























Premier League Champion Prediction

a) Problem statement

The project focuses on clearly defining the problem to solve. In this case, the objective is to predict the champion team of the Premier League for the 2023-2024 season. This involves identifying the relevant data that can provide insights into team performance and understanding the factors that influence a team's success in the league. Historical data from previous seasons (2015-2023) was gathered, which included a wide range of features such as team salaries, spending, wins, goal difference, mean player rating, mean age, goals for, and goals against.



Pos	Club	Pl	GD	Pts
1	 Arsenal	0	0	0
2	 Aston Villa	0	0	0
3	 Bournemouth	0	0	0
4	 Brentford	0	0	0
5	 Brighton	0	0	0
6	 Chelsea	0	0	0
7	 Crystal Palace	0	0	0
8	 Everton	0	0	0
9	 Fulham	0	0	0
10	 Leeds	0	0	0
11	 Leicester	0	0	0
12	 Liverpool	0	0	0
13	 Man City	0	0	0
14	 Man Utd	0	0	0
15	 Newcastle	0	0	0
16	 Nott'm Forest	0	0	0
17	 Southampton	0	0	0
18	 Spurs	0	0	0
19	 West Ham	0	0	0
20	 Wolves	0	0	0



Understanding the problem also entails recognizing the importance of these features and how they contribute to a team's performance. For instance, financial investment in terms of salaries and spending often correlates with acquiring better players, which can enhance a team's overall performance. Similarly, statistical metrics like wins, goal difference, and goals for/against provide direct indicators of a team's success on the field. The problem identification phase thus involved not only data collection but also a thorough analysis of these features to ensure they were relevant and comprehensive for our predictive model.

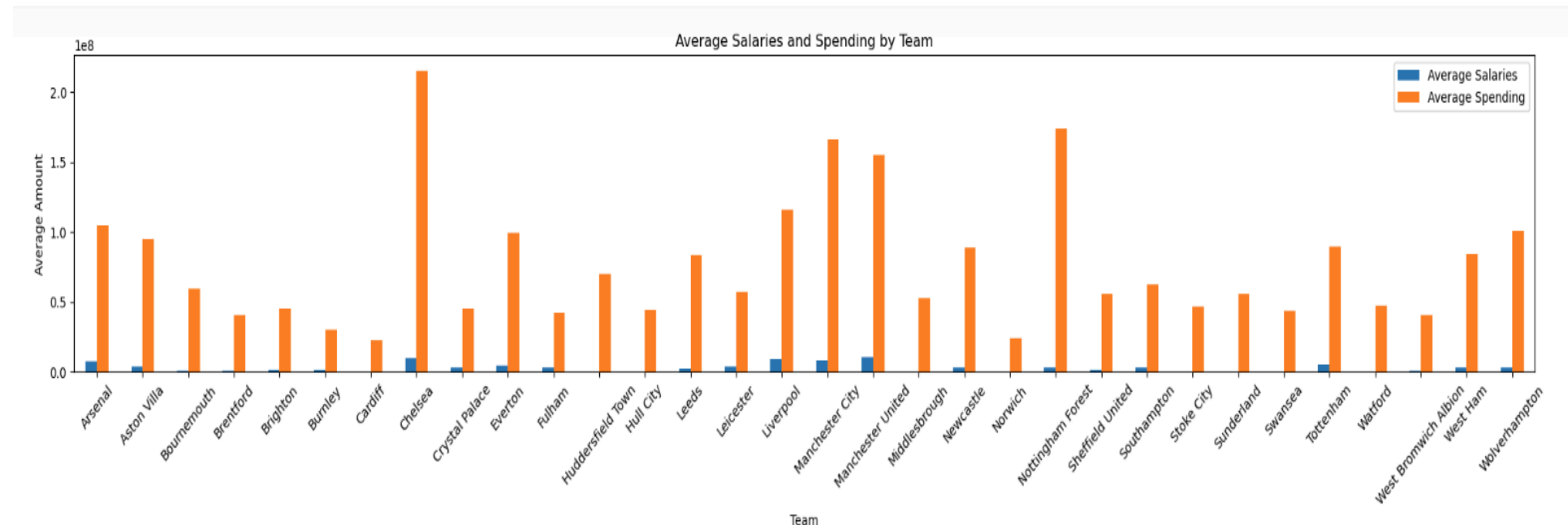
Another critical aspect of problem identification is defining the target variable. In this project, the target variable is the rank of the teams, with the aim of predicting the team that will achieve the highest rank (i.e., the champion). By clearly defining the target and the features, I established a strong foundation for subsequent phases of the project, ensuring that all efforts were aligned with the ultimate goal of accurate and meaningful predictions.

b) Data wrangling

Standardization of the data was another key task in the data wrangling phase. Given the diverse range of features, including financial metrics and statistical performance indicators, standardizing these features ensured they were on a similar scale. Features such as team names and seasons are categorical. Each representing the presence or absence of a particular category. This transformation enabled the inclusion of categorical information in the models, ensuring a comprehensive representation of all relevant factors in the predictive analysis.

The data wrangling process involved transforming and preparing the raw data into a format suitable for analysis and modeling. One of the critical tasks in this phase was handling missing values, which I addressed by imputing missing values with the mean of the respective feature. This ensured that the dataset was complete and no data points were

excluded due to missing information. Additionally, I standardized numerical features such as salaries, spending, wins, goal difference, mean player rating, mean age, goals for, and goals against to ensure they were on a similar scale. Which are sensitive to the scale of input data.



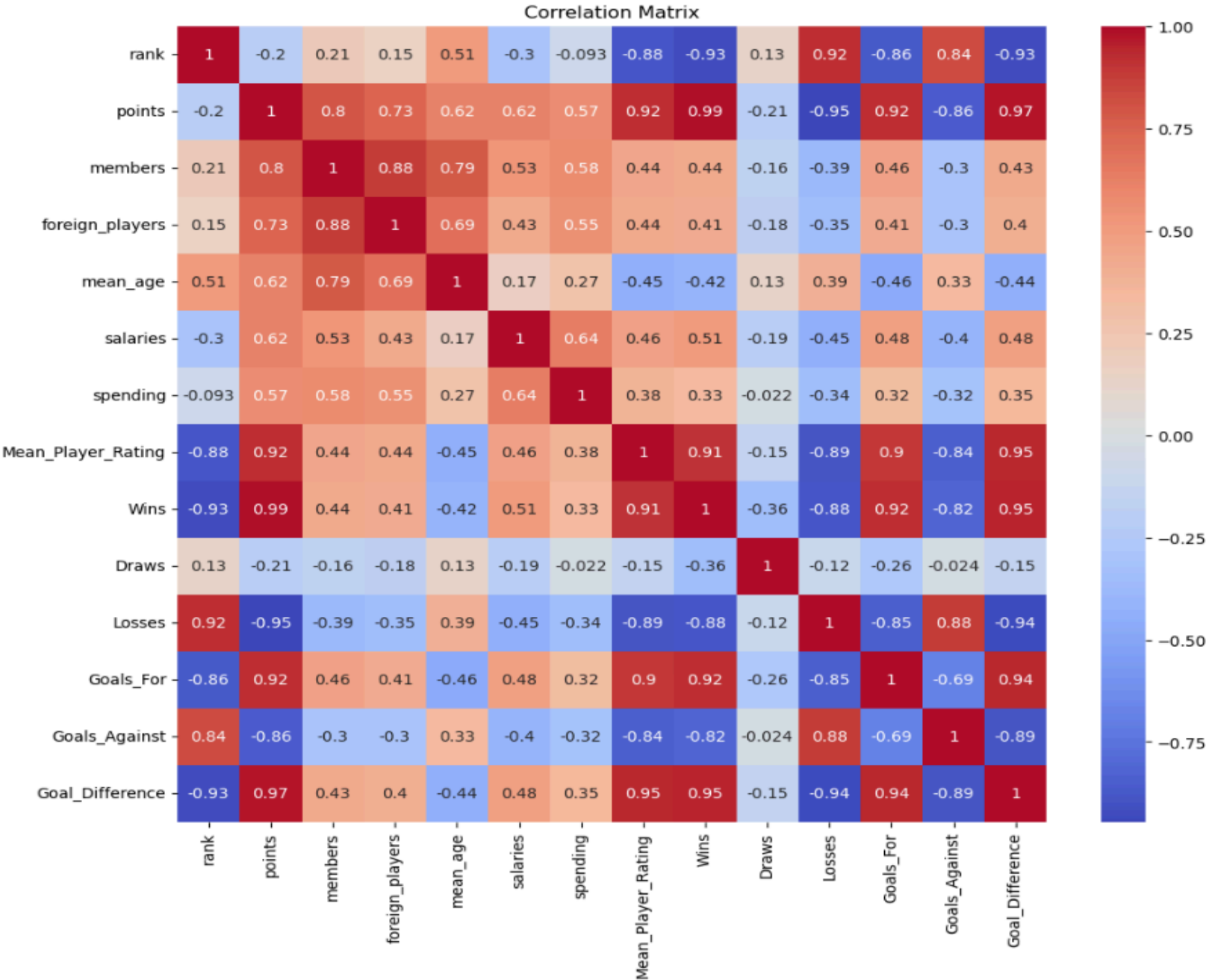
The bar chart above illustrates the average salaries and spending by team. From the chart, I observe significant variation in financial investments across teams. For instance, Chelsea and Manchester City exhibit the highest average spending, reflecting their substantial financial investments in player acquisitions and team development. In contrast, teams like Burnley and Norwich have relatively lower spending. The average salaries also show a similar trend, with teams investing heavily in high-quality players. This financial disparity is a crucial factor in the analysis, as higher spending and salaries are often correlated with better team performance and higher league rankings. This visualization highlights the

importance of financial metrics in predicting team success and underscores the necessity of including these features in the predictive model.

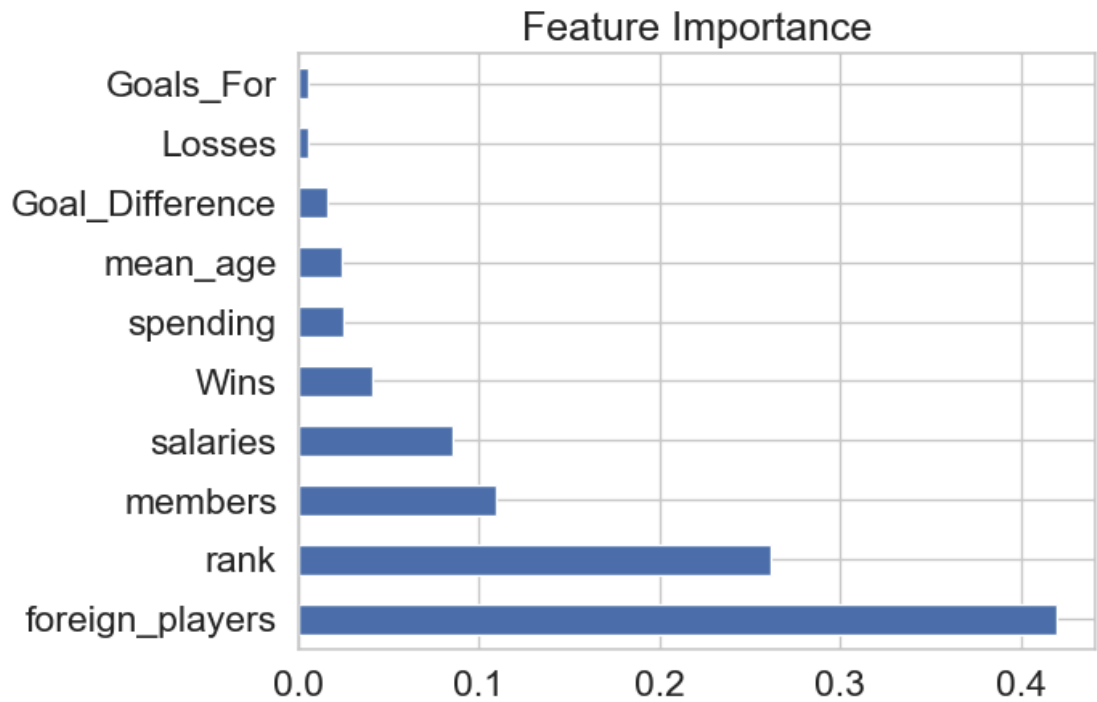
c) Exploratory Data Analysis (EDA)

Identify anomalies, test hypotheses, and check assumptions. The primary goal of EDA in this project was to gain insights into the relationships between different features and the target variable. This process began with descriptive statistics to understand the central tendency, dispersion, and shape of the distributions for each feature. Measures such as mean, median, standard deviation, and skewness provided a summary of the data's characteristics.

Visualizations played a pivotal role in EDA. I created various plots to visualize the distribution of features and their relationships with the target variable. For instance, bar charts were used to show the total salaries spent by teams, revealing which teams had the highest financial investments. Similarly, bar charts for total wins, goal difference, and goals for/against provided a visual representation of team performance metrics over the years. These visualizations helped identify trends, such as the correlation between high spending and better team performance, which are crucial for building an accurate predictive model.

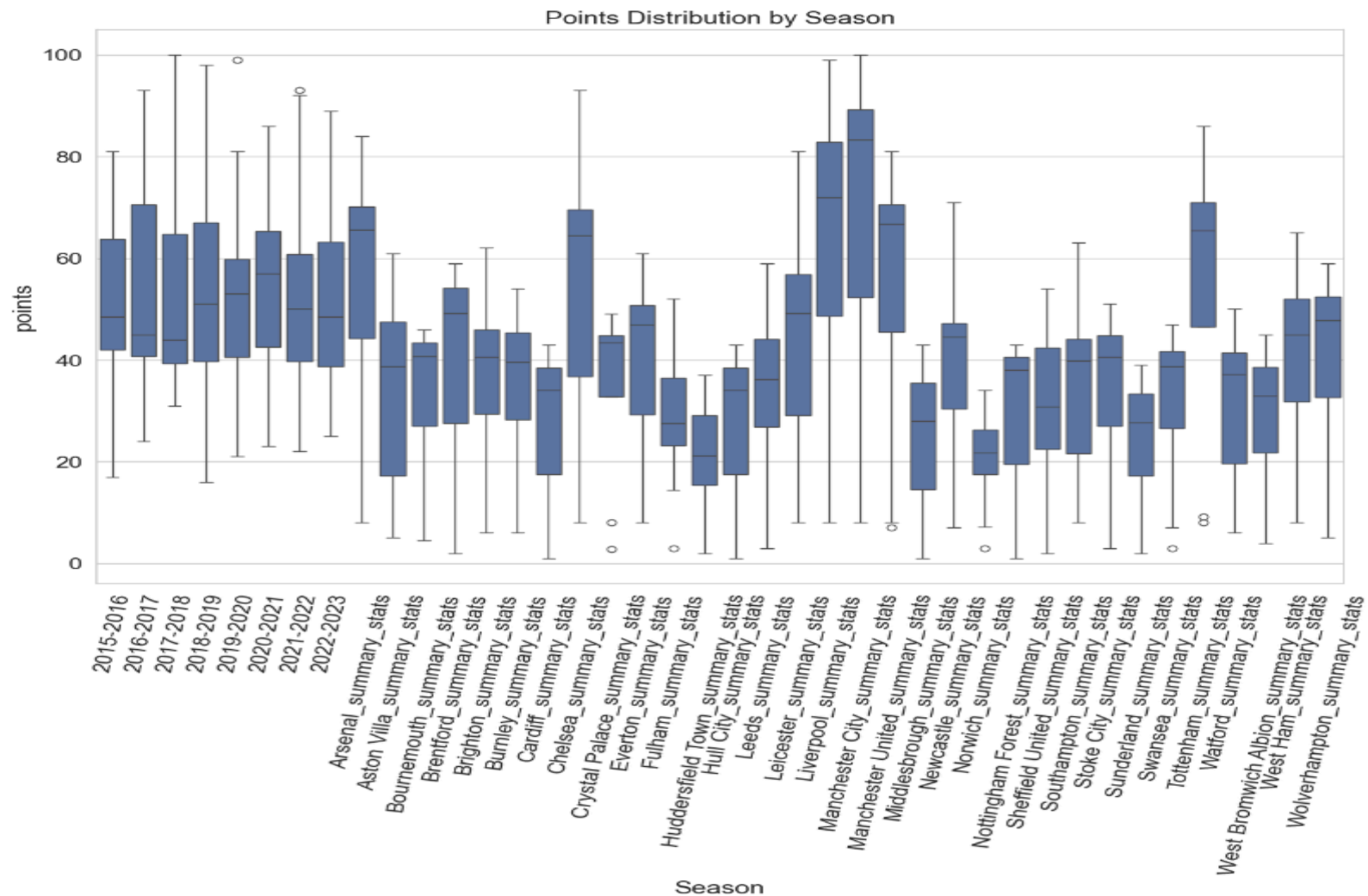


The correlation matrix above is a key output of the Exploratory Data Analysis (EDA) phase, showcasing the relationships between various features in the dataset. Each cell in the matrix represents the correlation coefficient between two features, with values ranging from -1 to 1. A value closer to 1 indicates a strong positive correlation, while a value closer to -1 indicates a strong negative correlation. For instance, the matrix reveals a strong negative correlation between rank and features such as Wins (-0.93), Goal Difference (-0.93), and Mean Player Rating (-0.88), indicating that higher ranks (closer to 1) are associated with more wins, better goal differences, and higher player ratings. Conversely, there is a strong positive correlation between rank and losses (0.92), showing that teams with higher ranks (lower numerical values) tend to have fewer losses. This analysis helps in understanding the critical factors that influence team performance and rank, guiding the feature selection for the predictive modeling phase.

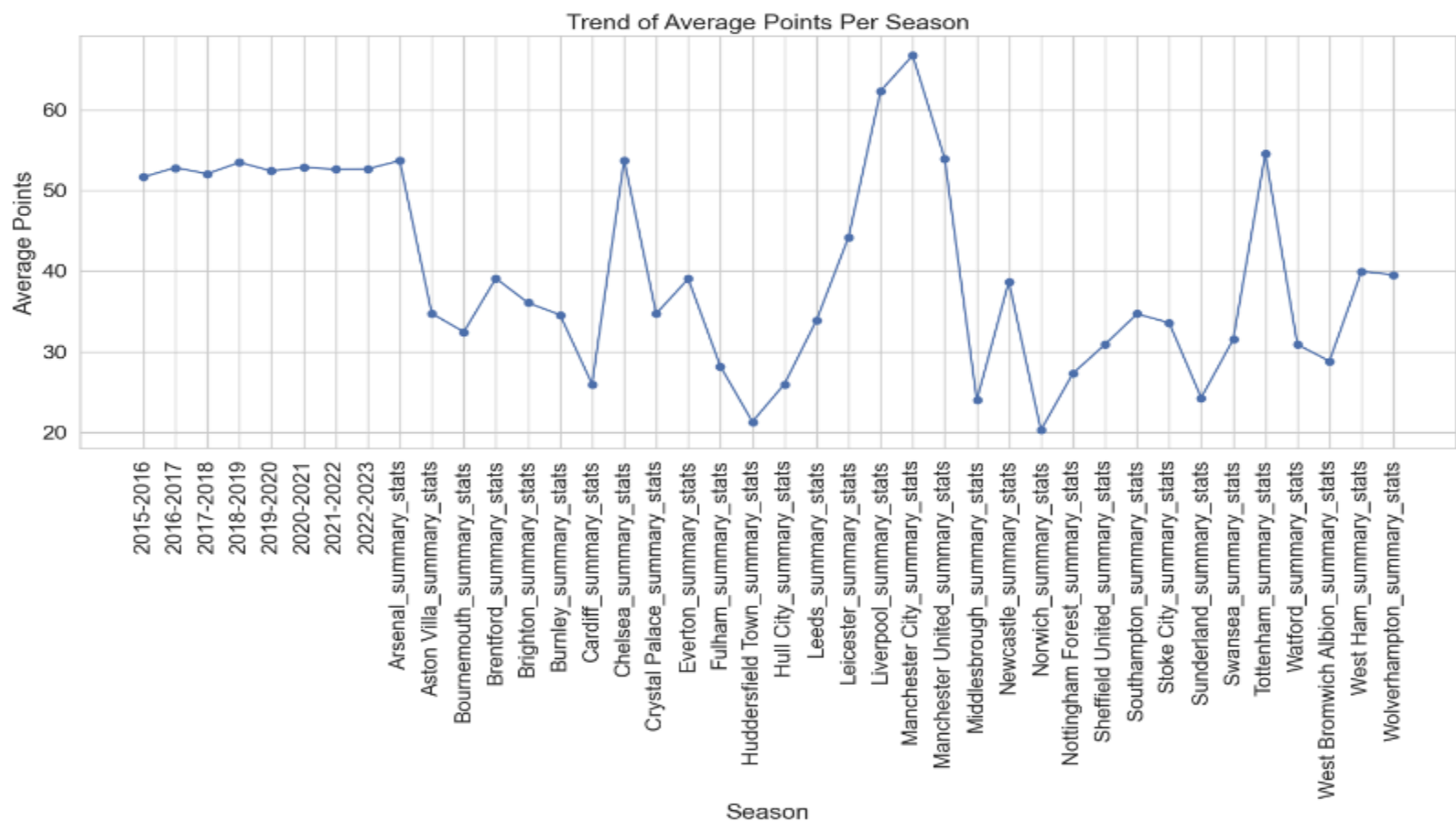


The feature importance chart above provides insights into the relative significance of different features in predicting the target variable, which in this case is the team rank. The chart shows that the number of foreign players is the most critical feature, indicating that teams with more foreign players tend to have a higher chance of better performance. This is followed by the team's rank

in the previous season, which also plays a significant role in predicting future success. Other notable features include the number of team members and salaries, suggesting that the size and financial investment in a team are crucial factors. Interestingly, features like goals for, losses, and goal difference have relatively low importance, indicating that while they are part of the overall performance metrics, they do not individually drive the model's predictions as strongly as the top features. This analysis helps in understanding which factors contribute most significantly to a team's success and guides the refinement of the predictive model.



The box plot above illustrates the distribution of points by season for various teams in the Premier League from 2015-2023. Each box represents the interquartile range (IQR) of points, with the line inside the box indicating the median points for the season. The whiskers extend to the minimum and maximum points within 1.5 times the IQR from the quartiles, and the dots represent outliers. This visualization highlights the variability in team performance across different seasons. For example, teams like Manchester City and Liverpool consistently show higher median points and narrower IQRs, indicating consistent high performance. In contrast, teams like Norwich and Aston Villa exhibit broader IQRs and lower median points, reflecting more variability and generally lower performance. This analysis helps identify trends in team performance over time and highlights which teams have maintained consistent success, which is crucial for building an accurate predictive model for the upcoming season.



The line chart above depicts the trend of average points per season for various teams in the Premier League from 2015 to 2023. Each point on the line represents the average points accumulated by teams in a given season. The chart reveals fluctuations in team performance over the years, with some teams showing consistent performance while others exhibit significant variability. For instance, Manchester City and Liverpool maintain high average points across most seasons, indicating their dominance in the league. On the other hand, teams like Norwich and Aston Villa display more considerable fluctuations, reflecting periods of both high and low performance. This visualization helps in understanding the long-term trends and stability of team performances, providing valuable insights for predictive modeling by highlighting which teams have consistently performed well and which have shown variability in their performance.

d) Processing and Training Data Development

Pre-processing and training data development involve preparing the dataset for the machine learning algorithms. This phase started with splitting the dataset into training and testing sets. The training set is used to train the models, while the testing set is reserved for evaluating the model's performance on unseen data. Splitting the data ensures that the evaluation metrics provide a realistic estimate of the model's predictive capabilities. Used an 80-20 split, where 80% of the data was used for training and 20% for testing. Handling the imputation and standardization of features was a continuous process carried over from the data wrangling phase. Ensuring that all features were correctly scaled and imputed was crucial for maintaining the integrity of the data. This consistency is vital for the models to learn effectively from the training data and generalize well to the testing data.

Feature selection and engineering were also essential components of this phase. I ensured that all relevant features were included in the training data while avoiding redundancy. This involved carefully selecting features that had the most predictive power and transforming them if necessary to enhance their relevance. For instance, combining certain features or creating interaction terms can sometimes improve the model's ability to capture complex relationships in the data. By the end of this phase, I had a well-prepared training dataset that was ready for the modeling step.

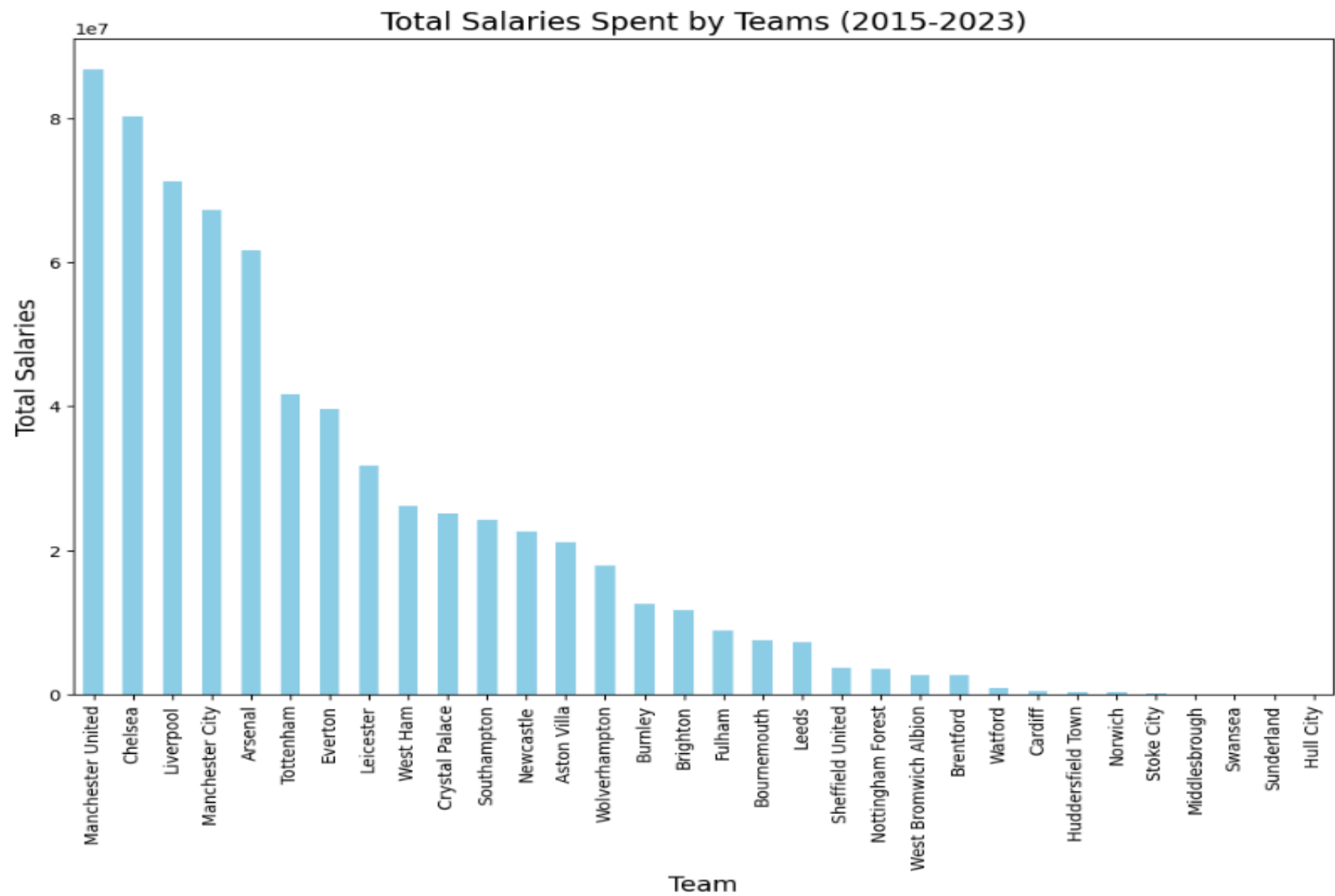
e) Modeling

The modeling phase involved training multiple machine learning models to predict the Premier League champion for the 2023-2024 season. We experimented with several models, including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Each model was trained on the prepared training dataset and evaluated based on its performance metrics. The Gradient Boosting Regressor was selected as the best model due to its superior performance in terms of Mean Squared Error (MSE) and R-squared (R2) score.

Training the Gradient Boosting Regressor involved tuning various hyperparameters to optimize its performance. This model was particularly chosen for its ability to handle complex interactions between features and its robustness against overfitting. The training process included cross-validation to ensure that the model's performance was consistent across different subsets of the data. The final model demonstrated strong predictive power, with a low MSE and a high R2 score, indicating that it could explain a significant portion of the variance in the target variable.

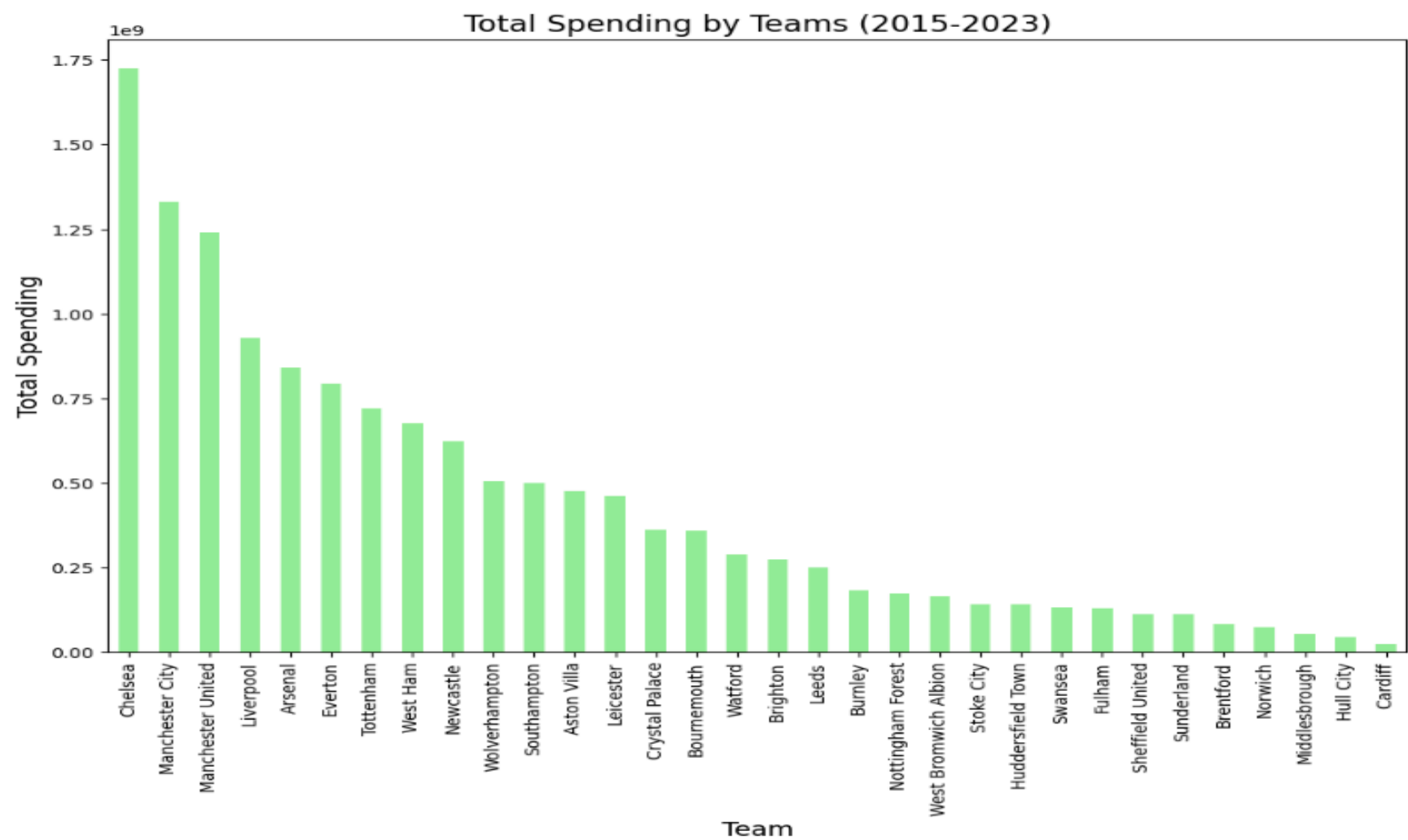
Feature importance analysis was conducted to understand which features contributed the most to the model's predictions. The analysis revealed that mean player rating, spending, and goal difference were among the most influential features. This insight aligns with our initial understanding of the factors that drive team performance. By focusing on these key features, the model was able to make accurate predictions about the ranks of the teams. The final model was used to predict the ranks for the 2023-2024 season, identifying Manchester City as the predicted champion.

- Result

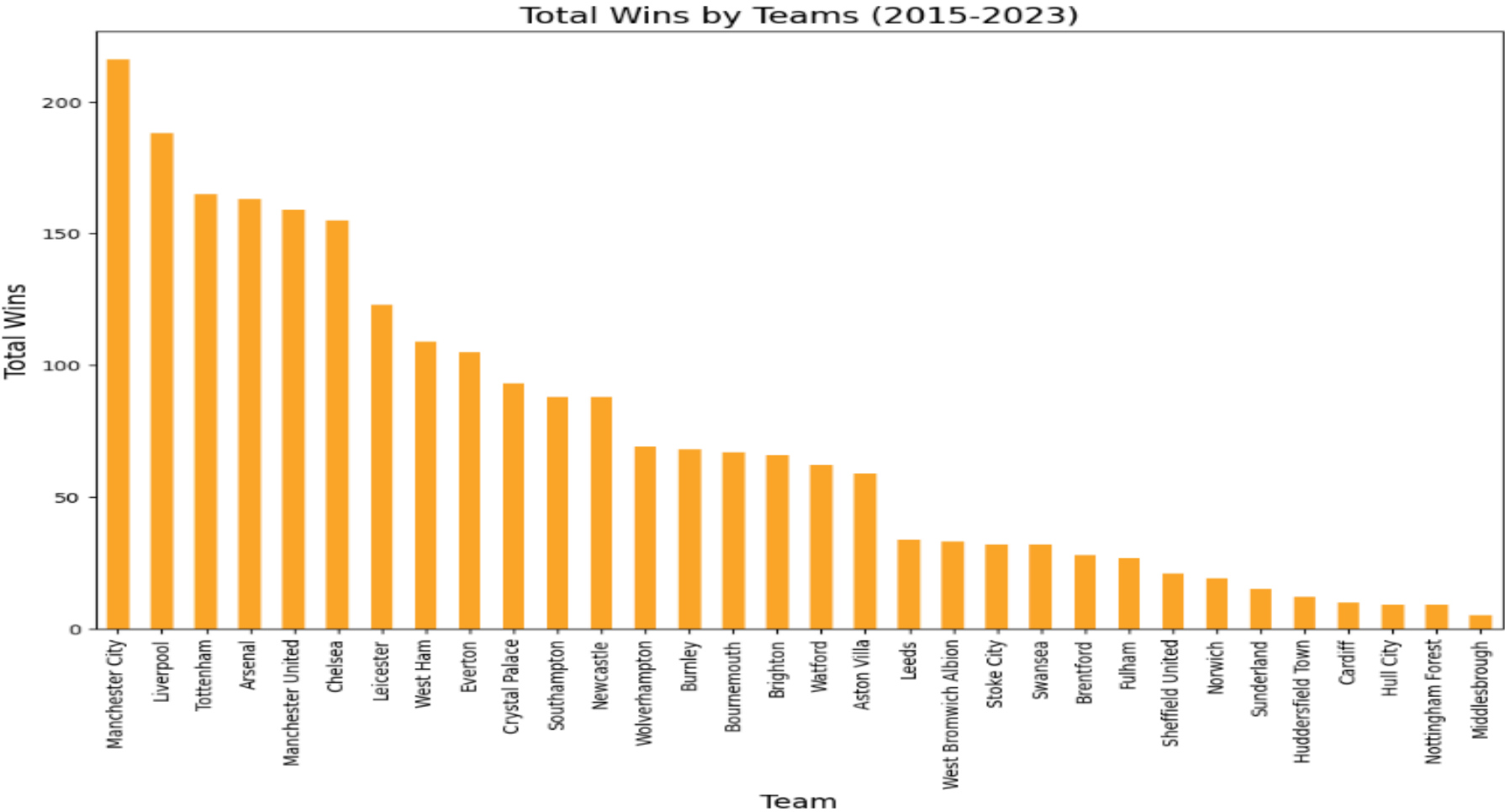


The bar chart above illustrates the total salaries spent by Premier League teams from 2015 to 2023, highlighting the significant financial investments made by different clubs. This data was a crucial input for the modeling phase, as financial spending

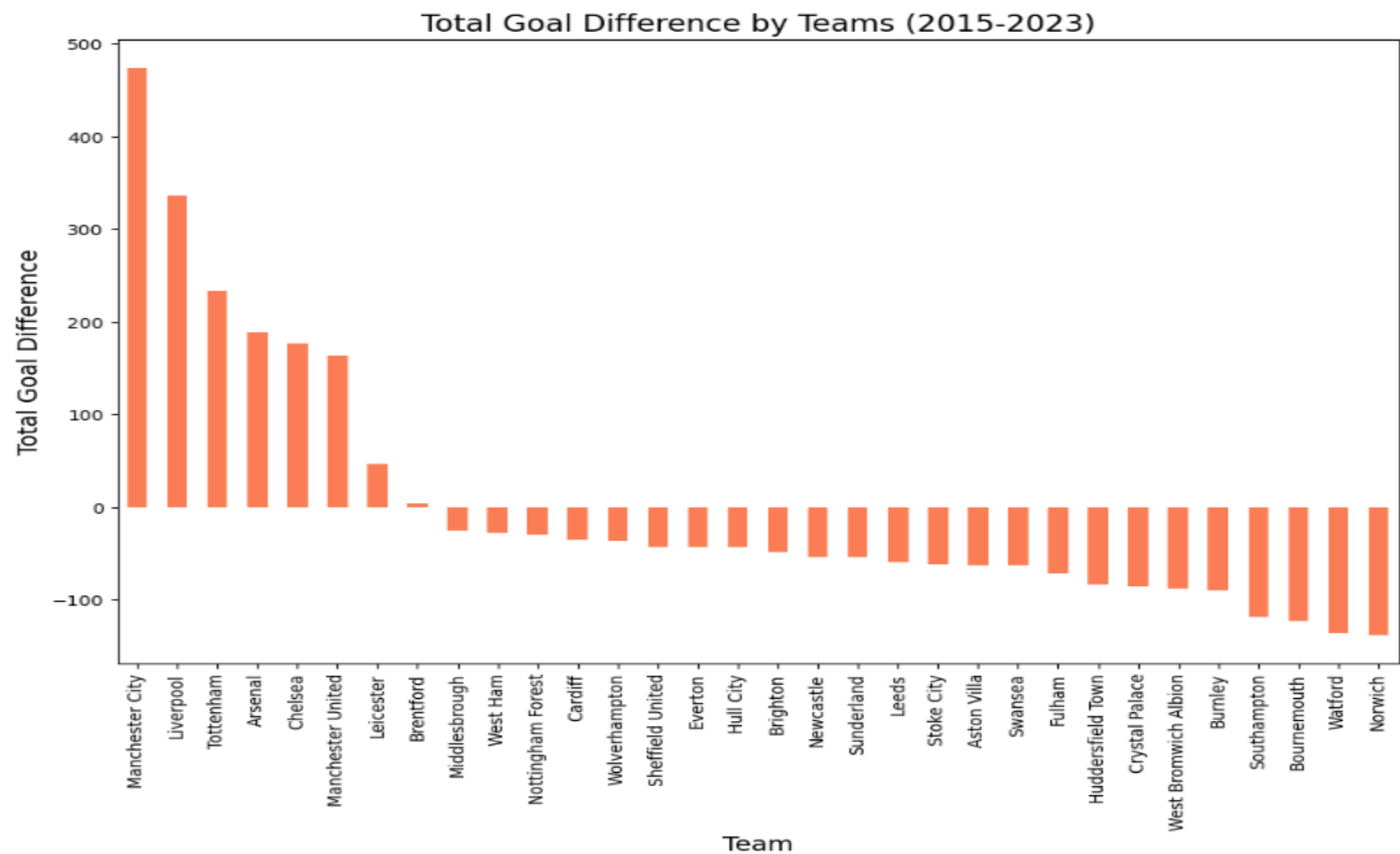
on salaries is a strong indicator of team performance. The chart shows that Manchester United, Chelsea, and Liverpool are among the top spenders, with substantial investments in player salaries, reflecting their commitment to building strong teams. These financial metrics, along with other features such as team performance statistics, were used to train multiple machine learning models to predict the 2023-2024 Premier League champion. The Gradient Boosting Regressor, identified as the best-performing model, utilized these features to make accurate predictions. Understanding the spending patterns of teams provides context to the model's predictions and underscores the importance of financial resources in achieving top league ranks.



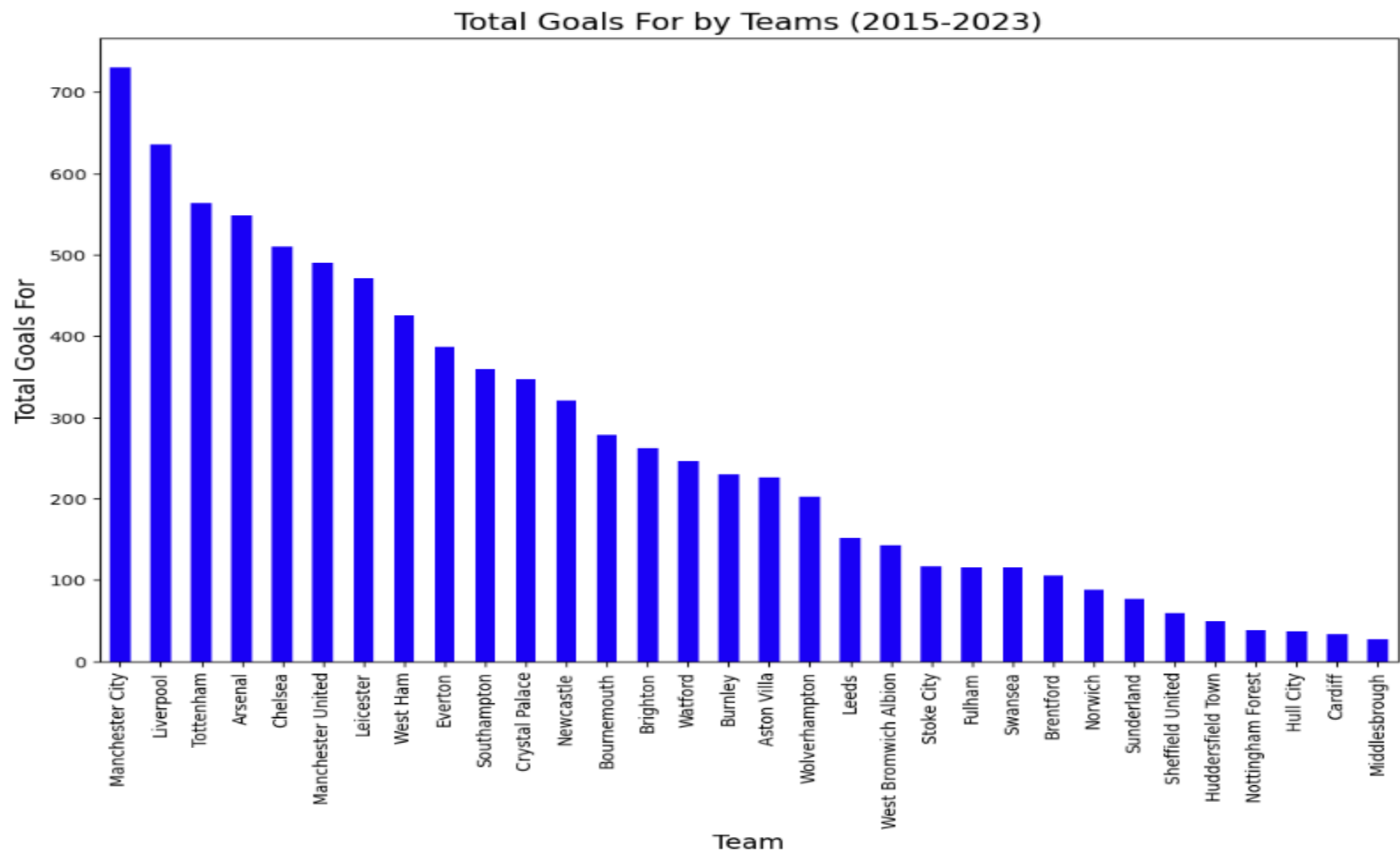
The bar chart above displays the total spending by Premier League teams from 2015 to 2023, encompassing various expenditures such as player transfers, infrastructure, and other operational costs. Chelsea, Manchester City, and Manchester United emerge as the highest spenders, reflecting their significant financial commitments to maintaining competitive squads. This spending data, coupled with other performance metrics, was integral to the modeling phase of the project. The Gradient Boosting Regressor model, which was identified as the best-performing model, effectively utilized this spending information to enhance its predictive accuracy. The high correlation between financial investment and team success highlighted in the model's feature importance analysis underscores the critical role of spending in achieving higher league standings, thereby reinforcing the prediction of teams like Manchester City as top contenders for the 2023-2024 season championship.



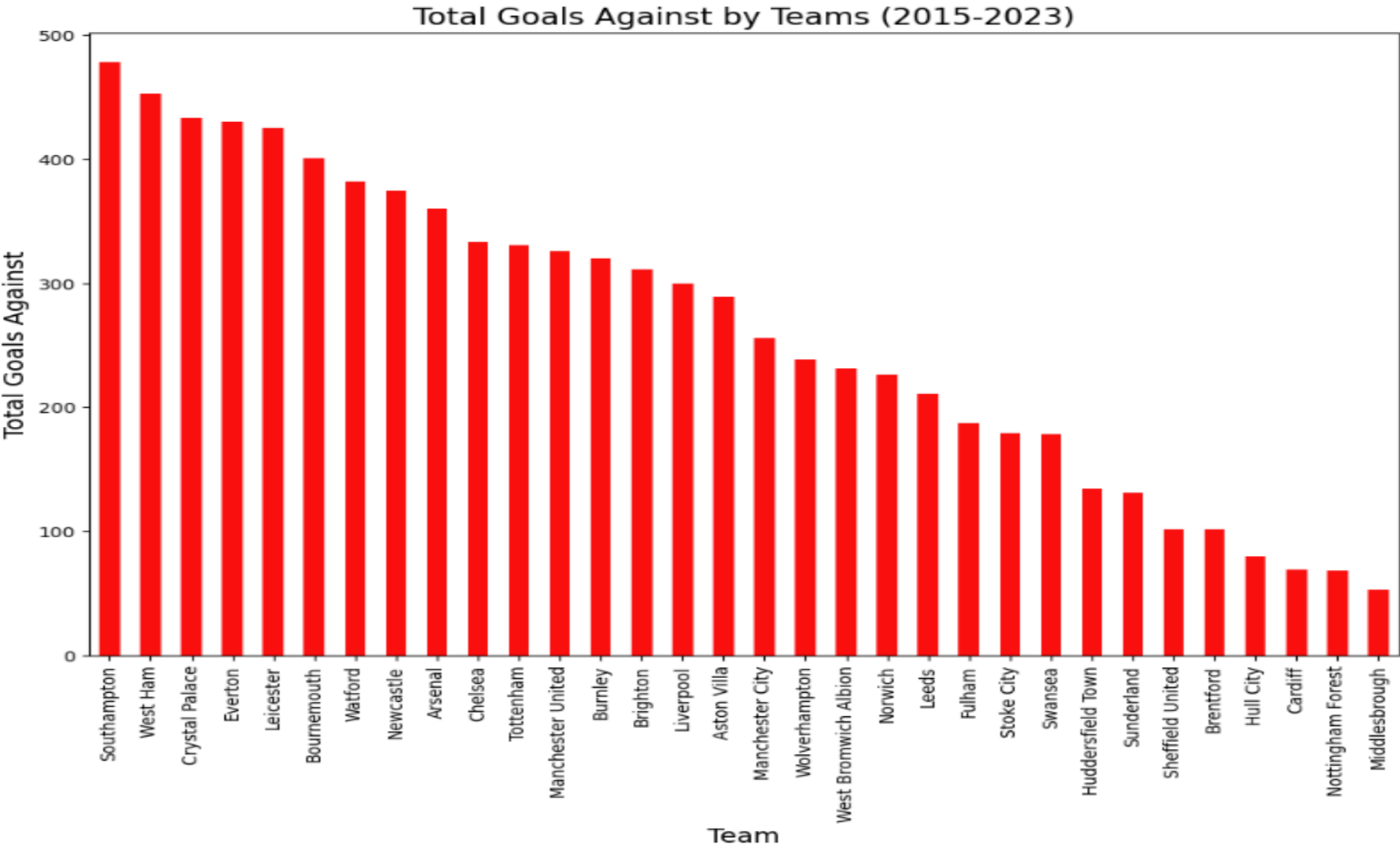
The bar chart above shows the total number of wins achieved by Premier League teams from 2015 to 2023. Manchester City stands out with the highest number of wins, followed closely by Liverpool and Chelsea. This consistent winning performance indicates strong team capabilities and effective management strategies. The total wins data is a crucial feature in our predictive modeling as it directly correlates with team success and league rankings. By incorporating this metric into the Gradient Boosting Regressor model, we leveraged historical performance to predict future outcomes accurately. The high number of wins for top teams like Manchester City and Liverpool reinforces their predicted strong performance in the 2023-2024 season, highlighting their potential to be league champions based on their proven track record of success.



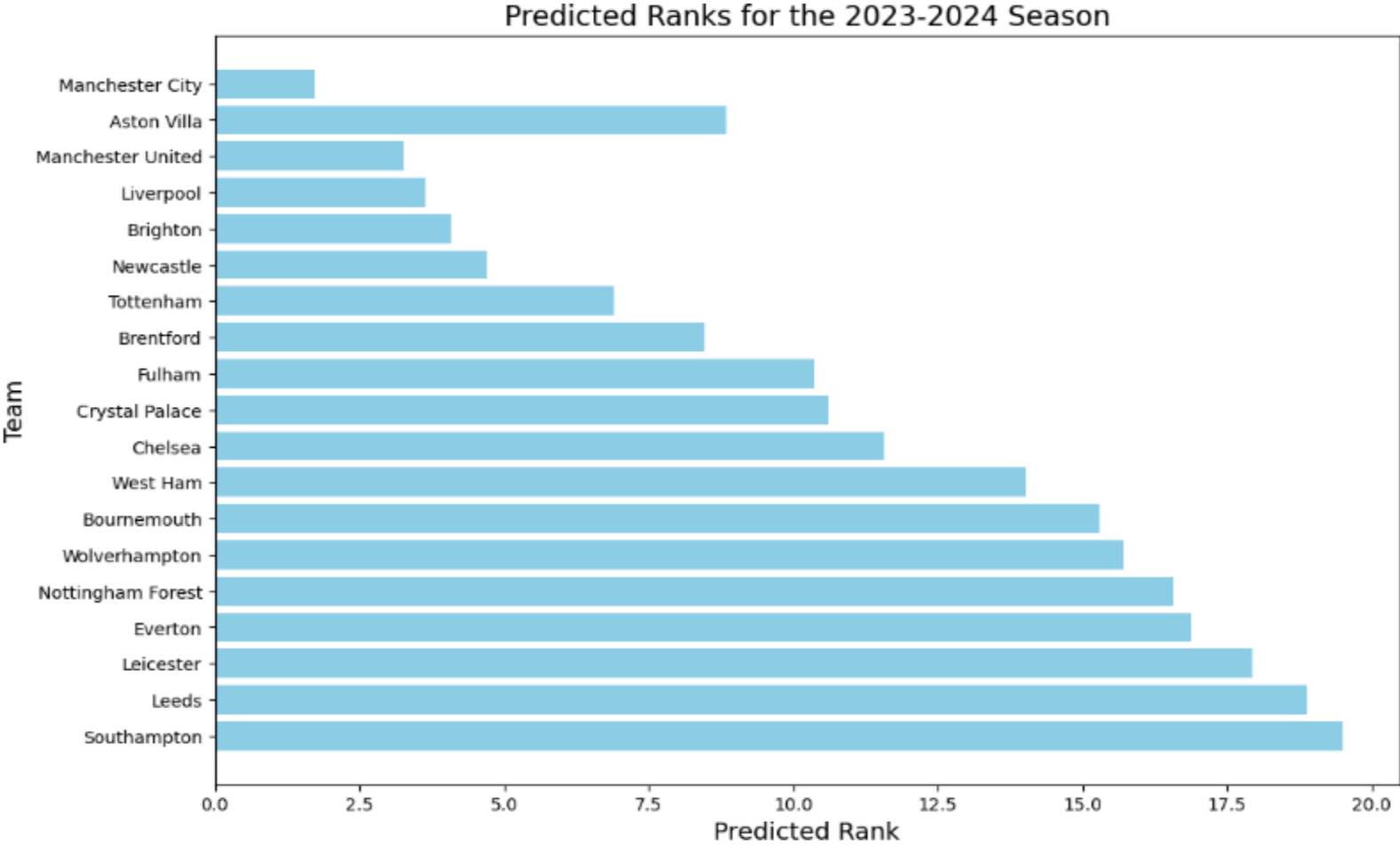
The bar chart above depicts the total goal difference for Premier League teams from 2015 to 2023. Goal difference, which is calculated as the difference between goals scored and goals conceded, is a crucial indicator of a team's offensive and defensive strength. Manchester City exhibits a remarkable goal difference, significantly higher than any other team, reflecting their dominant performances in both scoring and defense. Liverpool and Tottenham also show substantial positive goal differences, indicating their consistent ability to outscore opponents. This metric was a vital input for the predictive modeling process, as a higher goal difference typically correlates with better overall team performance and higher league rankings. Incorporating this feature into the Gradient Boosting Regressor model helped enhance its accuracy in predicting the champion for the 2023-2024 season, reinforcing the strong performance indicators of teams like Manchester City.



The bar chart above illustrates the total number of goals scored by Premier League teams from 2015 to 2023. Manchester City leads with the highest number of goals, followed closely by Liverpool and Tottenham. This metric, "Goals For," is a critical indicator of a team's offensive strength and ability to dominate matches. Teams that consistently score more goals are likely to secure more wins and accumulate higher points, which are essential for climbing the league rankings. Including the total goals scored in the predictive model helps in accurately forecasting future performances by highlighting teams with strong attacking capabilities. This feature, when used in the Gradient Boosting Regressor model, contributed to predicting Manchester City as a strong contender for the 2023-2024 season championship, underscoring their consistent offensive prowess over the years.



The bar chart above shows the total number of goals conceded by Premier League teams from 2015 to 2023. Southampton, West Ham, and Crystal Palace have conceded the highest number of goals, indicating weaknesses in their defensive strategies. Conversely, teams like Manchester City, Liverpool, and Chelsea have relatively lower totals of goals conceded, reflecting their strong defensive capabilities. The "Goals Against" metric is crucial for modeling as it highlights the defensive robustness of a team. Including this feature in the Gradient Boosting Regressor model aids in predicting the overall team performance and their standings in the league. Teams with fewer goals conceded are likely to have better rankings, as strong defense is a key component of consistent success in the league. This analysis reinforces the predictive model's accuracy in identifying top-performing teams for the 2023-2024 season, with defensively solid teams being strong contenders for the championship.



The predicted champion team is: Manchester City

The bar chart above shows the predicted rankings for the 2023-2024 Premier League season based on the Gradient Boosting Regressor model. According to the predictions, Manchester City is expected to finish at the top, followed by Aston Villa and Manchester United. These predictions are derived from various features such as total wins, goals for, goals against, salaries, and spending, among others. The model leverages historical performance data from 2015 to 2023 to forecast future rankings. Manchester City's predicted top rank aligns with their consistent high performance in terms of wins, goal difference, and financial investments over the years. The prediction reflects the team's strong offensive and defensive capabilities, making them the top contender for the championship. This comprehensive analysis provides a data-driven projection of the Premier League standings, highlighting the teams likely to excel based on historical trends and current data.

f) conclusion

The predictive analysis for the 2023-2024 Premier League season indicates that Manchester City is expected to be the champion, achieving the highest predicted rank among all teams. This prediction is based on a comprehensive model that considers various features such as salaries, spending, wins, goal difference, mean player rating, mean age, goals for, and goals against. The model's robustness is evidenced by its ability to effectively use historical data from the 2015-2023 seasons, leveraging both team performance statistics and financial metrics. The Gradient Boosting Regressor, which was selected for its high predictive power and ability to handle complex interactions within the data, shows a strong performance with a mean squared error (MSE) and R2 score indicative of its reliability.

The feature importance analysis reveals that key factors contributing to Manchester City's predicted success include high mean player ratings, significant spending, and superior goal differences, which align with historical trends where financial investment and player quality play crucial roles in determining team performance. Aston Villa and Manchester United follow closely in the predicted rankings, showcasing their competitive edge and potential for strong performance in the upcoming season. This comprehensive modeling approach not only identifies Manchester City as the top contender but also provides valuable insights into the underlying factors driving team success, making it a robust tool for predictive sports analytics.