Course Project#2

Title: Club Soccer Predictions

Project Description:

This project focuses on predicting future football team performance by analyzing historical Soccer Power Index (SPI) data. SPI is a comprehensive rating system that evaluates teams based on offensive and defensive capabilities, providing insights into their expected performance against average opponents. Initially developed by FiveThirtyEight in 2009 for international soccer and expanded to club soccer in January 2017, SPI now covers numerous leagues globally. The model has been continuously refined, incorporating data from over 550,000 matches sourced from ESPN's extensive database, GitHub repositories, and detailed play-by-play records since 2010. This expansion has enhanced predictive accuracy and allowed for more detailed global rankings, especially for UEFA club soccer.

The core dataset for this project, obtained from FiveThirtyEight's publicly available spi_matches.csv file, offers a rich historical record of match-by-match ratings and forecasts dating back to 2016. It includes essential metrics such as offensive and defensive ratings, expected goals scored, and goals conceded. These metrics will enable a comprehensive analysis of how team performance evolves over time and how various factors influence outcomes.

1.1. 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_matches.csv

Problem Statement:

- *How does SPI change over time?*

I intend to look at the top ten ranked clubs (as of 12/2/19) based on SPI, and then also look at a slice of mid-range clubs.

- *Comparing different leagues?*

There will be too many different leagues. It may be difficult to clean any useful information from the graphs because there will be a lot of information packed into one pair plot.

Dataset Details:

The dataset uses several different metrics to track performance and make predictions. The main metric calculated by dataset is referred to as SPI(Soccer Power Index). SPI is calculated by looking at a team's expected goals scored compared to the same team's expected goals conceded. Both of these predictive metrics are based on expected performance against an average team at a neutral venue. SPI is then determined as the percentage of games won against an average team at a neutral venue, given the team's expected goals for and goals against. A top European team would have an SPI value above 70, meaning that they would beat an average team 70% of the time at a neutral location. SPI would allow us to compare two teams based on general quality, however SPI does not take into account factors such as injuries or location. Throughout the season, SPI changes based on a team's performance, specifically based on a weighted total goals and overall result. The dataset is extremely unbalanced. Even a "null" classifier which always predicts class=0 would obtain over 99% accuracy on this task. This demonstrates that a simple measure of mean accuracy should not be used due to insensitivity to false negatives.

Method:

The methodology for this project is to systematically process and analyze football team performance data to predict team rankings for the year 2029. The process begins with an in-depth analysis of the dataset to identify relevant features contributing to team performance, such as offensive and defensive ratings, SPI ratings, historical performance data, and match outcomes. Irrelevant or redundant features are discarded, and the data is cleaned to address missing values, outliers, and inconsistencies. Data normalization and standardization are applied where necessary to prepare for effective model training.

Once the data is prepared, it is divided into training, validation, and test sets to ensure accurate and unbiased model evaluation. Feature engineering techniques are employed to enhance the predictive capability of the dataset, including creating composite metrics and adjusting for home/away performance variations. Suitable machine learning algorithms, such as Random Forest, Gradient Boosting, or Neural Networks, are selected for model training, with hyperparameters optimized through cross-validation. Offensive and defensive ratings, along with SPI ratings, serve as key input variables to simulate match outcomes and season projections.

The trained model is utilized to simulate the 2029 season, predicting match results and calculating the probability of each team winning the league, qualifying for international competitions, or facing relegation. These outcomes are aggregated to generate projected team rankings for 2029. To support this, comprehensive exploratory data analysis (EDA) is conducted to identify trends, patterns, and key factors impacting team performance. Visualizations are created to highlight the relationships between offensive/defensive ratings and overall SPI ratings, providing clear insights into performance drivers.

Additionally, upcoming matches are assessed based on their quality—calculated using the harmonic mean of both teams' SPI ratings—and their importance, measured by the impact on a team's seasonal outlook. Various match formats, including league matches, home-and-away ties, and cup finals, are simulated to understand their influence on team rankings. Finally, model predictions are validated against historical data and known trends, providing actionable insights to support strategic decision-making for future seasons. This structured methodology ensures the project effectively utilizes data-driven approaches to deliver accurate and insightful predictions for football team rankings in 2029.

Ethical Problem:

Analyzing football team performance data raises several ethical concerns that must be carefully addressed. Data privacy is paramount, requiring strict adherence to data protection regulations to ensure that sensitive and personal information is securely handled and not misused. Bias within the dataset is another critical issue; historical disparities between teams or leagues could lead to skewed or unfair predictions if not properly mitigated. Ensuring fairness in model development through balanced data representation and implementing bias detection methods is essential. Additionally, transparency in how predictions are generated is vital to build trust among stakeholders. The responsible use of model outcomes is also necessary to prevent misuse in areas like gambling, manipulative media narratives, or unfair decision-making in team management. Ethical safeguards must be established to ensure the analysis serves positive and constructive purposes within the sports industry.

Challenges/Issues:

Several challenges and issues may arise during the analysis and prediction of football team performance. One major challenge is the availability and quality of data. Incomplete, inconsistent, or outdated data can hinder model accuracy and reliability. Additionally, integrating data from various sources may introduce compatibility issues, requiring significant preprocessing and normalization efforts. Another challenge is managing the inherent unpredictability of sports, as unforeseen factors like player injuries, managerial changes, or unexpected match conditions can drastically affect outcomes and are difficult to model accurately. Model overfitting is also a concern, where the model may perform well on historical data but fail to generalize to future scenarios. Computational complexity and resource constraints may further impact the efficiency of large-scale simulations. Lastly, addressing biases in the data and ensuring fair and transparent predictions pose ongoing challenges that must be managed carefully throughout the project.

References:

1. FiveThirtyEight, "ABC News FiveThirtyEight," ABC News, Accessed: Sept. 24, 2024. [Online]. Available: https://abcnews.go.com/538
2. Kaggle. (n.d.). *Club soccer prediction* [Python notebooks]. Retrieved from https://www.kaggle.com/search?q=club+soccer+prediction+notebookLanguage%3APython
3. FiveThirtyEight. (n.d.). *Soccer SPI data* [Data set]. GitHub. Retrieved from https://github.com/fivethirtyeight/data/tree/master/soccer-spi
4. Santos, V. (2019, October 23). *Machine learning algorithms for football prediction using statistics from Brazilian Championship.* Towards Data Science. Retrieved from https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8
5. Basu, S. (2020, August 10). *Predicting real soccer matches using fantasy game scouts.* Level Up Coding. Retrieved from https://levelup.gitconnected.com/predicting-real-soccer-matches-using-fantasy-game-scouts-a3b388edb8aa