SocceR: An Analysis of Europe Soccer Database Using R

20204880_ErnChern, 20170906_AssemZhunis
20180745_MaiTungDuong, 20204805_JoshuaJulioAdidjaja

1 Introduction

"You win by effort, by commitment, by ambition, by quality, by expressing yourself individually but in the team context."

— Jose Mourinho, professional soccer manager

Soccer is one of the most popular sports in the world because of its simple rules, but behind these rules of the game, there is an interesting complexity. Evaluating performance in soccer games is a very challenging task, as it is a team sport. All 11 players are essential to gain victory. However, a team also relies on some very best players to produce moments of magic. Some talented players can worth much more than others. On the other hand, team attributes can also shape individual players' performance. In this project, we provide a quantitative analysis of the European soccer database by first checking the significant players' and teams' attributes in different leagues. Then, a regression model was constructed to predict the victory in a match. We analyzed the strategy of different teams and cover some of the myths in soccer. Finally, we also investigated whether individual talents are directly translated to team success, or they imply the strong existence of factors other than individual talent contributing to the team victory.

2 Research Questions

As there are 3 main datasets about teams, players, and matches, we first want to do some direct analysis on these datasets:

RQ1. What are the most important features for a successful player?

RQ2. What are the most important features for a successful team?

RQ3. Can we predict the probability of a victory in a match?

We would like to also to have some deep-dives into the data with more insightful research questions:

RQ4. Does individual talent directly translate to team success?

RQ5. How did teams differ in their strategies?

RQ6. What is the role of home advantage and how it changed over time?

3 Data

We used the European Soccer Database from the Kaggle platform (2). It contains data for 11 European Countries with their lead championship and Players and Teams' attributes sourced from EA Sports' FIFA video game series, including the weekly updates. Data is collected from the 2008/09 to 2015/16 seasons with a total size of 298.59 MB. The detail of this database is shown in Appendix.

4 Methodology

For player and team attributes, all negative values and outliers (mean ± 2 standard deviation) were removed. The numeric values were scaled and the categorical values were converted to dummy variables. For players, as there are several attributes for each year for each player, we aggregated the values by their means.

To calculate the winning rate, we gave 1, 0 and 0.33 points for winning, losing and draw respectively for teams in each match in the database. We used these numbers as they are used in the real soccer scoring system. For each season we calculated the mean of points and defined the result to be the winning rate of the team.

We used the linear regression to analyze players' attributes with players' overall rating as the dependent variable. On the other hand, we applied the principal component analysis (PCA) for the teams' attributes. The features were extracted from the variables in regression model based on 0.95 significance level.

To predict the winning probability for the home team, we performed the logistic regression with the dependent variable being the match result (home team wins: 1, home team does not win: 0). The independent variables include the home and away team's features, along with the average individual features of the players in each team. We realized that many of these features are actually highly correlated, hence we decide to keep only 1 features in each correlated group. Hosmer-Lemeshow (HL) test is also utilized to observe the goodness of fit.

To investigate individual role of players in the team's success we conducted the Spearman's correlation test. First, we calculated average overall

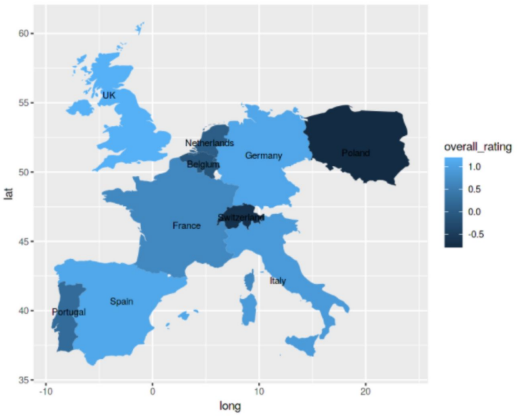


Figure 1: European Players' Overall Rating

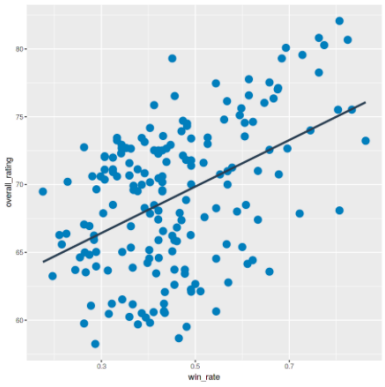


Figure 2: Winning rate vs Overall rating

Countries	Significant features	R^2
England	**height,** heading_accuracy, long_passing, ball_control, acceleration, reactions, **stamina, long_shots, aggression,** positioning, gk_positioning	0.81
Germany	heading_accuracy, short_passing, long_passing, acceleration, reactions, **balance**, strength, positioning, gk_diving, defend_medium	0.87
Spain	heading_accuracy, short_passing, **free_kick_accuracy**, long_passing, ball_control, **sprint_speed**, reactions, strength, positioning, **marking**, gk_diving, gk_positioning, **gk_reflexes, defend_high**, defend_medium	0.83

Table 1: Significant Players' Features

rating of each team's players and ranked teams according to this value. Second, we normalized these two rankings, winning rate and average player's overall rating, and, finally, performed the Spearman's test.

There are many approaches to define the home advantage (5; 6). In our work, we estimated the home advantage rate for each country by using the following equation:

$$HomeAdvantage = \frac{HomeWins - AwayWins}{HomeWins + AwayWins} \times 100\%$$

Here $HomeWins$ is the number of matches where the home teams got a higher number of goals than the away teams and draws are ignored. Same applies for the $AwayWins$. Therefore, the higher this value is, the higher is the home advantage and since it is normalized we can use it to compare the countries. Compared to the prior work, we want to see how home advantage differs across the countries and how it's contribution to team's success changed over the time.

5 Result and Discussion

5.1 Players' Performances Across European Countries

We plotted the map for average player rating for each country as shown in Fig. 1. Note that brighter color means the higher rating. The level of each country is vastly different and countries like England, Spain, and Germany really come out with the most advanced level of soccer. Hence, we decided to treat each country separately and consider England, Spain, and Germany.

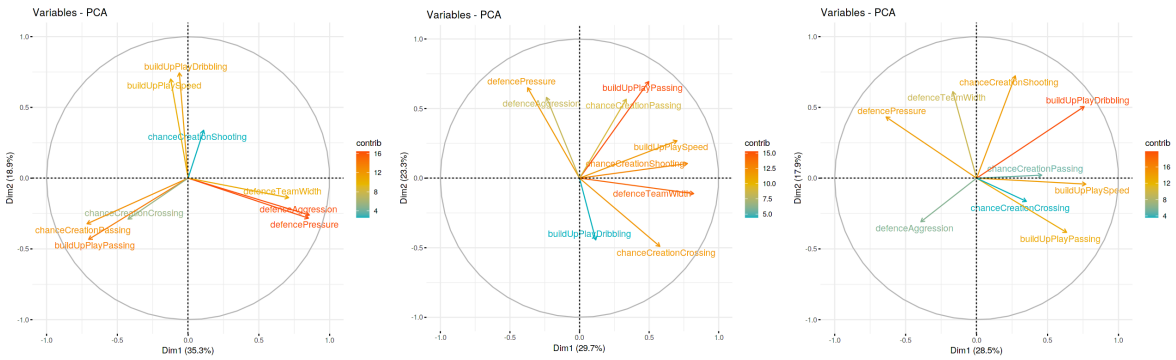
5.2 Individual Talent and Team Success

We got the Rho value of **0.39** with p-value less than 0.01. We interpreted this as a weak monotonic correlation. We may conclude that individual talent, although very important, does not guarantee the team success. Refer to the Fig. 2 to see the correlation. So further investigations on the strategies of teams are needed. We will further address this task with the PCA results in Section. 5.4.

5.3 Players' Features

We reported the importance features (using p-test score) of linear regression on players in Table. 1. We observed that there are common important features shared by all countries, such as long passing, ball control, positioning, etc. Apart from them, the bold ones are the one that is unique for each country. This reveals some insight about each country's style of soccer, for example, England appears to have an aggressive, physical style with the importance for height, aggression, stamina while Spain seems to have a more technical, skill-focused playing style.

One of the interesting insights from the linear regression model is that the 'height' variable in England is highly significant (p-value = 0.007) and has



(a) England (b) Germany (c) Spain

Figure 3: PCA results for team features

a negative coefficient (-0.1005). This implies that players in England who have good performances tend to have shorter height. This can be because of several advantages as a short soccer player, including a lower center of gravity that provides better stability and a strong counter-torque (3). Although there are still merits for a tall soccer player, such as better defense, it depends on the team strategies, as we discussed.

Another interesting insight is that Spanish league La Liga's *'defend medium'* and *'defend high'* variables are highly significant (p-values are 0.018 and 0.00024 respectively), but they have negative coefficients (-0.19 and -0.24 respectively). This is out of our expectations, as it suggests a better defense can bring a disadvantage in the league. By investigating further, it is actually a long debate among fans whether La Liga is a weak defensive league and focuses more towards technical attacking ability (4). Our finding can support this idea that the league focuses on having more goals rather than being defensive.

5.4 Teams' Features

In the England Premier League (Fig. 3a), team features can clearly be divided into 3 clusters. Teams in the upper cluster show their focus towards the individual player's abilities, such as dribbling and speed. Meanwhile, the lower-left cluster put emphasis on passing and crossing or we could say as teamwork. On the lower right cluster lies the teams that have a defensive playing style.

From Fig. 3, we can see that almost all features have a high contribution to explaining the distribution of the data, except one, "chance Creation Shooting". Teams in the England Premier League have low variance in the shooting feature. It could either mean shooting is not an important feature in this league, or shooting is really important so that every team has a high level of shooting ability. We then used a linear regression model between these features with each team's winning rate. The result shows the importance of buildUpPlayPassing, buildUpPlayDribbling, and defenseAggression (p-value < 0.1). On the other hand, chanceCreationShooting has a low impact on the winning rate, therefore, not an important feature.

However, there are no clear patterns or clusters in Germany and Spain compared to England. But using a linear regression model in Germany, we noticed that buildUpPlayDribbling has a significant relation with the winning rate (p-value < 0.01). Meanwhile, in Spain, the PCA shows high variance in buildUpPlayDribbling, but the linear regression model shows a quite significant relation with the winning rate. Overall, we can infer that buildUpPlayDribbling is the most important feature across these three leagues.

5.5 Win prediction

Predicting the probability of victory of a match for home teams was done by logistic regression. We achieved 73% accuracy in Spain with the confusion matrix shown in Fig. 4. HL test reveals the goodness of fit about 0.7.

However, other countries' results are not as good. We concluded that it is not trivial to predict the match result based on the teams and players quality alone. There are many external factors that can affect the match result. Spain turns out to be the most predictable leagues, while England and Germany is quite hard to make an educated guess. We conjectured that the leagues in England and Germany are more competitive and the teams are actually

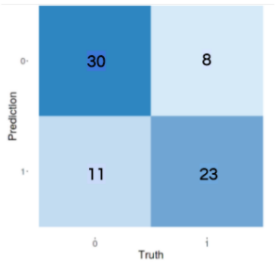
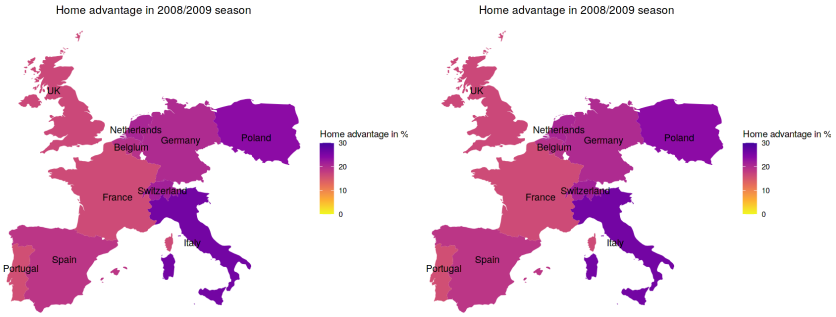
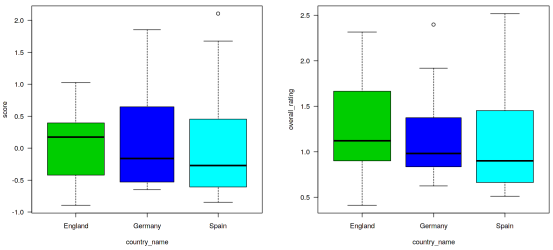


Figure 4: Confusion matrix for Spain win prediction



(a) 2008/2009 (b) 2015/2016

Figure 5: Home advantage map in two seasons



(a) Score (b) Overall rating

Figure 6: Distribution of teams score and overall rating

closer to each other. We tested the conjecture by looking at the variance of score and overall rating of the team in each country. The plots are shown in Fig. 6. We can see that Spain has significantly bigger variance (wider quality gap) in score compared to England, and in overall rating compared to Germany. This confirms our conjecture that Spain has generally wider gap between team, thus more predictable.

5.6 Home Advantage

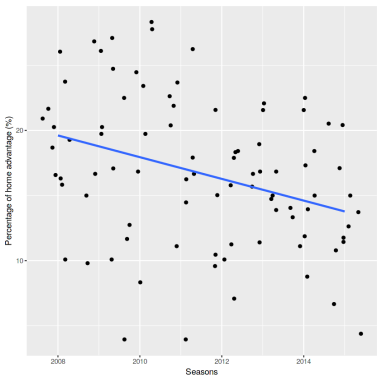


Figure 7: Home advantage in Europe from 2008-2016

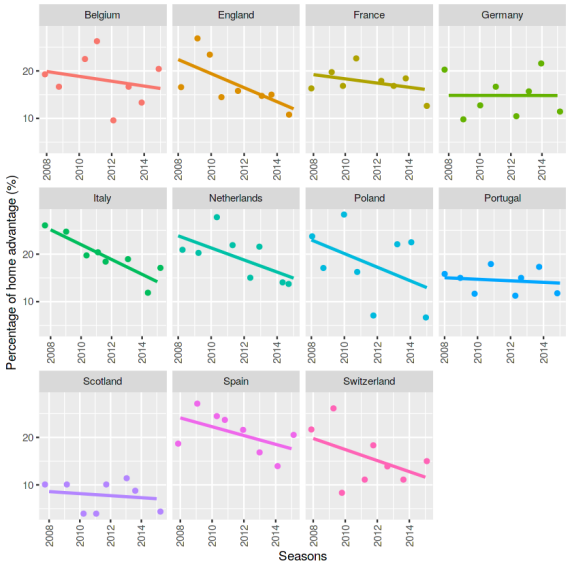


Figure 8: Home advantage in each country

We plotted the home advantage as a dependent variable over time in all countries from 2008/09 to 2015/16 seasons. We obtained a negative slope of -0.8339 with a p-value less than 0.01 (Fig. 7). We can also see the negative trend from the maps in the Fig. 5. The darker color in the map shows the higher home advantage. We can clearly see that colors are becoming brighter after 8 years. So, we may conclude that over time home advantage in Europe has been decreasing, meaning that teams are becoming more professional and are not affected much by home advantage.

We visualize how home advantage changes in each country in Fig. 8. We can see that most of the countries faced a dramatic drop in home advantages like Switzerland and Poland, while countries like Portugal and Germany did not show any change in the 8-year period.

6 Limitations and Future work

One of the limitations is that our data spans only from 2008 to 2016, so we are lacking more recent data. Therefore, our prediction model might not be applicable for nowadays matches.

Another drawback is that we consider only top 3 leagues from the dataset, which limits the applicability of our prediction model. We think that it would be interesting to conduct same analysis on the leagues in the bottom of the ranking. We may identify the differences between the top and bottom teams and come up with the further strategies to improve results of the weak teams.

7 Conclusion

As we expected, the Spearman's correlation test implies that the individual talent of players is important but does not guarantee the success of a team. By analyzing the important features of players in deciding their overall rating, and the important features of teams in deciding their winning rate, we provided an insight on how strategies are different in three different European leagues. We also discussed some of the myths and debates among soccer fans, including home advantage and height of the players. Overall, we are able to find interesting results and insights in soccer games, by applying some techniques in R, but there are still a lot more to be learned by investigating further on different strategies of teams over different seasons.

8 Demo

A readily-runnable demo of our method is available at https://www.kaggle.com/maitungduong/soccer

References

[1] Project repository with analysis codes and outputs. https://github.com/assemzh/socceR

[2] Mathien, H. (2016). European Soccer Database. Kaggle. https://www.kaggle.com/hugomathien/soccer

[3] Parrish, R. (2013). The Advantages of Short Soccer Players. SportsRec. https://www.sportsrec.com/1006527-advantages-short-soccer-players.html

[4] Atkinson, T. (2013). Is La Liga Really a Weak Defensive League? Bleacher Report. https://bleacherreport.com/articles/1527423-is-la-liga-really-a-weak-defensive-league

[5] Home advantage in soccer: A retrospective analysis https://doi.org/10.1080/02640418608732122

[6] Comparison of Home Advantage in European Football Leagues https://doi.org/10.3390/risks8030087

Leagues	Matches	Teams	Players	
1. Belgium Jupiler League 2. England Premier League 3. France Ligue 4 .Germany 1. Bundesliga 5. Italy Serie A 6. Netherlands Eredivisie 7. Poland Ekstraklasa 8. Portugal Liga ZON Sagres 9. Scotland Premier League 10. Spain LIGA BBVA 11. Switzerland Super League	Number: 25,979 Attributes: 1. date 2. home_team_api_id 3. away_team_api_id 4. home_team_goal 5. away_team_goal	Number: 299 Attributes: Build Up Play Attributes 1. Speed 2. Dribbling 3. Passing 4. Positioning Class Chance Creation Attributes: 5. Passing 6. Crossing 7. Shooting 8. Positioning Class Defence Attributes: 9. Pressure 10. Aggression 11. TeamWidth 12.Defender Line Class	Number: 11,060 Attributes: 1. overall_rating 2. potential 3. preferred_foot 4. attacking_work_rate 5. defensive_work_rate 6. crossing 7. finishing 8. heading_accuracy 9. short_passing 10. volleys 11. dribbling 12. curve 13. free_kick_accuracy 14. long_passing 15. ball_control 16. acceleration	17. sprint_speed agility 18. reactions balance 19. shot_power jumping 20. stamina strength 21. long_shots aggression 22. interceptions 23. positioning vision 24. penalties marking 25. standing_tackle 26. sliding_tackle 27. gk_diving 28. gk_handling 29. gk_kicking 30. gk_positioning 31. gk_reflexes

Table 2: European Soccer Database description