Name: Abdul Basit
Date: 30th May 2020
Course: DSC540 – Data Preparation
Assignment: Milestone 5 – Week 11 & Week 12
Title: Club Soccer Predictions – Lessons Learned

Introduction

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.

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.

Data Sources

1) **CSV file: spi_matches.csv** contains match-by-match SPI ratings and forecasts back to 2016.
2) **website:** https://projects.fivethirtyeight.com
3) **API:** https://projects.fivethirtyeight.com/soccer-api/club/

SPI field which is common in all the three datasets will be my reference value.

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.

Below are some steps I intend to perform in different milestones:

1) *Milestone 1* – Clean the datasets and filter only fields of interest. Remove NAs or duplicates. Calculate domestic team Soccer Power Index (SPI) ratings with respect to time.
2) *Milestone 2* – As a sample, plot the SPI score of couple top teams in order to illustrate the performance of each team since past 3 years.
   Look at each inter-league match from the past five years and calculate the expected score of the match based purely on each team's domestic rating at the time.

3) *Milestone 3* – Calculate a simple Defense and Offence Index and visualize data in such way to display that higher score should represent better performance.
   Take the difference between the expected score of the match and the actual score and run these results to find a rating for each league, expressed in how many goals better or worse than the global average that league is.
4) *Milestone 4* - Regress these calculated ratings toward market-value based ratings, lighted by how many inter-league matches are there for each league.
5) *Milestone 5* - Incorporate league strengths into the predictions for any inter-league matches to improve the final team ratings.
   Along the way, develop some heat maps or scatter plots to display score comparison of teams and how a team outperforms other teams.

MODEL TWEAK – LESSONS LEARNT

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.

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. 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.

To understand the magnitude of these importance adjustments, consider a match that is equally important to the two teams, where the home team has a 50 percent chance of winning the match, the away team has a 25 percent chance of winning the match, and the remainder is the chance of a draw.

If, instead, I assume that it is a hugely important match for the home team, and a meaningless match for the away team, the home team's chances of winning would go up to 58 percent, and the away team's chances would go down to 18 percent.

On the other hand, if the match was meaningless to both teams, the home team's chances of winning would go down to 43 percent, and the away team's chances would go up to 30 percent.

The improvement I saw in our match forecasts when incorporating match importance was about one-third the size of the improvement I saw when I added expected-goals metrics in 2016, and about one-half the size of the improvement I saw when I incorporated market values in preseason ratings for 2017.8