Course Project#2

Title: Club Soccer Predictions

Business Problem:

The primary business problem addressed in this project is to improve the predictive accuracy of football team performance using historical Soccer Power Index (SPI) data. Accurate predictions can provide valuable insights to various stakeholders, including team managers, league organizers, sports analysts, and bettors. However, understanding how SPI evolves over time and varies across leagues poses a significant challenge. Specifically, the project aims to analyze the stability and fluctuations of SPI for top-ranked and mid-range clubs, identifying key factors influencing these changes. Additionally, the project seeks to compare SPI trends across different leagues to uncover meaningful patterns, despite the complexity introduced by the vast amount of data and diverse league characteristics. Solving these problems will enable stakeholders to make more informed decisions regarding team strategy, league performance, and overall resource allocation in the competitive world of football.

Problem Statement

This project seeks to address key questions related to the evolution and variability of Soccer Power Index (SPI) data over time and across leagues. Specifically, the focus is on understanding:

1. **How SPI changes over time**: Analyzing trends for the top ten ranked clubs (as of 12/2/19) alongside mid-range clubs to identify stability and fluctuations in team performance and the factors driving these changes.

2. **Comparing different leagues**: Evaluating SPI trends across various leagues to uncover meaningful insights about league characteristics, such as team balance, strategy, and overall strength. The large volume of data and the diversity of leagues present challenges in visualizing and extracting actionable information.

By addressing these questions, the project aims to enhance the understanding of team and league performance dynamics, contributing to improved predictive modeling and informed decision-making for stakeholders.

Background/History

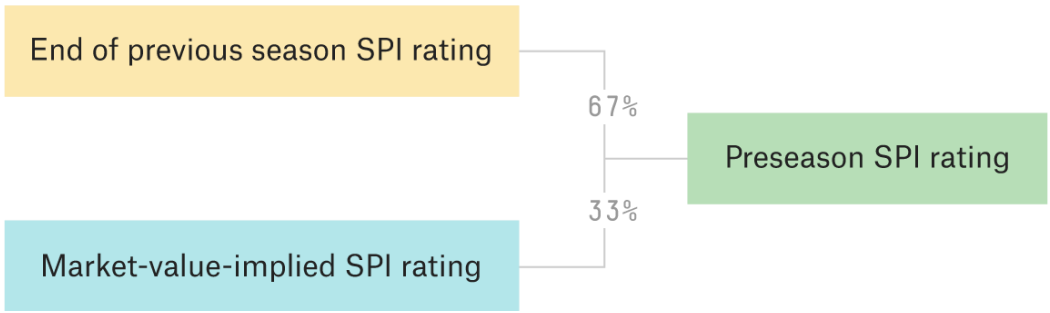
Soccer Power Index (SPI) is a comprehensive rating system initially developed by FiveThirtyEight in 2009 to evaluate the performance of international soccer teams. The SPI model assigns ratings to teams based on their offensive and defensive capabilities, offering insights into expected performance against average opponents. In January 2017, FiveThirtyEight expanded the scope of SPI to include club soccer, covering six leagues initially. Since then, the platform has broadened its coverage to encompass numerous leagues globally, continuously refining its predictive model and enhancing interactive visualizations to provide more accurate forecasts and detailed rankings, particularly for UEFA club soccer.

The SPI model integrates data from over 550,000 matches, sourced from ESPN's extensive database, GitHub repositories, and play-by-play records dating back to 2010. The model has been adapted over time to account for the nuances of club soccer, leveraging detailed metrics such as offensive and defensive ratings, expected goals scored, and goals conceded. The dataset for this project, specifically the spi_matches.csv file from FiveThirtyEight, provides a rich historical record

of match-by-match ratings and forecasts dating back to 2016. This historical data forms the foundation for understanding team and league performance trends, enabling predictive insights that inform strategic decisions in the football ecosystem.

Dataset Explanation

The dataset utilized in this project is derived from FiveThirtyEight's Soccer Power Index (SPI) dataset, which tracks team performance and predicts outcomes using a variety of metrics. The primary metric, SPI, serves as a measure of a team's overall quality and is calculated based on expected goals scored (offensive capability) versus expected goals conceded (defensive capability). These expectations are modeled as if the match is played against an average team at a neutral venue. SPI reflects the percentage of games a team would win under such conditions. For example, a top European team with an SPI value above 70 would defeat an average team 70% of the time in a neutral setting. However, SPI does not account for external factors such as injuries or location, which can influence actual match outcomes.



The dataset dynamically updates SPI values throughout the season, reflecting a team's performance as influenced by results and weighted totals of goals scored and conceded. The dataset includes columns that provide granular details about each match, including the season, date, league, teams involved, SPI ratings (spi1, spi2), probabilities of winning, losing, or drawing

(prob1, prob2, probtie), projected scores (proj_score1, proj_score2), and actual scores (score1, score2). Additional metrics, such as adjusted scores and expected goals (xg1, xg2), offer deeper insights into match performance.

Despite its rich detail, the dataset is highly imbalanced. For instance, a naive classifier that always predicts a single class (e.g., class=0) would achieve over 99% accuracy due to the skewed distribution of results. This imbalance highlights the need for careful evaluation using metrics beyond mean accuracy, ensuring sensitivity to false negatives and capturing the full complexity of the data.

- **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

Methodology

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.

- **Data Preparation**

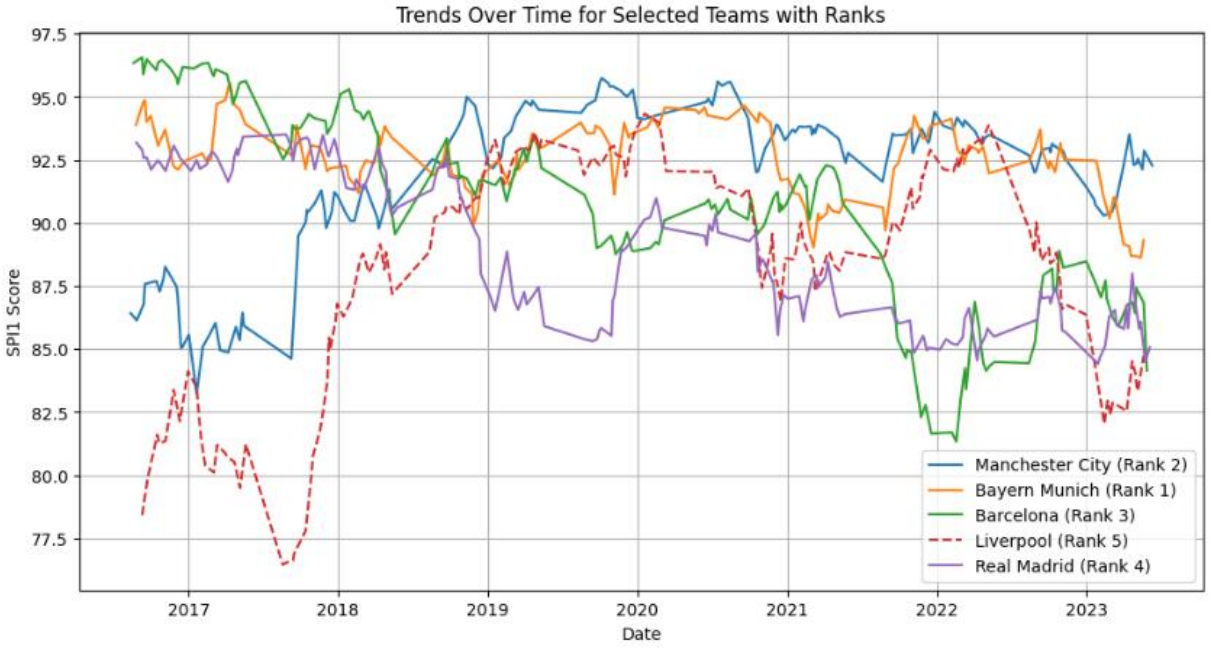
  The data preparation involved several steps to ensure the data was clean and ready for modeling. The steps taken are as follows:

- **Data Cleaning:** The initial dataset was filtered to remove any missing or inconsistent values, ensuring the quality of the data used for modeling. We focused on the key columns: team1, date, and spi1.

- **Selection of Top 30 Teams**: The average SPI1 scores were calculated for each team, and the top 30 teams based on these averages were selected for further analysis. This helped in focusing on the most competitive teams.

- **Time Series Formatting**: The data was organized into a time series format for each team, with spi1 values averaged by date to ensure consistency in the modeling process.

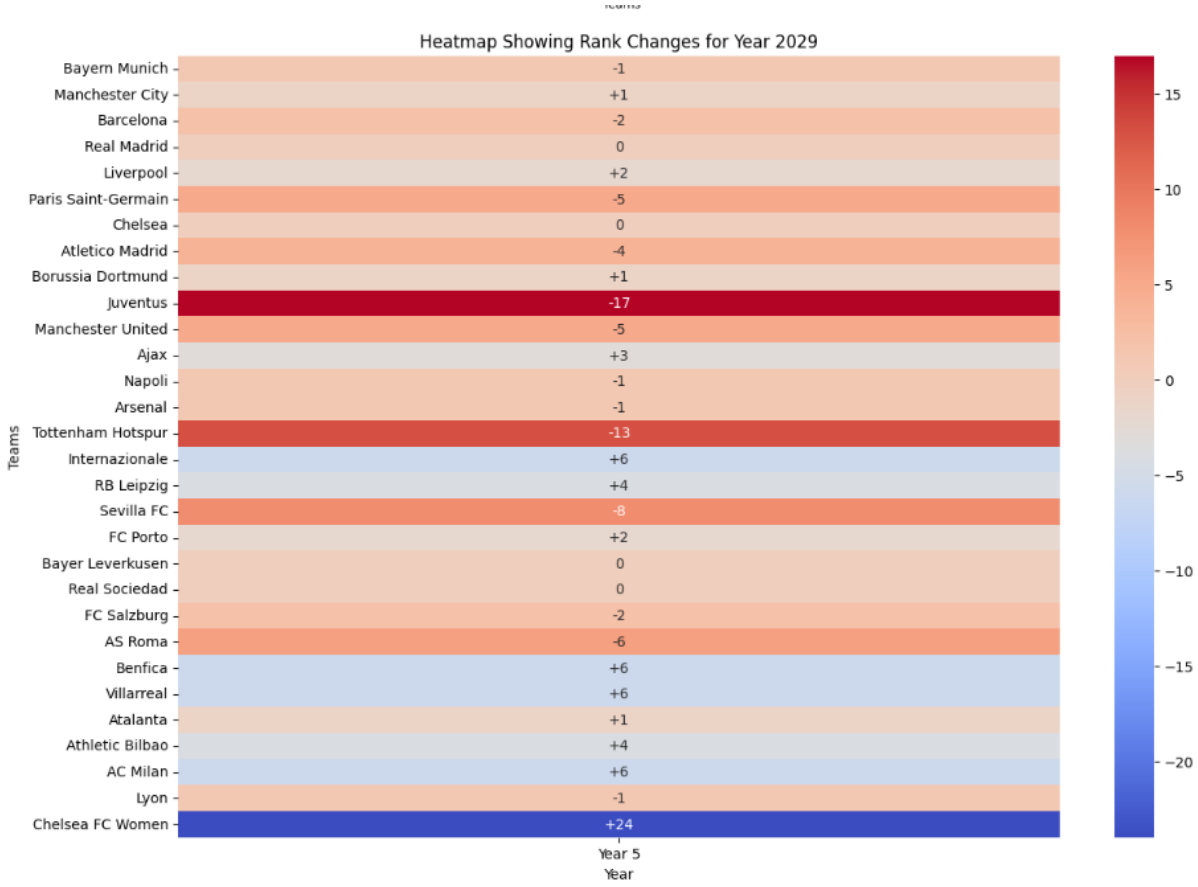
- **Feature Engineering for Offense and Defense Index:** The goals scored and conceded were aggregated for each team to compute the offense and defense indices. These metrics were standardized, and the defense index was multiplied by -1 to ensure higher values indicate better performance.

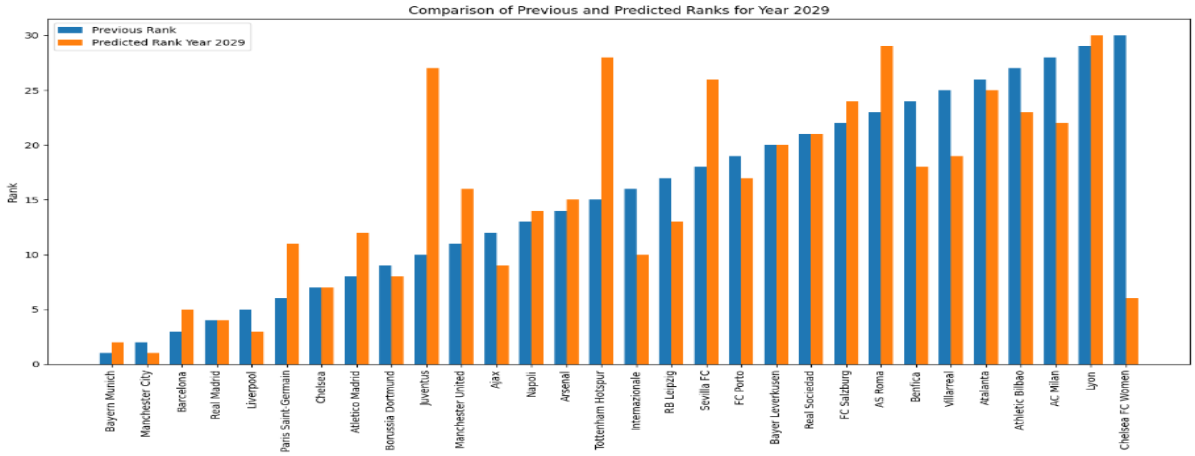


- **Rank Prediction:** The ranks of the top 30 teams were predicted for the year 2029 using Random Forest and ARIMA models. The training data spanned from the earliest available data up to five years before 2029, with the last five years set aside for evaluation.
- **Random Forest Regressor:** This model was trained to predict future SPI values based on historical data.
- **ARIMA Model:** Using the auto_arima function, the model was tuned to fit the best parameters for forecasting future SPI values.
- **Performance Evaluation:** The models were evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), with comparable errors indicating reliable predictions. The heatmap of predicted rank changes over five years visualized team dynamics, while offense vs. defense scatter plots provided insights into team strengths.

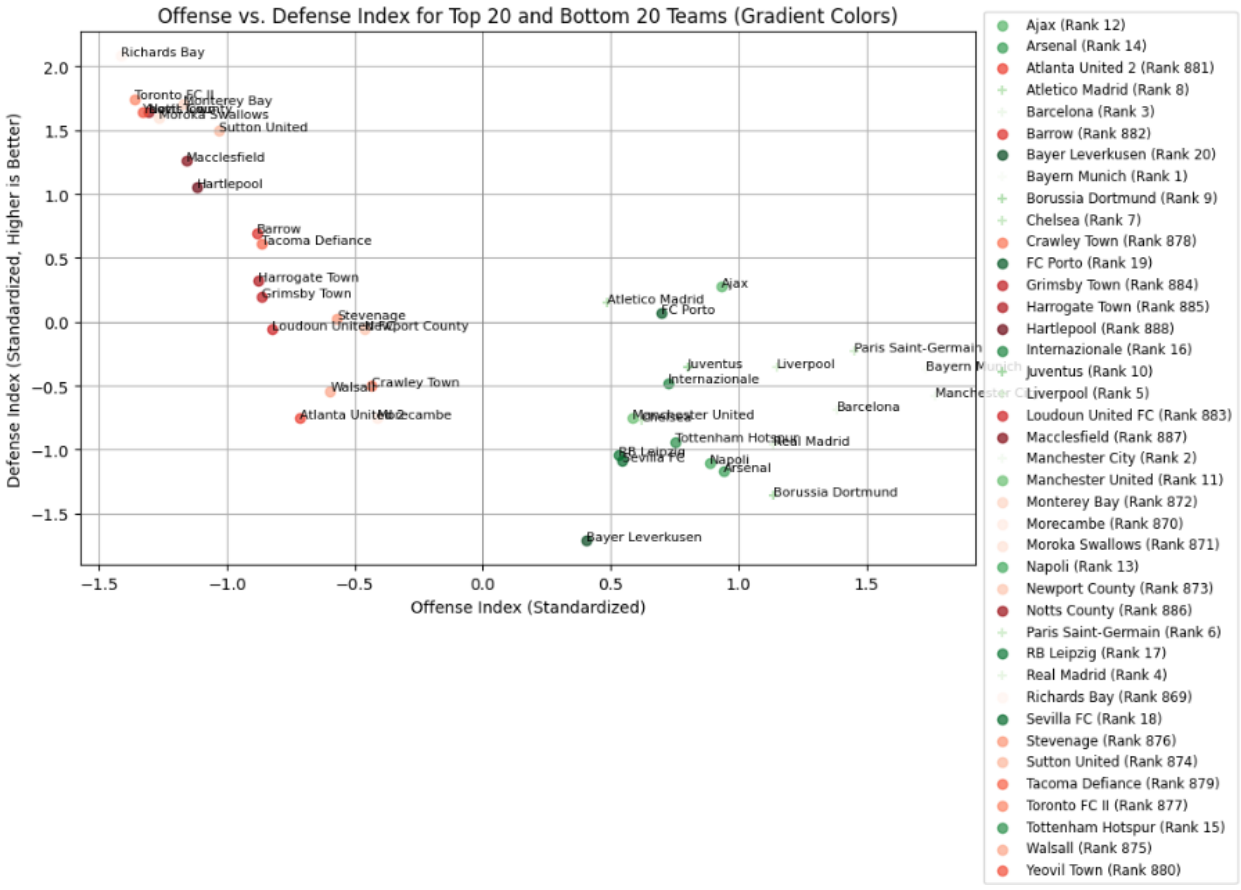
Analysis:

- **SPI Trends Over Time:** Teams like Bayern Munich and Manchester City showed consistent rank stability, whereas Juventus experienced a notable decline. Chelsea FC Women demonstrated significant improvement over the prediction period.

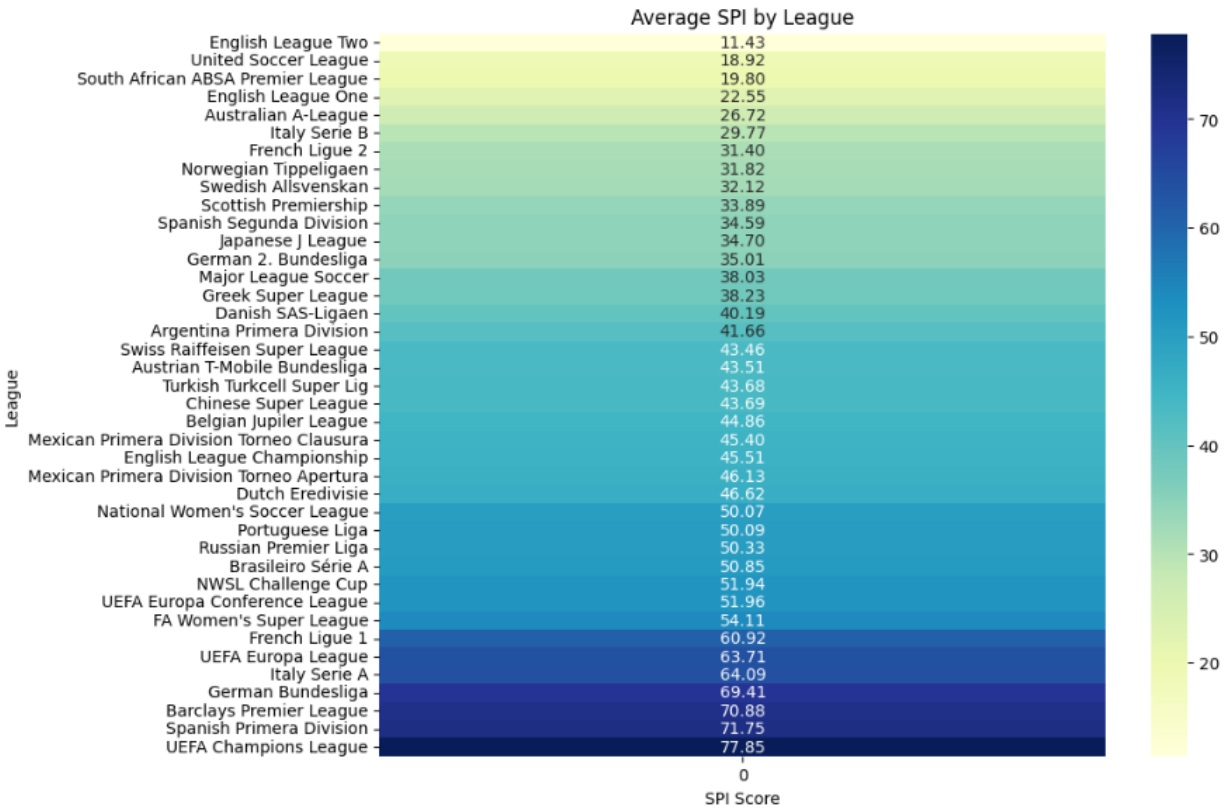




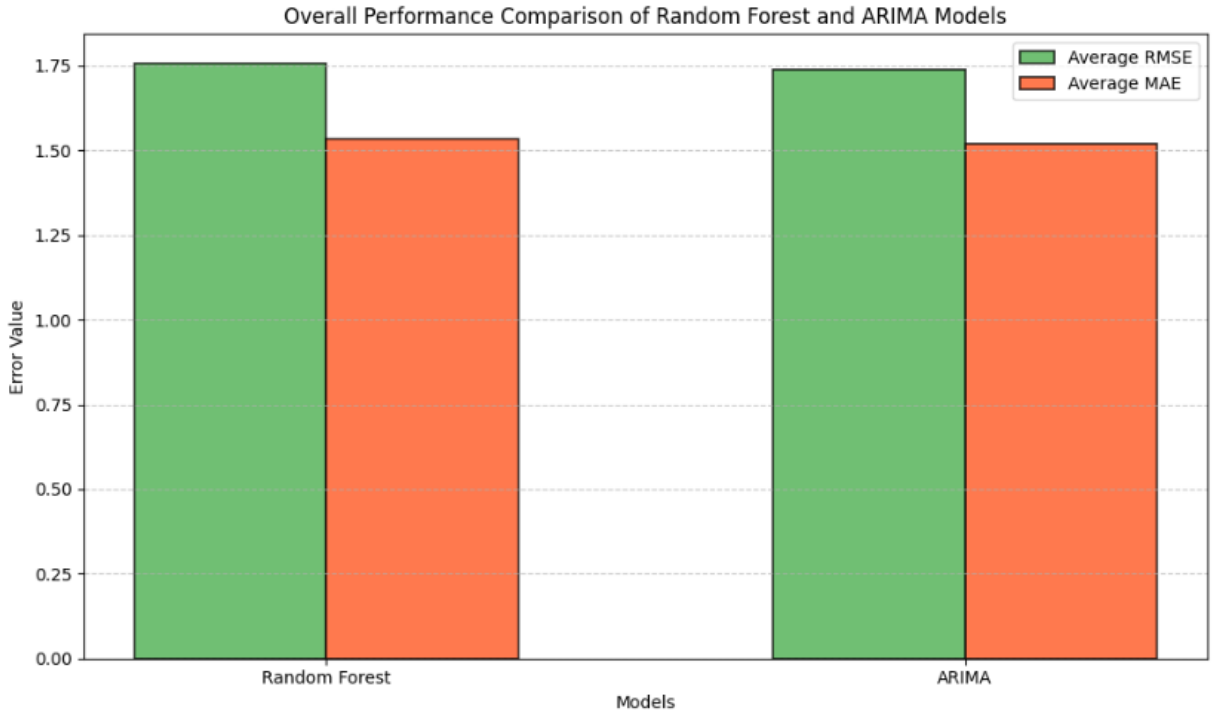
- **Offense vs. Defense Insights:** Quadrant analysis of teams based on offense and defense indices revealed that top-performing teams (e.g., Ajax, Bayern Munich) balanced both attributes well, whereas weaker teams (e.g., Crawley Town) struggled in both areas.



- **Comparing Different Leagues**: A heatmap showed disparities in average SPI across leagues, with top leagues (e.g., UEFA Champions League, Spanish Primera Division) having higher average SPI scores, reflecting stronger competition. Lower-tier leagues displayed either offensive or defensive weaknesses.



- **Model Performance:** The Performance of ARIMA as well as random forest is as below:



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.

Assumptions

The methodology for this project relies on several key assumptions to ensure the predictive models operate effectively and yield meaningful results:

1. **Data Completeness and Quality**: It is assumed that the cleaned dataset accurately represents team performance and includes all necessary historical data without significant omissions or errors.
2. **Stability of Historical Trends**: The models assume that historical SPI trends and team performance dynamics remain relatively consistent over time, allowing for reliable forecasting of future rankings.
3. **Top 30 Teams Represent Competitiveness**: The selection of the top 30 teams based on average SPI values is assumed to sufficiently represent the most competitive teams in global soccer, providing meaningful insights while excluding less impactful teams.
4. **Predictive Power of Offense and Defense Indices**: It is assumed that the engineered offense and defense indices are robust predictors of overall team performance and adequately reflect the balance between scoring and defensive capabilities.
5. **Random Forest and ARIMA Model Assumptions**: For the Random Forest Regressor, it is assumed that historical data features sufficiently capture the factors influencing SPI changes. For the ARIMA model, the time series data is assumed to be stationary after

necessary transformations, with past trends and patterns being indicative of future outcomes.

6. **Impact of Aggregated Data**: The averaging of SPI values and aggregation of goals scored and conceded by date are assumed to smooth out short-term fluctuations without losing essential patterns or trends.
7. **Evaluation Metrics**: The use of RMSE and MAE as evaluation metrics is assumed to provide a reliable measure of model performance, capturing errors in both magnitude and direction.

Limitations:

This project has several limitations that may impact its findings. The dataset's imbalance, particularly in match outcomes, poses challenges for model performance and fairness. The SPI metric does not account for external factors such as injuries, managerial changes, or home-field advantage, which can significantly influence match results. Additionally, the assumption of historical trend stability may not hold true in dynamic sports environments where unexpected events occur. Finally, the reliance on specific models like Random Forest and ARIMA may limit the exploration of alternative techniques that could offer better predictive accuracy.

Conclusion:

The analysis revealed that SPI trends for top clubs remained relatively stable over time, especially for consistently high-ranking teams. Teams with more fluctuations in SPI experienced larger prediction errors, indicating that team dynamics, transfers, and management changes significantly affect performance. The comparison across leagues indicated that stronger leagues

had higher SPI values and balanced team attributes. In contrast, leagues with lower SPI scores showed greater variability in team strategies, focusing more on either offense or defense.

The results suggest that while predictive modeling can offer reliable insights into future team performance, continuous changes in team composition and strategies need to be accounted for to improve prediction accuracy.

How Does SPI Change Over Time?

- **Rank Predictions:**
    - For the year 2029, predicted ranks for teams like Bayern Munich and Manchester City showed minor changes, indicating consistent performance.
    - Some teams like Juventus showed significant rank drops, while others like Chelsea FC Women improved significantly.
- **Prediction Models (Random Forest vs. ARIMA):**
    - Both models gave comparable average errors (RMSE and MAE), indicating reliable predictions.
    - Teams with more consistent SPI over time had smaller prediction errors.
- **Insights from Heatmap:**
    - The heatmap of rank changes showed that some teams experienced significant shifts over five years, reflecting changes in team form, transfers, or managerial changes.

Comparing Different Leagues

- **League SPI Distribution:**
    - The heatmap showed clear differences in average SPI across leagues.
    - Top leagues such as the UEFA Champions League and Spanish Primera Division had higher average SPI scores, indicating stronger competition.
    - Lower-tier leagues like English League Two and United Soccer League had the lowest average SPI.
- **Impact on Teams:**
    - Teams from stronger leagues (higher average SPI) generally performed better in terms of offensive and defensive indices.
    - The scatter plot for offense vs. defense showed that top league teams (Ajax, Bayern Munich) exhibited balanced performances.

Offense vs. Defense Insights

- **Quadrant Analysis:**
    - Teams in the upper right quadrant (e.g., Bayern Munich, Ajax) showed strong performance in both offense and defense.
    - Teams in the lower left quadrant (e.g., Crawley Town, Walsall) struggled both offensively and defensively.
- **Top 20 vs. Bottom 20 Teams:**
    - Gradient colors helped distinguish top 20 teams (green) from bottom 20 teams (red), revealing performance patterns across rankings.

- Marker styles differentiated top 10 teams, indicating their superior balance between offense and defense.

Future Work

Future work for this project could focus on enhancing the predictive models and expanding the scope of analysis. Incorporating additional features, such as player-level statistics, team budgets, and injury data, could improve the accuracy of predictions. Exploring advanced machine learning models like deep learning or ensemble methods may provide better performance compared to traditional models. Expanding the analysis to include more teams and leagues globally would allow for broader insights into competitive dynamics. Additionally, integrating real-time updates into the models could make predictions more adaptive to sudden changes in team composition or strategy. Finally, applying the methodology to other sports or extending it to predict individual player performance could open new avenues for application.

Questions an audience may ask

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

   Yes, with historical data, metrics like SPI, and machine learning models, we can predict game outcomes with reasonable accuracy, though external factors like injuries and weather may reduce precision.

2) How much does a player's performance affect the results of the match?

   A player's performance significantly impacts match outcomes, especially for key players. Metrics like goals, assists, and defensive contributions are critical, but the team's overall cohesion and strategy also play a role.

3) Can we model a Machine Learning algorithm to predict the game's final result?

   Yes, machine learning algorithms, such as Random Forest or Neural Networks, can predict results using features like SPI, team form, and historical match data, providing probabilistic forecasts.

4) How far can the predictive power of Artificial Intelligence get?

   AI can predict outcomes with high accuracy given sufficient quality data, but its limitations include unpredictable factors like injuries, referee decisions, and psychological elements of the game.

5) How does SPI change over time?

   SPI evolves throughout a season based on a team's performance, weighted by match outcomes, goals scored and conceded, and expected versus actual results.

6) How can we compare different leagues?

   Leagues can be compared using aggregated SPI values, offensive and defensive metrics, and variability in team performances, with stronger leagues showing higher average SPI and more balanced attributes.

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