Contents

1 Introduction

Football, or soccer as it is known in some parts of the world, has long been one of the most popular sports globally. Since the early 21st century, with an increase in globa broadcasting, interest in football has been rising all around the world. A survey of 18 major markets across the Americas, Europe, the Middle East and Asia shows the sport garnering powerful interest in more than 40% of the population, well ahead of its nearest rival sports [1]. Football is a sport that transcends gender all over the world. According to Nielsen SportsDNA global research[2], football is the most popular sport among women worldwide, and a recent study found that 70% of women find the men's FIFA World Cup "very appealing," while 58% find the women's FIFA World Cup "very appealing."

This rapidly growing interest has made data analytics a crucial component of football, allowing analysts to identify the critical factors that determine the outcome of matches. By leveraging advanced statistical models and machine learning techniques, people can make betting decisions, coaches can develop team strategies, and other important decisions in the sport. One of the significant breakthroughs in football analytics has been the compilation of extensive datasets covering various leagues such as the Premier League, Champions League, and League One in England, etc. In total 20 teams fight for the domestic league title year-round, each time playing 38 games, meeting the same team at home and again at away fixtures. Head-to-head wins usually are considered highly important at the end for determining the ultimate winner.

Today, in live matches, data is collected in many ways and under different categories. Some of these categories includes- performance data (team performance, goals, assists, shots, and tackles), tactical data (team formations, player positions, and playing styles), physical data (physical attributes, such as height, weight, speed, and injuries), biometric data (heart rate, breathing rate, and muscle fatigue), and environmental data (weather conditions, altitude, and stadium conditions). In this report, we discuss how we use the open source performance data available to examine the results of Head-to-Head games and the impact of home-field advantage, using a range of statistical measures and visualization techniques. Overall, football analytics represents an exciting and rapidly evolving field, with enormous potential for improving our understanding of the sport and making more informed decisions.

2 Data set Source and Detailed Description

2.1 Dataset Description

The football-data.co.uk England dataset is a collection of English football matches from 1993-1994 to 2022-2023 for various leagues including Premier League, Champion League, League1, etc. The dataset includes detailed information on each match such as the date, time, location, teams, and final score. In addition, it also includes data on the goals scored by each team, shots on target, fouls committed, cards given, and other relevant statistics.

2.2 Source

The website is a free resource that provides comprehensive football data from various leagues and competitions around the world. The data is available as CSV files with each CSV file containing data pertaining to each season. For our analysis, we are focusing on the Premier League because it is the most popular league and has sufficient data points to perform the analysis. Dataset – https://www.football-data.co.uk/englandm.php

2.3 Sample Size

We scraped the latest 12 seasons data i.e., 2011-2012 to 2022-2023 which contains data on 4451 matches with 21 variables of interest as described in *Appendix C-1* with our response variable as FTR with three categories (H: home team win, A: away team win, D: draw).

3 Methodology

3.1 Exploratory Data Analysis

A comprehensive EDA methodology was utilized to thoroughly examine and analyze the dataset.

3.1.1 Descriptive Statistics and Data Cleaning

We computed descriptive statistics such as mean, median, quantiles, minimum and maximum values for each variable to gain a basic understanding of their central tendencies and dispersion. Consequently, we carefully examined our dataset for any missing values. We conducted a thorough assessment of each variable individually and identified that only one row had missing data. Since our dataset contained a total of 4451 rows, we determined that removing this single row would not have a significant impact on our analysis. Hence, we proceeded to remove this row and continued our analysis with the remaining 4450 rows. Furthermore, we also performed data validation checks and converted categorical data into factors for further analysis.

3.1.2 Data Visualization and Outlier Analysis

To examine the quality of our data, we utilized various visualization techniques, including bar plots, box plots, scatter plots and histograms. To assess the distribution of categorical variables, we used bar plots, which are effective visualizations that display frequency of categories using rectangular bars. For continuous variables, we employed histograms, which are useful for exploring the spread and shape of data. In addition to examining the distribution of data using bar plots and histograms, we also checked for outliers using box plots. These visualizations allowed us to identify potential outliers and assess their impact on our data analysis. By thoroughly examining the distribution, outliers, and skewness of our data, we ensured the quality and reliability of our analysis and subsequent modeling approaches.

We also utilized two-dimensional plots to investigate whether there were any discernible patterns or evidence of predictors effectively separating the three classes.

3.1.3 Correlation Analysis

Furthermore, we performed correlation analysis to examine the correlations between variables (both dependent and independent) to identify potential patterns or relationships, and determining the strength and direction of these correlations. This step is crucial because highly correlated predictor variables can lead to unstable parameter estimates, high standard errors, and reduced predictive performance and therefore, need to be removed.

3.1.4 Feature Engineering

The variables *HomeTeam* and *AwayTeam* are identifiers of the teams that partake in a particular match. We possess historical data of 37 teams, and the use of one-hot encoding to these variables would introduce 36 additional variables per team (for both Home and Away teams). The *HomeTeam* and *AwayTeam* variables hold vital information about the prowess of the teams that engage in the match, and their significance in predicting the match outcome cannot be overemphasized. For this reason, we have opted to convert these categorical variables into numerical ones. This conversion will result in the creation of new variables that reflect the strength of each team in the event. To circumvent data leakage, we randomly selected one home and one away match per team for each season. This data was eliminated from the final dataset used in the modeling phase. Moreover, we are utilizing data from the Premier League, where a win attracts 3 points, while a draw results in 1 point being awarded to each participating team. Therefore,

$$\text{Home Strength} = 100 * \frac{(\text{Total wins at home}) * 3 + (\text{Total draws at home}) * 1}{(\text{Total matches played at home}) * 3}$$

$$\text{Away Strength} = 100 * \frac{(\text{Total away wins}) * 3 + (\text{Total away draws}) * 1}{(\text{Total matches away matches played}) * 3}$$

The entire EDA process was iterative and data-driven, involving hypothesis generation, testing, and refinement to draw meaningful insights and conclusions from the data.

To ensure that our model generalizes well on unseen data, we performed a 80-20 split (using random sampling) on the dataset which resulted in Training set with 3294 rows and Testing set with 863 rows and we performed all further analyses on training data.

3.2 Main Data Analysis

Our goal for the main data analyses was to predict the full time result (FTR) of the match and identify the significant predictor variables that are associated with the outcome of a football match.

For the main data analysis we employed three different models to analyze our data: Multinomial Logistic Regression, Decision Trees, and Random Forest. These three models were selected based on their suitability for our data and research objectives. By utilizing a combination of these models, we aimed to gain a comprehensive understanding of the relationship between our predictor variables and the multi-class categorical response variable, and to make informed and robust model for our data analysis.

3.2.1 Multinomial Logistic Regression

Firstly, given our response variable is a multi-class categorical variable, we utilized Multinomial Logistic Regression. It is an extension of the binomial regression model where the response variable takes more than two categories. Given, a category j = 1 as baseline, a multinomial logit model links the probabilities p_{ij} to the predictors as follows:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1i} + ... + \beta_{(p-1)j}x_{(p-1)i} = x_i^T\beta_j = \log p_{ij}/p_{i1}$$

So that we get, $p_{i1} = 1 - \sum_{j=2}^{J} p_{ij}$, and $p_j = \frac{\exp(\eta_{ij})}{1+\sum_{j=2}^{J}\exp_{\eta ij}} for 2 \leq j \leq J$

Here, $\eta_{i1} = \log p_{i1}/p_{i1} = log 1 = 0$

It can be used to classify new observations by predicting the probability of each possible category. The category with the highest probability is then assigned to the observation. In this way, the multinomial logistic regression model can be used as a classifier to predict the category of a new observation based on its predictor variables.

We changed the reference level to *Draw* category to understand the impact of predictors on a win for either *Home* or *Away* Team.

3.2.1.1 Model Selection For model selection, we started with a model that includes all the relevant predictors and used step wise selection using AIC metric to eliminate the insignificant predictors. AIC (Akaike Information Criterion) is calculated by taking into account the goodness-of-fit of a model and the complexity of the model. AIC penalizes models for having a higher number of parameters, aiming to strike a balance between model performance and complexity. A lower AIC value indicates a better-fitting model, as it considers both model fit and parsimony.

3.2.1.2 Goodness-of-fit Test To test the goodness of fit we performed a Chi Squared test. The null hypothesis assumes that there is no significant difference between the observed and expected frequencies, indicating that the model fits the data well.

$H_0 : P_1 = E_1, P_2 = E_2, ..., P_k = E_k$, where P_1, P_2, ..., P_k are the observed probabilities in each category of the response variable, and E_1, E_2, ..., E_k are the expected probabilities predicted by the model.

3.2.1.3 Interaction Analysis Furthermore, we utilized interaction plots to visualize the interaction effects between two predictors on the outcome variable. It displays the relationship between the outcome variable and one predictor, separately for different levels or categories of another predictor.This analysis provides us with visual evidence of the presence or absence of interactions.

3.2.1.4 Model Diagnostics Lastly, to assess the validity and appropriateness of the multinomial model, we performed model diagnostics which involves evaluating the linearity and outlier assumptions and performance of the model to ensure that it is reliable and accurate in capturing the underlying patterns and relationships in the data.

3.2.2 Decision Trees

Next, we employed Decision Tree Classifier, a non-parametric supervised learning technique that can be used for classification and regression tasks. They are a type of predictive model that uses a tree-like structure to represent and classify observations based on sequence of binary decisions. The decision tree algorithm recursively splits the data into different branches based on the predictor variables and their corresponding values, ultimately leading to terminal nodes (i.e., leaves) that represent the predicted response variable category for a given set of predictor variable values. The decision tree can then be used to classify new data points into one of the categories based on the paths followed in the tree. Some of the advantages of decision trees are that they are easy to interpret and visualize, can handle both categorical and continuous data, and capture non-linear relationships between predictors and response variable.

3.2.2.1 Hyper-parameter Tuning We also performed hyper-parameter tuning using various parameters. One of the key parameters we used was complexity parameter(cp) which saves computing time by pruning off splits that do not decrease the overall lack of fit by a factor of cp.

3.2.2.2 Pruning However, decision trees are prone to over fitting, especially when the tree is deep and complex. To overcome this issue we performed pruning on the decision tree model using the complexity parameter value obtained at the least relative error. Pruning refers to the process of selectively removing certain branches or nodes from a decision tree in order to improve its performance by reducing complexity.

3.2.3 Random Forests

Finally, we implemented Random Forest classifier, an ensemble learning technique which involves constructing multiple decision trees using different subsets of the training data and randomly selected features at each split point. These decision trees are built independently, without correlation or information sharing during the training process. Once the decision trees are constructed, the random forest algorithm aggregates their predictions to make the final prediction. This approach helps to improve the accuracy and robustness of the classifier, as the ensemble of decision trees can compensate for individual tree biases and reduce overfitting by incorporating diversity in the feature selection and sample subsets used in each tree.

An advantage of random forests is that they handle high-dimensional data and can capture complex non-linear relationships between the input features and the target variables. They are less prone to over fitting than single decision trees.

3.2.3.1 Hyper-parameter Tuning We performed hyper-parameter tuning to control the number of trees (*ntree*) using grid search. A higher number of trees may increase the model's accuracy but also increase the computation time.

3.2.4 Model Evaluation and Comparison

3.2.4.1 Confusion Matrix We constructed confusion matrix for all the models as it displays the predicted class labels against the actual class labels, providing a summary of the model's performance in terms of true positive, true negative, false positive, and false negative predictions. It helped us assess the accuracy of the model, providing insights into the classification performance and identifying any misclassifications or errors made by the model.

3.2.4.2 AUC Score We employed several evaluation metrics to assess and compare the performance of the three models: multinomial logistic regression, decision trees, and random forest. The primary comparison metric chosen for our analysis was the area under the receiver operating characteristic curve (AUC), which is a widely used measure of a model's ability to accurately classify data points into their respective classes. The AUC provides a comprehensive overview of the model's discriminatory power across all classes, with higher values indicating better performance.

3.2.4.3 One-vs-Rest Curve (ROC) Additionally, we plotted one-vs-rest ROC curves for the multinomial data, which allowed us to assess the performance of each class compared to the rest of the classes. This approach provided insights into the model's ability to discriminate between different classes individually, which can be particularly useful when dealing with imbalanced datasets or when specific class performance is of particular interest. For our 3-class response variable we got 3 different one-vs-rest scores and we averaged them to get a final one-vs-rest model score.

3.2.4.4 Kappa Statistics Furthermore, we computed the Kappa statistic, which is a measure of agreement between the predicted and observed classes, taking into account the possibility of chance agreement. The Kappa statistic ranges between -1 to +1 and helps us understand the level of agreement between the model's predictions and the ground truth, with higher values indicating higher agreement.

The use of multiple evaluation metrics allowed for a thorough comparison of the models' performance and facilitated the selection of the best-performing model for our analysis.

4 Results

4.1 Exploratory Data Analysis

The exploratory data analysis (EDA) revealed important insights into the dataset. Descriptive statistics provided a comprehensive summary of the data as show in *Appendix C-2* and didn't highlight any incorrect values or any immediate need for scaling.

From distribution plots *(refer to Appendix B-1 and Appendix B-2)*, we found some predictor variables - Away Corners (AC), Away Shots on Target (AST), Home Corners (HC), Home Shots on Target HST) to be right skewed but the skewness was not significant to perform transformations. While outlier analysis using box plots *(refer to Appendix B-3)* identified outliers in some variables, the variables were retained in the analysis as outliers were not deemed as erroneous values. As shown in *Fig 1*, higher shots on target by the home team (left panel) and away team (right panel) were associated with improved chances of winning for each respective team which indicated they might be important predictors for full time result.

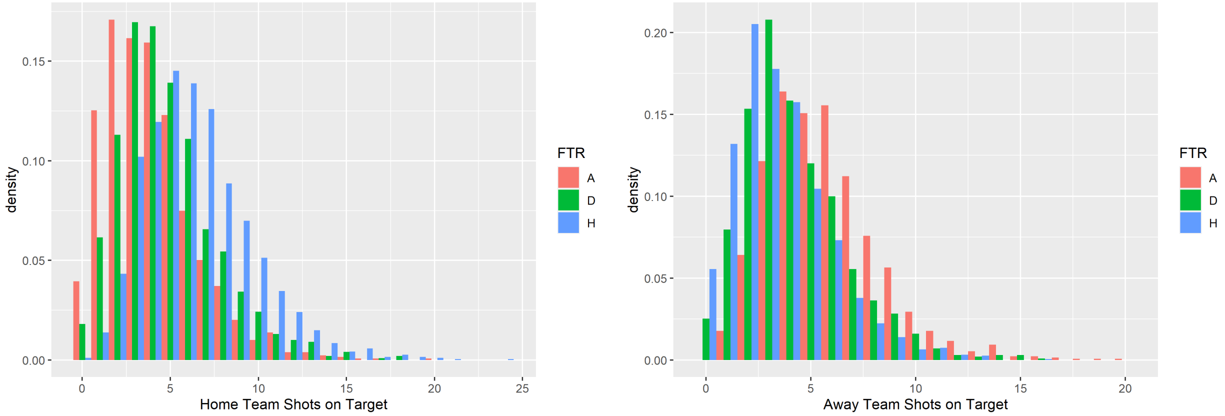

Figure 1: Distribution plots for categorical variables

Correlation analysis uncovered significant relationships between variables. As highlighted in *Fig 2* below, four variables were dropped from further analysis due to their high positive correlations: FTHG (full-time home team goals) with HTHG (half-time home team goals), FTAG (full-time away team goals) with HTAG (half-time away team goals), HS (shots taken by the home team) with HST (number of shots on target by the home team), and AS (shots taken by the away team) with AST (number of shots on target by the away team).*(refer to Appendix B-4 for full correlation plot)*

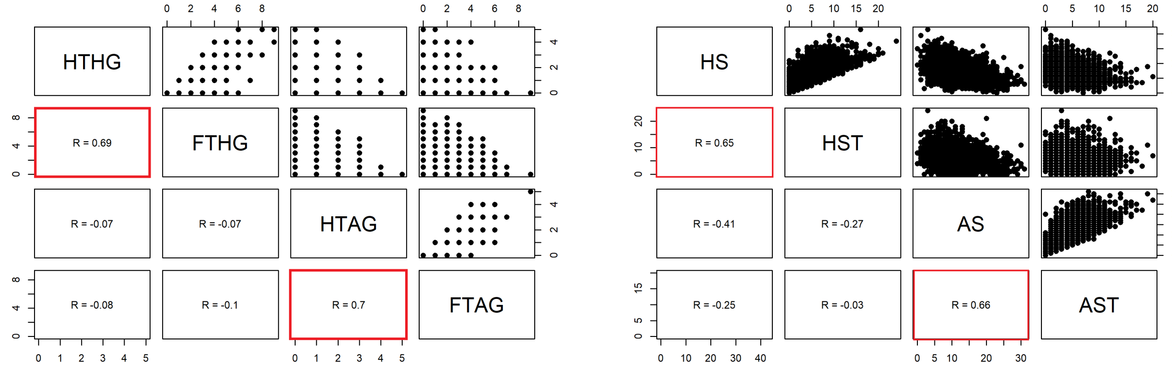

Figure 2: Pearson correlation plots for removed variables (high correlations highlighted in red)

Additionally, the engineered variables *Home Strength* and *Away Strength* provided insights into team standings presented in *Fig 3 (on next page)*, which highlights the relative performance of teams. The first quadrant (top-right) highlights the strongest teams whereas second quadrant highlights the team strong on *Away Ground*. Similarly, third quadrant highlights overall weak teams and fourth quadrant highlights team strong on Home Grounds.

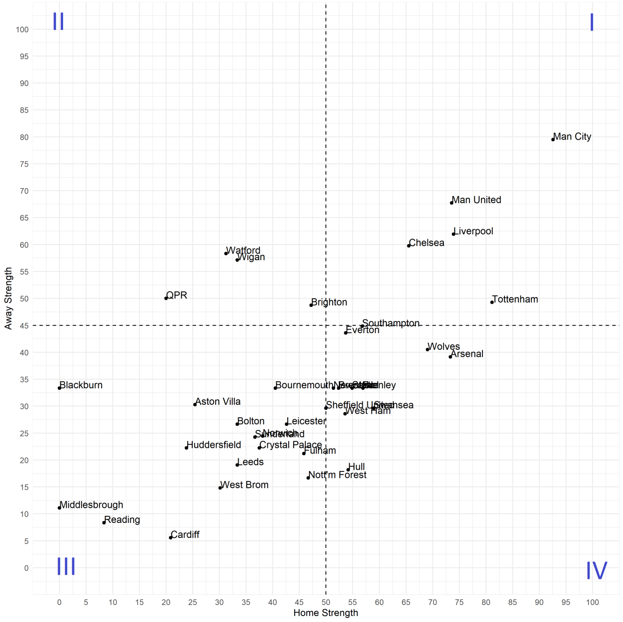

Figure 3: Team Strength

From all 37 teams, *Man City* and *Tottenham* ranked in top two at home grounds and *Man*

City and *Man United* ranked in top two at away grounds *(refer to Appendix B-5 for plot with full list).*

Overall, the results of the EDA have enriched our understanding of the dataset, identified patterns and correlations, and informed subsequent steps in the analysis and interpretation of the results. These findings contribute to the foundation of the research and provide a robust basis for further analysis and inference.

4.2 Main Data Analysis

4.2.1 Multinomial Regression

Stepwise selection using AIC criteria was used to select the statistically significant predictors and to obtain the final model.

4.2.1.1 Interpretation of Coefficients According to the findings presented in *Table 1* below, the likelihood of the away team winning or drawing a match increases significantly by a factor of approximately 3 for every unit increase in the number of goals scored by the away team in the first half. Similarly, the odds of the home team winning or drawing the game increases by a factor of 2.88 for each unit increase in the number of goals scored by the home team in the first half.

Furthermore, the study indicates that the odds of the away team winning or drawing a match increase by a factor of 1.23 for every additional shot on target by the away team. Likewise, the probability of the home team winning or drawing the match increases by a factor of 1.21 for each additional shot on target by the home team.

In addition, the research highlights that the odds of the away team winning or drawing a match increases by a factor of 1.94 for each red card awarded to the home team. Conversely, the likelihood of the home team winning or drawing the game increases by a factor of 1.66 for every red card awarded to the away team.

Lastly, the study indicates that the odds of the away team winning or drawing a match decrease by a factor of 0.99 for every unit increase in the strength percentage of the home team. Similarly, the probability of the home team winning or drawing the game decreases by a factor of 0.99 for each unit increase in the strength percentage of the away team. These findings provide valuable insights into the factors that influence the outcomes of football matches and could be useful for sports analysts and bettors alike.

Table 1: Coefficient Estimates from Multinomial Model

Winner	HTHG	HTAG	HST	AST	HC	AC
AwayVsDraw	0.37	3.04	0.86	1.23	1.02	0.96
HomeVsDraw	2.88	0.40	1.21	0.83	0.94	1.05

Table 2: Coefficient Estimates from Multinomial Model

HY	AY	HR	AR	Home_Strength	Away_strength
0.91	0.93	1.94	0.57	0.99	1.01
0.89	0.93	0.48	1.66	1.02	0.99

4.2.1.2 Goodness of fit Based on the results of the Chi square test, the obtained p-value was found to be less than the predetermined significance level of 0.05. Therefore, it is appropriate to reject the null hypothesis that the model fits the data well. Consequently, the evidence indicates that there is a lack of fit in the model, which suggests that the model is not sufficient to explain the relationship between the predictors and the response variable adequately. As a result, further investigation is done to identify the reasons for the lack of fit.

4.2.1.3 Interactions Based on our domain knowledge and investigation, we have explored the possibility of interactions among a set of variables, as evidenced by the graphs presented below. Our analysis indicates that there is no substantial evidence to support the existence of interactions among these variables.

These findings suggest that the relationships between the variables can be adequately explained by their individual effects, and no additional explanatory power is gained by considering their interactions

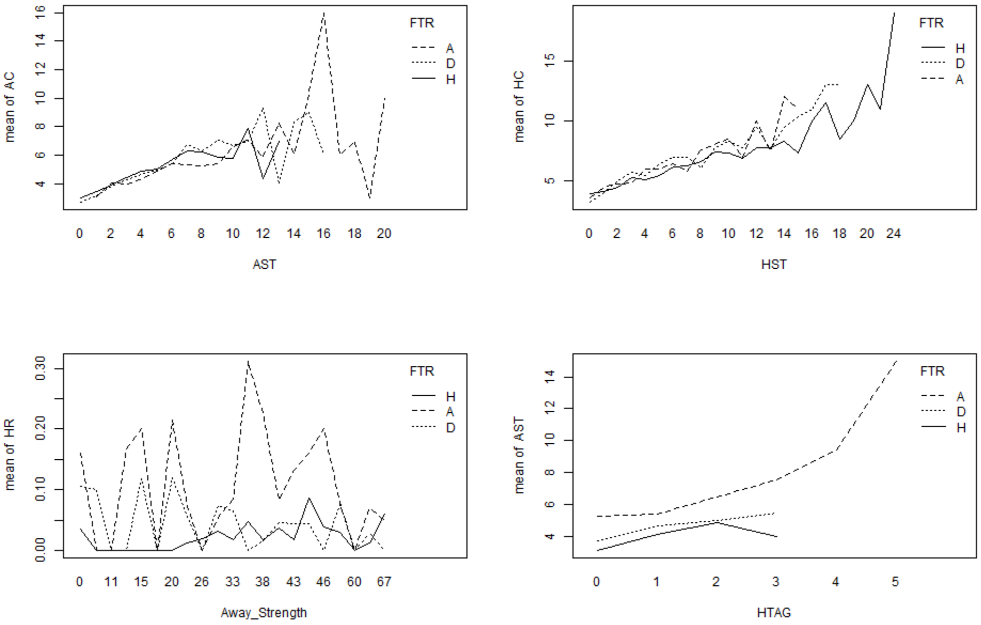

Figure 4: Checking for interaction amongst predictors

4.2.1.4 Model Diagnostics Upon examining the scatter plots presented below, which depict the residuals (y-axis) plotted against the fitted values (x-axis), we have found no discernible pattern. Specifically, the residuals appear to be randomly scattered around the horizontal line at zero for both models (Away vs Draw and Home vs Draw).

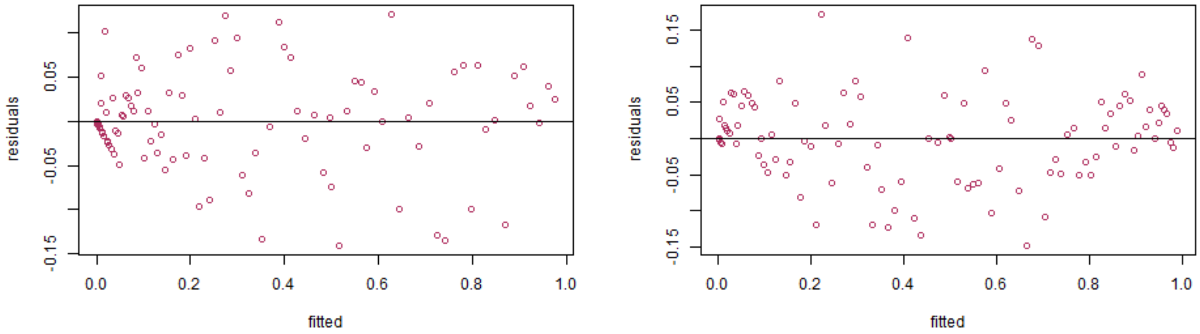

Figure 5: Residuals vs fitted value plots. Left panel: away vs draw and Right panel: home vs draw classification.

Upon examining the halfnormal plot presented below, we have found no substantial evidence to suggest the presence of outliers in the dataset.

This finding is significant as the presence of outliers can significantly influence statistical analyses and lead to erroneous conclusions. By ruling out the possibility of outliers, we can be more confident in the validity and reliability of our results.

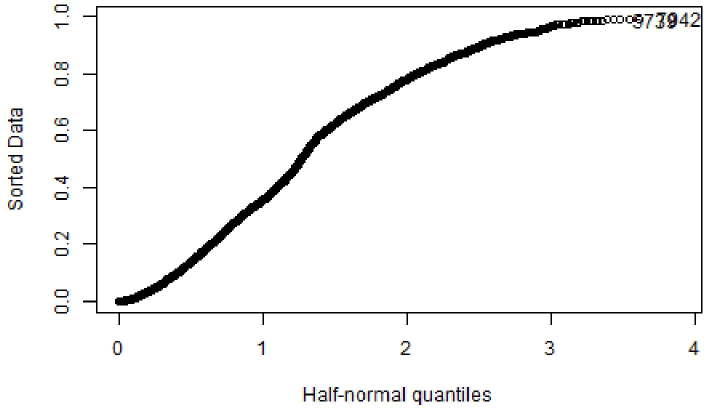

Figure 6: Half-Normal Plot

4.2.2 Performance Metrics

The ROC (Receiver Operating Characteristic) curve, generated using the One vs Rest method for this model, is illustrated in the plot presented below.

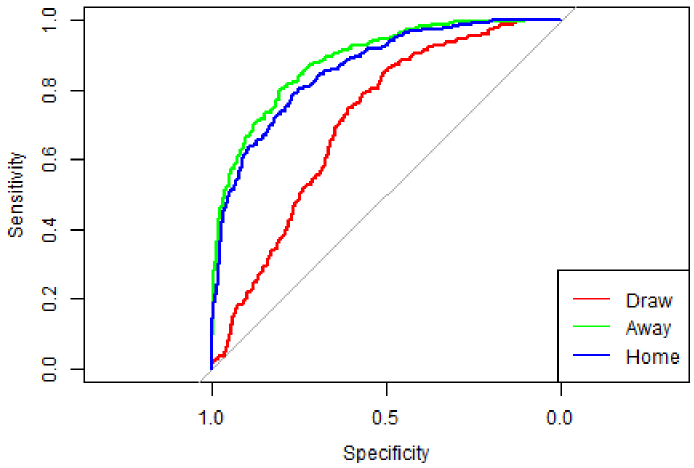

Figure 7: ROC curve for One-Vs-Rest classification

Upon examining the performance of the model, we have found that it struggles to predict the match outcome as a DRAW. However, it performs well in predicting the home win or away win outcomes. To further investigate this matter, we have computed the AUC (Area Under the Curve) scores presented below.

Table 3: AUC scores for each outcome of the match

Metric	Draw	Away	Home
AUC	0.71	0.89	0.86

We have used Accuracy and Kappa statistic to test the model on the training and testing dataset and following results were obtained:

Table 4: AUC scores for each outcome of the match

Dataset	Accuracy	Kappa
Train	67.36	0.48
Test	64.19	0.43

The training accuracy of the model is moderately high at 67.36%, while the test accuracy is slightly lower at 64.19%. This suggests that the model is performing reasonably well on new, unseen data. The kappa statistic, which measures the level of agreement between the model predictions and the true outcomes, is moderate for both training (0.48) and testing (0.43) datasets. This indicates a reasonable level of agreement between the model predictions and true outcomes for both the training and testing data.

4.3 Decision Trees

After fitting the decision tree to the training dataset, we applied pruning *(for tree structure, refer to Appendix B-6)* to the tree to identify the optimal value of the complexity parameter, which was found to be 0.0022 *(for relative error vs complexity parameter plot, refer Appendix B-7)*.

4.3.1 Variable Importance:

Based on the variable importance plot in *Fig 8* below, we can observe that the Half time Away goals variable is the most important feature for predicting the outcome of the match. It is closely followed by the shots on target by the team variable, which also has a significant impact on the outcome of the match. The high importance of the Half time Away goals variable suggests that it is a strong predictor of match outcomes. This is consistent with previous research that has shown that early goals in a match can have a significant impact on the final outcome. Similarly, the importance of the shots on target variable suggests that teams that are more accurate in their shooting are more likely to win matches.

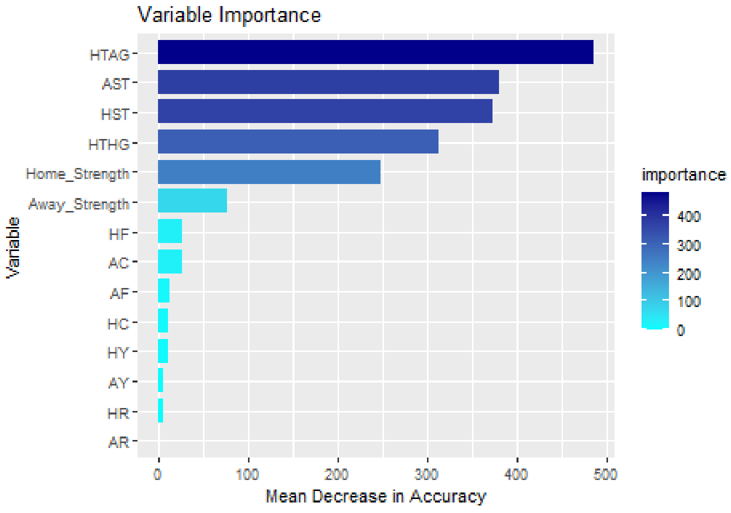

Figure 8: Variable Importance Plot for Decision Trees

4.3.2 Performance Metrics

The ROC curve using One vs Rest method for this model is plotted below (*Fig 9*):

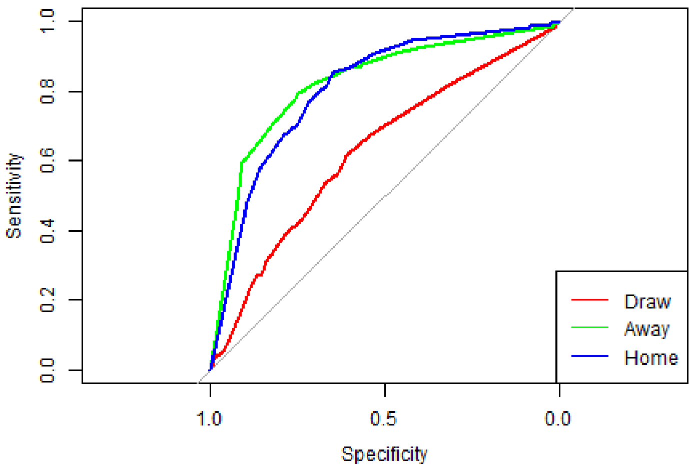

Figure 9: ROC curve for One-Vs-Rest classification

Our analysis revealed that the model exhibited suboptimal performance when predicting match outcomes as draws. Conversely, it demonstrated a commendable ability to accurately predict either a home or away team win. To gain a more comprehensive understanding of this phenomenon, we can delve deeper into the matter by examining the AUC scores presented below in *Table 5*.

Table 5: AUC scores for each outcome of the match

Metric	Draw	Away	Home
AUC	0.63	0.82	0.81

In order to evaluate the effectiveness of the model, we employed two commonly used statistical measures, namely Accuracy and Kappa statistic, to assess its performance on both the training and testing datasets. The resultant outcomes are presented as follows:

Table 6: Performance metrics for Training and Testing data

Dataset	Accuracy	Kappa
Train	69.28	0.51
Test	62.46	0.41

The training accuracy of the model is moderately high, at 69.28%. However, the test accuracy is slightly lower at 62.46%, indicating that the model may not be generalizing well to new, unseen data.

The kappa statistic, which measures the level of agreement between the model predictions and the true outcomes, is moderate for training at 0.51. However, it is slightly lower for testing at 0.41, indicating a lower level of agreement on new data.

4.4 Random Forests

Based on the grid search results, we determined that the optimal number of trees in the forest for this model is 2000 *(for dot plot refer to Appendix B-8)*. Subsequently, we trained the model using this value on the training dataset.

4.4.1 Variable Importance

The predictor importance plot (*Fig 10*) reveals significant insights into the predictors that have a high impact on determining the outcome of the match. It is evident from the plot that Shots on target and Half-time goals are the most important predictors in determining the match outcome. In addition to these predictors, the strength of the teams competing in the match also plays a crucial role in determining the result. These findings indicate that accurate modeling of the match outcome requires the consideration of a combination of predictors, including both on-field performance indicators and off-field factors such as team strength.

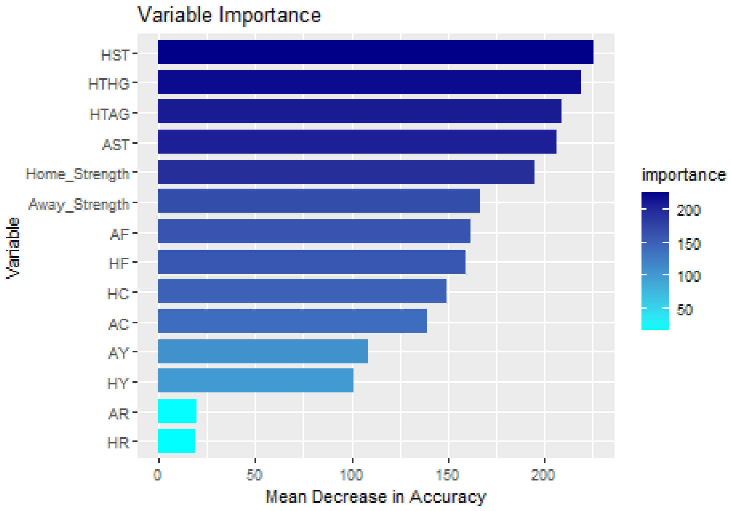

Figure 10: Variable Importance Plot for Random Forests

4.4.2 Performance Metrics

The ROC curve using One vs Rest method for this model is plotted below in *Fig 11*:

Our analysis indicated that the model exhibited limited proficiency in accurately predicting match outcomes as draws. However, it demonstrated a noteworthy level of accuracy when forecasting either a home or away team victory. Further examination of the AUC scores in *Table 7* can provide deeper insight into this matter.

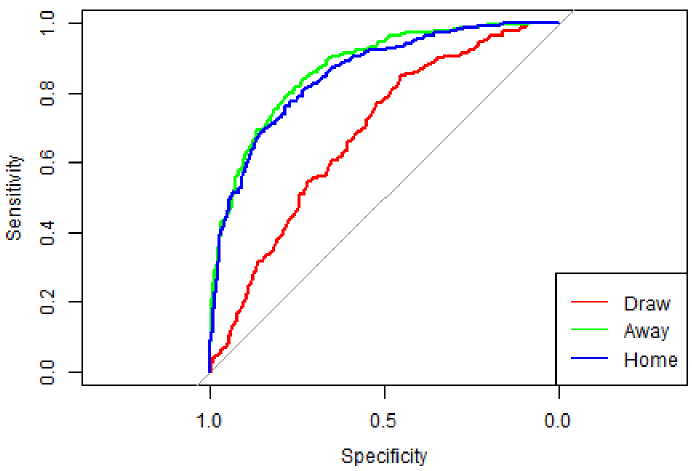

Figure 11: ROC curve for One-Vs-Rest classification

Table 7: AUC scores for each outcome of the match

Metric	Draw	Away	Home
AUC	0.68	0.87	0.85

To evaluate the model's efficacy, we utilized two widely recognized statistical measures, namely Accuracy and Kappa statistic, to scrutinize its performance on both the training and testing datasets. Subsequently, we obtained the ensuing outcomes:

Table 8: Performance metrics on Training and Testing dataset

Dataset	Accuracy	Kappa
Train	65.10	0.44
Test	63.96	0.43

The model seems to have a similar performance on both the training and testing datasets, as indicated by the comparable accuracies of 65.1% and 63.96% respectively. The kappa statistic also shows moderate agreement between the predicted and true outcomes for both the training and testing datasets. This suggests that the model is not overfitting on the training data and is generalizing well to new data. However, there is still room for improvement in the model's performance as the accuracies are not high.

4.5 Model Comparison

The Multinomial model outperforms other models with an accuracy of 64.19 on the test dataset. However, the models' performance in predicting draws is a concern as they mostly misclassify them as home wins. This misclassification can be attributed to several factors such as home advantage or the strength of the home team, which is common in soccer matches. The inability to predict draws accurately can affect the overall performance of the models as draw outcomes are not uncommon in soccer matches. The confusion matrix below (*Fig 10*, *Fig 11*, *Fig 12*) of all the three models shown below that the models are prone to misclassifying draw outcomes. This suggests that the models' ability to predict draws needs to be improved for better overall performance.

Table 9: AUC scores for each outcome of the match

Metric	Multinomial	DecisionTree	RandomForest
Mean_AUC	0.82	0.75	0.8

Table 10: Confusion Matrix for Multinomial Model

TruthVsPredicted	Draw	Away	Home
Draw	40	63	112
Away	39	200	35
Home	28	32	314

Table 11: Confusion Matrix for Decision Tree

TruthVsPredicted	Draw	Away	Home
Draw	54	59	102
Away	40	194	40
Home	35	33	306

Table 12: Confusion Matrix for Random Forests

TruthVsPredicted	Draw	Away	Home
Draw	51	64	100
Away	36	192	46
Home	37	39	298

5 Conclusion

Our analysis has revealed that several variables have a significant impact on the outcome of soccer matches, with shots on target and half-time score emerging as particularly influential. We found that matches in which a team has a higher number of shots on target are more likely to result in a win for that team. This is likely since shots on target are a good indicator of a team's attacking prowess, and teams that can create more scoring opportunities are more likely to come out on top. Similarly, we found that the half time score is a strong predictor of the outcome of the match. Matches in which one team has a lead at half time are more likely to result in a win for that team. This is likely due to a combination of factors, including the psychological boost that comes from being ahead, as well as the tactical adjustments that teams can make at half time to defend their lead or mount a comeback.

Our findings highlight the importance of these variables in predicting the outcome of soccer matches and suggest that they should be given careful consideration when building predictive models. By including variables such as shots on target and half-time score in our models, we can improve their accuracy and ensure that they are capturing the most important factors that contribute to the final outcome of a match. Ultimately, this can help us make more informed decisions and better understand the complex dynamics that underlie the sport of soccer.

We also suggest that home team strength and away team strength should be calculated based on recent data rather than historical data because team performance can change over time, and recent performance is likely to be more indicative of current form and ability. Using recent data allows the models to capture changes in team performance, such as the impact of injuries or changes in coaching staff. While historical data can be useful for understanding long-term trends and patterns, it may not reflect the current state of the team. By incorporating recent data into the models, we can get a more accurate picture of each team's current strength and make more accurate predictions of match outcomes. Overall, it's important to continually refine the models and incorporate new data to improve their accuracy and ensure that they are accounting for the most up-to-date information about each team.

6 References

[1] (2018, June) Fan Favorite: The Global Popularity of Football is Rising, Nielsen, https://www.nielsen.com/insights/2018/fan-favorite-the-global-popularity-of-football-is-rising/

[2] (2018, June) World Football Report, Nielsen, https://www.nielsen.com/insights/2018/world-football-report/

[3] (2022, Feb) Trevisan, Vinícius. Multiclass classification evaluation with ROC Curves and ROC AUC, Towards Data Science https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a/

[4] (2023) The Multinomial Distribution and the Chi-Squared Test for Goodness of Fit, UC Berkley, https://www.stat.berkeley.edu/~stark/SticiGui/Text/chiSquare.htm/