Pisco Med Publishing

Original Research Article

Analysis and Prediction of Soccer Games: An Application to the Kaggle European Soccer Database

Wuhuan Deng*, Eric Zhong

Chengdu Foreign Language School, Chengdu 611731, Sichuan, China. E-mail: 1029431715@qq.com

Abstract: The study of soccer game data has many applications for both fans and teams. The effective analytical work can not only help the teams to improve their offensive and defensive skills and strategies, but also could assist the fans to make a bet. In this work, the authors study the European League Dataset with statistical methods to analyze the game data. Moreover, machine learning techniques are designed to predict the game results based on in-game performance and pre-game odds provided by bookmakers. With rational feature engineering and model selection, our model results in an overall 95% accuracy.

Keywords: Soccer; Python; Data Science; Artificial Neural Network; Statistics; Poisson Distribution

1. Introduction

Soccer is one of most popular sports in the world, especially in Europe and South America. Numerous people are enthusiastic about this sport. More than a billion people watched the final match of World Cup in 2018. Another fact of soccer is that it is quite excellent for betting at different levels. The outcome of a soccer game depends on many factors, like the home-away ground effect, the physical and psychological conditions of key players. On the other hand, for each team, a good understanding and analysis of their past games is effective to help improving their skills, strategies and training methodologies.

In literature, Balogun O. and Ogunseye AA[1] used Artificial Neural Network (ANN) to predict the scoreline of Manchester United matches against opposing teams for matches played in the English Premier League in 2019. Their data spans a period if nine years from 2009 to 2018. 331 of the data set were used as the training data set, while 12 was used for validation. Their artificial neural network model has 6 input layers 5 hidden layers, and 2 output layers. And their model gave an accuracy of 73.27% for goals scored by Manchester United. And Yoel F. Alfredo and Sani M. Isa[2] also published a similar research paper in 2019. Their data comes from football data.co.uk which is a common data set to be used in conducting research in football match predictions. There are 71 attributes in the data, and they can be divided into two categories: football match statistics and bookmaker odds prediction. Only 14 attributes are selected by them, which they think can have good impacts on prediction and result accuracy. They use 3 models to predict the result: C5.0 Model, Random Forest Model, and Extreme Gradient Boosting Model. And each model has a high accuracy of prediction.

2. Dataset description

The open dataset for study in this work is acquired from www.kaggle.com, built by Hugo Mathien, includes soccer match data of 11 European countries with their

doi: 10.18282/i-s.v6i1.332

lead championship from seasons 2008 to 2016. The database covers 24,637 matches and over 10,000 players. Betting odds from up to 10 bookmakers are included. Detailed match events like goal types, possession, corner, cross, fouls, cards are also recorded in the database. Moreover, this dataset integrated players and teams attribute ratings sourced from EA Sports' FIFA video game series.

The dataset is a SQL database file contains 7 tables of "Country", "League", "Match", "Player", "Player_Attribute", "Team", "Team_Attributes". For the match table, 134 attributes are recorded, some of which are significant for predicting game results, while appropriate feature engineering based on relative soccer knowledge may also be conducted in the later prediction process.

3. Statistical analysis

3.1 Best offensive and defensive teams

Match data provide supporting evidence to reveal the offensive and defensive performance for each team. Hence, the first investigation conducted on this dataset is to study the game performance for each team from 2008 to 2016 based on goal events. The attributes applied include "home_team_goal", "away_team_goal" and another two designed features, "shot_efficiency" and "goal_efficiency". The feature "shot_efficiency" is the radio of the number of shot-ons and the number of total shots, and "goal_efficiency" is the radio of the number of goals and the number of shot-ons. With simple feature engineering, the two introduced new attributes demonstrate the efficiency of team performance of shooting and scoring, which are significant to evaluate team offensive capabilities.

By calculating the averaged attributes from 2008 to 2016 for all teams, the top away and home teams are demonstrated in **Figure 1** and **Figure 2.**

Away team goal performance top list

Rank	away_team	away_team_goal
1	FC Barcelona	2.33
2	Real Madrid CF	2.22
3	FC Zurich	2.13
4	Ajax	2.11
5	PSV	2.07
6	Celtic	2.01
7	FC Bayern Munich	1.99
8	SL Benfica	1.99
9	FC Porto	1.98
10	Rangers	1.93

Away team defensive performance top list

Rank	away_team	home_team_goal
1	Grasshopper Club Zurich	0.63
2	Rangers	0.75
3	FC Porto	0.77
4	SL Benfica	0.84
5	Juventus	0.84
6	FC Bayern Munich	0.85
7	FC Barcelona	0.86
8	Celtic	0.88
9	Sporting CP	0.9
10	Legia Warszawa	0.97

Away team shot efficiency top list

Rank	away_team	away_team_shot_efficiency
1	Valenciennes FC	80%
2	Karlstruher SC	75%
3	FC Utrecht	69%
4	Novara	68%
5	SC Freiburg	68%
6	Vitesse	67%
7	FC Barcelona	66%
8	Legia Warszawa	65%
9	Paris Saint-Germain	65%
10	SC Heerenveen	64%

Away team goal efficiency top list

Rank	away_team	away_team_goal_efficiency
1	DSC Arminia Bielefeld	32%
2	Valenciennes FC	25%
3	CD Numancia	22%
4	SV Darmstadt 98	22%
5	ADO Den Haag	22%
6	Rangers	21%
7	Xerez Club Deportivo	20%
8	Zaglebie Lubin	20%
9	Paris Saint-Germain	20%
10	AS Monaco	19%

Figure 1. Away team attributes ranking.

Home team goal performance top list		
Rank	home_team	home_team_goal
1	BSC Young Boys	3.38
2	Real Madrid CF	3.32
3	FC Barcelona	3.26
4	FC Bayern Munich	2.81
5	PSV	2.72
6	Ajax	2.65
7	SL Benfica	2.59
8	Celtic	2.56
9	Manchester City	2.4
10	FC Porto	2.38

Home team defensive performance top list		
Rank	home_team	away_team_goal
1	FC Zurich	0.13
2	FC Porto	0.52
3	Ajax	0.57
4	Celtic	0.58
5	SL Benfica	0.66
6	FC Barcelona	0.66
7	FC Bayern Munich	0.71
8	FC Vaduz	0.71
9	Legia Warszawa	0.73
10	Rangers	0.74

Home team shot efficiency top list		
Rank	home_team	home_team_shot_efficiency
1	Paris Saint-Germain	67%
2	FC Barcelona	66%
3	Monchengladbach	66%
4	FC Twente	65%
5	N.E.C.	65%
6	Bournemouth	65%
7	Real Madrid CF	65%
8	Ajax	65%
9	Hercules Club	64%
10	Wolfsburg	64%

Home team goal efficiency top list		
Rank	home_team	home_team_goal_efficiency
1	Korona Kielce	25%
2	Paris Saint-Germain	23%
3	FC Barcelona	23%
4	Livomo	22%
5	Ajax	22%
6	Celtic	22%
7	Real Madrid CF	21%
8	Hecules Club	21%
9	Wolfsburg	20%
10	Reading	20%

Figure 2. Home team attributes ranking.

The above two figures demonstrate that FC Barcelona appears 7 times, and real Madrid CF appears 5 times, which match with the practical game performance of these two teams during that time period.

3.2 Goal distribution

Numbers of goals for both away team and home team in each of the matches are collected in order to study the distribution of goals. Two Poisson distribution model are built to fit the goal distribution. Demonstrated in Figure X to Figure X, the Poisson distribution is appropriate to accurately model the true distributions of home team goals and away team goals. Therefore, as the two variables home goal and away goal follow the Poisson distribution, the difference between them, which is the net goal, should fit the Skellam distribution. Figure X shows the Skellam distribution of net goals accurately fit the true distribution of the net goals calculated by taking the difference of home and away goals for each match.

The total goal and net goal distribution are shown in the **Figure 3** and **Figure 4.**

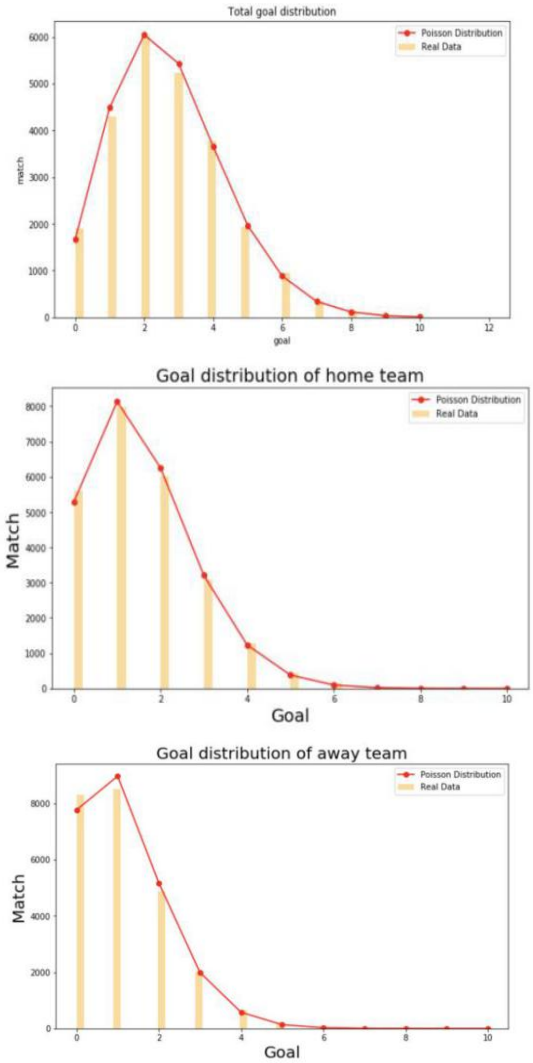

Figure 3. Goal distribution of away team and home team.

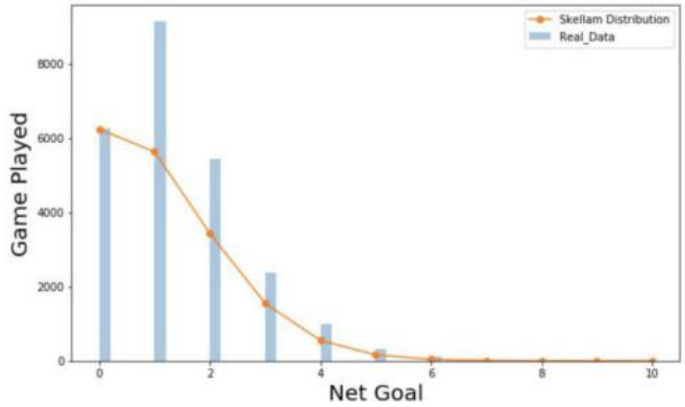

Figure 4. Net goal distribution.

4. Game result predictions

4.1 Description

This prediction of this dataset is slightly different from normal predictions of match results. Normally, the prediction is focused on the result of matches which have not happened yet, based on some pre-game features, such as squad, and historical match results. The prediction in this work is to foretell the match results based on pre-game bet radios and in-game performance. The objective of this prediction is to analyze the key features that mostly affect the final match results, which could be useful to help improving team training and strategy making.

4.2 Feature engineering

There are total 144 attributes in the match table of the database file, 28 of which are selected, including bet radios provided by different bookmakers and in-game features like possession (the time of controlling the ball in percentage), shot-efficiency (on target shot/total shot), goal-efficiency (goal/on target shot), etc.

Featuring engineering is a significant and necessary process to appropriately help improving the prediction accuracy. Sufficient background knowledge of the practical application is essential to help understand the prediction problem. For soccer games, the offensive efficiency plays an important role in evaluating the game performance, and strongly related to the final results. Therefore, additional features including shot efficiency and score efficiency are designed to characterize the offensive efficiency.

4.3 Models

Four models are built in the work, including Logistic Regression, Decision Tree, Random Forest and Deep Neural Network. The Logistic Regression model is appropriate for this problem since the dependent variable win-draw-lose is categorical. It is easy to implement, and computational efficient. For the Decision Tree model, each branch of it represents a possible decision, outcome, or reaction, and the last branch of the tree represents the final result. In our Decision Tree model, the criterion is information entropy: a mathematical measure of the degree of randomness in a set of data, with greater randomness implying higher entropy and greater predictability implying lower entropy. The advantage of Decision Tree is that data processing is simple or unnecessary. However, when the types of data increase, the accuracy will decrease, since the over-fitting problem occurs. For the Random Forest model, literally, it gives a higher accuracy than the Decision Tree model since the problem of over-fitting is ameliorated, with the cost of increasing computational complexities. The last model we generate is the Deep Neural Network model, which is suitable to describe the non-linear relationship between the match features and results. In this work, a neural network with 5 fully-collectly dense layers, 5 activation layers, 2 dropout layers and 1 batch normalization layer were designed. Softmax transformed function is applied in the model, and the loss of the model is based on sparse categorical crossentropy. The batch size of the model is 32, and the epochs of the model is set to as 500.

4.4 Results

The four models are all implemented to predict the match results based same training and testing data. Comparison of the prediction results for the four models are provided in **Figure 5.**

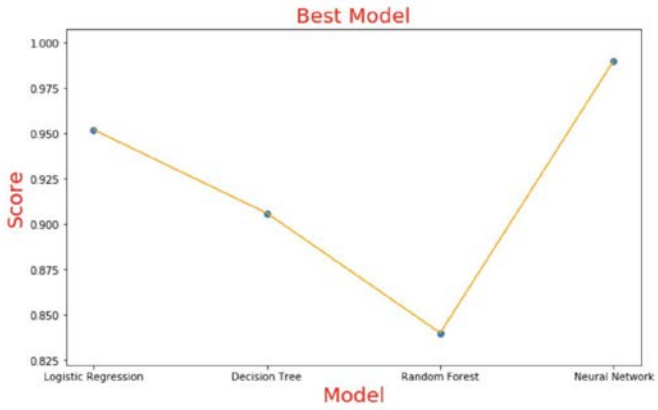

Figure 5. Four models' prediction result.

Among which, the Deep Neural Network model gives the highest accuracy of 0.99. The Logistic Regression, Decision Tree and Random Forest models results in 0.95, 0.91 and 0.84. Moreover, we implement the best model on the 5 top leagues of Europe exclusively, which are English Premier League (England), Laliga (Spain), Bundesliga (Germany), Serie A (Italy) and Ligue1 (France), and the results are shown in **Figure 6.**

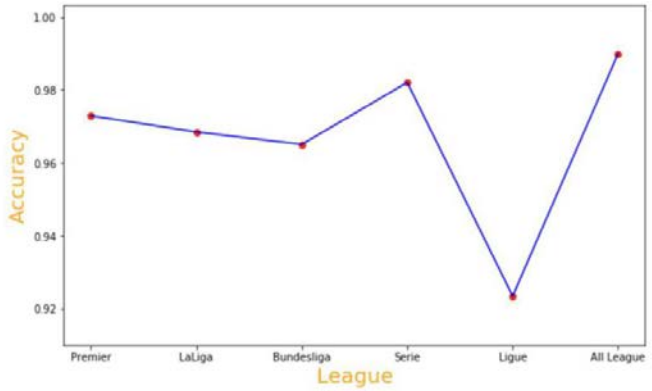

Figure 6. Top 5 leagues' prediction result.

For analyze the most import features for soccer match prediction. 4 most important features including possession, total shot, shot efficiency and goal efficiency are compared. Each of the 4 features is solely applied for the prediction for the 5 leagues. The results in **Figure 7** to **Figure 10** indicate that the possession, total shot and shot efficiency are most important for Laliga, and goal efficiency is most significant for Ligue.

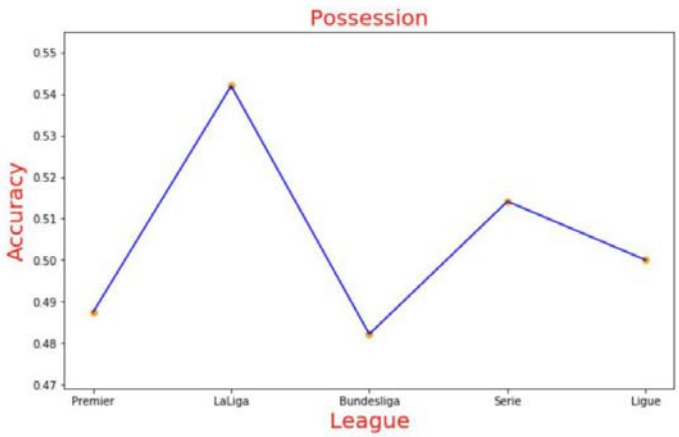

Figure 7

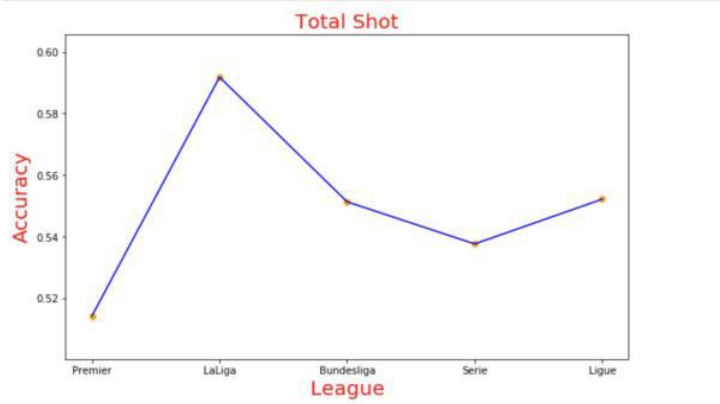

Figure 8

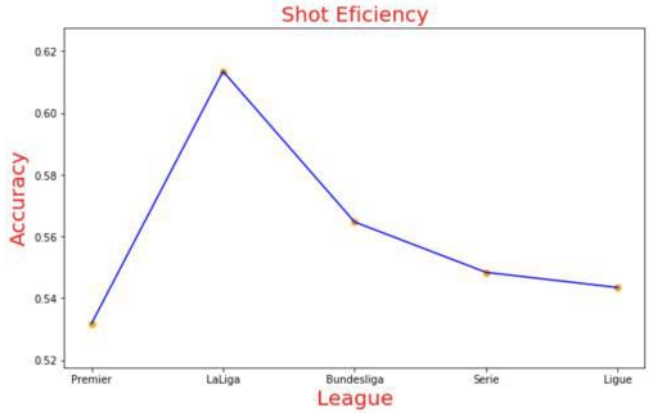

Figure 9

Figure 10

5. Conclusion

In this work, a novel analysis and prediction on an open soccer game database are proved. By using simple statistical methods, the best offensive and defensive teams are successfully evaluated, and the Poisson and Skellam Distribution is verified for fitting the goals. Moreover, feature engineering and four prediction models are conducted to foresee the match outcomes. The results indicate that our feature engineering and the designed Neural Network is effective for the match result prediction.

References

1. Ogunseye AA. Artificial neural network approach to football score prediction. Journal of Artificial Intelligence 2019; 1.
2. Alfredo YF, Sani MI. Football match prediction with tree based model classification. International Journal of Intelligent Systems and Applications 2019; 11(7): 20–28.