Further Investigatian of Football Data Analysis

Ozan Berk Bitirgen
TOBB ETU
obitirgen@etu.edu.tr

Furkan Kucuk
TOBB ETU
f.kucuk@etu.edu.tr

Izzet Baris Ozturk
TOBB ETU
izzetbarisozturk@etu.edu.tr

ABSTRACT

The goal of this project is to predict match outcomes of European football matches more accurately than bookkeepers, thereby beating the odds and to, in the end, generate a positive return on investment. Based on the dataset provided on kaggle.com that includes basic match data, FIFA player statistics and bookkeeper data, we aim to implement various mining techniques to our datasets and built a model to predict the probability of each match outcome as win, draw, or defeat.

KEYWORDS

football data analysis, data mining, sports feature engineering, result prediction

ACM Reference Format:
Ozan Berk Bitirgen, Furkan Kucuk, and Izzet Baris Ozturk. 2019. Further Investigatian of Football Data Analysis. In *Proceedings of BIL 573 - Data Mining Term Project (TOBB ETU - BIL 573 - 2019 / SUMMER).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Football has always been the most popular sport in the world. People enjoy playing it or watching amateur or professional games, and teams have massive supporter groups. Moreover, it also gathers some attention from betters, and has a share of multi-billion dollar with the advent of online betting[1]. Because of this reason both academic research groups and industrial organizations who look to profit from potential market inefficiencies set their sights on football odds. Since there are many factors can influence the football matches outcome, such as teamwork, skills, weather, home advantage and many others like red cards, injuries during matches or simple luck like unexpected slips by players or goals, the prediction in football has become a challenging research problem.

Achieving better prediction results of sports matches is an active research area. However, an acceptable result is hard to achieve since there is lots of factor effecting result. One of main reason makes it hard to obtain these factors is, they are not numerical factors. For example, there should be some solid numerical method to score each players abilities. However, especially in team sports, it's hard to obtain each players abilities as a numerical output. Luckily, there are some evaluations that can be considered as accurate fairly e.g. scout scores. Since video game called "FIFA" has the most complete and realistic data for teams and football players, it is very good choice for football matches prediction. Besides it's accuracy, there is a clear methodology for them to how they score each player.

The remainder of this paper is organized as follows; Related Works section provides some background information about the progress have been made so far. Dataset section gives some information about the dataset used in this paper. The Explanatory Data Analysis section sheds light on to dataset and its attributes in order to help gain some insights about historical statistics and outputs of European League results. As this project will move on to next stages, there is a Future Work section which will give some information about what will be done in the near future.

2 RELATED WORKS

Predicting the results of the outcome of football matches is started as early 1977 by [3]. In [3], a model called least squared model has developed which rate both strength of the home and the away team using matrix of goal scoring distribution. In [4], a Bayesian network is proposed in order to take account the variation of time for all attributes simultaneously which also known as Dynamic Bayesian networks(DBNs). Consequently, home and away team's offensive and defensive strength will be varied over time.

Next, in [5], a complex framework for predicting football matches results is presented. This complex framework known as FRES system consists of two main components: rules based theorem and Bayesian network component. Thus, FRES system is a compound of two techniques which cooperate together for predicting football matches results. FRES system also was implemented in-game time-series approach which make the prediction more realistic. However, expert knowledge is needed for FRES system in order to run it well. In [2], Bayesian hierarchical model for the prediction of football results is suggested. The number of goals scored by the two teams in each match have been used to developed Bayesian hierarchical model. In [6], a football prediction model called pi-rating is presented to generate forecasts about the football matches outcome whether home win,draw or away win for English Premier League matches during seasons 2010/2011 which incorporate objective information and subjective information such as team strength, team form, psychological

TOBB ETU - BIL 573 - 2019 / SUMMER,
,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/1122445.1122456

impact and fatigue. In [7], a statistical model for the analysis and predicting of football match results is developed which assumes a bivariate Poisson distribution with intensity coefficients that changes randomly over time. However, [7] claimed that the work was based on classical perspective and suggested the used of Bayesian Networks for better account of parameter uncertainty. In conclusion, it is shown that Bayesian networks has significant value for predicting football matches results.

Among the related works in the literature, this paper takes off from [8], who presented an approach to forecasting results in which the Bayesian Networks provided a means for representing, displaying, and predicting the results of expert knowledge in a football game. Their results showed that the Bayesian networks is generally superior to the other techniques such as the MC4, a decision tree learner, naive Bayesian learner (NB), and k-nearest neighbour learner for this domain in terms of predictive accuracy. The Bayesian networks proposed by [8] successfully gained predictive accuracy of 59.21% which outperformed other machine learning techniques by 41.72% (MC4), 47.86% (NB) and 50.58% (KNN). [8] used the presence and absence of 3 key players, the home advantage, opposition team quality based on position in league table and the position of main key player named Wilson whether he played as midfield or not as the attributes for predicting the football matches results.

3 DATASET

The current dataset consists of historical football data gathered from a web API[9]. The current attributes of the dataset are as the following.

- Countries
- Leagues (First division leagues in each country)
- Seasons
- Teams
- Matches and their outcomes
- Match statistics

Historical data ranges from 2008 to 2016. This list may to be updated in future. Explanatory data analysis is to be done in the following section of this paper. Some other attributes also will be added in future.

The attributes to be added are:

- Betting odds
- Player attributes from FIFA game[10]
- Lineup informations
- Weather report for the match date
- etc...

4 EXPLANATORY DATA ANALYSIS

An explanatory data analysis (EDA) is conducted in order to obtain some intuition about the dataset. This intuition might give one a strong clue about which attributes are the most important for making some predictions. The preliminary outcome of this section is understanding some most obvious and strong features to use. Despite using these attributes as features, a feature generation study also can be conducted.

Figure 1: League locations

The intuition obtained from EDA can be use in this feature generation studies. Understanding data is also important for deciding how to apply data mining techniques on the current dataset. EDA gives some information about the structure of the database, gives hint about how to preprocess the dataset before applying data mining techniques and may even leads to customizing state-of-the-art techniques for that specific task.

First of all, one can want to know that which countries are consisting in the dataset. Countries can be enumerated as; Belgium, England, France, Germany, Italy, Netherlands, Poland, Portugal, Scotland, Spain and Switzerland.

As mentioned before, there is only one league from each country in the dataset. The dataset and their countries are as in the Figure 1.

Each league may have different number of teams or number of weeks. Hence, they differ how much matches played in each season, and in total. It can be seen in Figure 2.

It is a known fact that being home or away team has a huge effect on the winning chances as it is a huge psychological factor. However, it can be shown that it differs for each league. In some leagues this feature may lose its dominance among other possible features. It can be seen in Figure 3.

Certain teams hold some greater advantages like having better players. This point leads them to score higher number of goals than other teams both in away and home matches. This could be also another strong feature to make some predictions. The analysis can be seen in Figure 4.

One may want to see the correlation between the number of matches played in a specific league and the total number

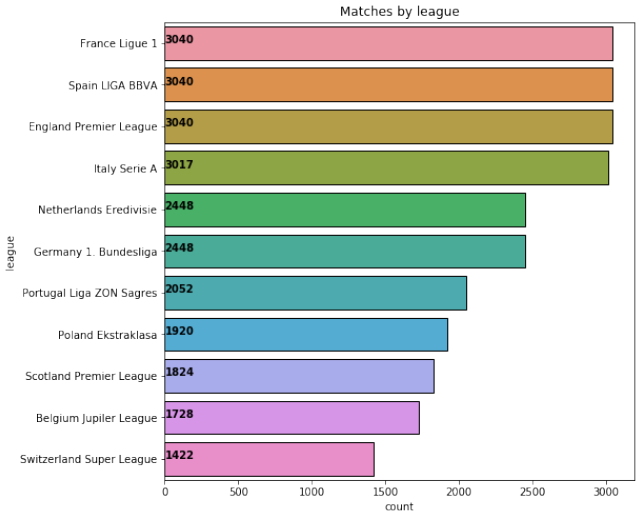

Figure 2: Matches played by league

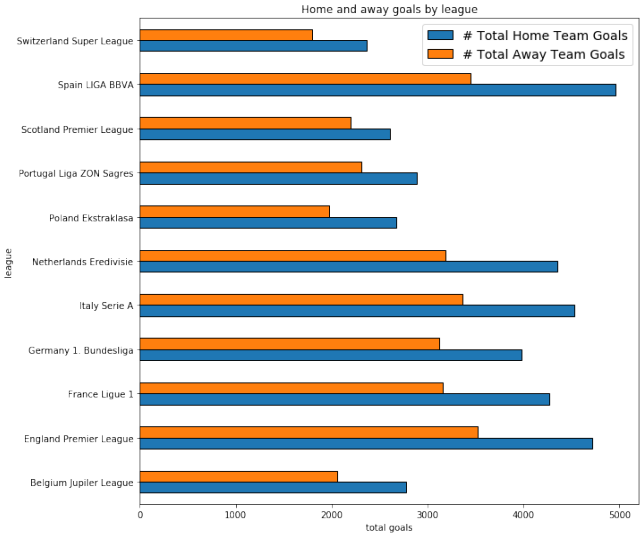

Figure 3: Home goals vs Away goals by league

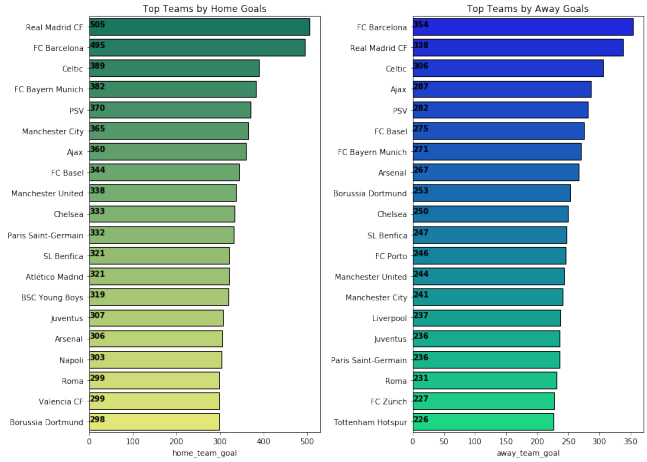

Figure 4: Top scoring teams by home and away goals

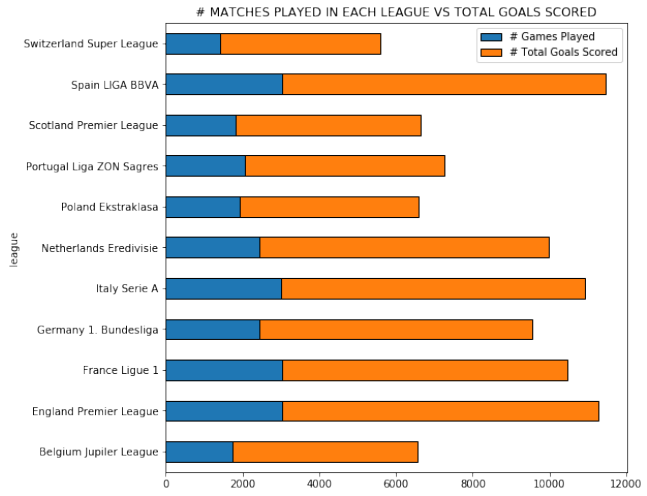

Figure 5: Comparison of match played vs total goals scored

of goals scored. As it can be seen from Figure 5, there is a direct correlation between matches played and goals scored. It's an expected behavior from real world exercises, and indicates that number of goals scored by each team, at least in it's non-customized form, is not a reliable feature to use in predictions.

Despite the chaotic features which could be engineered to predict match outcomes, analysis of goals by season shows a consistent graphic as it can be seen in Figure 6. If outliers ignored, it can be seen that away goals shows a little more volatility throughout the season than home team goals. However, other statistical calculations like interquartile ranges, median, maximum and minimum stays the same for every years.

To conclude the preliminary EDA process, one may infer some intuitions about useful features that effects heavily one the match results. Obviously, some teams are achieving consistent and stable good results comparing to other teams. Being on the home or away side also has a huge effect on results. However, these alone does not conclude the feature engineering process. Further feature engineering may required in order to make robust predictions with high accuracy.

5 FUTURE WORK

Since football matches are complex systems that are not easy to model, in this project, first of all, the detailed modeling of football matches will be studied. The expected result from this modeling is the prediction of match results with a high percentage. However, only physical modeling will not be enough to predict the outcome of football matches. Because,

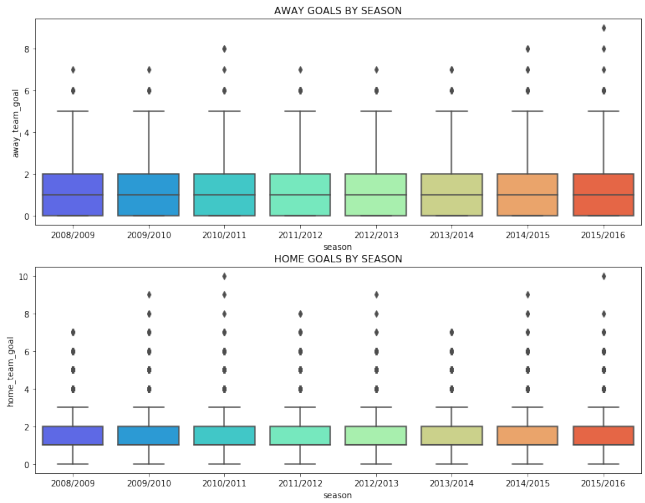

Figure 6: Box plots of goals by seasons

the factors that determine the outcome of football matches include chance, team/player morale and match-fixing parameters that are very difficult to predict and model. Within the scope of this project, it is aimed to model and predict complex factors such as chance, psychology, match-fixing and to make an expert system that predicts football match results with good success. In addition, betting on football matches with this system is expected to achieve a certain level of positive returns.

In order to better model the effects of psychological factors, it is considered that the results of the matches will be classified into 5 categories: big loss, small loss, draw, small win and big win.

In some cases, some teams have to win certain matches (such as teams that have a relegation). Or sometimes some teams before big matches appear on the court with more substitutes and rest their star players. In this case, the loss rate of this match increases. Team psychology affected by the fixture is also included in the research topics of this project.

In this project, the features will tried to be derived for the parameters that affect the results of the matches such as team psychology, team formation, style of the managers.

As mentioned earlier, as a result, all modelled parameters and derived features will be tried for different machine learning models and hereby a good forecaster for the match results will be tried to obtain.

REFERENCES

[1] Constantinou, A. C. & Fenton, Norman E. (2013). Profiting from arbitrage and odds biases of the European football gambling market. The Journal of Gambling Business and Economics, Vol. 7, 2: 41-70
[2] Baio G and Blangiardo M 2010 Journal of Applied Statistics 1–13
[3] Stefani R 1977 IEEE Transactions on Systems, Man, and Cybernetics 7 117–121
[4] Rue H and Salvesen O 2000 Journal of the Royal Statistical Society: Series D (The Statistician) 49 399–418
[5] Min B, Kim J, Choe C, Eom H and (Bob) McKay R I 2008 Knowledge-Based Systems 21 551–562
[6] Constantinou A C, Fenton N E and Neil M 2012 Knowledge-Based Systems 36 322–339
[7] Koopman S J and Lit R 2015 Journal of the Royal Statistical Society: Series A (Statistics in Society) 178 167–186
[8] Joseph A, Fenton N E and Neil M 2006 Knowledge-Based Systems 19 544–553
[9] http://football-data.mx-api.enetscores.com/
[10] http://sofifa.com/