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 accurately. Based on the scraped data provided by www.sofifa.com that includes basic match data and FIFA player statistics, 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. 2018. Further Investigatian of Football Data Analysis. In *TOBB ETU '19: Data Mining Project, July 27–07, 2019, Sogutozu, Ankara.* ACM, New York, NY, USA, 5 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 is 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 its accuracy, there is a clear methodology for them to how they score each player. The remainder of the paper is organized as follows. In Feature Engineering section, we will describe the engineering of the relevant features in order to perform the predictive analysis. Feature Selection section explains the different feature selection techniques that are em- played for finding the best performing features. Predictive Modelling section describes the machine learning techniques that are used for performing the predictive analysis. Results and Conclusions section contains a thorough analysis of the results obtained by the machine learning algorithms. Finally, we will describe what more can be done as future work.

2 RELATED WORK

Predicting the results of the outcome of football matches is started as early 1977 by [?]. In [?], 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 [?], 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 teamâĂŹs offensive and defensive strength will be varied over time.

Next, in [?], 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 [?], 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 [5], 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 impact and fatigue. In [6], 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, [6] 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 [?], who presented an approach to forecasting results in which the Bayesian Networks provided a means for representing,

Data Mining '19, July 07–27, 2019, TOBB ETU, AN

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

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 [**?**] successfully gained predictive accuracy of 59.21% which outperformed other machine learning techniques by 41.72% (MC4), 47.86% (NB) and 50.58% (KNN). [**?**] 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 FEATURE ENGINEERING

We scraped our data from a www.sofifa.com which is the one of the best data providers for Historical Football Data. We used data from 2008 to 2016, spanning 8 seasons. Although we also had access to data from much earlier seasons, we chose not to use them due to the limited match statistics available in these earlier seasons.

Previous research (Joseph et al., 2006; Owramipur et al., 2013) has shown that the quality of the results in this prediction task is associated directly with the quality of the feature set used for modelling the system. Both our understanding of the problem domain and our predictive analysis suggest that the selection of the correct set of features is of paramount importance. Thus, one of the most crucial aspects of our research was the engineering of features that would be likely to help in predicting the outcomes of a given football match. The features engineered will be explained thoroughly. Both prior research and our own intuition suggested that the home/away factor is an important characteristic of the problem.Thus, we treat the home/away factor as a globa l characteristic and compute the feature values for both the home and away teams for every feature that we engineer. Some of the models used previously have incorporated features such as the relative strengths of the attack and defense for the home and away teams. Influenced by these, we have also incorporated the Attack, Midfield, Defense and Overall ratings for the respective teams. We collected the data for the Rating statistics from an online database (https://www.worldfootball.net) that has a collection of in dividual team ratings (Attack, Midfield, Defense, Overall) for each season that are generated by the football video game series FIFA, which is released annually by Electronic Arts. The database also included the season-wise (static throughout a season) Rating statistics for each team, which account for variation in a teams strength across seasons. The ratings used are the ones that are determined by the algorithm used by EA Sports for their video game series FIFA. We scraped the Ratings (Attack, Midfield, Defense, Overall) statistics from the database directly, rather than using any kind of team player rating aggregation of our own to compute the different Rating statistics for each team. However, the introduction of these features raised the issue of their non-Gaussian distribution over the dataset, which dampened the results of both probabilistic and lin ear models. We toyed with various different approaches for dealing with this problem, and ultimately found that the optimal solution was to consider the differential of the Rating statistics.

For Ri ∈ Attack, Midfield, Defense, Overall :
Ri = Ri(home) - Ri(away)
where R(Home) = Home Team Rating and R(Away) = Away Team Rating.

The next feature that we have considered is the Goal Difference. The Goal Difference is of pivotal importance when building predictive models for football. The pi-rating system introduced by Constantinou and Fenton (2013) pro vides empirical proof that the Goal Difference works well as a feature for forecasting football match results. We used the traditional unweighted Goal Difference, which is a running sum of the numbers of goals scored differenced by the running sum of the numbers of goals conceded by each team before coming into a new match. The differential features like these had a much better univariate distribution, which is quite useful for some statistical learning techniques and by encoding the information held by two variables in one, we were able to reduce the dimensions of the problem hyperspace, which makes our models less likely to get stuck in a local minimum.

Since having such a large number of features increases our feature space dimension greatly, we aimed to select the best performing and most relevant features by performing feature selection. We tested our feature set in 'Table 1' with the Random Forest method and used Information Gain for feature selection. We used all methods in Weka and found that the best accuracy comes with this. After that, we get the extracted features shown in 'Table 2'. We observed that differential features are better fit to the Gaussian probability density function without losing the vital information. Some examples are in Figure 1 and Figure 2.

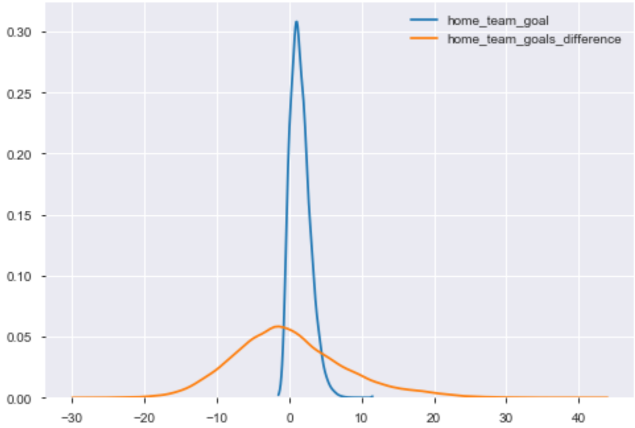

Figure 1: The graph shows how derived parameter home_team_goal_difference fits the gaussian distribution

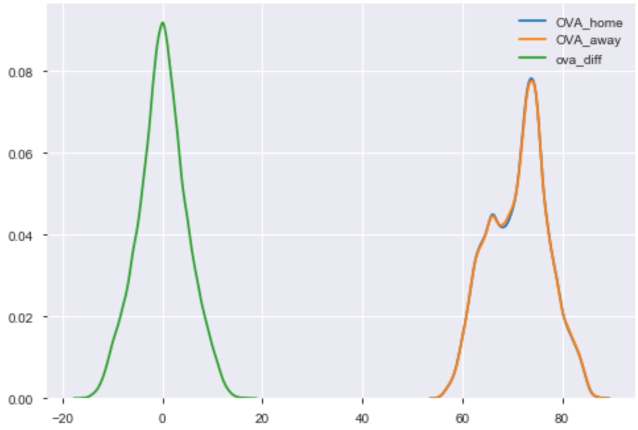

Figure 2: The graph shows how derived parameter ova_diff fits the gaussian distribution

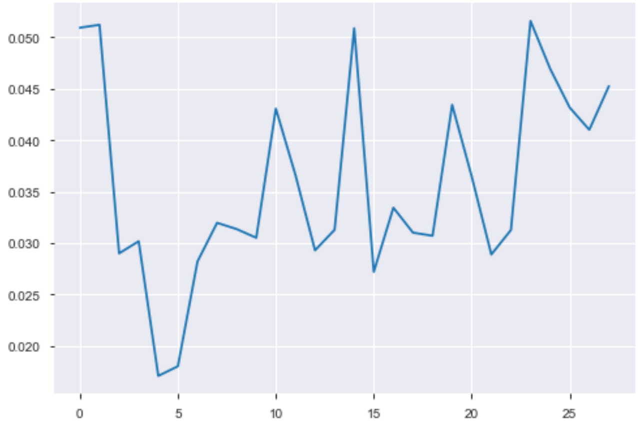

Figure 3: The graph shows the feature importance

Table 1: Feature Description

Feature name	Feature abbreviation
Home team goals difference	home_team_goals_difference
Away team goals difference	away_team_goals_difference
number of k games won by home team	games_won_home_team
number of k games won by away team	games_won_away_team
games won against opponent	games_against_won
games lost against opponent	games_against_lost
Home team overall strength	OVA_home
Home team attack strength	ATT_home
Home team midfield strength	MID_home
Home team defence strength	DEF_home
Home team transfer budget	TransferBudget_home
Home team speed rate	Speed_home
Home team dribbling rate	Dribbling_home
Home team pass building rate	BuildPassing_home
Home team position building rate	BuildPositioning_home
Home team crossing rate	Crossing_home
Home team passing creation rate	ChancePassing_home
Home team pass building rate	Shooting_home
Home team position creation rate	ChancePositioning_home
Home team aggression style	Aggression_home
Home team pressure style	Pressure_home
Home team width style	TeamWidth_home
Home team defence style	DefenderLine_home
Home team domestic prestigue	DP_away
Home team international prestigue	IP_home
Home team first 11 overal age	SAA_home
Home team overall age	TAA_home
Away team overall strength	OVA_away
Away team attack strength	ATT_away
Away team midfield strength	MID_away
Away team defence strength	DEF_away
Away team transfer budget	TransferBudget_away
Away team speed rate	Speed_away
Away team dribbling rate	Dribbling_away
Away team pass building rate	BuildPassing_away
Away team position building rate	BuildPositioning_away
Away team crossing rate	Crossing_away
Away team passing creation rate	ChancePassing_away
Away team pass building rate	Shooting_away
Away team position creation rate	ChancePositioning_away
Away team aggression style	Aggression_away
Away team pressure style	Pressure_away
Away team width style	TeamWidth_away
Away team defence style	DefenderLine_away
Away team domestic prestigue	DP_away
Away team international prestigue	IP_away
Away team first 11 overal age	SAA_away
Away team overall age	TAA_away
Overall rating statistic	ova_diff
Attack rating statistic	att_diff
Midfield rating statistic	mid_diff
Defence rating statistic	def_diff

Table 2: Extracted Features

Feature name	Feature abbreviation
Home team goals difference	home_team_goals_difference
Away team goals difference	away_team_goals_difference
number of k games won by home team	games_won_home_team
number of k games won by away team	games_won_away_team
games won against opponent	games_against_won
games lost against opponent	games_against_lost
Home team overall strength	OVA_home
Home team attack strength	ATT_home
Home team midfield strength	MID_home
Home team defence strength	DEF_home
Home team transfer budget	TransferBudget_home
Home team domestic prestigue	DP_away
Home team international prestigue	IP_home
Home team first 11 overal age	SAA_home
Home team overall age	TAA_home
Home team speed rate	Speed_home
Home team dribbling rate	Dribbling_home
Home team pass building rate	BuildPassing_home
Home team position building rate	BuildPositioning_home
Away team overall strength	OVA_away
Away team attack strength	ATT_away
Away team midfield strength	MID_away
Away team defence strength	DEF_away
Away team overall age	TAA_away
Overall rating statistic	ova_diff
Attack rating statistic	att_diff
Midfield rating statistic	mid_diff
Defence rating statistic	def_diff

4 PREDICTIVE MODELLING

We tested a basic pipeline without the derivated parameters. The preliminary results were between %35 and %40. The model that we used was the decision tree ensemble technique, random forest. Random forest is a robust machine learning algorithm which is capable of mapping complex nonlinear decision boundaries.It overcomes the high variance problem caused by decision trees through building a large number of trees by bootstrapping samples, and then taking the majority vote results in order to improve accuracy and limit over-fitting.

When building classification trees, the cost criterion for determining a split is generally either the Gini Index or Cross-Entropy. After tuning the hyper-parameters, we found that the best results were obtained using the Gini Index (G) and forest size as 50. Since random forest automatically uses Information Gain for feature selection, we extracted the most valuable features based on Information Gain. And before the predictive pipeline, we implemented PCA and reduced our feature space from 28 to 7.

Our main idea was finding anomalies and labeling them as big win and big lose according to their labels. We thought that this may give insight to our predictive model for the unpredicted psychological effects on football matches. For this purpose, we run DBSCAN on our dataset and relabeled the unclustered matches as big win and big lose. Then used the same predictive modelling to predict the unclustered events.

5 RESULTS AND CONCLUSIONS

We splitted our dataset as %80 training data and %20 test data. Then we run the predictive pipeline to see the results. It seems that derivated features are greatly increasing the accuracy as its %60.

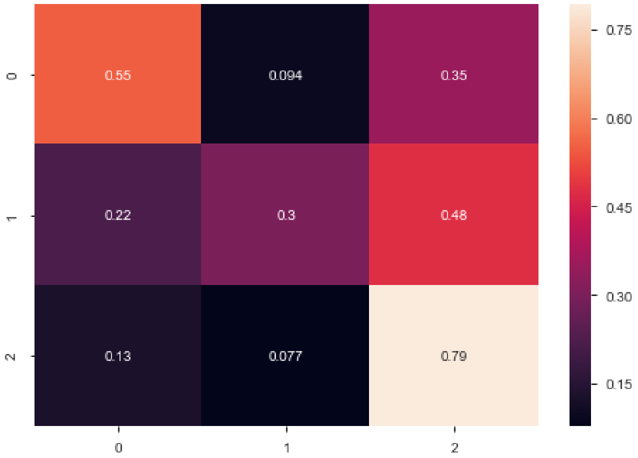

Figure 4: Confusion matrix

It is seen that engineering new features without information loss based on both intiution of dataset and probabilistic models had superior effects on the accuracy of the predictive model. Various machine learning methods were employed to make predictions. However, since the dataset was high dimensional and yet consist a low number of samples in it, tree based methods appears to perform best among others. This fact might show us that tree based methods tend to perform better for datasets like this one.

6 FUTURE WORK

The work this paper describes is rather an introduction for sports prediction. Since there are a vast number of factors coming from different distributions effecting the results, further analysis on the current datasets and gathering even more samples coming from different distributions that have effect on match outcomes would be certainly beneficial. This would enable using more advanced techniques for both predictive analysis and feature engineering, e.g. sequential pattern mining, employing more powerful machine learning pipelines etc. Psychological effects were briefly considered in this work, however, it is a known fact that these effects may influence the outcomes heavily. The 'big win' and 'big loss' factors also were touch upon in this work, since these kinds of outcomes may have the biggest impacts on the psychological effects, this could be one of future engineering improvements. Furthermore, since it is s an active research area that tempts large communities, an optimization technique can be employed to make descent predictions on not all games, but the chosen ones.

REFERENCES

[1] 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] 2010 Journal of Applied Statistics 1-13

[3] tefani R 1977 IEEE Transactions on Systems, Man, and Cyber netics 7 117-121

[4] 2000 Journal of the Royal Statistical Society: Series D (The Statistician) 49 399-418

[5] Constantinou A C, Fenton N E and Neil M 2012 Knowledge-Based Systems 36 322-339

[6] Koopman S J and Lit R 2015 Journal of the Royal Statistical Society: Series A (Statistics in Society) 178 167-186

[7] in B, Kim J, Choe C, Eom H and (Bob) McKay R I 2008 Knowledge-Based Systems 21 551-562

[8] ww.sofifa.com

[9] ww.worldfootball.net