2023

Club Soccer Prediction – Final White Paper

WEEK 8 – PROJECT 2 – MILESTONE 3

ABDUL BASIT

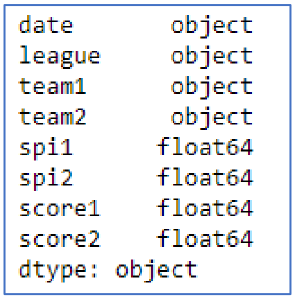
Business Problem

The Club Soccer Predictions Ire first published in January 2017 by https://fivethirtyeight.com with only six leagues. Since then, they have steadily expanded the number of leagues they forecast, added features to their interactive graphics, tweaked their predictive model to perform better and published their global United European Football Association (UEFA) club soccer rankings. [1]

The forecasts are based on a substantially revised version of ESPN's **Soccer Power Index (SPI)**, a rating system originally devised by FiveThirtyEight in 2009 for rating international soccer teams. They have updated and adapted SPI to incorporate club soccer data (for more than 550,000 matches in all) that they collected from ESPN's database and the GitHub repository, as Ill as from play-by-play data that has been available since 2010.

Dataset Explanation

- CSV file: spi_matches.csv contains match-by-match SPI ratings and forecasts back to 2016.
- website: https://projects.fivethirtyeight.com
- API: https://projects.fivethirtyeight.com/soccer-api/club/
- SPI field which is common in all the three datasets will be my reference value.



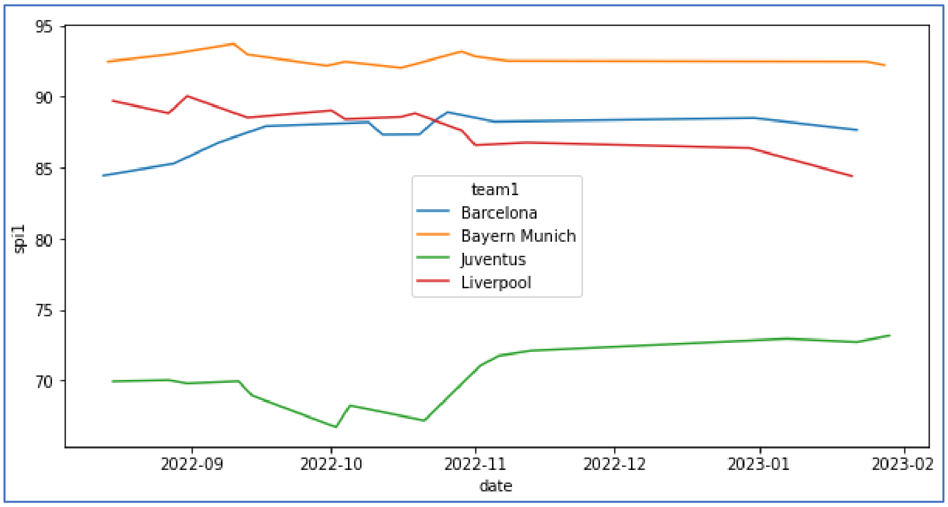
```
date         object
league       object
team1        object
team2        object
spi1        float64
spi2        float64
score1      float64
score2      float64
dtype: object
```

Methodology

Dataset shows that every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating, which represents the percentage of available points — a win is worth 3 points, a tie worth 1 point, and a loss worth 0 points - the team would be expected to take if that match Ire played over and over again.

Given the ratings for any two teams, I plan to project the result of a match between them in a variety of formats — such as a league match, a home-and-away tie or a cup final — simulate whole seasons to arrive at the probability each team will win the league, qualify for the UEFA Champions League or be relegated to a lower division. [2]

The sheer volume of matches taking place at some times of the year can be paralyzing. With that in mind, so it was necessary to split the rate upcoming matches on their quality and importance.



Importance is a measure of how much the outcome of the match will change each team's statistical outlook on the season. This outlook considered different factors depending on which league the match is being played in; for some leagues, the outlook only considered winning the league, while other leagues incorporate the possibility of being promoted or relegated, or qualifying for the Champions League. To calculate the importance of a match to a team, I generated probabilities for each factor conditional on winning (or losing) the match, and then found the difference between those two possible numbers. I used the factor with the maximum range of difference for each team and scale the result to between 0 and 100. Finally, I averaged the match's importance to both teams to find the overall match importance. All leagues were treated equally when calculating importance, so a match to decide the winner of the Swedish League would rate just as high as a match to decide the winner of the English Premier League.

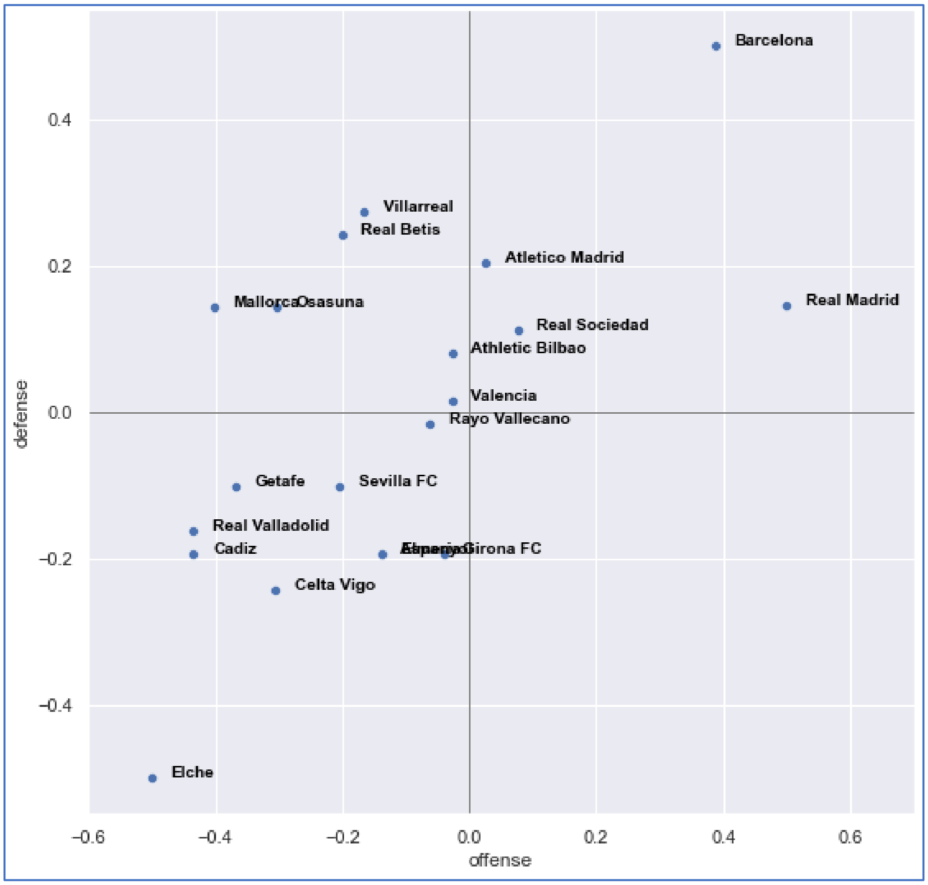
As of 2020, match predictions incorporated importance in two ways

1) When a match was more important to one team than the other, that team tended to outperform expectations, with its boost in performance relative to how much more important the match was to them.
2) If a match was not important to either team, uncertainty in the outcome of the match increased.

Analysis

Dataset shows that every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating, which represents the percentage of available points — a win is worth 3 points, a tie worth 1 point, and a loss worth 0 points - the team would be expected to take if that match Ire played over and over again. [3]

Given the ratings for any two teams, I plan to project the result of a match between them in a variety of formats — such as a league match, a home-and-away tie or a cup final — simulate whole seasons to arrive at the probability each team will win the league, qualify for the UEFA Champions League or be relegated to a lower division. [4]



Conclusion

My goal in working with this data was to get some more practice with pandas and seaborn, so from that perspective, this project was a complete success. In the future I look forward to exploring more of the data used by 538 in conjunction with the statistics and programming skills that I am currently learning.

Future Work

Further studies on this topic can extend the research by addressing several relevant questions. For example, there is a lack of information regarding to official youth matches since the positioning tracking systems have been used only in senior professional teams. This study intended to overcome this issue; however, formal competitive environments should be used in future studies to provide a step forward insight. The present exploratory data-approach should also be applied to a wide range of contextual variables, such as game status, different teams' tactical formations, playing home vs. playing away, etc. These different contexts may afford different collective behaviors understanding which, in turn, will enrich the performance programs development.

Answers to Questions an audience may ask

- Can we predict the results of a football game before it starts?

  Yes, by using the appropriate models like Multinomial Classification model and Generalized Linear Model (GLM) we can transform the datasets into an H2O environment and transform the label column of the results into a factor.
- How much does a player's performance affect the results of the match?

  This can be achieved by developing a Machine Learning model that is based on the performance data of each player, generated by Globo's Fantasy Game named CartolaFC, and the conditions that affect each match.
- Can we model a Machine Learning algorithm to predict the game's final result?

Yes, using good predictive power, with satisfactory training and validation performance showing indications that it can be refined and used to predict soccer matches using the Fantasy game data as input.

- How far can the predictive power of an Artificial Intelligence get?

  H2O package in Python is a very powerful library prepared only for the development of projects in the area of Artificial Intelligence.

- How does SPI change over time?

  Finding out the top 10 ranked clubs based on SPI and then also look at a slice of mid-range clubs we can expect to see a lot of them staying near the top over time.

- How can we compare different leagues?

  To make a more cohesive visualization, we can limit the teams to members of the top 5 leagues, as opposed to the top 50 teams overall. While this would add more teams in total, it would also produce a much clearer pair plot, making it much easier to understand the visualizations.

References

[1] Boice, J. (2019, June 5). *How Our Club Soccer Projections Work*. FiveThirtyEight. https://fivethirtyeight.com/features/how-our-club-soccer-projections-work

[2] AFootballReport. (n.d.). *Predicting the World Cup with the soccer power index (SPI)*. Football Predictions, Tips and Stats. Retrieved January 28, 2023, from https://afootballreport.com/post/88490446802/predicting-the-world-cup-with-the-soccer-power

[3] Published by Statista Research Department, & 6, J. (2023, January 6). *Soccer power index - best soccer teams 2023*. Statista. Retrieved January 28, 2023, from https://www.statista.com/statistics/808025/best-soccer-club-teams-worldwide/

[4] Draths, T. (2021, August 11). Predicting soccer team strength - version II. Medium. Retrieved January 28, 2023, from https://towardsdatascience.com/predicting-soccer-team-strength-version-ii-11b5c66cf9d8