DATA ANALYSIS AND STATISTICAL MODELING OF FOOTBALL MATCH DATA

(Surekha Peethambaram Muralidhar, Siddhartha Vaddempudi, Mani Chandana Alle)

Research question: How effectively can the number of shots on target by the home team predict the number of full-time home goals in football matches within the top 5 European football leagues, considering other key factors?

INTRODUCTION

In the context of football, many factors can influence the number of goals scored by a team. Among these, shots on target often reflect a team's offensive capabilities and their ability to create goal-scoring opportunities. Our research question centers on exploring the predictive power of shots on target for the full-time goals scored by the home team in football matches within the top 5 European leagues (England, Spain, Germany, Italy, France).

We initially aim to understand whether there is a statistically significant relationship between the number of shots on target and the number of goals scored by the home team at the end of the match. This will involve building a Simple Linear Regression (SLR) model where the predictor variable is the number of shots on target by the home team, and the response variable is the total number of goals scored by the home team.

However, football is complex, and many factors can impact goal scoring. The SLR model may not fully capture these complexities, which could lead to violations of its underlying assumptions, such as independence, linearity, constant variance, and normality of residuals. To address this, we can expand our analysis to consider additional variables, building a Multiple Linear Regression (MLR) model. This more complex model can account for other contributing factors like possession percentage, number of corners, or other team and player statistics.

Yet, if the MLR model also violates assumptions or does not adequately capture the non-linear and interactive effects between these variables, we can turn to more flexible models such as

Random Forests. Random Forests, a type of ensemble learning, can handle complex, non-linear relationships and interaction effects among predictor variables, providing a more robust prediction model for goal scoring.

DATA DESCRIPTION

The dataset sourced from (https://www.football-data.co.uk/data.php)comprises detailed match data from the top two divisions of five major European football leagues: England, Spain, Germany, Italy, and France. Spanning from the 1993-94 season to the 2022-23 season, the dataset encompasses over 110,000 matches, equating to approximately 20,000 hours of gameplay. Each CSV file in the dataset represents a specific season and league, structuring the data across numerous rows and columns that detail various aspects of each match.

DIMENSIONS

The dataset contains over 110,000 rows and approximately 40 columns, each representing different attributes of the matches.

VARIABLES

RESPONSE VARIABLE

Full Time Home Goals (FTHG) - This variable represents the total number of goals scored by the home team by the end of the match and serves as our primary variable of interest for predicting outcomes based on shots on target.

POTENTIAL PREDICTOR VARIABLES

- Home Shots on Target (HST) - Main predictor for our initial models.
- Away Shots on Target (AST) - Could help adjust predictions by considering the attacking strength of the opposition.
- Total Shots (HS and AS) - May provide insights into the overall attacking volume.
- Fouls Committed (HF and AF) - As indirect indicators of game aggression or defensive pressure.
- Corners (HC and AC) - Could indicate attacking pressure or opportunities created.

- Possession Percentage (if available) - Higher possession might correlate with more shots and goals.

Researchers and football enthusiasts can leverage this dataset to explore a wide range of questions, such as identifying teams with the best or worst performance when holding a lead, examining the efficiency of teams in front of goal, analyzing defensive and offensive capabilities, and studying the impact of referees on match outcomes. The dataset also allows for cross-league comparisons, providing insights into different playing styles, strategies, and trends in European football.

DESCRIPTIVE STATISTICS AND GRAPHICAL ANALYSIS

To better understand the relationship between shots on target and goals scored, initial exploratory data analysis will include:

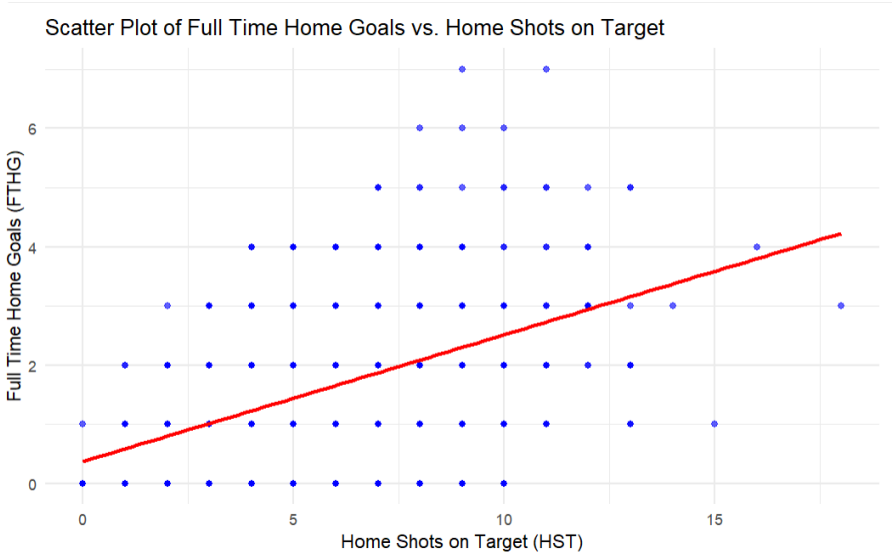

Fig. 1: Scatterplot of Home Shots on Target (HST) vs. Full Time Home Goals (FTHG).

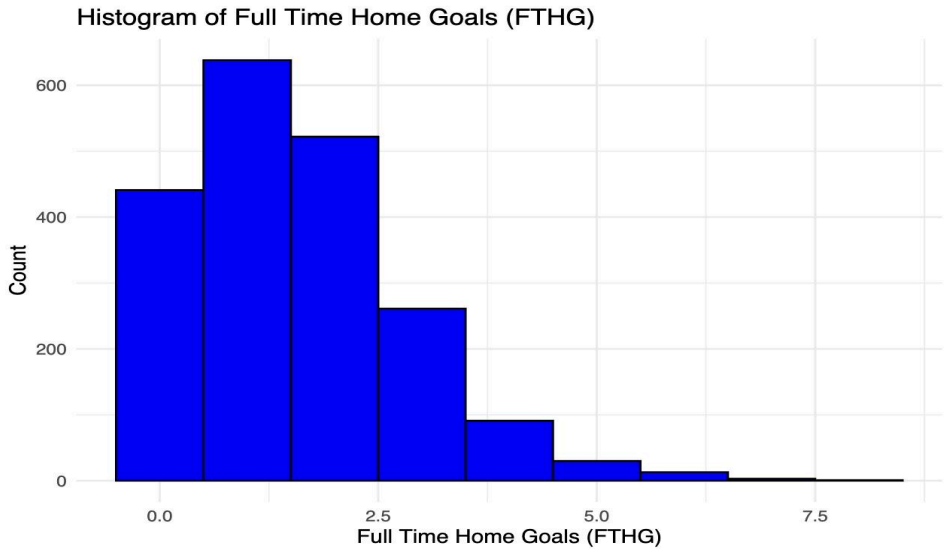

Fig. 2: Histogram of Full Time Home Goals (FTHG)

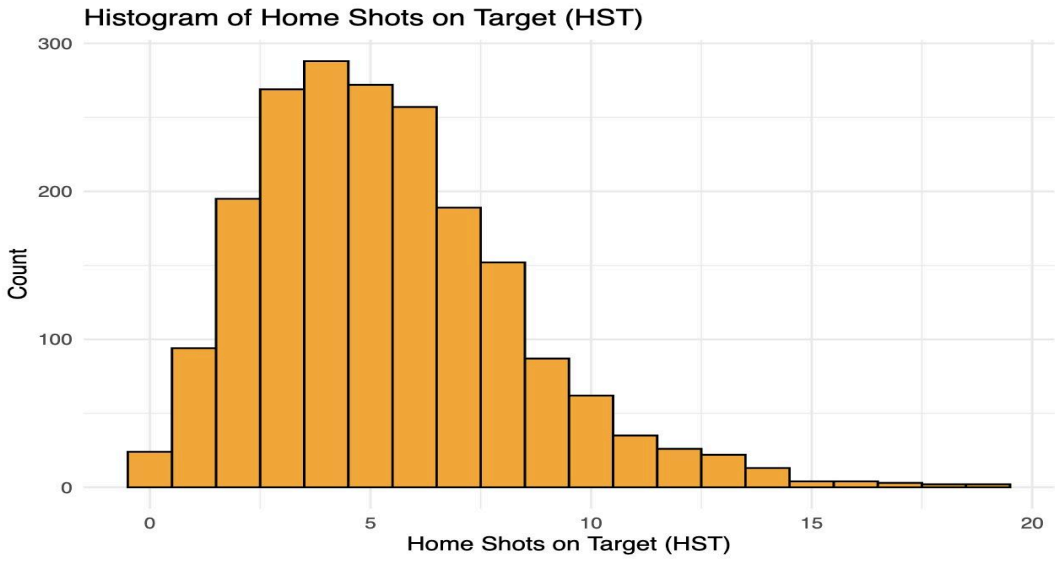

Fig. 3: Histogram of Home Shots on Target

Statistic	Median (IQR)	Distribution (%)
FTHG	1.00 (1.00, 2.00)	-
FTAG	1.00 (0.00, 2.00)	-
HTHG	-	0: 52%, 1: 34%, 2: 11%, 3: 2.5%, 4: 0.4%, 5: <0.1%, 6: <0.1%, 7: <0.1%
HTAG	-	0: 61%, 1: 30%, 2: 7.7%, 3: 1.4%, 4: 0.2%, 5: <0.1%, 6: <0.1%
HS	13.0 (10.0, 16.0)	-
AS	11.0 (8.0, 14.0)	-
HST	5.00 (3.00, 6.00)	-
AST	4.00 (2.00, 5.00)	-
HF	13.0 (10.0, 16.0)	-
AF	13.0 (10.0, 16.0)	-
HC	5.0 (3.0, 7.0)	-
AC	4.00 (3.00, 6.00)	-
HY	2.00 (1.00, 3.00)	-
AY	2.00 (1.00, 3.00)	-
HR	-	0: 91%, 1: 8.2%, 2: 0.4%, 3: <0.1%
AR	-	0: 89%, 1: 11%, 2: 0.7%, 3: <0.1%, 9: <0.1%

Table 1: Descriptive statistics for key variables (mean, median, mode, and standard deviation for FTHG, HST, AST, HS, AS, HF, AF).

These plots help in visualizing the relationship between the number of shots on target and the goals scored, providing an initial check on the feasibility of linear regression modeling.

METHODS AND RESULTS

DATA CLEANING AND PREPROCESSING

Before proceeding with the analysis, we performed data cleaning and preprocessing steps. Initially, we removed any rows containing missing values (NA) using the “drop_na()” function

from the ‘dplyr’ package. This step ensures that our analysis is based on a complete dataset without any missing information.

REGRESSION MODELING

We employed various regression modeling techniques to analyze the relationship between the predictor variables and the response variable (number of full-time home goals (FTHG)).

1. **SIMPLE LINEAR REGRESSION (SLR):** Linear Regression is a commonly used type of predictive analysis. A statistical method used to understand the relationship between a single predictor variable and a single response variable.

$$\text{Model Specification: } FTHG = \beta 0 + \beta 1(HST) + \varepsilon$$

- FTHG is the response variable, and HST is the primary predictor.

i) ASSUMPTIONS

- Linearity,
- Independence
- Constant Variance
- Normality of residuals.

ii) MODEL FITTING: We fit a simple linear regression model using the lm() function in R, with FTHG(Full time home goals) as the response variable and HST (Home shots on target) as the predictor variable to estimate coefficients β0 and β1.

iii) MODEL DIAGNOSTICS: To assess the validity of the linear regression model assumptions, we generated diagnostic plots and analyzed the residuals. The diagnostic plots, including the residuals vs. fitted values, normal Q-Q plot, scale-location plot, and residuals vs. leverage plot, were examined for any patterns or deviations from the assumptions of linearity, normality, constant variance and independence.Additionally, we performed the Shapiro-Wilk test to cross-check the normality of the residuals.

TRANSFORMATIONS AND MODEL REFINEMENT

Based on the violations in diagnostics and residual analysis, we explored potential transformations to improve the model fit and address any violations of assumptions.

i) LOG TRANSFORMATION: We applied a log transformation to both the response and predictor variables to account for potential non-linearity in the relationship.

To incorporate additional predictor variables and investigate their impact on the number of full-time home goals, we constructed multiple linear regression models.

2. **MULTIPLE LINEAR REGRESSION (MLR):** It is the most common form of Linear Regression. Multiple Linear Regression basically describes how a single response variable Y depends linearly on a number of predictor variables.

Model Expansion: $FTHG = \beta0 + \beta1(HST) + \beta2(AST) + \beta3(HS) + \beta4(HC) + \varepsilon$

- Additional predictors AST, HS, and HC are included.

i) VARIABLE SELECTION: Stepwise regression using AIC for model optimization.

ii) ASSUMPTIONS: Re-check assumptions of simple linear regression including multicollinearity using the Variance Inflation Factor (VIF).

iii) MODEL FITTING AND DIAGNOSTICS: Ordinary least squares, residual analysis, and model validation.

Also, we performed a transformation to ensure to fit the model accurately with respect to satisfy the assumptions.

BOX-COX TRANSFORMATION: To determine the optimal transformation for the response variable, we employed the Box-Cox transformation technique.

As an alternative to linear regression where the assumptions were violated even after the power transformations, we explored the use of random forest regression, a powerful ensemble learning technique, to model the relationship between the predictor variables and the number of full-time home goals.

3. RANDOM FOREST REGRESSION:

i) MODEL SPECIFICATION: Ensemble of decision trees for improved predictive accuracy and overfitting control.

ii) PARAMETER TUNING: Number of trees (ntree), maximum features (mtry) determined by cross-validation.

iii)MODEL TRAINING: Bootstrapped samples, randomly excluded predictors for each tree.

iv) MODEL VALIDATION AND VARIABLE IMPORTANCE: Out-of-Bag (OOB) samples for accuracy estimation, variable importance scores.

CONCLUSION:

This study explored the effectiveness of using home team shots on target to predict full-time home goals in the top 5 European football leagues. Our findings revealed,

1. PREDICTIVE VALUES OF SHOTS ON TARGET

- Simple Linear Regression (SLR) confirmed a positive relationship between home shots on target and home goals, suggesting a foundational correlation.

2. ENHANCED UNDERSTANDING THROUGH MLR

- Multiple Linear Regression (MLR), incorporating additional factors like away shots and home corners, provided deeper insights and improved prediction accuracy over SLR.

3. SUPERIORITY OF RANDOM FOREST

- The Random Forest model outperformed linear models in handling complex, non-linear relationships, making it more suitable for predicting football match outcomes.

4. MODEL ROBUSTNESS

- Extensive diagnostic checks ensured the robustness of our models, and highlighted the need for model adjustments to improve assumption compliance.

5. FUTURE DIRECTIONS

- Future research could incorporate player-specific statistics and in-game events to enhance model accuracy. Further exploration of advanced machine learning techniques could also be beneficial.

In summary, while shots on target are a strong predictor of goals, incorporating multiple factors and employing advanced modeling techniques like Random Forest provides a more accurate predictive framework. This research enriches football analytics by demonstrating how various statistical models can be effectively used to predict match outcomes.

CODE APPENDIX

The github link for the project is

https://github.com/manichandana8/Football-match