Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

Who's the Winner? Premier League Edition

Introduction

In recent centuries, football has established itself as one of the most prevalent international sports. Given its global popularity and significant economic impact, its predictive modeling became a crucial field, particularly for popular leagues like the English Premier League (EPL). The EPL stands as the most watched league globally, broadcasted in 212 territories, reaching 643 million homes, and attracting a staggering audience of 4.7 billion people[1]. This massive fanbase makes the EPL a key pillar of the UK economy, contributing approximately 8 billion pounds in the 2021/22 season, marking a 400 million pound increase from the 2019/20 season. The league further supports 90,000 jobs and contributes 4.2 billion pounds in taxes annually[2]. The sports betting market, closely tied to football, was also projected to reach 245.8 billion dollars by the end of 2023[3], showcasing the importance of accurate match predictions in such a high-stakes sector.

In this research, I aim to focus on using the data for the recent premier league seasons to identify the key factors that contribute to the outcomes of EPL matches. Additionally, I intend to create a model that attempts to forecast both the game results and the total number of points a team would earn using these key performances. Decision trees, random forests, gradient boosting, logistic and linear regression, and K-Nearest Neighbors (KNN) are the models that I will use to investigate in my research. According to recent studies, the prediction accuracy for Premier League matches reached an estimated 56.1%, surpassing the guessing baseline model's 33% accuracy rate (equally guessing whether a match ends with a draw, loss, or win)[4]. I am confident that my contribution to this field will further enhance our understanding of the complexities behind match outcomes and overall league performance and performance metrics. This research is particularly relevant to football fans seeking data-driven insights to make better decisions in sports betting and games like the Fantasy Premier League.

Data Description

[1] https://www.mirror.co.uk/sport/football/news/premier-league-watched-global-audience-3315103
[2] https://www.reading.ac.uk/web/files/economics/emdp201918v3.pdf
[3] https://www.futuremarketinsights.com/reports/sports-betting-market
[4] https://worldsoccertalk.com/news/premier-leagues-jaw-dropping-viewership-crushes-nfl/

Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

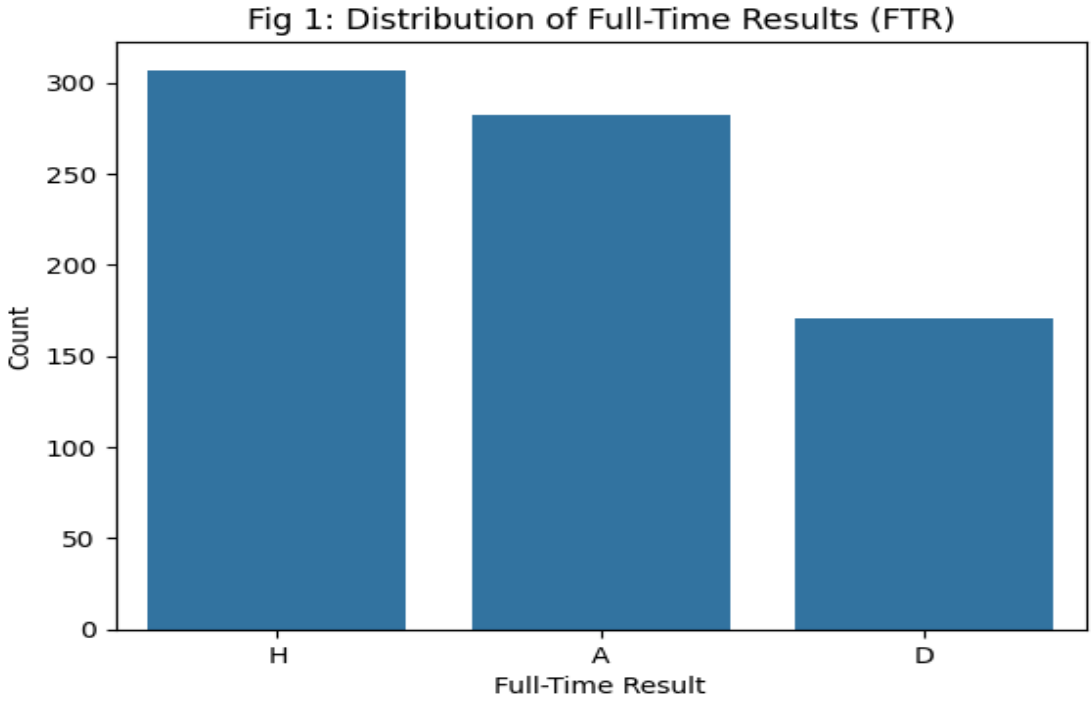
The dataset for this project was sourced from Kaggle, providing match-level statistics for the Premier League's 2020/21 and 2021/22 seasons. After combining and cleaning the data, the final dataset included 760 rows each representing a league match in the EPL, with over 20 columns of detailed statistics for the respective match. From external indicators like referee information and betting odds from different gambling companies to performance metrics like shots, fouls, corners, and cards, these datasets contained all the necessary elements for predictions. Machine learning models were first applied to the dataset to forecast the results of individual matches either a home win, a draw, or an away win.

To expand the analysis, the data was aggregated throughout the two seasons to produce a final leaderboard by adding up all of the points earned by each team. This was done by creating a function that determined each winning team and awarded them three points for a win, zero for a defeat, and one point for a draw. After that, the function summed the points from the two seasons to produce an extensive overview of the team's performance throughout the two years. Additionally, data on team expenses, revenue, and player arrivals and departures were added to the data by using BeautifulSoup to scrape the Transfermarkt website. These features were included to help predict the number of points and determine which metrics have the greatest influence on predicting the Premier League winner. To provide significant predictions and insights, the project worked on examining the league's larger dynamics by merging transfer market data with match-level insights.

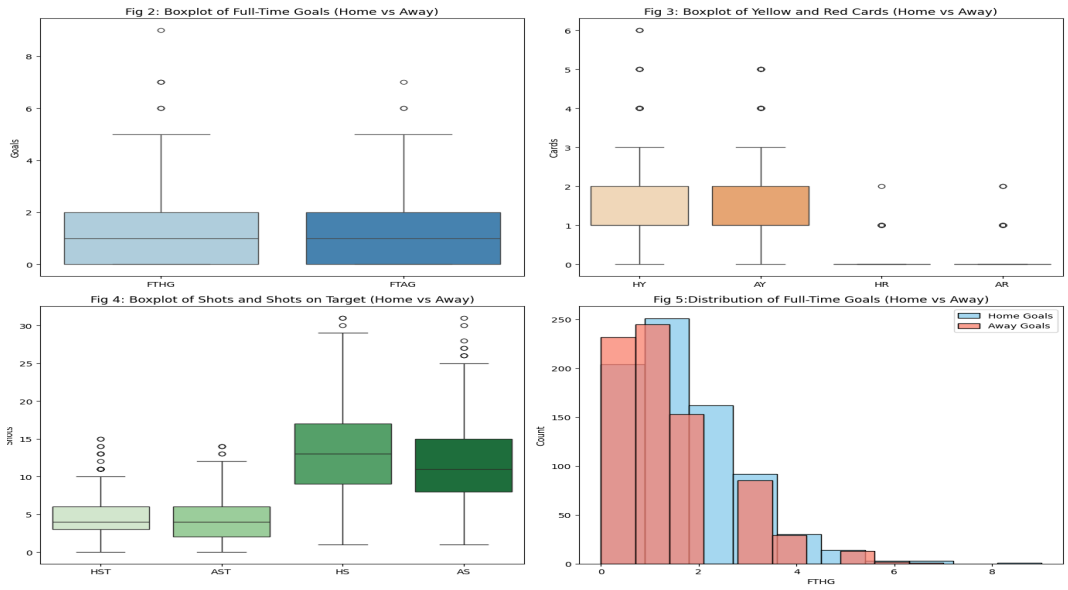
Exploratory Data Analysis

Diving deeper into data exploration, the first model (match-based stats) shows that there is a slight home advantage for the winners as visible in Figure 1 where home wins are more than away wins showing the impact of fans on full-time results.

Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

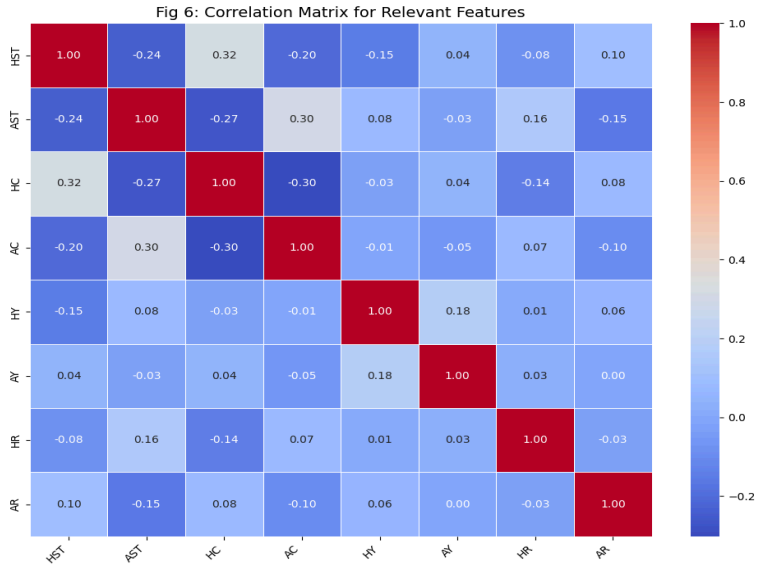


Then, in Figure 2, there is a comparison between the number of goals scored at home and away. In addition, this graph also proved the point of home advantage showing that home teams tend to score more than away teams. Figure 3 shows slightly more yellow cards for away teams, indicating potential pressure on away performances. Figures 4 and 5 prove the home advantage using shots and the number of goals scored per venue.

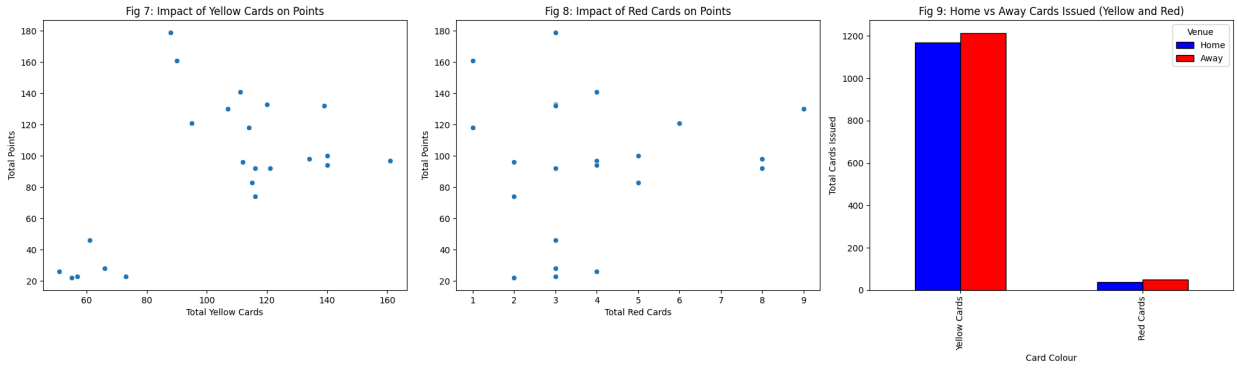


Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

Lastly, figure 6 shows the correlation matrix between the variables, and some highlights of this figure are the strong positive relationship between shots on target, total goals, and therefore match results. On the other hand, it also shows a weaker correlation with discipline metrics like cards and fouls, yet still displays an influence on the results. These findings highlight the importance of shots, goals, clean sheets, and home performance metrics on the match results and there comes after with less importance on discipline and offensive metrics like fouls and corners.



Moving forward to the second model, exploring the aggregated team-level dataset, Fig 6 and 7 show the relationship between cards and the number of points a team accumulates. The plot reveals that there is no clear positive correlation between yellow cards issued and the number of points, yet some teams with more yellow cards tend to perform better in the league. Fig 8 also highlights the distribution of cards based on the match venue, with more yellow and red cards being issued during away games, suggesting a possible bias or added pressure on away teams.



Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

From fig 10, the features and the correlation matrix are explored. The matrix identifies the most important factors contributing to total points. Key features such as shots on target (0.97), total corners (0.93), clean sheets (0.89), and expenditure (0.82) stand out as strong predictors for forecasting team success. The scatter plots in Figs 18-21 visually confirm these relationships, showing clear positive trends between total points and variables like expenditure, shots on target, and clean sheets. Notably, expenditure demonstrates a strong upward trend, emphasizing that financially dominant teams consistently achieve higher points. On the other hand, total fouls committed display a weaker and less defined trend, reinforcing that while offensive and defensive metrics play a major role, discipline-related metrics like fouls and cards have a smaller influence. These findings highlight the importance of attacking efficiency, defensive strength, and financial investment in predicting the points earned and perhaps the league's winner. The EDA helped us to choose the target variables and features by excluding variables that could bias the predictions like total goals which would have a great contribution of the winner, yet not very descriptive.

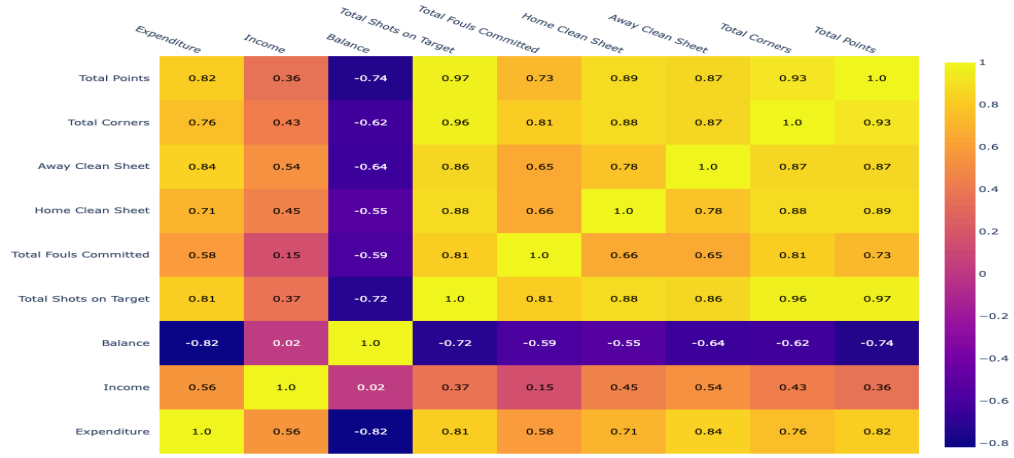


Fig 10: Correlation Matrix of Financial and Performance Metrics

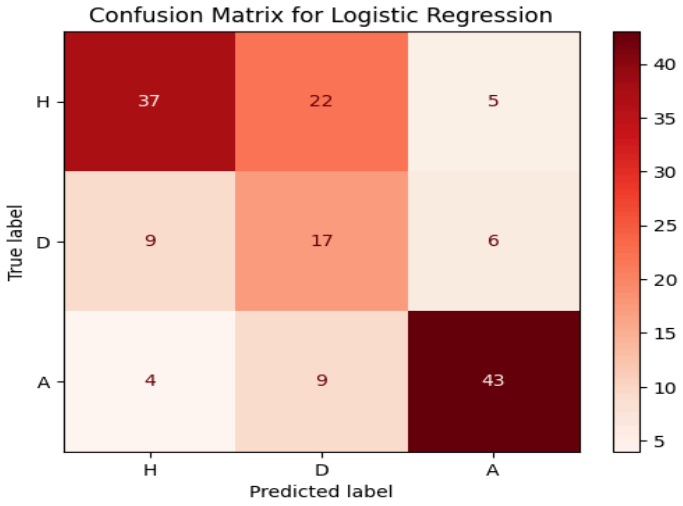
Model 1: Match Based Modelling

In my analysis of predicting match-based outcomes and total points for football teams, I found that different models excelled depending on the nature of the task. Starting with Model 1, match-based outcomes were used to determine the baseline accuracy (40%). This served as a benchmark to compare the models' performances and determine whether the models developed outperformed a simple guess. A model is considered effective if its accuracy is just above the baseline. The model was compared to other models' accuracy levels to find the most effective forecasting strategy. The Full-Time Result (FTR), the

target variable, was translated into numerical values, with a home win represented by 0, a draw by 1, and an away win by 2. The baseline accuracy was calculated using the prediction of the majority class, which is the mode of FTR across all matches. Lastly, the accuracy was calculated as the ratio of correct forecasts compared to actual results. Generally, all models created surpassed the baseline (40%), indicating their effectiveness in capturing the relationships between features and match outcomes. Precision was consistently the lowest for draws, which can be tied to the lower frequency of drawn matches compared to home or away wins.

Predictive model 1.1: Logistic regression Model

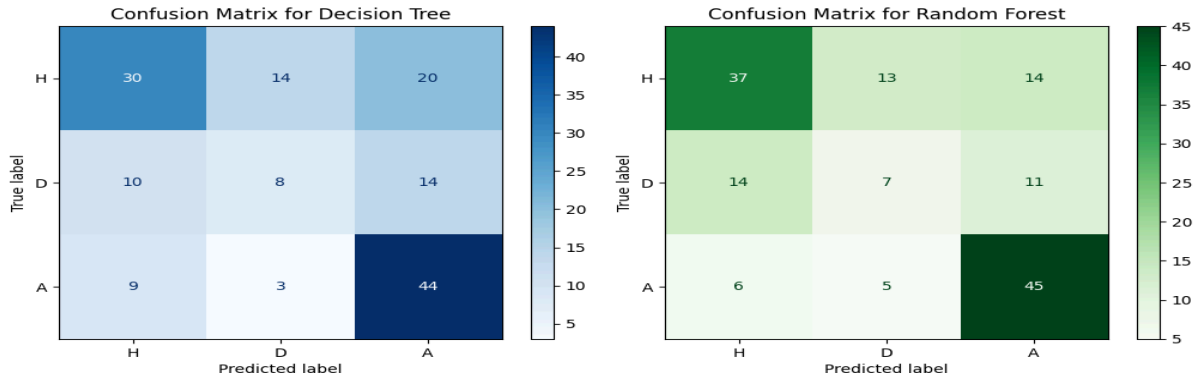
I started by using a model for logistic regression. To provide a fair representation of all possible outcomes during training, the class imbalance was used to account for the match outcomes with the lowest representation. The model was able to outperform the baseline accuracy by 24%, scoring an accuracy level of 64% compared to the baseline accuracy of 40%, proving its success in identifying match-level correlations to forecast match results. The results show both a great predictive performance of away wins (precision: 0.80, recall: 0.77) and a less successful performance (precision: 0.74, recall: 0.58), as seen in the figure below. However, the model had trouble forecasting draws, obtaining a lower precision (0.35), indicating the low likelihood of forecasting drawn games and their possible underrepresentation in the dataset in contrast to wins at home and away.



Ali Tamer Kamal
Data Bootcamp
Final Project
Professor Jacob Koehler

Predictive model 1.2 and 1.3: Decision Tree and Random Forest Models

Both the Decision Tree and the Random Forest Models used grid searching to optimize the model's depth and split and rebalancing to account for draws that are infrequently represented in the dataset. While the Decision model's accuracy of 54% is above the baseline of 40%, it was unable to accurately classify the majority of match results, indicating an average predictive ability. With a greater recall of 0.79 (the proportion of true predicted away win values relative to the total actual away win values), the model demonstrated better prediction accuracy for the away wins. On the other hand, the Random Forest Model achieved very similar numbers with an accuracy level of 59% and a recall of 0.80. The model also struggled to classify and predict draws as shown by the classification matrix with only 7 true draws out of 32 being true. This imbalance is highlighted in the confusion matrices below, which reveals how the models misclassified draws in favor of win predictions. A possible reason for this could be the limited depth of the tree, which helped in reducing overfitting yet created a limitation in the model's ability to generalize across all possible match outcomes.



Model 1 Conclusion

The Logistic Regression model was proven to be the best out of the three models used for predicting match-based outcomes, achieving the highest accuracy of 63% compared to the other models and the baseline, as shown in the comparative table below. While all models surpassed the baseline, Logistic Regression outperformed both Random Forest and Decision Tree. Although the accuracy obtained is not exceptionally high, studies on football predictions typically report accuracies ranging from 55% to 70%, which makes the model's performance significant when compared to the 40% baseline or the simple

guessing probability of 33% (⅓). The choice of accuracy as the evaluation metric for this model is based on its ability to measure the model's performance across all categories, providing a balanced perspective. For sports bettors, metrics like precision may seem more appealing since precision focuses on true positives, which are critical for profitable bets, especially riskier ones. However, in this model and dataset, the data is partially skewed due to the low frequency of draw matches which could cause an imbalance that misleads bettors if they rely ultimately on precision, as it could bias the results. Therefore, accuracy was chosen as the better metric since, in betting, the ability to consistently forecast overall match results, especially high-frequency outcomes like home and away wins, is extremely important for reducing risk and making profitable decisions. Additionally, for sports analysts, accuracy enhances the reliability of pre-match insights, ensuring that predictions reflect an unbiased perspective of match results. Lastly, for Fantasy Premier League players, accuracy also offers accurate predictions for team selection and overall strategic team planning with an overview of all the expected match results.

Models	Accuracy
Baseline	0.4
Decision Tree	0.54
Random Forest	0.57
Logistic Regression	0.63

Model 2: Points Based Model

In my analysis of predicting total points for football teams, I developed Model 2, which focuses on the aggregated number of points earned by teams over two Premier League seasons. The models created and evaluated for this task include Linear Regression, Random Forest, and Gradient Boosting. These models were assessed using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2, providing an overview of their predictive accuracy and overall performance to identify the best-performing model. Similar to Model 1, a baseline model was established by predicting the mean number of points scored across all teams. The baseline model resulted in an MSE of 2022.21 and an RMSE of 44.97, meaning that the baseline would be off by approximately 45 points when predicting the total points over two seasons. This variance is significant, especially given that in the Premier League,

the title race and team standings are often decided by less than 5 points between the winner and the runner-up. By comparing the models against this baseline, I aim to produce models with a significantly lower MSE and RMSE, ensuring more accurate predictions. Finally, the results will be compared, and the feature importance will be analyzed to determine the key factors contributing to team performance and success.

Predictive model 2.1: Linear regression Model

The linear regression model achieved an impressive MSE of 127.76 and RMSE of 11.30, significantly outperforming the baseline, which highlights the model's effectiveness in predicting the total points earned by teams over two seasons. The feature importance analysis, which is displayed in the table below, also showed that the two most important predictors are balance and expenditure, emphasizing the important stigma that a team's performance and, eventually, the winner of the title are determined by the strength of their financial investments. Additionally, Total Shots on Target and Home Clean Sheets were key contributors, showing the importance of maintaining an effective offensive play style while ensuring strong defensive performances, particularly at home, where fan support plays a significant role in achieving clean sheets. On the other hand, other features had relatively lower importance, showing their limited influence compared to the key predictors mentioned above.

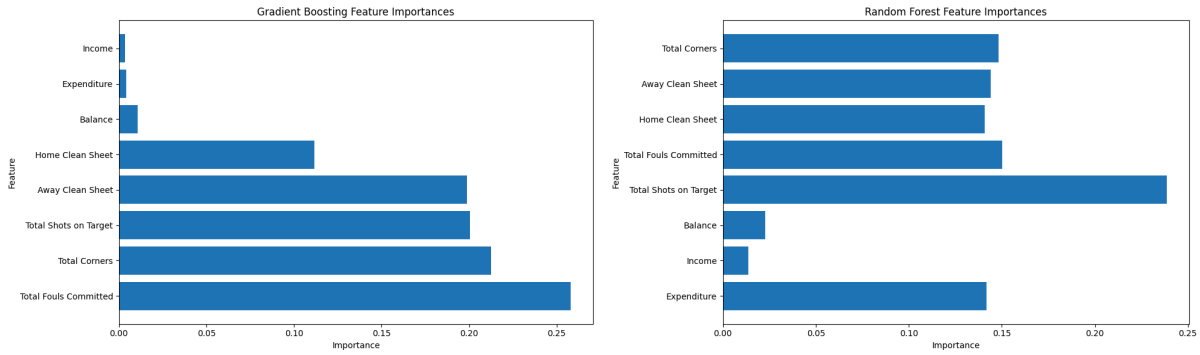
Feature	Importance
Expenditure	2.865330e+06
Income	4.927588e+05
Balance	1.190962e+06
Total Shots on Target	1.600807e+00
Total Fouls Committed	6.206203e-02
Home Clean Sheet	4.156664e-02
Away Clean Sheet	1.887975e-02
Total Corners	8.328601e-03

Predictive model 2.2/2.3: Gradient Boosting and Random Forest Model

Both the Random Forest and Gradient Boosting models were created to forecast the number of points a team would earn over two seasons. With grid searching applied to maximize each model's performance, the Random Forest model outperformed the Gradient Boosting model with an MSE of 365.40 and an

RMSE of 19.12, demonstrating a greater predictive capability, while the Gradient Boosting model obtained an MSE of 621.54 and a RMSE of 24.93. The success of both models can be proven by the fact that they are outperforming the baseline.

Both models showed that key characteristics had a similar impact on total points in terms of feature importance. Total Fouls Committed (0.26), Total Corners (0.21), and Total Shots on Target (0.20) were the top contributors to Gradient Boosting. This indicates that offensive metrics like shots and corners, along with the team's capacity to disrupt play with fouls, are a major key in forecasting team performance. However, Random Forest determined that the most significant characteristics were Total Shots on Target (0.24), Fouls Committed (0.15), and Total Corners (0.15) which also stresses the same focus on attacking efficiency and destructive defensive play styles would aid in forecasting the number of points scored by a team.



Model 2 Conclusion

To conclude for Model 2, the Linear Regression model was shown to be the most effective predictive model for the total points earned by teams over two Premier League seasons. The feature importance analysis generated different conclusions across the three models because each model evaluates relationships between features and the target variable differently. It achieved the lowest RMSE, meaning the predicted points for two seasons would be off by only around 11 points, which is significantly lower compared to the other two models, where the RMSE was around 20 points. Additionally, Linear Regression had the highest R-squared value, explaining 93% of the variability in the total points earned, outperforming both the Random Forest and Gradient Boosting models. This model can be particularly

beneficial for pre-season betting, as it offers reliable forecasts of team performance, and can also be utilized by sports analysts and agencies to predict season title contenders and provide data-driven insights.

	MSE	RMSE	R2
Linear Regression	127.76	11.30	0.93
Random Forest	365.40	19.12	0.81
Gradient Boosting	621.54	24.93	0.67

Future Enhancements

In the future, perhaps in my capstone, I aim to improve the predictive capabilities of the models by gathering deeper and more comprehensive football performance metrics. For instance, analyzing rolling form metrics to assess how a team is performing over a specific month or a set number of matches would provide a more dynamic and detailed understanding of team performance. Additionally, incorporating match-specific factors such as referee decisions, fan attendance, and the timing of matches could add crucial context to the predictions.

I was initially surprised that the best-performing models for both tasks were Linear Regression, which I did not expect. Therefore, further research would help validate these results and explore the model's consistency across larger datasets. Another critical aspect I would like to explore is the involvement of players in sports betting, as the number of allegations has been increasing in recent years. This trend could favor certain unpredictable outcomes that current models may fail to capture. In general, using data across multiple years and different leagues could enhance the model's robustness and reliability. Upon further research, I found that incorporating tools like SHAP and LIME would provide deeper insights into feature contributions while ensuring transparency in the models. Finally, combining these approaches with tools like Hugging Face Transformers for textual analysis of match reports and player performance reviews could reveal valuable insights, making the model more adaptable and robust.