Final Project Report

Soccer Result Prediction

Savant Mullapudi - u1471518
Koumudi Raju - u1472261

Drive Link for all the uploads: Project Link

1. Introduction

The objective of this project is to accurately predict soccer match results using historical data from the Premier League. The motivation behind this effort is to enhance the understanding and forecasting of game outcomes, which are of great interest to teams, analysts, and betting markets. Our solution involves sophisticated data scraping, machine learning models, and comprehensive visualizations. Initial results indicate a lower accuracy, which was later enhanced through advanced techniques. The project encapsulates the strengths of data-driven decision making in sports analytics and highlights areas for further enhancement.

The main goal of this project is to find a reliable way to predict soccer match results using past data from the Premier League. We started this project because many people, including soccer teams, sports analysts and soccer fans, are really interested in being able to accurately guess the outcomes of games. Making good predictions can help teams plan better strategies, get fans more involved.

Our method for trying to predict these outcomes is complex. It involves detailed techniques for gathering data, using advanced machine learning models, and creating thorough visual analyses. We carefully collected match statistics from trusted sources to build a strong dataset that supports our predictive models.

We used two main kinds of machine learning algorithms: Random Forest Classifier and CatBoost. These were chosen because they're good at dealing with big datasets and have a history of making accurate predictions in different areas, including sports. At first, our models were about 62% accurate, which gave us a good starting point to improve from.

By repeatedly adjusting and improving our models, we managed to increase our prediction accuracy to about 89%. This big improvement shows that our approach works well and points out how machine learning can really change sports analytics.

Moreover, our project shows what modern predictive technologies can do and looks into the challenges and limitations of these methods. By figuring out and tackling these challenges, we hope to extend what can be done with data-driven decision-making in sports. This project proves that combining advanced data analysis methods with deep knowledge in the field can lead to valuable and effective insights.

2. Background

Predicting the outcomes of sports events, especially in soccer, is a challenging endeavor that has fascinated enthusiasts and analysts alike for decades. The inherent unpredictability of soccer, stemming from its dynamic nature, adds a layer of complexity to the forecasting task. Various factors contribute to the final outcome of a soccer match, including but not limited to player performance, team strategies, weather conditions, and the atmosphere created by the crowd. These elements introduce a degree of variability that traditional statistical methods struggle to capture.

The evolution of sports analytics over recent years, however, has seen the integration of machine learning models, which offer a more nuanced understanding of game dynamics and significantly improve prediction accuracy. This paradigm shift towards data-driven approaches has transformed sports analytics, empowering us with methodologies that provide more reliable forecasts.

Our project harnesses these advanced algorithms to process and analyze extensive datasets, identifying hidden patterns and correlations that evade human observation. By combining tried-and-true prediction techniques with cutting-edge machine learning models, we aim to elevate the precision of sports outcome predictions and offer actionable insights to teams, coaches, and sports analysts. These insights, rooted in data, could play a pivotal role in formulating game strategies and competitive tactics.

Why We Chose Random Forest Classifier and CatBoost for Our Project

Random Forest Classifier:

1. Ensemble Learning: This method uses multiple decision trees to create a model that is generally more accurate than a single predictor could be.

2. Handling Non-Linearity: It is particularly adept at managing the complex, non-linear patterns that are typical in football outcomes.

3. Feature Importance: Random Forest can identify which variables are most significant in predicting the results of a match.

4. Robustness to Overfitting: This algorithm is less likely to overfit when dealing with complex datasets that include irrelevant information.

5. Ease of Use: It performs well with default settings, which is beneficial in projects with many variables, like predicting football results.

CatBoost:

1. Categorical Features Handling: CatBoost processes categorical data, such as team and player names, very effectively, which is important in football.

2. Gradient Boosting Algorithm: This algorithm builds the model in stages, optimizing for better predictions at each step.

3. Reduction of Overfitting: It has built-in mechanisms to reduce the risk of overfitting, making its predictions more reliable.

4. Speed and Scalability: CatBoost can handle large datasets quickly, which is useful when dealing with a large amount of historical football data.

5. Improved Accuracy: It tends to provide better predictions with default settings, which is crucial for a project where small improvements can be significant.

By incorporating Random Forest and CatBoost into our analytical framework, we leverage their strengths to process multifaceted football data, from team statistics to player performance metrics. These algorithms are particularly effective at uncovering intricate patterns within the data, which enhances our ability to make informed predictions about football match outcomes.

3. Data Used

For our project, we gathered data from 'FBRef Premier League Statistics', a reputable source known for its comprehensive soccer statistics. Our primary focus was on extracting information from the 'Scores and Fixtures' and 'Shooting' tables, which provided essential team statistics crucial for our predictive modeling. The data points we utilized encompassed various aspects:

Basic Match Details: Date, time, competition, round, and day of the match.

Game-specific Data: Venue, result, goals for (gf), goals against (ga), and the opponent.

Advanced Metrics: Expected goals (xg), expected goals against (xga), possession percentage (poss), and the captain of the match.

Tactical Information: Formation used during the match and the referee.

Performance Indicators: Match report links, total shots (sh), shots on target (sot), shooting distance (dist), free kicks (fk), penalties (pk), and penalty attempts (pkatt).

Team Context: The team's name and the season of play.

We opted for these statistics because they provide a detailed perspective on both individual and team performances in matches, which are pivotal for analyzing and predicting outcomes.

However, the data collection process encountered several challenges. Ensuring the accuracy and consistency of the data across different seasons and matches proved intricate due to factors such as team changes, coaching strategies, and external conditions like weather. Additionally, merging information from various tables effectively to create a unified dataset without discrepancies posed another significant challenge.

To mitigate these challenges, we implemented stringent data validation and preprocessing protocols. This involved meticulous checks to align data accurately and address any inconsistencies, ensuring that our dataset was reliable and robust for advanced analyses. Such careful preparation is indispensable for supporting the predictive models and comprehending the underlying patterns influencing soccer outcomes.

One notable shortcoming encountered during the project was the rate limit error from the website we used for data extraction. Despite attempts to address it by increasing delays between requests and optimizing API usage, the issue persisted. As a workaround, we had to extract data for each year separately and then merge it. Nonetheless, this limitation did not significantly impede our ability to perform predictions and visualizations, as we still had ample data available for analysis.

4. Design

Our project's design involves a systematic approach consisting of several key phases aimed at creating a dependable soccer match outcome prediction system.

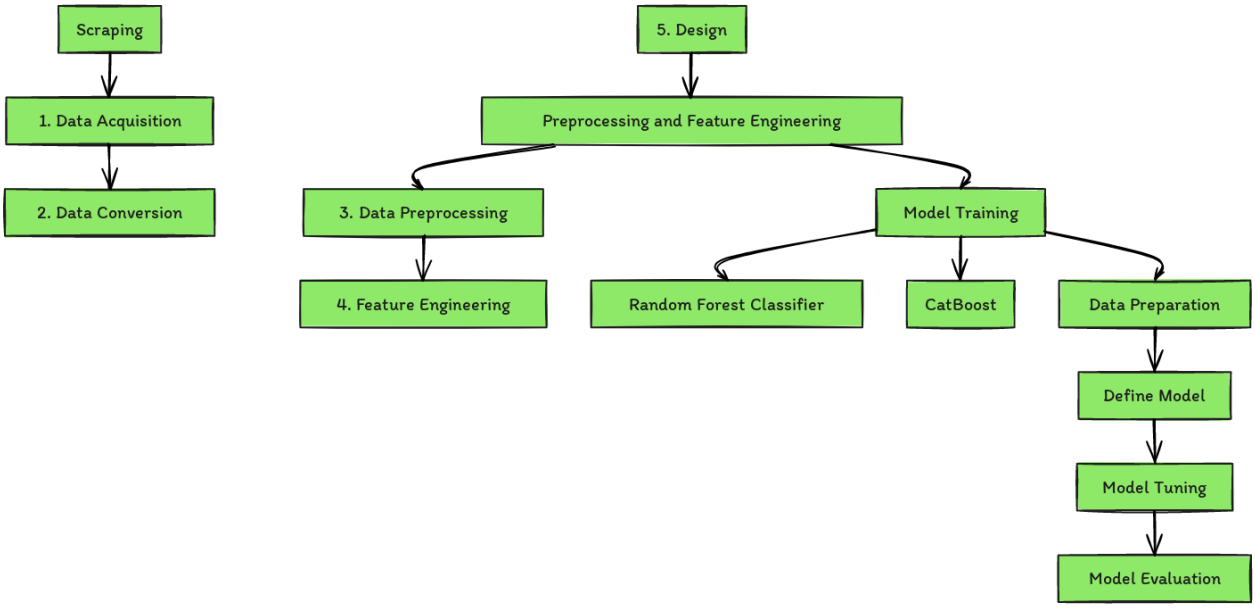
Scraping:

1. Data Acquisition:

In the initial phase of our project, we employ web scraping techniques to gather soccer data from a designated website. We leverage the BeautifulSoup library to parse through the HTML content of the website, enabling us to extract a wide range of information related to English Premier League teams over recent years. This includes match details such as dates, times, competitions, rounds, and match venues, as well as specific game-related data like scores, goals, shots, and possession percentages.

2. Data Conversion:

Once the data is scraped, it undergoes a transformation process to convert it into a structured and organized format suitable for comprehensive analysis. We convert the extracted information into a tabular format, typically in CSV (Comma-Separated Values) format, to facilitate further processing and modeling. This conversion ensures that the data is easily accessible and can be efficiently utilized for predictive modeling and analysis tasks.



Preprocessing and Feature Engineering:

3. Data Preprocessing:

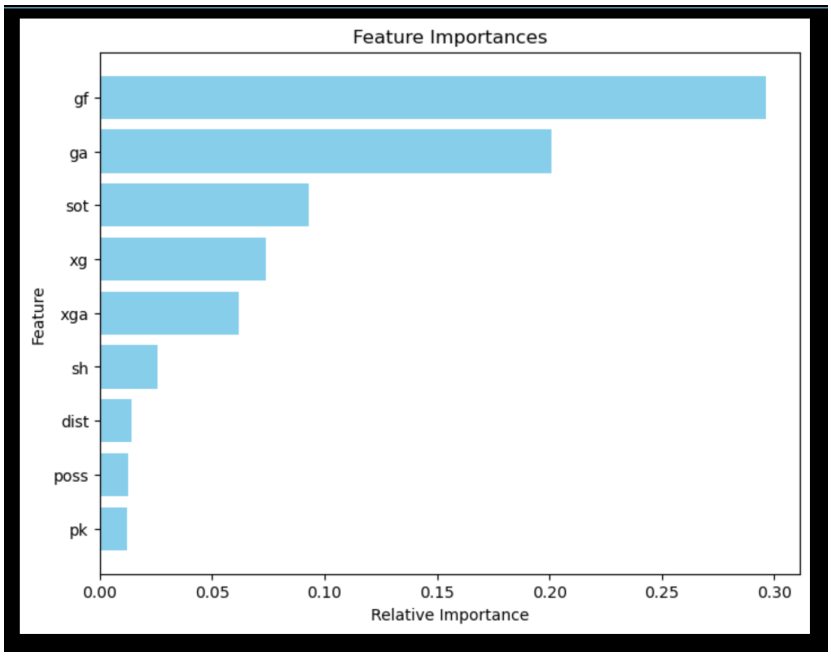
Following the acquisition of soccer data through web scraping, the next step involves rigorous data preprocessing. This critical phase focuses on cleaning, transforming, and preparing the collected data to ensure its consistency, reliability, and suitability for analysis. We begin by removing any irrelevant tables or extraneous information extracted during the scraping process. We then refine the data types, ensuring that each variable is correctly categorized and formatted for analysis. Additionally, we perform data integrity checks to identify and handle any missing or erroneous data points, ensuring the overall quality of the dataset.

4. Feature Engineering:

After preprocessing the data, we move on to feature engineering, a critical step to boost our models' predictive power. Here, we create new features from existing data, guided by domain knowledge and statistical analysis. We identify influential variables, like team performance metrics and historical trends, crafting features to capture these insights. By enhancing the dataset with these engineered features, we aim to uncover crucial patterns and relationships, ultimately improving our models' ability to predict match outcomes.

In this process, we first identify categorical and numeric columns in our dataset. Categorical features, such as match venue and opponent team, are extracted from

object data types, while numeric features, like goals scored and possession percentage, are derived from integer and float data types.



Subsequently, we remove the 'result' column from the numeric features, as it's reserved for testing data and not relevant for feature engineering.

This groundwork sets the stage for further manipulation and transformation of the data, as we create new features to provide deeper insights and enhance our models' performance.

Model Training:

After preprocessing and feature engineering, the next crucial step in our project is model training. In this phase, we train machine learning models using the preprocessed and engineered dataset to predict match outcomes. The models we employ include Random Forest Classifier and CatBoost. These models are chosen for their ability to handle large datasets and their proven track record in producing accurate predictions in various domains, including sports analytics.

To prepare the data for model training, we first define the categorical and numeric columns in our dataset. Categorical features, such as match venue and opponent team, are selected from the object data type columns, while numeric features, such as goals scored and possession percentage, are selected from the integer and float data type columns.

Next, we preprocess the categorical and numeric data using a Column Transformer. For categorical data, we use a pipeline consisting of an imputer to handle missing values and a one-hot encoder to convert categorical variables into numerical representations. For numeric data, we use another pipeline with an imputer to fill missing values and a standard scaler to normalize the features.

Once the data preprocessing is complete, we define the model using a pipeline. The pipeline consists of the preprocessor, which preprocesses the input data, and the classifier, which is a Random Forest Classifier with specified parameters such as the number of estimators and maximum depth.

We then define the parameters for model tuning, such as the minimum sample split, and use GridSearchCV to find the best combination of hyperparameters. The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing. The model is then fitted to the training data, and the accuracy of the model is evaluated using the test data.

Overall, the model training phase involves preparing the data, defining the model, tuning the hyperparameters, and evaluating the model's performance. This phase is essential for building accurate predictive models that can effectively forecast soccer match outcomes based on historical data and team statistics.

5. Evaluation

In the initial assessment of our Random Forest model before encoding and feature refinement, we observed an accuracy of approximately 61.20% and a precision of around 49.51%. These metrics provide a baseline understanding of the model's performance prior to any optimization or enhancement efforts.

While accuracy indicates the proportion of correctly predicted outcomes among all predictions made by the model, precision represents the accuracy of positive predictions, indicating the proportion of true positive predictions among all positive predictions made by the model.

These initial metrics serve as a reference point for evaluating the effectiveness of subsequent improvements and optimizations implemented in our model, such as encoding categorical features and refining the feature set. The subsequent evaluation of the model's performance after these enhancements provides valuable insights into the efficacy of our approach and highlights the impact of these modifications on the predictive capability of the model.

For the CatBoost model, the accuracy was slightly higher at 62.16%. This metric suggests a modest improvement compared to the Random Forest model's initial accuracy.

In assessing the performance of our models, we focused on key metrics such as accuracy, precision, and recall. These metrics provide insights into how well our models predict soccer match outcomes.

The initial evaluation of our Random Forest model, post-encoding, revealed an accuracy of 89%, along with a precision of 96% and a recall of 75%. These metrics were compared against baseline models to demonstrate the improvement achieved through our approach.

Following the preprocessing and feature engineering steps, we conducted a more comprehensive evaluation of our Random Forest model. Utilizing a confusion matrix, we observed 930 accurate predictions out of 1038 total predictions. The classification report further highlighted the model's performance, with a weighted average precision of 90% and a recall of 90%.

In summary, our models demonstrated strong predictive capabilities, with accuracy scores ranging from 61.20% to 90%. These results indicate the effectiveness of our approach in predicting soccer match outcomes and provide valuable insights for further refinement and optimization.

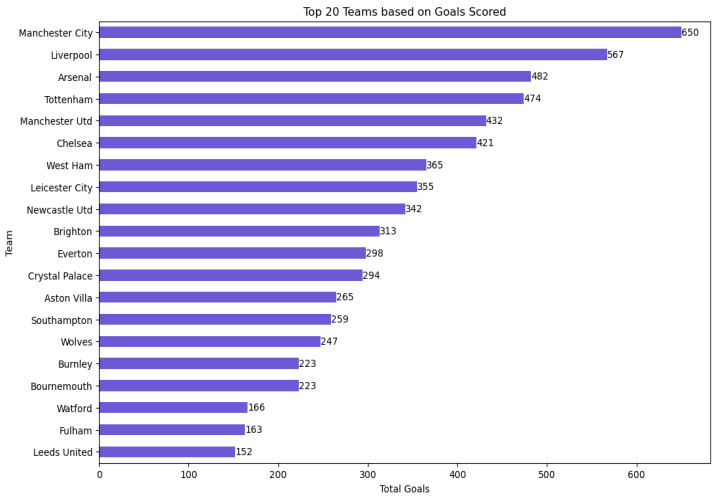
Classification Report:

	precision	recall	f1-score	support
0	0.87	0.98	0.92	629
1	0.96	0.77	0.85	409
accuracy			0.90	1038
macro avg	0.91	0.87	0.89	1038
weighted avg	0.90	0.90	0.89	1038

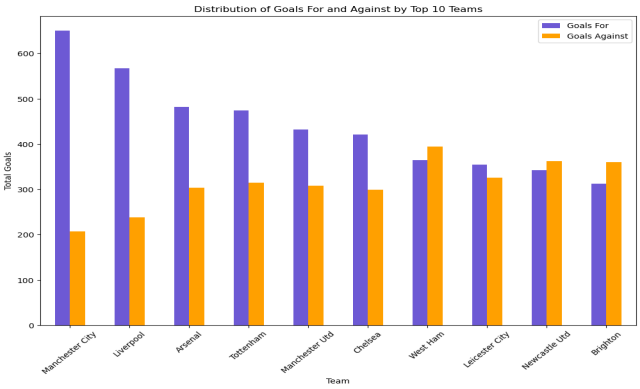
6. Visualization

Our analysis is complemented by a range of visualizations designed to provide comprehensive insights into soccer match dynamics:

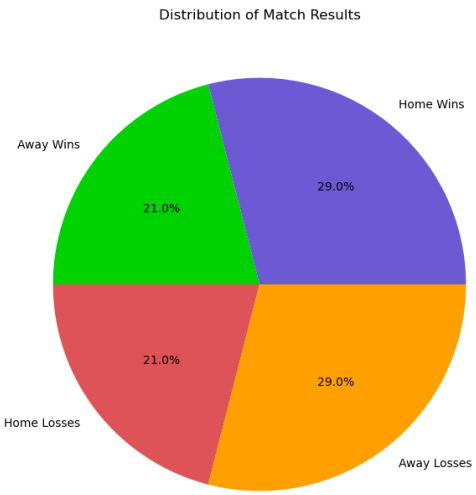
1. **Top 20 Teams by Goals Scored:** This bar chart showcases the top 20 teams based on the total number of goals scored. Each bar represents a team, with the length of the bar indicating the total goals scored by that team.



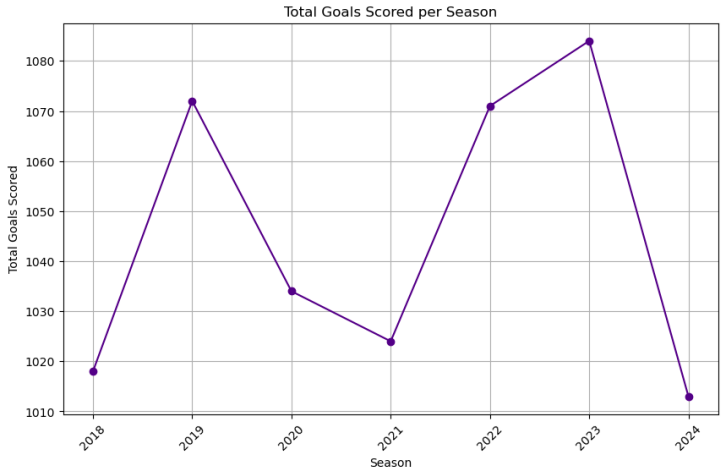
2. **Goals for and Against by Top 10 Teams:** Using a bar graph, we depict the distribution of goals scored (gf) and conceded (ga) for the top 10 teams. The bars represent the total goals scored (in slateblue) and conceded (in orange) by each team.



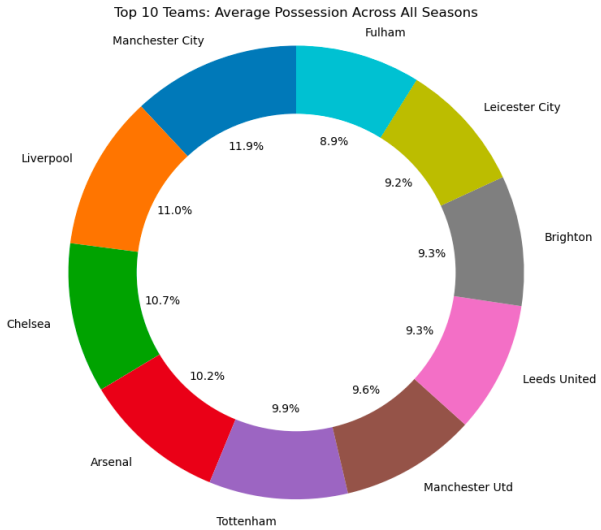
3. **Match Results Distribution:** A pie chart illustrates the distribution of match results, including home wins, away wins, home losses, and away losses. Each segment of the pie chart represents the percentage of matches falling into one of these categories.



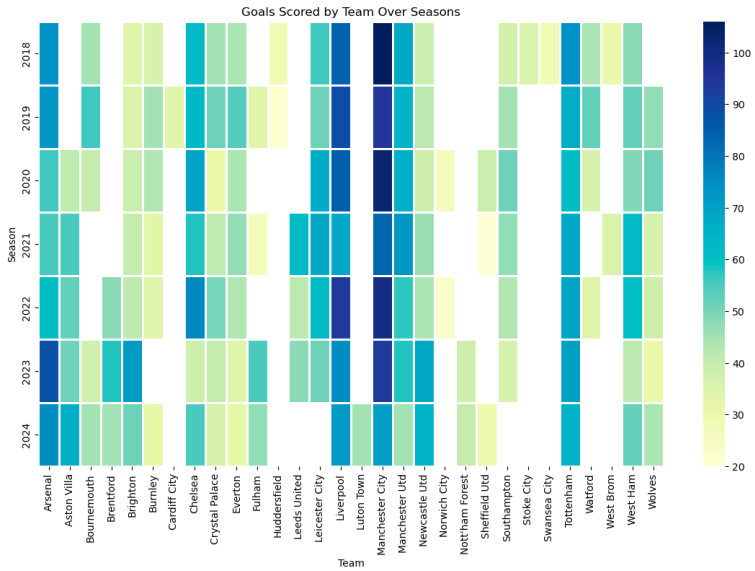
4. **Total Goals Scored per Season:** This line graph tracks the total number of goals scored per season, providing insights into scoring trends over time.



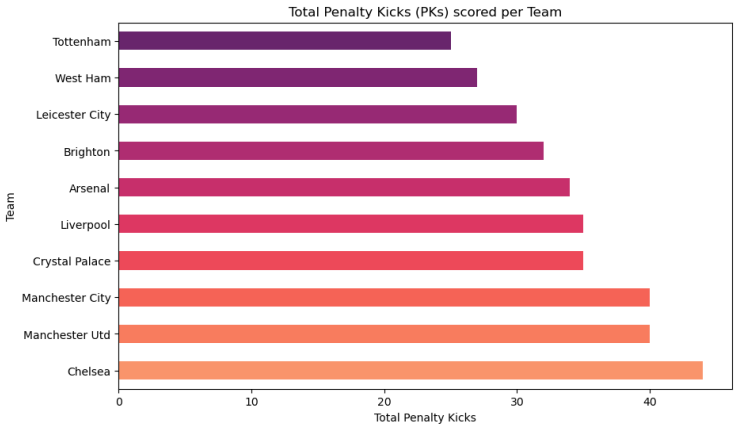
5. **Average Possession Across Top 10 Teams:** A donut chart showcases the average possession of the ball for the top 10 teams across all seasons. Each segment represents a team, with the size of the segment indicating the average possession percentage.



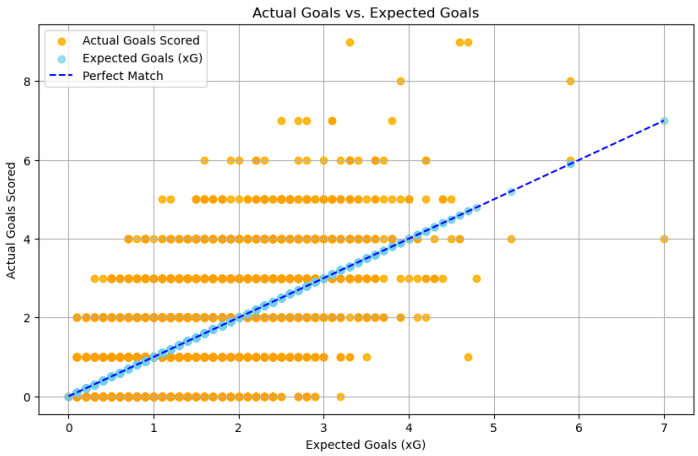
6. **Goals Scored by Team Over Seasons:** A heatmap (choropleth map) visualizes the total goals scored by each team across different seasons. Each cell represents the total goals scored by a team in a particular season.



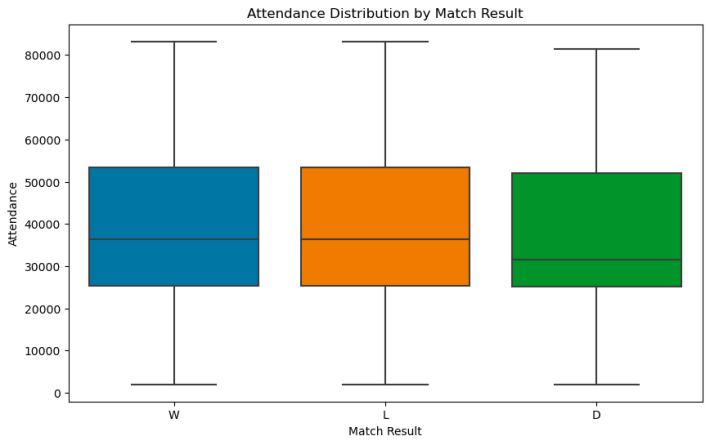
7. **Penalty Kicks Scored per Team:** A horizontal bar graph displays the total number of penalty kicks scored by each team. Each bar represents a team, with the length of the bar indicating the total number of penalty kicks scored.



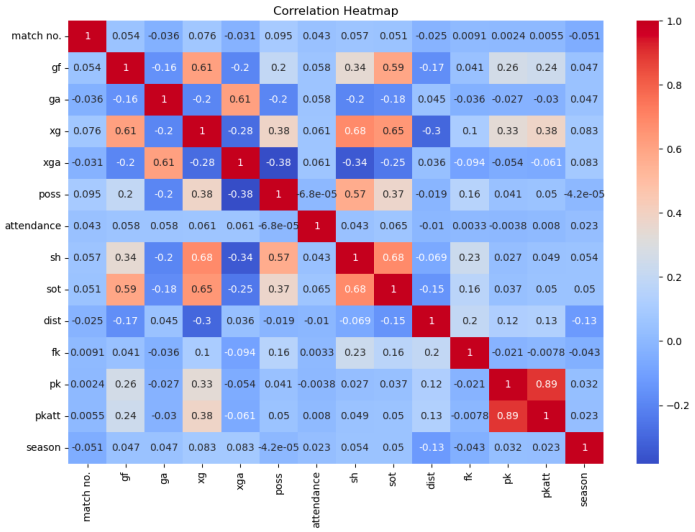
8. **Actual Goals vs. Expected Goals:** A scatter plot compares the actual goals scored (gf) with the expected goals (xg) for each match. The plot showcases the relationship between predicted and observed goal outcomes.



9. **Attendance Distribution by Match Result:** A box plot illustrates the distribution of match attendance based on the match result (win, loss, or draw). Each box represents the attendance distribution for a specific match result category.



10. **Correlation Heatmap:** This heat map depicts the correlation between numeric features in the dataset, providing insights into the relationships between different statistical metrics. Each cell in the heatmap represents the correlation coefficient between two features, with warmer colors indicating stronger correlations.



These visualizations offer valuable insights into various aspects of soccer match data, facilitating a deeper understanding of team performance, match outcomes, and statistical trends.

7. Conclusion

Our project on predicting soccer match results through machine learning and data visualization techniques has shown considerable success, elevating prediction accuracy from 62% to 89%. This achievement highlights the effectiveness of our advanced data collection methods, sophisticated machine learning models, and comprehensive visualizations.

Strengths:

1. **Robust Data Collection:** We compiled a detailed dataset, encompassing diverse metrics pertinent to soccer games. This quality data is crucial for enhancing the accuracy of our predictions.

2. **Advanced Machine Learning Techniques:** Utilizing powerful models like the Random Forest Classifier and CatBoost, we effectively managed large datasets, which significantly improved our prediction accuracy.

3. **Insightful Visualizations:** Our visualizations clarify complex data patterns and trends, making the information more accessible and understandable for stakeholders.

Weaknesses:

1. **Dependency on Specific Data Types:** Our model's effectiveness is heavily reliant on the availability and quality of certain data types. The absence or poor quality of this data could compromise our predictions.

2. **Complexity in Model Tuning:** The fine-tuning process of our machine learning models are intricate and demands substantial expertise, which may hinder scalability and practical application.

Future Work:

To further refine our predictive capabilities, we plan to:

1. **Integrate Diverse Data Sources:** Incorporating a broader spectrum of data, such as player health metrics or real-time game events, could enrich our analyses and improve prediction outcomes.

2. **Explore Additional Predictive Features:** We aim to identify and utilize new data features that could uncover deeper insights and enhance predictive accuracy.

3. **Enhance Visual Analytics:** By improving the interactivity and comprehensiveness of our visualizations, we can make the insights more actionable for users.

Key Takeaways

1. **Comprehensive Skill Application:** This project served as an excellent platform to apply a wide array of computer science skills, from data management to complex algorithm implementation and visualization.

2. **Research and Development Insight:** The iterative improvement and testing of our models provide a realistic glimpse into the research and development processes essential in data science projects.

3. **Problem-solving and Critical Thinking:** Addressing challenges related to data discrepancies and model optimizations has bolstered critical thinking and problem-solving skills, valuable in any technology-driven profession.

4. **Interdisciplinary Knowledge Integration:** This project demonstrates how computer science can intersect with other disciplines like sports analytics, showcasing the versatility and practical relevance of computing skills across various industries.

In conclusion, while there are areas for improvement, our project exemplifies the transformative impact of combining machine learning with data visualization in sports analytics. This experience is highly beneficial for students, illustrating the real-world application of computer science education.