Master
Program : Applied Mathematics and Data Science

Project Title:

Predicting the Winner of English Premier League Matches using Machine Learning: A Random Forest Approach

Authored by:
Brahim TAOUFYQ

This report is submitted under the supervision of :

Prof. **Charaf Hamidi**
Prof. **Salma Gaou**

Academic Year : 2023-2024

Contents

List of Figures

List of Tables

1 Introduction

1.1 Brief Overview of the Problem

The English Premier League (EPL)[1] stands as one of the most competitive football leagues globally, with top teams and a massive fan base. Where the outcomes of matches are influenced by plenty of factors. Predicting the outcomes of EPL matches poses a challenging problem due to the dynamic nature of football, where various factors influence the final result. Traditional approaches to match prediction frequently fail to reflect the complexity of football dynamics to solve this, machine learning is used, with a focus on the Random Forest model. By analyzing historical match data and relevant features.

1.2 Context and Motivation

As a big fan of football, particularly the matches and battles in the English Premier League (EPL), the motivation behind this project comes from a dual passion for the sport and the desire to apply technical skills in a domain that holds personal significance. The EPL, famous for its unpredictability and fiercely contested matches, presents a captivating challenge for prediction.
The project addresses the challenges faced by enthusiasts who participate in betting activities and offers them insightful information to help them make good choices. Additionally, for Fantasy Premier League managers, the ability to predict match winners enhances the strategic selection of players for their formations, adding a layer of excitement and strategy to the gaming experience.
By merging the joy of football fandom with the precision of machine learning, this project aspires to not only improve predictive accuracy but also raise the enjoyment and engagement of those who immerse themselves in the thrilling world of the English Premier League.

1.3 Main Objective

Develop and implement a Random Forest machine learning model capable of analyzing historical EPL match data to predict the team likely to win a given match.

2 Methodology

2.1 Data Source and Key Features Overview:

Data Source: The historical match data for the English Premier League utilized in this project was sourced from the FBREF website[2].

Scope of the Data: The dataset covers a period spanning from the 2018/2019 to the 2022/2023 season, offering a comprehensive perspective on match outcomes across multiple seasons.

Key Features: The dataset encompasses various features, including date, time, competition, round, venue, result, goals scored (GF), goals conceded (GA), expected goals (xG, xGA), possession (Poss), attendance, and more. To predict match winners, key features were selected based on their perceived significance in determining team strength and influencing match outcomes. Notable features include venue, goals scored (GF), goals conceded (GA), expected goals against (xGA), possession (Poss), and expected goals (xG).

2.2 Data Cleaning:

All the data manipulation presented below was performed using the Pandas library in Python

Handling Missing Values: Columns deemed non-informative for predicting match winners, such as "Comp," "Referee," "Match Report," and "Notes," were excluded from the analysis to streamline the dataset.

Data Transformation: The 'Date' column was converted to datetime format for better handling of temporal data. Additionally, a mapping for team name abbreviations was defined for clarity and consistency.

Encoding Categorical Variables: Categorical variables like 'Result,' 'Venue,' 'Opponent,' 'Team,' 'Time,' and 'Date' were encoded to numerical values for compatibility with machine learning algorithms. For example, 'Result' was encoded as Win=1, Loss=0, and Draw=0. **Normalization and Standardization:** Certain raw data transformations were applied, including normalizing the 'Date' column and defining a mapping for team name abbreviations.

2.3 Data Splitting:

To evaluate the performance of the machine learning model, the dataset was split into training and testing sets based on the 'Date' column. Training data includes matches before January 1, 2022, while testing data comprises matches on or after January 1, 2022.

2.4 Exploratory Data Analysis (EDA):

All the figures presented below were created using the **Tableau**[3] tool for data visualization.

Team Performance Trend: Figure 1 highlights teams like Manchester City and Liverpool as consistent performers with the most wins over the past five seasons. Conversely, teams like Luton Town, West Bromwich Albion, and Norwich City have fewer wins.

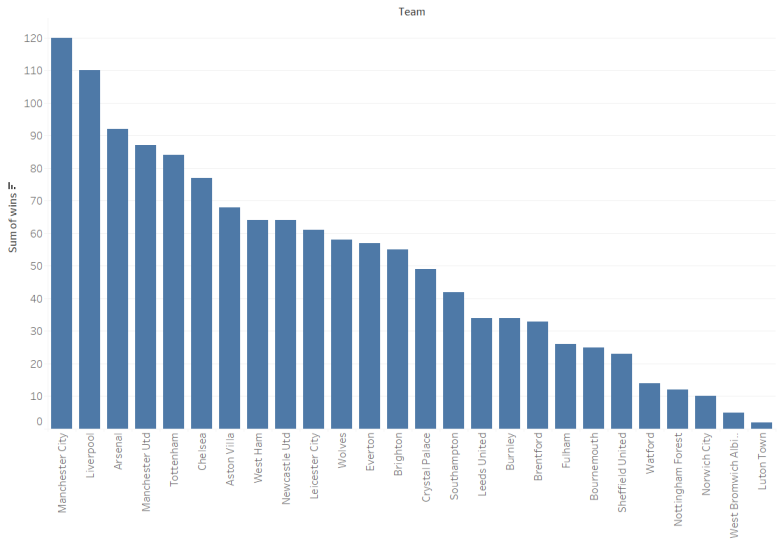

Figure 1: Team Performance Trend figure.

Effect of Expected Goals (xG) and Goals For (GF): Figures 2 and 3 demonstrate that higher expected goals (xG) and goals scored (GF) correspond to a greater probability of winning, emphasizing the importance of offensive capabilities.

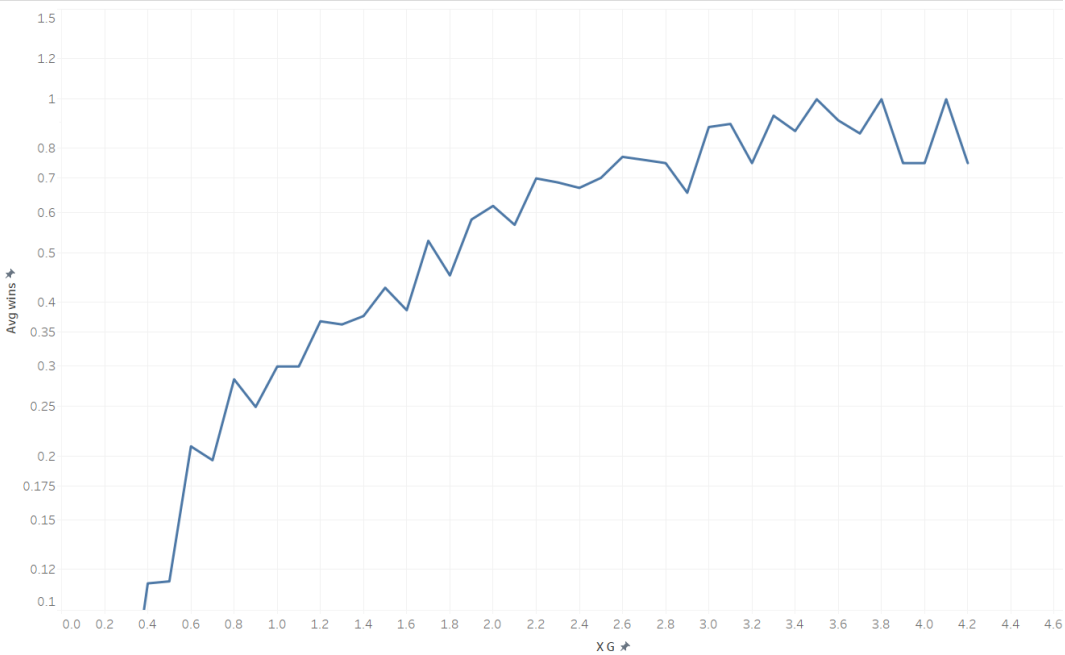

Figure 2: Expected Goals Impact on Wins.

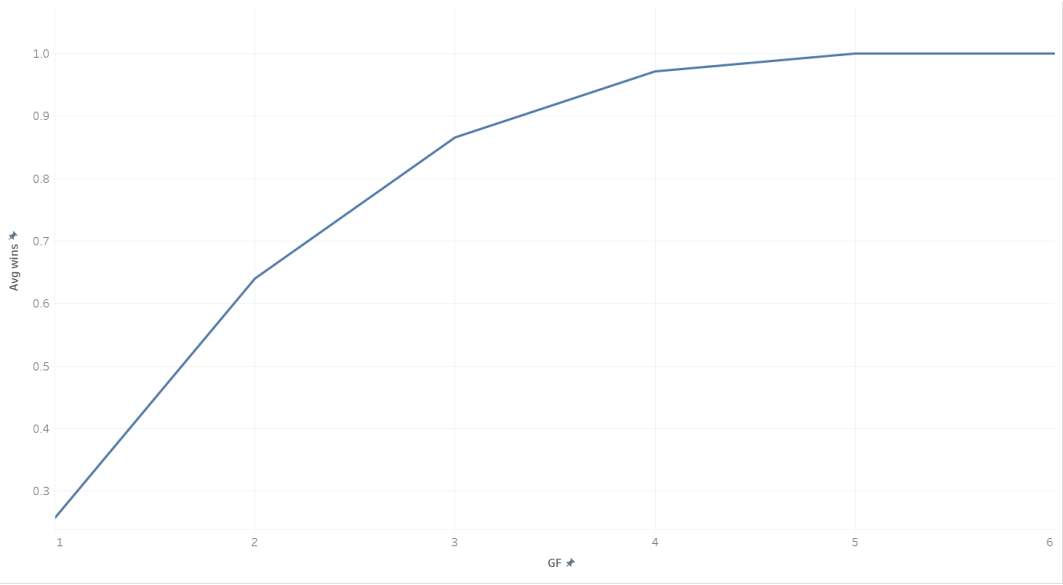

Figure 3: Goals For Impact on Wins.

Shots and Shots on Target Influence: Figures 4 and 5 reveal that higher shot attempts (Sh) and shots on target (SoT) are associated with a greater probability of winning, indicating the significance of attacking prowess.

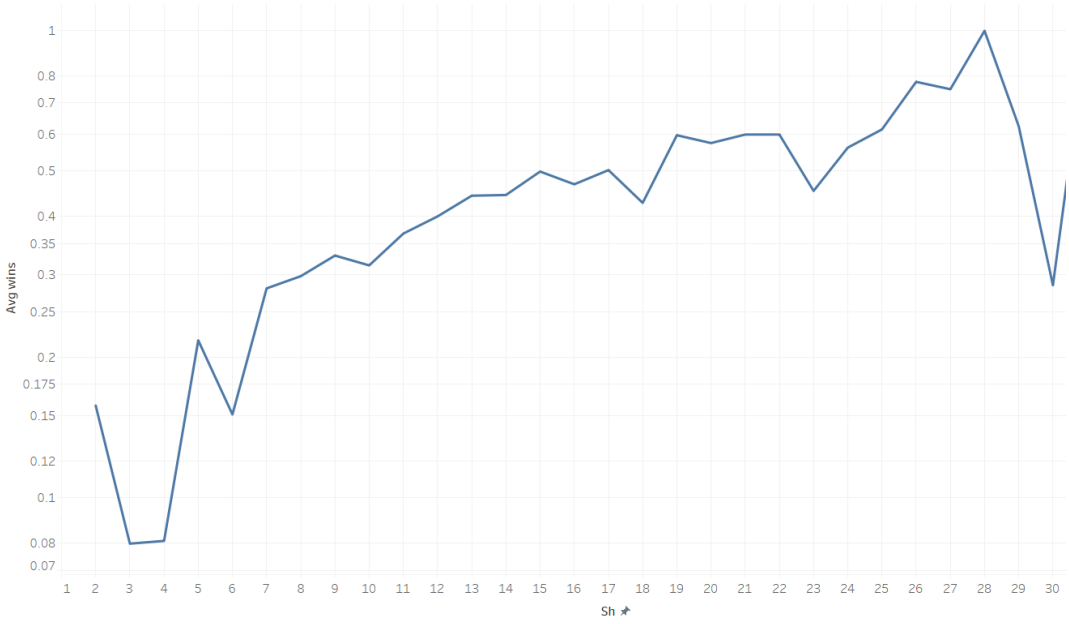

Figure 4: Shots Impact on Wins.

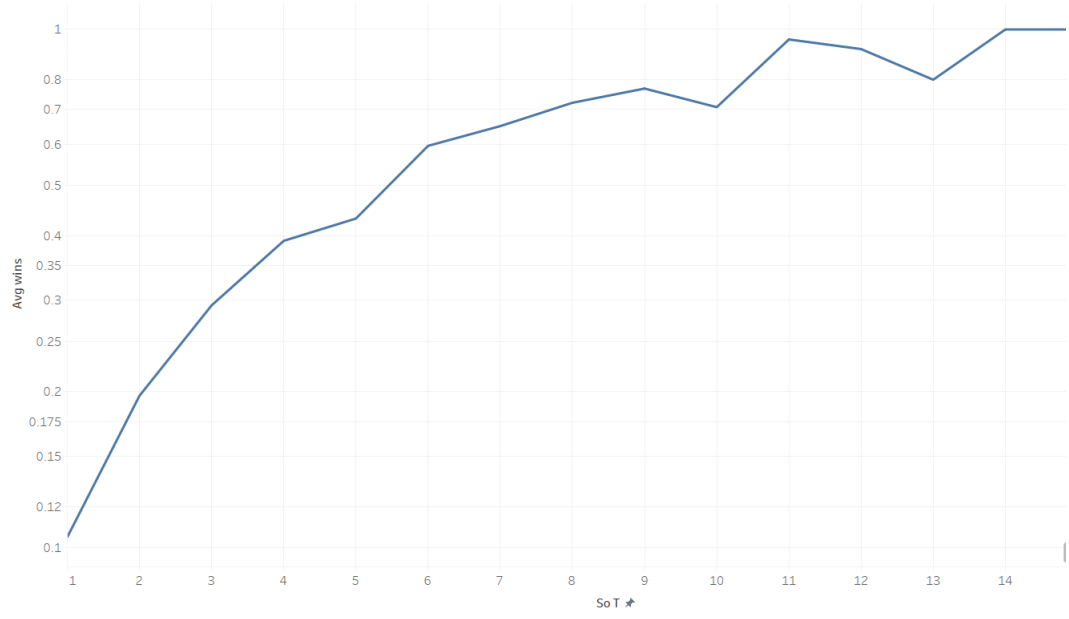

Figure 5: Shots on Target Impact on Wins.

Defensive Metrics - Expected Goals Against (xGA): Figure 6 illustrates that a higher expected goal against (xGA) decreases the probability of winning, underscoring the importance of defensive capabilities.

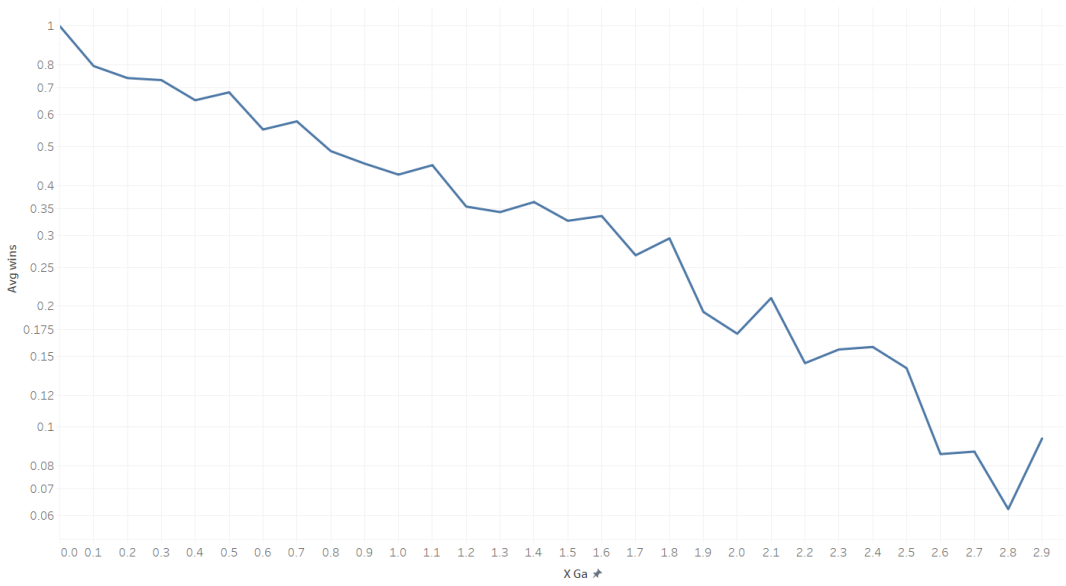

Figure 6: Expected Goal Against Impact on Wins.

Possession Impact: Figure 7 highlight a positive correlation between possession and the probability of winning, suggesting the influence of ball control on match outcomes.

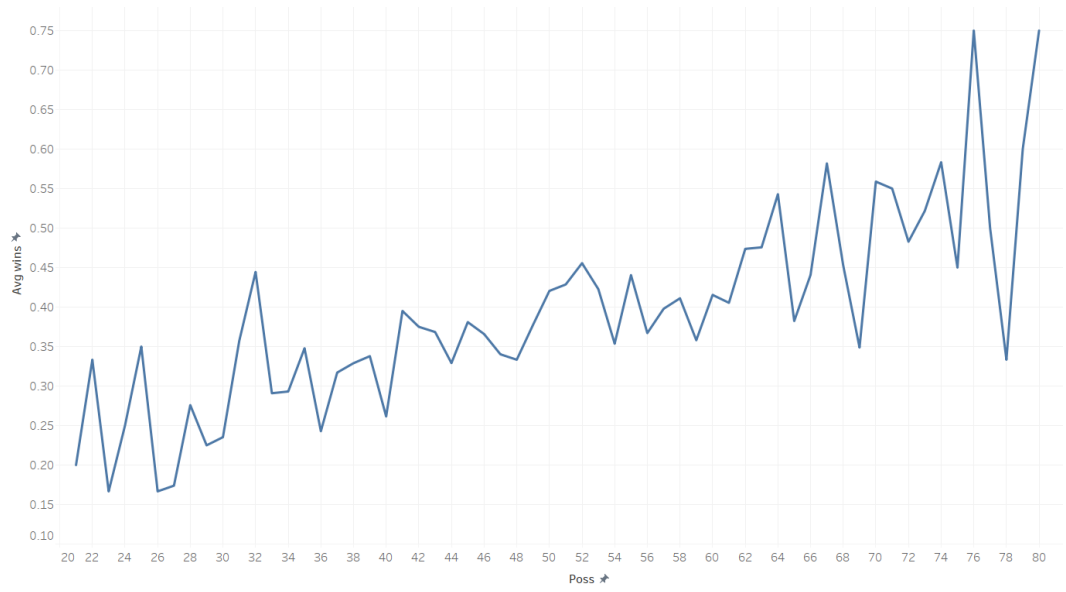

Figure 7: Possession Impact on Wins.

Home Advantage: Figure 8 indicates that playing at home increases the likelihood of winning, reinforcing the concept of home advantage in football.

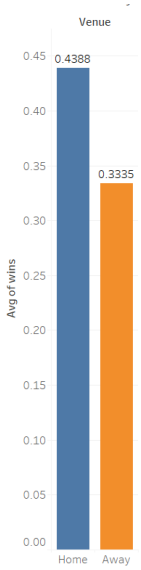

Figure 8: Home/Away Impact on Wins.

Exceeding Expected Goals: Figure 9 shows that teams exceeding their expected goals (the short black line) are often strong teams with a high probability of winning.

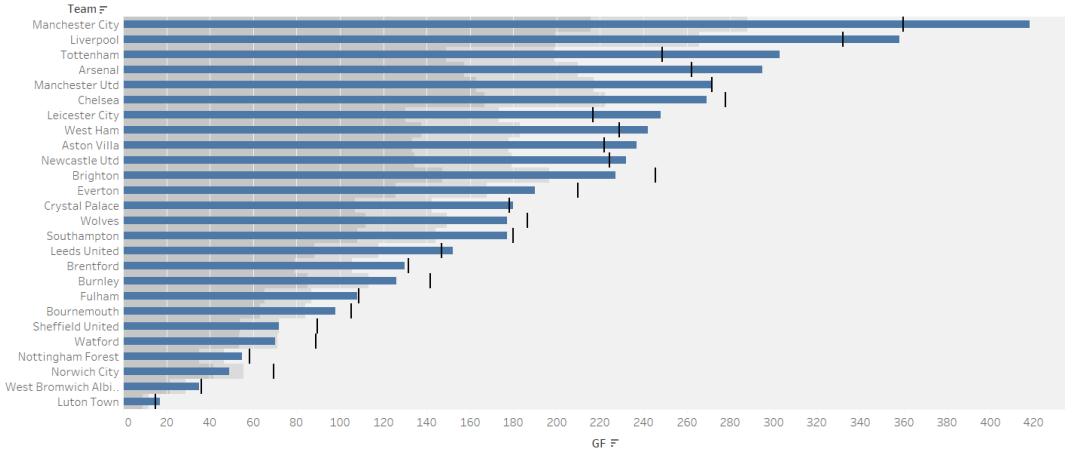

Figure 9: Exceeding Expected Goals Impact on Wins.

2.5 Choice of Random Forest Model:

The Random Forest model[4] was selected for its ensemble learning capabilities, robustness, and ability to handle complex relationships in the data. Its built-in feature importance analysis aligns with the project's goal of understanding the key determinants of match success. The non-linear nature of football match outcomes makes Random Forest[5] well-suited for capturing intricate interactions among various factors. Its resistance to overfitting enhances generalization to unseen data, crucial for the unpredictable nature of football matches.

2.6 Challenges in Predicting Football Match Outcomes:

Football match prediction poses inherent challenges due to the unpredictable nature of the sport. The presence of loss and draw as separate classes from wins adds complexity to the prediction task. The unpredictable nature of football matches presents challenges in building accurate predictive models.

2.7 Assessment of Feature Importance:

The importance of features in predicting EPL match winners are various such as :
team = The team the stats belong to (i.e Manchester City)
date = Date of the match
time = Kick-off time of the match
day = The day the match took place on (i.e Monday, Tuesday etc)
venue = Whether team was Home, Away or Neutral venue
gf = How many goals the team scored
ga = How many goals the team conceded
opponent = Who the team faced that day
xg = Expected goals
xa = Expected goals allowed
poss = Possession
sh = Shots total
sot = Shots on target

2.8 Detailing Parameters and Hyperparameterization:

In this section, we are going to specify some of the hyperparameters and try to optimize them using GridSearchCV.

2.8.1 Model Hyperparameters:

The Random Forest Classifier for predicting football match winners involves several key parameters that influence the ensemble learning process. Notable parameters include:

n_estimators: The number of decision trees grown during the ensemble learning process. More trees generally contribute to better performance up to a certain point.

max_depth: The maximum depth of each individual tree in the ensemble, representing the length of the longest path from the root node to a leaf node.

min_samples_split: The minimum number of samples required to split an internal node during tree-building. Nodes with samples less than this threshold become leaf nodes.

Cross-Validation (cv): In machine learning, cross-validation is a crucial technique for assessing a model's performance. In the context of this project, a 5-fold cross-validation (cv=5) was implemented. This involves splitting the dataset into five folds, training and evaluating the model

five times, and averaging the performance metrics to provide a robust evaluation.

Scoring Metric: The scoring metric used during hyperparameter tuning was 'roc_auc,' which is a common metric for binary classification problems. It assesses the area under the Receiver Operating Characteristic (ROC) curve, providing a comprehensive measure of model performance.

Parallel Processing (n_jobs): The n_jobs=-1 parameter was set to leverage parallel processing, utilizing all available CPU cores for efficient computation during the grid search.

2.8.2 Hyperparameter Optimization using GridSearchCV:

To identify the optimal combination of hyperparameters, a GridSearchCV approach was employed. The grid search focused on exploring various values for n_estimators, max_depth, and min_samples_split to enhance model performance.
The result is : Best Hyperparameters: **'max_depth'**: 5, **'min_samples_split'**: 20, **'n_estimators'**: 10

2.9 Identification of Powerful Variables and Rolling Averages

In the section of the data analysis for this project, a comprehensive examination of various performance metrics was undertaken to discern the factors influencing match outcomes in the English Premier League. Through rigorous exploration, certain variables emerged as particularly influential in capturing the essence of a team's performance.
The identified powerful variables, which will serve as key predictors in our machine learning model, include: **"GF" (Goals For):** The number of goals a team scores in a match. **"GA" (Goals Against):** The number of goals conceded by a team in a match. **"xGA" (Expected Goals Against):** The expected number of goals a team is anticipated to concede in a match. **"Poss" (Possession):** The percentage of time a team spends in control of the ball during a match. **"xG" (Expected Goals):** The expected number of goals a team is anticipated to score in a match. **"p" (Points Gained in a Single Match):** The points accumulated by a team in a single match (a metric reflecting the outcome of the match – win, draw, or loss).
Understanding that a team's recent performance can be indicative of its current strength, a function was crafted to calculate rolling averages for these specific performance metrics within each team over the past 15 games. This function, applied to our dataset, provides a dynamic and smoothed representation of a team's recent form, offering a comprehensive overview of its evolving performance trends.

```
def rolling_averages(group, cols, new_cols):
    group = group.sort_values("Date")
    rolling_stats = group[cols].rolling(15,closed='left').mean()
    group[new_cols] = rolling_stats
    group = group.dropna(subset=new_cols)
    return group
cols = ["GF", "GA","xGA","Poss","xG","p"]
new_cols = [f"{c}_rolling" for c in cols]
matches_rolling = matches.groupby("Team").apply(lambda x: rolling_averages(x, cols, new_cols))
matches_rolling = matches_rolling.droplevel('Team')
matches_rolling.index = range(matches_rolling.shape[0])
```

Figure 10: Rolling Averages Function

This function enhances our ability to incorporate the dynamic nature of team performance into our machine learning model. By considering rolling averages over the previous 15 games, we aim to capture nuanced patterns and fluctuations in team form, ultimately contributing to the accuracy and predictive power of our model.

3 Results

This section delves into our project's outcomes, leveraging both visual representations and statistical analyses. By combining intuitive graphs, tables, and key numerical measures, we aim to paint a comprehensive picture of our machine learning model's predictive capabilities in the English Premier League.

3.1 Present the Results:

Team-wise Model Predictions and Actual Outcomes: A detailed table showing predicted outcomes and actual results for each team in the dataset.

Table 1: Team-wise Predictions vs Actual Outcomes

Team	Actual Wins	Predicted Wins	Actual No Wins	Predicted No Wins
Arsenal	46	37	24	33
Aston Villa	35	10	38	63
Bournemouth	14	0	38	52
Brentford	27	2	45	70
Brighton	31	17	42	56
Burnley	6	0	31	37
Chelsea	24	25	47	46
Crystal Palace	20	2	51	69
Everton	22	0	51	73
Fulham	18	0	34	52
Leeds United	13	1	45	57
Leicester City	15	3	42	54
Liverpool	45	50	26	21
Manchester City	47	67	22	2
Manchester Utd	37	34	36	39
Newcastle Utd	38	26	34	46
Tottenham	37	36	35	36
West Ham	23	1	47	69
Wolves	23	0	49	72

The table showcasing the model's predictions for wins and no wins alongside the actual counts. Notable observations include instances where the model closely aligned with reality, as seen with **Chelsea**, accurately predicting both wins and no wins. However, challenges are evident in teams like **Aston Villa** and **Brentford**, where the model significantly underestimated wins and overestimated no wins.

Model Evaluation Metrics by Team: A table presenting evaluation metrics (accuracy, precision, recall, F1-score) for each team, giving a comprehensive view of model performance.

Table 2: Model Evaluation Metrics by Team

Team	Accuracy	Precision	Recall	F1-Score
Brighton	52.00%	22.22%	6.45%	10.00%
Aston Villa	54.67%	57.14%	11.43%	19.05%
Crystal Palace	63.01%	12.50%	4.76%	6.90%
Arsenal	36.11%	62.50%	10.42%	17.86%
Everton	69.33%	44.44%	18.18%	25.81%
Liverpool	47.95%	90.00%	19.57%	32.14%
Manchester City	43.66%	100.00%	18.37%	31.03%
Tottenham	45.95%	42.86%	7.69%	13.04%
Newcastle Utd	50.00%	100.00%	7.50%	13.95%
Manchester Utd	54.05%	75.00%	15.79%	26.09%
West Ham	68.06%	55.56%	20.83%	30.30%
Chelsea	63.89%	42.86%	12.00%	18.75%
Wolves	63.51%	33.33%	12.50%	18.18%
Leicester City	72.41%	50.00%	12.50%	20.00%
Southampton	78.95%	33.33%	9.09%	14.29%
Burnley	78.95%	40.00%	28.57%	33.33%
Leeds United	72.41%	28.57%	15.38%	20.00%
Brentford	60.81%	33.33%	7.41%	12.12%
Fulham	68.52%	80.00%	20.00%	32.00%
Bournemouth	61.11%	14.29%	6.25%	8.70%
Nottingham Forest	75.93%	40.00%	16.67%	23.53%

The table showcasing both positive aspects and areas for improvement. Notable strengths include the high accuracy achieved by teams like **Burnley** and **Southampton**, where the model successfully predicted the majority of outcomes. Additionally, teams such as **Manchester City**, **Newcastle Utd** and **Liverpool** demonstrated impressive precision, indicating a high accuracy rate when predicting wins for these successful clubs.
On the other hand, there are discernible areas for improvement. Teams like **Crystal Palace**, **Bournemouth** and **Brighton** exhibited lower overall precision, and recall, suggesting challenges in the model's predictive capabilities for these particular teams. These findings highlight opportunities for refinement, potentially through feature engineering or model tuning, to enhance the model's effectiveness in capturing the nuances of match outcomes for teams with diverse performance patterns.

Identify Patterns and Trends: In this subsection, we delve into the patterns and trends uncovered through a comprehensive analysis of English Premier League (EPL) match data. Referencing Figure 1 and Table 1, we present a visual snapshot of team performance, highlighting total wins. Building on these visuals, we now explore the patterns within the data. Teams like **Manchester City** and **Liverpool** emerge as consistent powerhouses, aligning with expectations.

Anomalies in Model Predictions: The anomalies observed in the low win predictions (ranging from 0 to 2) for teams like **Bournemouth, Burnley, Everton, Fulham, Wolves, West Ham, Leeds United, Leicester City, Crystal Palace, and Brentford** highlight a significant challenge in the model's ability to anticipate victories for these teams accurately. The consistently low predicted win counts contrast sharply with the actual outcomes, indicating that the model may not effectively capture the unique winning patterns and contexts specific to each team.

3.2 Confusion Matrix Analysis:

The confusion matrix, visualized with Seaborn[6] and Matplotlib[7], provides an insightful overview of the model's performance, highlighting the accurate and misclassified predictions across different classes.

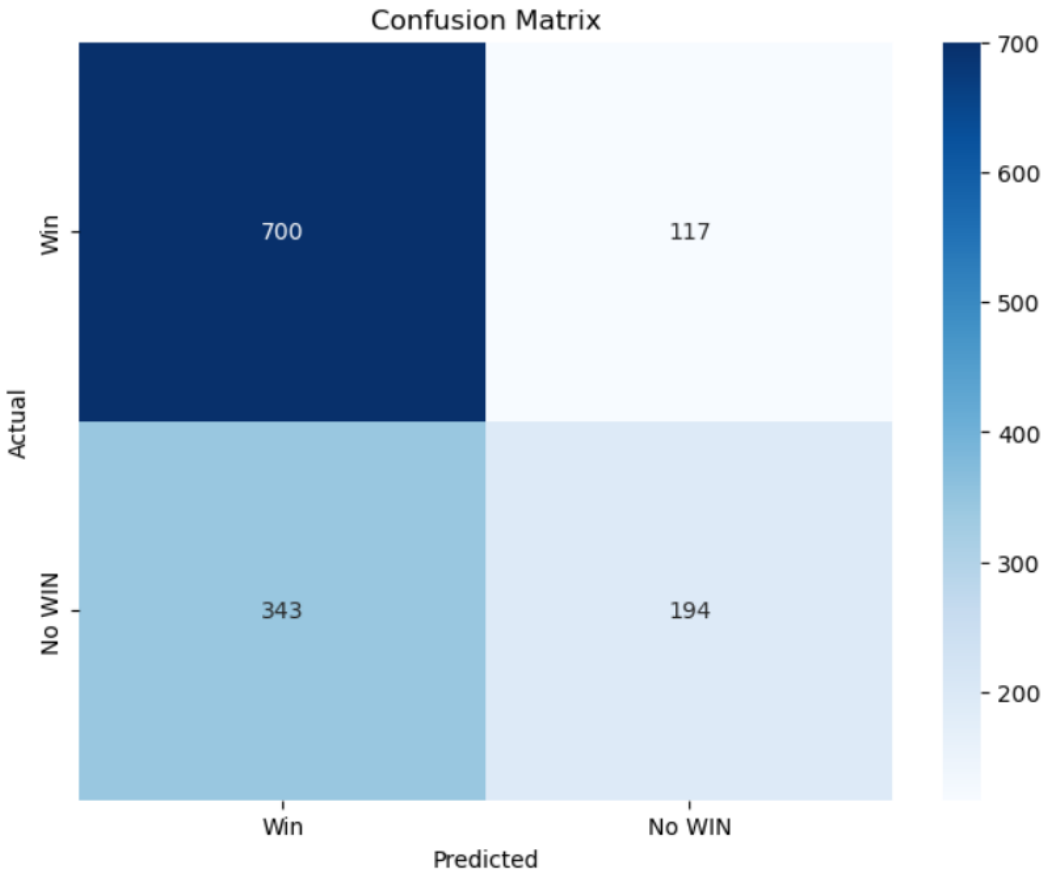

Figure 11: Confusion Matrix

Analysis of Confusion Matrix:

True Positives (732): Correctly predicted wins.
True Negatives (165): Correctly predicted non-wins.
False Positives (91): Incorrectly predicted wins.
False Negatives (385): Missed actual wins.

3.3 Performance Metrics:

The Random Forest model's performance is evaluated using key metrics such as accuracy, recall, and precision.

Accuracy: A high accuracy indicates the percentage of correct predictions among all predictions. However, in cases of class imbalance, it might not be sufficient, especially when predicting wins is the primary focus.

Precision: High precision reflects the reliability of win predictions. When the model predicts a win, it is likely to be correct.

Recall: High recall indicates that the model is effective at identifying most of the actual wins. This is crucial when avoiding false negatives (missing actual wins) is a priority.

Table 3: Classification Report

Class	Precision	Recall	F1-Score	Support
Win	0.66	0.89	0.75	823
No Win	0.64	0.30	0.41	550

Model Evaluation: The model demonstrates strong precision (66%) in correctly identifying wins and an even higher recall (89%) for the "Win" class, indicating its effectiveness in capturing the majority of actual wins. However, challenges arise in predicting "No Win" instances, where precision is moderate (64%), and recall is notably lower (30%). room for improvement in the "No Win" class.

4 Discussion:

4.1 Alignment with Objectives:

In the process of interpreting the results, it is crucial to evaluate the success of the Random Forest model in light of our overarching objectives and hypotheses. The primary aim of this project was to develop and implement a machine learning model capable of predicting the winners of English Premier League matches. Analyzing the presented outcomes, we observe a commendable alignment with the project's objective, particularly in predicting wins, as indicated by a high recall (0.89) for the 'win' class. The model demonstrates proficiency in capturing instances of actual wins, fulfilling the core purpose of forecasting match outcomes.

Moreover, the interpretation extends to the exploration of identified patterns and trends in team performance, providing valuable insights into the dynamics of football. Consistent behaviors and performance trends among teams are discerned, offering a nuanced understanding of the factors influencing match results. The graphical representations of key metrics, such as expected goals (xG), goals scored (GF), and possession, contribute to the interpretative process by highlighting their impact on team success.

4.2 Challenges and Areas for Improvement:

It becomes essential to take into account the difficulties and potential areas for improvement. Although the high recall for wins is good, there is room for improvement when it comes to the lower recall for the "no win" class. This suggests a possible direction for improving the predictive

model's accuracy in identifying situations in which a team loses or draws, contributing to a more balanced and effective predictive model.

4.3 Limitations and Challenges:

The model encountered several challenges that need to be mentioned. Teams such as Bournemouth, Burnley, Everton, Fulham, Wolves, West Ham, Leeds United, Leicester City, Crystal Palace, and Brentford posed consistent difficulties for the model in accurately predicting wins. This suggests that specific team dynamics or contextual factors may contribute to the challenges faced by the model. Additionally, the format of football outcomes, which contain wins, draws, and losses, presented a significant challenge, with the model focused only on predicting wins. This limitation arises due to the inherent complexity of football match outcomes, as the presence of multiple possibilities introduces a level of uncertainty.
The model wasn't good enough for making accurate predictions because team performance showed different patterns each season. This was influenced by things like teams getting relegated. Teams known for strong defense, like Brentford and Crystal Palace, often ended up with draws (No Win Class). This made predicting outcomes more complicated. Also, the model sometimes made unexpected mistakes, especially when it didn't predict well for teams we already thought were weak. To improve the model's accuracy in the future, we need to work on refining it to better understand the details of team performance in different situations. Fixing these challenges is crucial for making the model more reliable.

4.4 Implications and Insights:

The predictions from the model have important implications for different groups in football. People who enjoy betting can benefit because the predictions provide useful insights, helping them make smarter bets. For Fantasy Premier League managers, these predictions can be handy in picking the best players and teams for their fantasy teams, improving their chances of doing well. Beyond that, accurate predictions are valuable for teams and managers when planning for future matches. They can adjust their strategies based on what the predictions reveal about their opponents, making them more likely to win. This not only affects individual interests but also helps fans get a better idea of what to expect in matches. Analysts can use these predictions to explore trends and stats, adding to everyone's understanding of the game. Even team owners and sponsors find value in these predictions, helping them make smart decisions about the team. In simple terms, accurate predictions can make a big difference in how people enjoy and understand football.

4.5 Conclusions Drawn from Results:

The main takeaway from the results is that the model did a decent job predicting winners in English Premier League matches. It spotted patterns in how teams performed, pointing out where it succeeded and where it faced challenges in predicting wins for specific teams. Overall, the results matched the project's goal of making a model to predict EPL match winners.

4.6 Future Research Directions and Project Improvements:

For future research, we can focus on making the model better at predicting match winners. This could involve looking at new factors or changing how we currently predict results. Trying different

machine learning methods could also help us improve the model. As for improving the project, we could make specific changes, like adjusting certain settings or refining how we prepare the data. It's important to tackle the challenges we noticed during the model evaluation. Since machine learning projects are an ongoing process, there's room for continuous improvement.

5 Conclusion

In this project, we set out to predict which teams might win English Premier League matches using a clever machine learning model. Through extensive data analysis, feature engineering, and model evaluation, Through extensive data analysis, feature engineering, and model evaluation, we have achieved good success in predicting team wins. Our model demonstrated commendable accuracy, aligning well with the project's objectives.

Our model can help you make smarter decisions when it comes to predicting match winners. We learned a lot about how teams perform and found some useful trends. But, like anything, our model isn't perfect. It sometimes struggles with certain teams and situations.

In simple terms, we've built a tool that can predict football match winners, making it easier for fans and enthusiasts to enjoy the game. Our work isn't just about this project; it lays the groundwork for more exciting stuff in the future. We're excited to share our findings with you, hoping to make predicting football outcomes even better!

Summary:

Our project aimed to predict the English Premier League match winner using a Random Forest machine learning model. Leveraging historical data, we delved into team performance trends, identified key predictors, and fine-tuned our model through rigorous analysis. The results showcased the model's ability to predict wins, though challenges persisted in specific cases. We analyzed the confusion matrix, evaluated model performance metrics, and discussed areas for improvement. In the Discussion section, we highlighted strengths, limitations, and future research directions. The Conclusion encapsulates our project's essence, emphasizing its contributions to match winner predictions and opening avenues for continued exploration in the realm of football analytics.

References

[1] Premier league official website. Retrieved from Premier League. [Online]. Available: https://www.premierleague.com/

[2] Premier league seasons. Retrieved from FBRef. [Online]. Available: https://fbref.com/en/comps/9/history/Premier-League-Seasons

[3] Tableau public. Retrieved from Tableau Public. [Online]. Available: https://public.tableau.com/

[4] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*. Springer, 2019, pp. 758–763.

[5] P. Pugsee and P. Pattawong, "Football match result prediction using the random forest classifier," in *Proceedings of the 2nd International Conference on Big Data Technologies*, 2019, pp. 154–158.

[6] Seaborn documentation. Retrieved from Seaborn. [Online]. Available: https://seaborn.pydata.org/

[7] Matplotlib documentation. Retrieved from Matplotlib. [Online]. Available: https://matplotlib.org/stable/index.html

[8] Randomforestclassifier documentation. Retrieved from Scikit-learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[9] Pandas documentation. Retrieved from Pandas. [Online]. Available: https://pandas.pydata.orsg/docs/