Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220080

Football Result Prediction Based on Machine Learning

Wenbo YU[1]
Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, China

Abstract. By constructing features such as goal difference, FIFA team ratings and league points, and by using machine learning technologies such as logistic regression and support vector machine, appropriate models are constructed to make binary prediction of match results.

Keywords. Logistic regression, support vector machine, predictive model, machine learning.

1. Background Introduction

In recent years, more and more people pay attention to statistics, and football is no exception. Although football data analysis is still a developing field, with the rise of the information age in the 21st century, data is playing an increasingly important role in football. For example, there are academics who use increasingly sophisticated machine learning techniques to predict the outcome of games. In this project, objective data analysis is mainly used to predict the results of premier League matches by using logistic regression, SVM and other models, and also evaluate the performance of the model and predictive results. (The Premier League is one of the top five leagues in European football, with 380 games played each season.)

2. Literature Review

As a popular sport, football has been studied by many scholars. For example, Cao Weimin and Shi Zhishe [1] analyzed the 2000 European Football Championship finals by using the methods of literature, observation, statistics and comparative research. Hou Huisheng, Zang Hepeng, Li Fengqiao [2] used Q-type clustering, rank correlation analysis and multiple comparative statistical methods to analyze the index data of 11 attacks in 64 matches of the 2006 World Cup Finals. In addition, in terms of football data modeling, Karlis and Ntzoufras[3] also used binary Poisson distribution model to illustrate football data. Fenton and Neil[4] proposed the Pi-Football model based on Bayesian network to predict the results of Football matches. The research results of

[1] Corresponding Author, Wenbo YU, Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, China; E-mail: 1628463139@qq.com.

Gudmundsson & Wolle[5] are more commercialized. They developed a series of automatic football analysis tools, together constitute a complete football data analysis software.

3. Methodology

3.1. Dataset Description

Two football datasets were used for analysis. The first data sets the FIFA game data set (https://sofifa.com/players?r=210006&set=true&col=oa&sort=desc) in this website, we can find related to the players, the team and league FIFA game data, Examples include individual player ratings, numerical values, and overall offensive and defensive team ratings.

The second dataset used data from UK football betting websites. (http://football-Data.co.uk/englandm.php) On the second website, we can find football data related to different leagues, which can effectively help us get reliable league football data in the first time.

Data tables are provided on the site, which contain a range of attributes such as final match results, home and away teams, goals scored, shots times, and can be used to create new features to improve the performance of the model.

3.2. *Application Method*

Firstly, we will prepare the fitting data by creating available features through four steps of data selection, data cleaning, feature engineering and heat map, and use the logistic regression, SVM and XGBoost model provided by SKlearn to fit the data of premier League season 16-19 to achieve the prediction of match results. Finally, the F1 score and accuracy evaluation model are used.

3.2.1. Data Selection

Before creating the right model, we need to select the right data set to fit the model. We tried all data sets from 2000 to 2019, and determined that the selected year data should not be too far away from the predicted season, and the selected season data should not be too few, otherwise the model may not fit well and the generalization ability is not strong enough to have good prediction. After comparing the performance of the model with data sets of different seasons, we finally decided to use 16-19 seasons as the input data set of the model.

3.2.2. Data Cleaning

Missing data was found in the 2018 season dataset as one of the inputs and the data for this season was deleted.

3.2.3. Feature Engineering

(Note: Data from the first three weeks of each season was removed from the experiment, due to a lack of historical information to predict at the beginning of the season.)

Table 1. Feature created

Feature	Description
Goal difference for home and away teams H/ATGD	Goals scored - Goals conceded
Points for home and away teams H/ATP	Win = 3, draw = 1, lose = 0
Each team's last three game performances H/AM	Home team: win = win; draw or lose = lose Away team: win or draw = win; lose = lose
FIFA ratings for home and away teams H/AR	FIFA game ratings

From the table 1, four features are created and a matchweek feature is added to the data sheet, i.e. the matchweek in which the match took place. Goal difference and team points can be averaged to better fit the model. After creating the features, we looked at wins and losses in the Premier League and we found that from the table 2 the home team had a winning rate of 47.43%, proving that the Premier League does have a home advantage. As a result, the tag ratio is unbalanced on win, lose and draw issues. Therefore, we use a binary classification problem of whether the home team will win or not, which is also a way to solve the problem of unbalanced tag ratio.

Table 2. Home court situation

16 - 19 seasons PL data	
Total matches	1050
Total features	12
Home team win	498
Home team win rate	47.43%

3.2.4. Heat Map

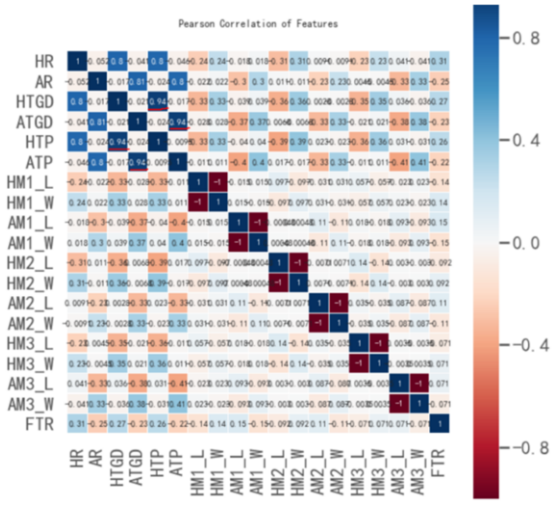

Figure 1. Heat map

The figure 1 above shows a strong correlation between HTP and HTGD, as well as between ATP and ATGD. The higher the average score, the higher the goal difference. In order to avoid multicollinearity, the H/ATP feature of league points was deleted while the H/ATGD feature of goal difference was retained. However, given that multicollinearity only appeared in logistic regression models, the league point H/ATP feature was continued to be used in other models to predict results.

3.3. Model Building

Table 3. Model building and parameters

API Interface	Parameters
Sklearn - Train_test_split	X_all, y_all, test_size=0.3
Sklearn - LogisticRegression	Solver = 'liblinear' class_weight = 'balanced'
Sklearn - SVC	Kernel = ' rbf ' class_weight = ' balanced '
XGBoost	booster = 'gblinear'
Sklearn - f1_score	target, y_pred, pos_label=1
Accuracy	(target == y_pred) / y_pred

From the table 3, we have built the models and evaluation methods, then tried to tune the XGBoost.

Sklearn - GridSearchCV	Using GridSearchCV to adjust parameters 'n_estimators':[80,90,100,110,120] 'max_depth': [3,4,5,6,7,8,9]

4. Experimental Results and Analysis

Table 4. Performances

Model	F1 scores in training set	Accuracy of training set	F1 scores in test set	Accuracy of test set	Sample size
Logistic regression	0.6573	0.6680	0.6601	0.6730	735
SVM	0.7041	0.7129	0.6465	0.6667	
XGBboost	0.4889	0.5310	0.5310	0.6635	
Enhanced XGB	0.4861	0.6231	0.5244	0.6603	

It can be seen from the table 4 that in the training set, SVM had the best performance in F1 score and accuracy, reaching about 70%, while in the test set, logistic regression had a better overall performance, maintaining an overall accuracy of about 67%. The accuracy of other models was also good, reaching about 66%. However, the results so far are only binary predictions in an idealized state. It is very difficult to predict the real results of a football match, there is no prediction model or institution that can achieve a predicted winning rate of more than 60%. Even the odds on betting websites tend to

come up with something unexpected, with so many random factors influencing the game that weak teams tend to pull off unexpected reversals.

The next four prediction models are used to briefly predict some premier League matches for the 2020-2021 season.

Match prediction 1: 2020/11/22 Tottenham Hotspur (home) VS Manchester City (away)

Table 5. Predictive result 1

Model	Predictive results
Logistic regression	Home didn't win
SVM	Home win
XGBoost	Home win
Tuning XGBoost	Home win

Final result: Tottenham hotspur 2-0 win for Manchester City, table 5 shows 3 models giving the correct answer.

Match prediction 2: 2020/11/23 Liverpool (home) VS Leicester City (away)

Table 6. Predictive result 2

Model	Predictive result
Logistic regression	Home win
SVM	Home win
XGBoost	Home didn't win
Tuning XGBoost	Home didn't win

Final result: Liverpool 3-0 to Leicester city, table 6 shows the 2 models giving the correct answer.

5. Conclusion

By using various objective data of football matches to create suitable features to fit more suitable models, it has been possible to achieve rough prediction of football match results. However, these models still remain in binary prediction, that is, the only prediction result is whether the home team win or not, and the prediction accuracy is about 60%. This probability is still quite different from the actual probability, so more available features and models need to be considered to improve the range and accuracy of prediction, and provide reference for subsequent research on football prediction.

References

[1] Shi Zhshe, & Cao Weimin, (2002). On the Relationship between shooting position and goal -- Statistical Analysis of shooting data in European Football Championship 2000. Hubei Sports Science and Technology, 021(002), 228-230.

[2] Hou Huisheng, Zang Hepeng, & Li Fengqiao. (2008). Comprehensive Evaluation of the Offensive Ability of soccer teams in the 2006 World Cup. Journal of Beijing Sport University, 031(001), 138-140,143.

[3] Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381-393.

[4] Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, *36*, 322-339.

[5] Gudmundsson, J., & Wolle, T. (2014). Football analysis by using spatio-temporal tools. *Computers, Environment and Urban Systems*, *47*, 16-27.