Club Soccer Prediction Analysis

DSC680-Appled Data Science

Project 2 - Milestone 3 – Final Presentation

Abdul Basit

abasit@my365.bellevue.edu

Background

- The Club Soccer Predictions were 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.
- 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 well as from play-by-play data that has been available since 2010.

Dataset

- 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
```

Data Exploration

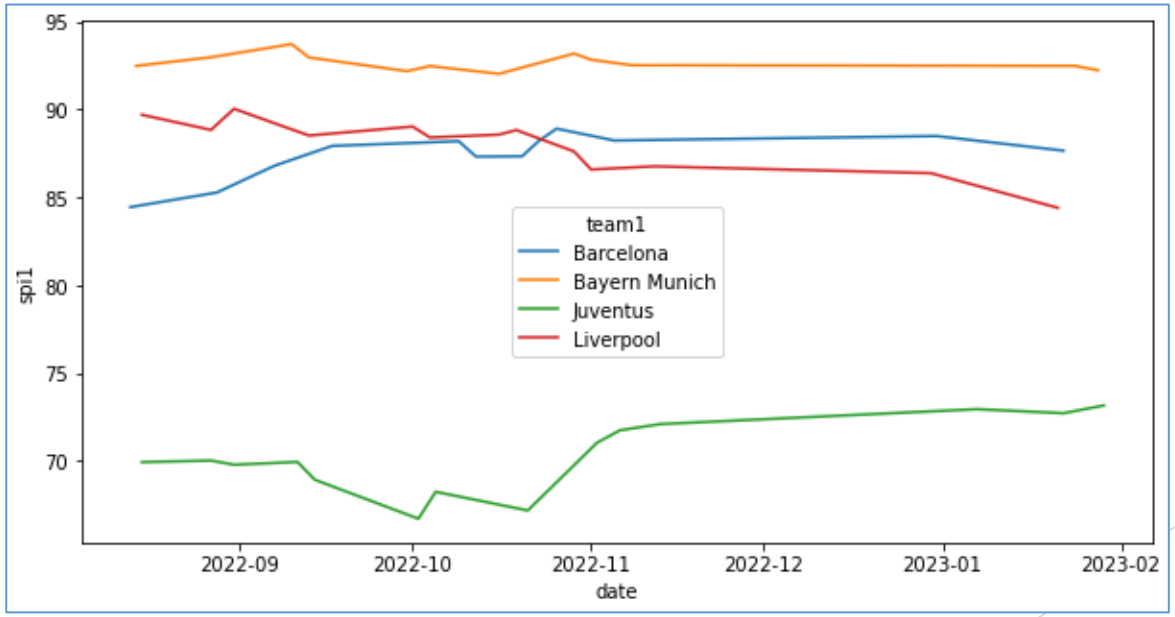
- Dataset shows that every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating, which represents the percentage of available points.

- 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.

	season	date	league_id	league	team1	team2	spi1	spi2	prob1	prob2	...	importance1	importance2	score1	score2	xg1	xg2	nsxg1	nsxg2	adj_score1	adj_score2
2534	2022	8/15/2022	1854	Italy Serie A	Juventus	Sassuolo	69.94	60.30	0.5443	0.2274	...	47.8	28.6	3.0	0.0	2.11	1.35	0.64	1.29	3.15	0.00
3009	2022	8/27/2022	1854	Italy Serie A	Juventus	AS Roma	70.03	73.60	0.3711	0.3461	...	54.2	61.3	1.0	1.0	0.55	1.08	0.59	0.58	1.05	1.05
3230	2022	8/31/2022	1854	Italy Serie A	Juventus	Spezia	69.79	46.93	0.6571	0.1278	...	42.3	36.3	2.0	0.0	1.65	0.17	1.22	0.57	1.58	0.00
3846	2022	9/11/2022	1854	Italy Serie A	Juventus	Salernitana	69.95	49.05	0.6402	0.1379	...	43.9	34.2	2.0	2.0	2.30	1.78	1.88	0.75	2.10	2.10
3936	2022	9/14/2022	1818	UEFA Champions League	Juventus	Benfica	68.94	76.93	0.3637	0.3661	...	100.0	100.0	1.0	2.0	0.95	2.51	1.47	2.19	1.05	2.10

Data Analysis & Visualization

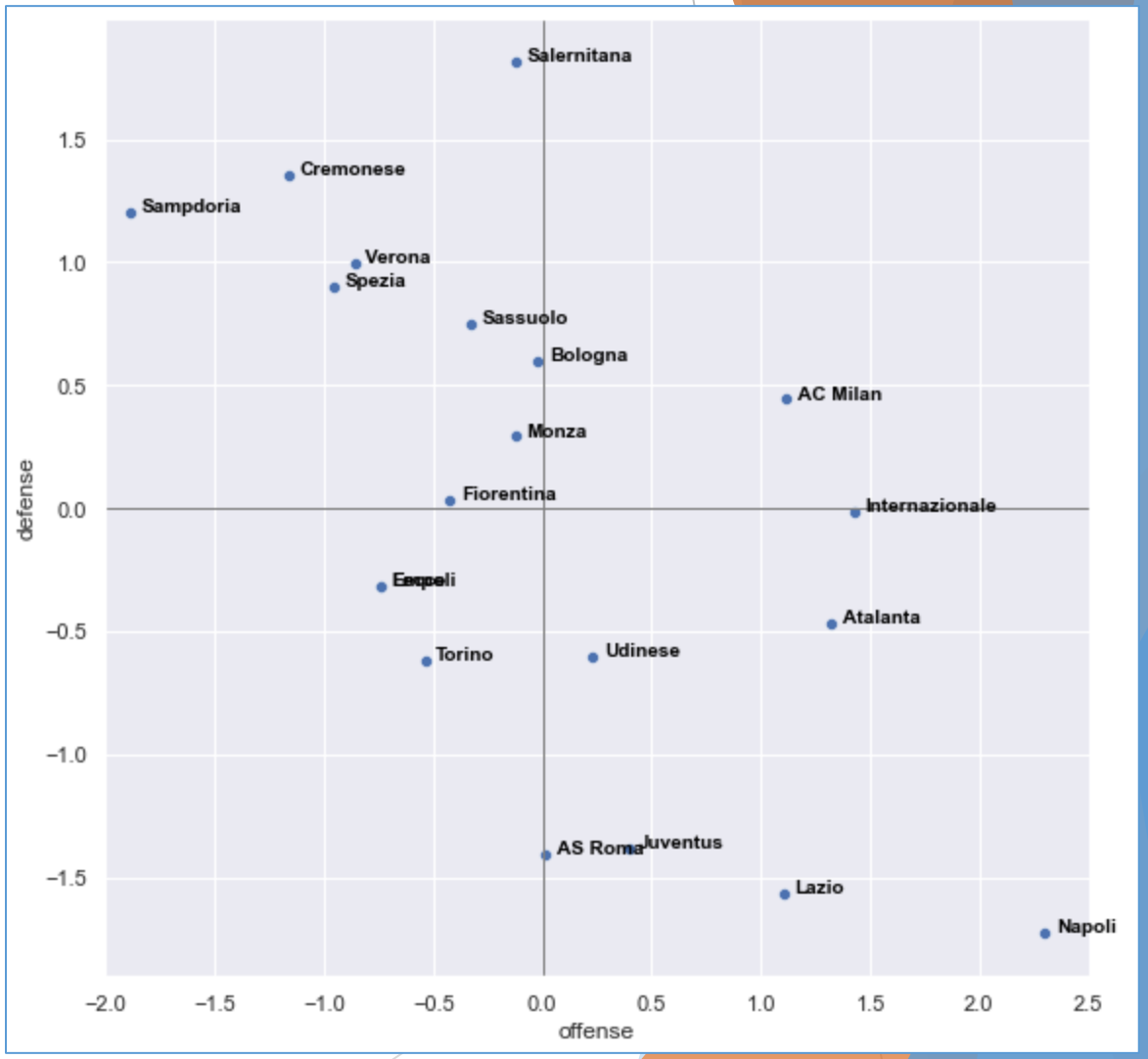
- *Quality* simply a measure of how good the teams are. Specifically, it's the harmonic mean of the two teams' SPI ratings because every team has an SPI rating between 0 and 100, match quality also ranges from 0 to 100.
- *Importance* is a measure of how much the outcome of the match will change each 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.



Methodology

- To better estimate each team's underlying quality of play, three metrics are being used to evaluate a team's performance after each match: *adjusted goals*, *shot-based expected goals* and *non-shot expected goals*.

- A team's composite offensive score for that match is an average of its performance across the three metrics, and its composite defensive score is an average of the three metrics for its opponent.



Conclusion

Almost all of the teams are clustered fairly low for expected goals and actual goals, but there are two teams that are among the top scoring teams overall. This could be a result of the top teams skewing the data by dominating all the other teams so much, or they could just be outliers that don't impact the other teams so much.

Warmer color means better score. The row will show the scores with each team. If the row shows a lot of warm color, means this team can usually out score the others. This graph indicates Barcelona, Athletico Madrid, and Real Madrid are the top teams, while Elche, Espanyol, and Getafe are on the other end.