**Name**: Abdul Basit
**Date:** 15TH January 2023
**Course:** DSC680 – Applied Data Sciences
**Assignment:** Milestone 1 – Week 5
**Action:** Project2 – Proposal and Data Selection
**Title:** Club Soccer Predictions

## 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.

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.

## Questions

1) Can we predict the results of a football game before it starts?
2) How much does a player's performance affect the results of the match?
3) Can we model a Machine Learning algorithm to predict the game's final result?
4) How far can the predictive power of an Artificial Intelligence get?
5) How does SPI change over time?
6) How can we compare different leagues?

## Challenges/Issues

The initial model presented 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 Cartola FC data as input. The model can be used for fun or even for betting. I will continue refining the model and testing the predictions in other matches. I am also available for tips, insights and questions. I hope I have helped in this journey through the Data Science and Machine Learning, here are some improvements that can be made to a future article.

## References


1) https://www.kaggle.com/search?q=club+soccer+prediction+notebookLanguage%3APython
2) https://github.com/fivethirtyeight/data/tree/master/soccer-spi
3) https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8
4) https://levelup.gitconnected.com/predicting-real-soccer-matches-using-fantasy-game-scouts-a3b388edb8aa