league4 has the moderate strength. The difference of test results before and after the removal of home/away for each league is counted, and the difference of accuracy of the four algorithms is averaged. The following table 8 lists the data:

Table 8: Average value of the difference in algorithm accuracy

	League 1	League 2	League 3	League 4
Accuracy diff	4 %	4.75 %	5.5 %	4.5 %

This indicates that the high-level performance of the strong team is relatively stable, and the degree of influence by home and away factors is relatively small. While the performance of the generally poor team fluctuates greatly, and the result of the game is affected by many more factors. The moderate strength teams are also affected to a certain extent. In addition, by analyzing the samples whose prediction results changed before and after the home/away removal, the teams with similar strength were more affected by the home/away factors.

The ROC curve[8] corresponding to the test results of the full sample dataset is given, and the corresponding single experiment is the one closest to the mean accuracy rate. The 8 pictures in Figure 1 correspond to the 8 experiments of the ALL dataset in table 7 respectively.



ISSN 2616-5805 Vol. 4, Issue 4: 51-56, DOI: 10.25236/AJMS.2023.040408

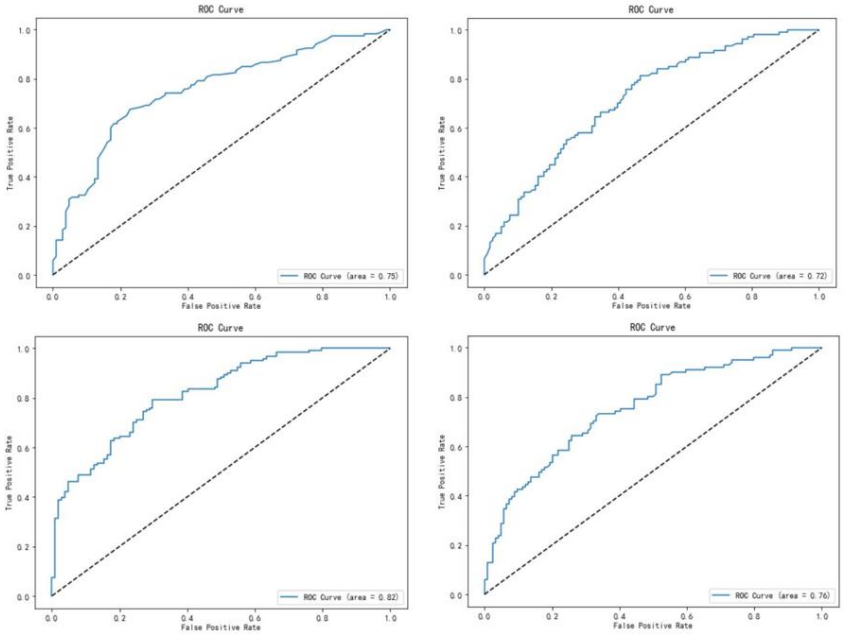


Figure 1: ROC Curve of 8 experiment

Obviously, the test results have good uniformity. It is worth noting that none of our tests were more than 80% accurate. On the one hand, it reflects the variability of sports competition, on the other hand, it urges us to think about adding and quantifying more on site factors in future research.

4. Conclusion

Through the AHP model and the four binary classification algorithms, the data analysis of the results of the English Premier League, Ligue 1, La Liga and Serie A in the 2021-2022 season can be concluded: in all games, paper strength has a dominant role in the victory rate, and the home and away and other temporary factors have an additional impact. The strong teams are less affected by home and away factors and can still achieve more goal difference in most cases. Similarly, the variables of home and away also have a small impact on the season results of the weak teams. For the team in the middle, when its strength is not stable, the influence factor of home and away factors is larger, and the team with low paper strength can also win in the favorable situation of home and away. For the overall league, if the overall strength of the league is strong, the high-level performance of the team will be more stable, while the weaker league will face the deviation in level, and be affected by the on-the-spot factors, making its win rate more volatile, the influence degree of home and away factors of other leagues is between the strong league and the weak league.

References

[1] Horvat, Tomislav, and Josip Job. "The use of machine learning in sport outcome prediction: A review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.5 (2020): e1380.
[2] Saaty, Thomas L. "How to make a decision: the analytic hierarchy process." European journal of operational research 48.1 (1990): 9-26 [J].
[3] Varela G, Novoa N, Jiménez M.F, et al. A pplicability of logistic regression (LR) risk modelling to decision making in lung cancer resection[J].Interactive Cardiovascular & Thoracic Surgery, (2003) (1):12-15.
[4] Liaw A, Wiener M .Classification and Regression by Random Forest [J]. R News, (2002), 23(23).
[5] Mccallum A, Nigam K. A comparison of event models for Naive Bayes text classification[J]. IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, (1998) 41--48.
[6] Rumelhart D E, Hinton G E, Williams R J .Learning Representations by Back Propagating Errors[J].Nature, (1986), 323(6088):533-536.
[7] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, (2014).
[8] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143.1 (1982): 29-36.