The Effect of the First Goal in the Football Match Outcome

Suram Bagratyan, Aram Grigoryan, Tigran Kostanyan, Leonid Sarkisyan, Silva Yeghiazaryan

Introduction

The aim of our project is to analyze football match data to evaluate the hypothesis: **Scoring the first goal increases the probability of winning the match.**

The data provides information about various events during the game starting from a simple goal and going to red card and foul. It has 941009 events from the biggest 5 European football leagues: England, Spain, Germany, Italy and France from 2011/2012 season to 2016/2017 season.

Objectives

Our objectives for the project are the following: - Explore and clean the dataset to extract relevant information. - Analyze the relationship between scoring the first goal and match outcomes. - Use visualizations to illustrate key insights. - Build a different models to quantify the impact of scoring first on winning probability.

Methodology

The methodolfy that we will use are the following: 1. Data Preparation: Filter and clean the data to focus on events like goals and match outcomes. 2. Visualization: Create charts to understand the timing and impact of the first goal. 3. Modeling: Build models to test the hypothesis quantitatively.

Data Processing

At first we load the data.

As the dataset includes unnecessary information as well, we will remove the columns unrelated to the analysis, such as player details or descriptive text to simplify the dataset.

The dataset includes information about different type of events. As we are only interested in the goals and also shots on target to later use for analyzis and model creation, we will remove the events other than these two.

The data does not include any information about how many goals were scored during each game, so we will extract that data from our dataset. This code calculates the total number of goals scored by the home and away teams in each game. By organizing the data this way, we can later determine the winner of each match and analyze the relationship between scoring first and match outcomes.

On the later analysis, we will need the number of the shots on target for each game. That is why we add a column that includes the total number of shots on target for each team during each game.

As we need to identify the first goal scoring teams, we filter the dataset to isolate the events where a goal was scored. By doing this, we focus on the necessary moments in the match that directly contribute to the outcome, helping us analyze the timing and impact of goals on winning probability.

As we are interested in only the first goal of the game game, we identify the first goal scored in each match. By extracting the first goal from each game, we can analyze how it impacts the outcome, particularly whether scoring first increases the likelihood of winning.

Based on the data that we have already extracted, we assign a winner to each game based on the number of goals scored by the home and away teams. The winner is determined by comparing the home goals and away goals. In case of a draw, those rows are filtered out. This helps in analyzing whether the team that scores first is more likely to win.

To have the first goal and game results information together, we merge the first_goals dataframe (which contains details about the team that scored the first goal) with the game_results dataframe (which contains information about the overall winner of the match).

There can be games, when for example the first goal is scored in the 90th minute of the game, so it is quite obvious that the likelihood of the team winning increases simply because there is little time remained for the opponent team to score. To extract this cases, we have thought about some threshold that we can use to test the hypothesis on, and we reached to a conclusion that 50th minute is quite good number because the team has the first half of the game + the first few minutes of the second half and we can say that even if the team scores first during the first 50 minutes the opponent team has quite some time to make a comeback. This is why we filter the dataset to only include matches where the first goal was scored before the 50th minute.

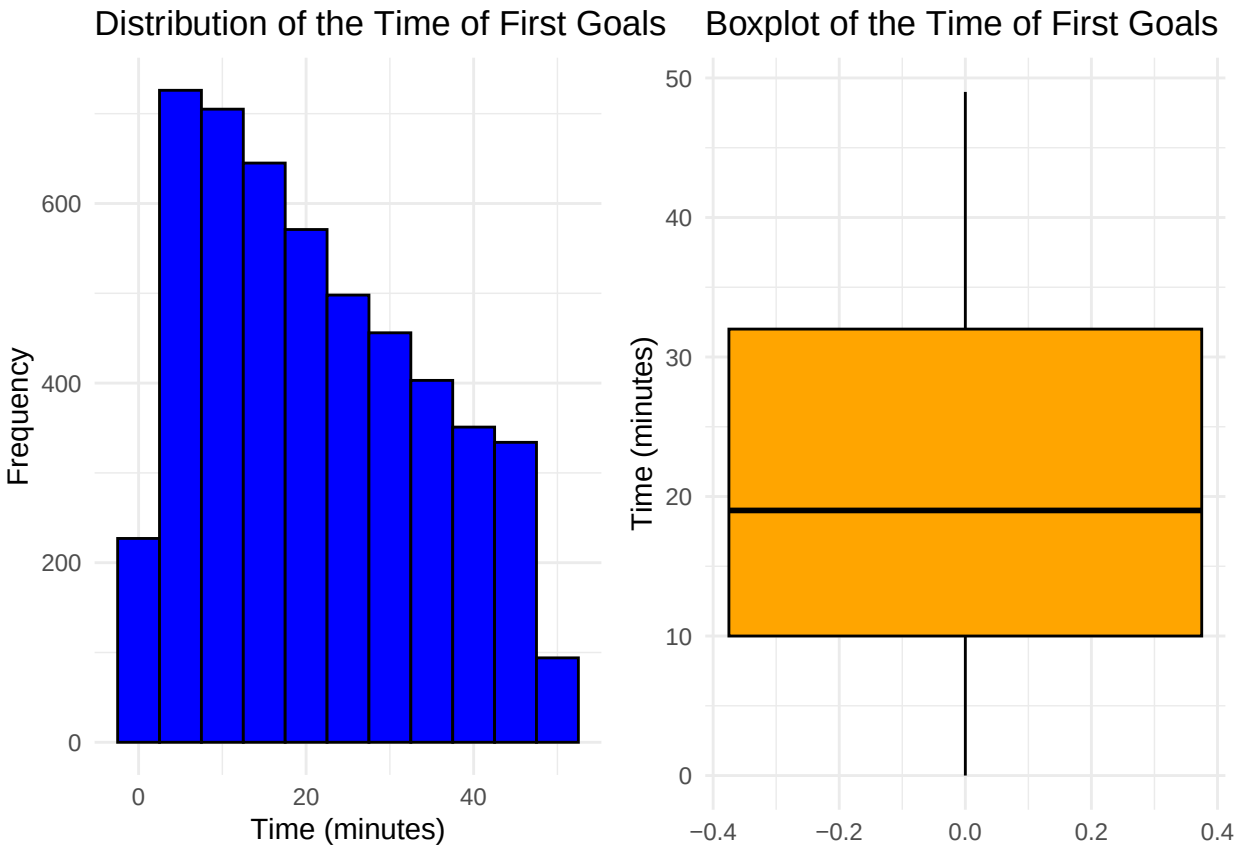
As we try to analyze the impact of the first goal on the winning percentage, we creates a new column first_goal_wins that indicates whether the team that scored the first goal went on to win the match. This step helps us directly analyze the relationship between scoring first and winning the match.

Another information that we nay need later is the goal_conversion_rate is created as the ratio of total goals to total shots on target. This metric will help assess the efficiency of teams in converting their shots into goals, which is a key aspect of our analysis.

We also add a new column is_home_team to the match_data dataset, where 1 indicates a home team and 0 indicates an away team. This categorization is essential for analyzing how home advantage may affect goal conversion rates.

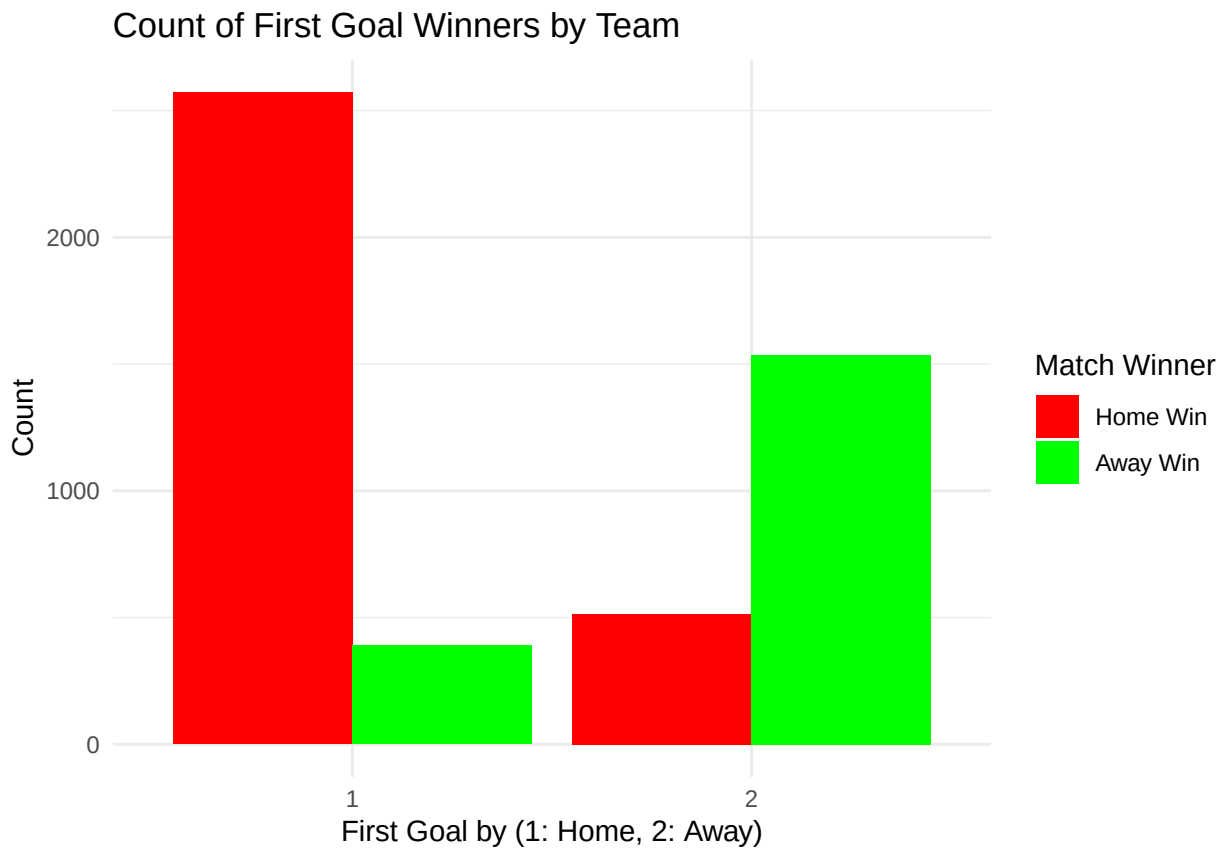
Exploratory Data Analysis (EDA)

Now let's explore the data, and see the Distribution of Time of First Goals, and try to detect potential outliers. As we can see, first goals are typically scored early in football matches, with the majority occurring within the first 20 minutes. The histogram highlights a declining frequency of first goals as time progresses, while the boxplot shows a median first goal time around 20 minutes, with most occurring between 10 and 30 minutes. This emphasizes the critical impact of early goals in shaping match outcomes.



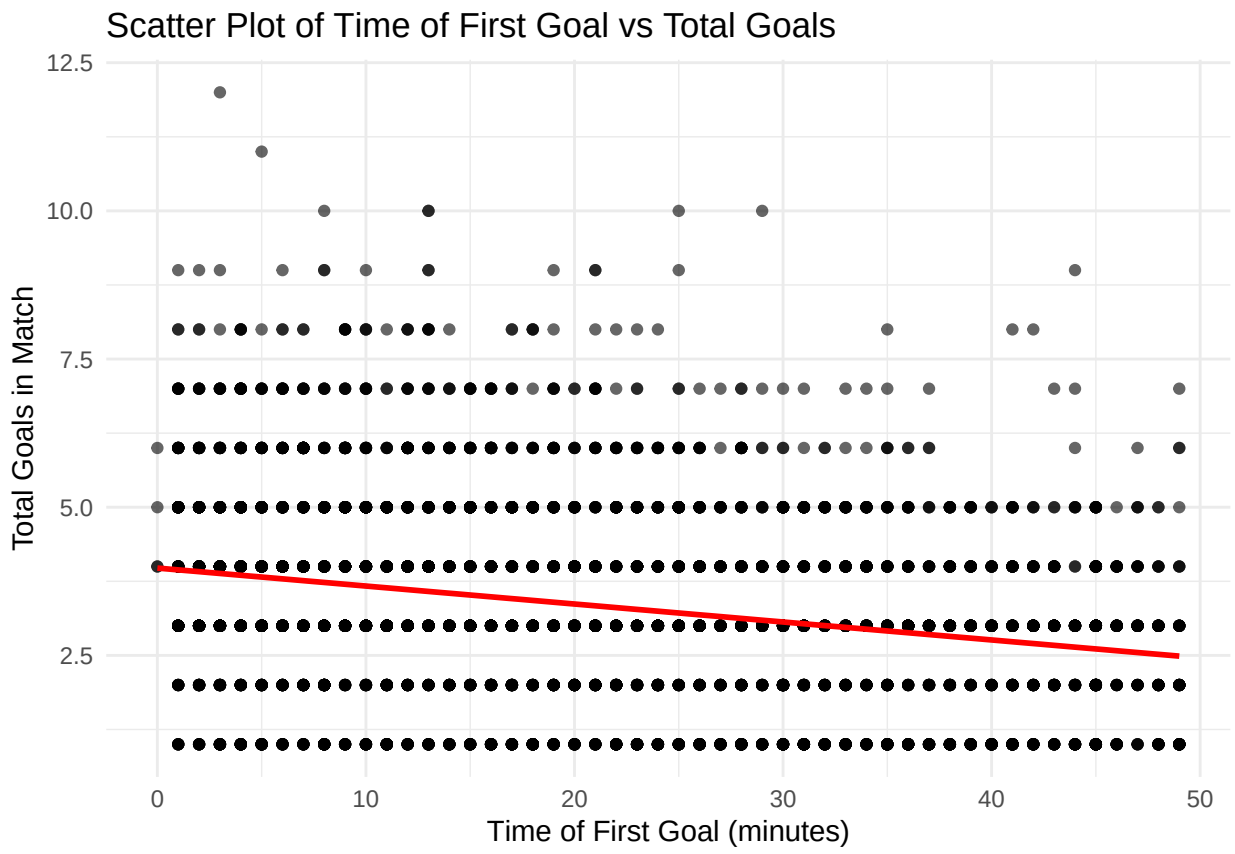
No Outliers Detected: The boxplot shows a clean range for the time of first goals. There are no data points significantly outside the whiskers (which represent 1.5 times the interquartile range). This implies that the time of first goals is consistently distributed within the observed range.

Now let's visualize the count of first goal winners by team type. This bar plot illustrates the relationship between the team scoring the first goal (home or away) and the match outcome. When the **home team scores first** (category "1"), they predominantly win, as shown by the tall red bar. Similarly, when the **away team scores first** (category "2"), they often secure victory, indicated by the green bar. The chart emphasizes the significant advantage of scoring the first goal, with a noticeable home advantage overall.



The scatter plot illustrates the relationship between the time of the first goal (in minutes) and the total number of goals scored in a match. Each dot represents a match, with the x-axis showing the time of the first goal and the y-axis indicating the total goals in the game.

The trend line suggests a negative correlation: as the time of the first goal increases, the total number of goals in the match tends to decrease slightly. This pattern indicates that earlier first goals may contribute to higher-scoring games, possibly because scoring early might lead to more aggressive play or open gameplay strategies. However, the overall distribution shows considerable variability, suggesting that while this trend exists, other factors also significantly influence the total number of goals in a match.



Hypothesis Testing

Chi Squared Test

The results of the Pearson's Chi-squared test with Yates' continuity correction indicate a statistically significant association between the two categorical variables being analyzed (the team scoring the first goal and the match outcome). The p-value of 0.001529 indicates a statistically significant relationship between scoring the first goal and the match outcome. This suggests that the two variables (scoring first and winning the match) are not independent. This indicates that teams scoring the first goal are more likely to win.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 10.044, df = 1, p-value = 0.001529
```

This logistic regression model incorporates additional predictors, such as the number of goals scored by both teams, to evaluate the likelihood of the team scoring the first goal winning the match. Intercept (3.81): The positive intercept indicates a strong baseline likelihood of winning for the first-goal scorer, assuming all other predictors are at their reference levels.

Side (-0.87): The negative coefficient suggests that when the away team scores the first goal, their chances of winning are significantly lower compared to the home team scoring the first goal. This reinforces the impact of home advantage.

Time (0.0101): The positive coefficient indicates that scoring the first goal later in the match slightly increases the likelihood of winning. This could imply less time for the opposing team to recover.

Home Shots on Target (0.0292): This predictor is not statistically significant ($p = 0.13$), indicating no strong evidence that home team shots on target directly affect the probability of the first goal leading to a win.

Away Shots on Target (0.0458): The positive coefficient suggests that more away shots on target slightly increase the likelihood of the first-goal scorer winning. This could reflect higher overall offensive effectiveness.

Most predictors are statistically significant, with p-values < 0.05, except for home shots on target ($p = 0.13$).

The logistic model suggests that scoring the first goal increases the likelihood of winning, but this probability is moderated by home advantage, the timing of the goal, and the final match scores. The number of goals scored by the opposing team significantly reduces the first-goal advantage, highlighting the importance of overall match dynamics.

```
##
## Call:
## glm(formula = first_goal_wins ~ side + time + home_shots_on_target +
##     away_shots_on_target + home_goals + away_goals, family = "binomial",
##     data = match_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           3.809751   0.212369  17.939  < 2e-16 ***
## side                 -0.871195   0.094056  -9.262  < 2e-16 ***
## time                  0.010057   0.003108   3.236  0.00121 **
## home_shots_on_target  0.029191   0.019316   1.511  0.13072
## away_shots_on_target  0.045793   0.021788   2.102  0.03557 *
## home_goals           -0.428984   0.037055 -11.577  < 2e-16 ***
## away_goals           -0.430216   0.040886 -10.522  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4724.0  on 5009  degrees of freedom
## Residual deviance: 4273.3  on 5003  degrees of freedom
## AIC: 4287.3
##
## Number of Fisher Scoring iterations: 5
```

With the logistic model, let's predict the probabilities of winning when scoring in the given time of the match. First we split data into training and testing sets for performance evaluation.

Then with our model, we make predictions on the test set.

We use metrics such as the Confusion Matrix to evaluate the model:

```
##          Actual
## Predicted    0    1
##         0   15   11
##         1  238 1239
```

```
## [1] 0.8343313
```

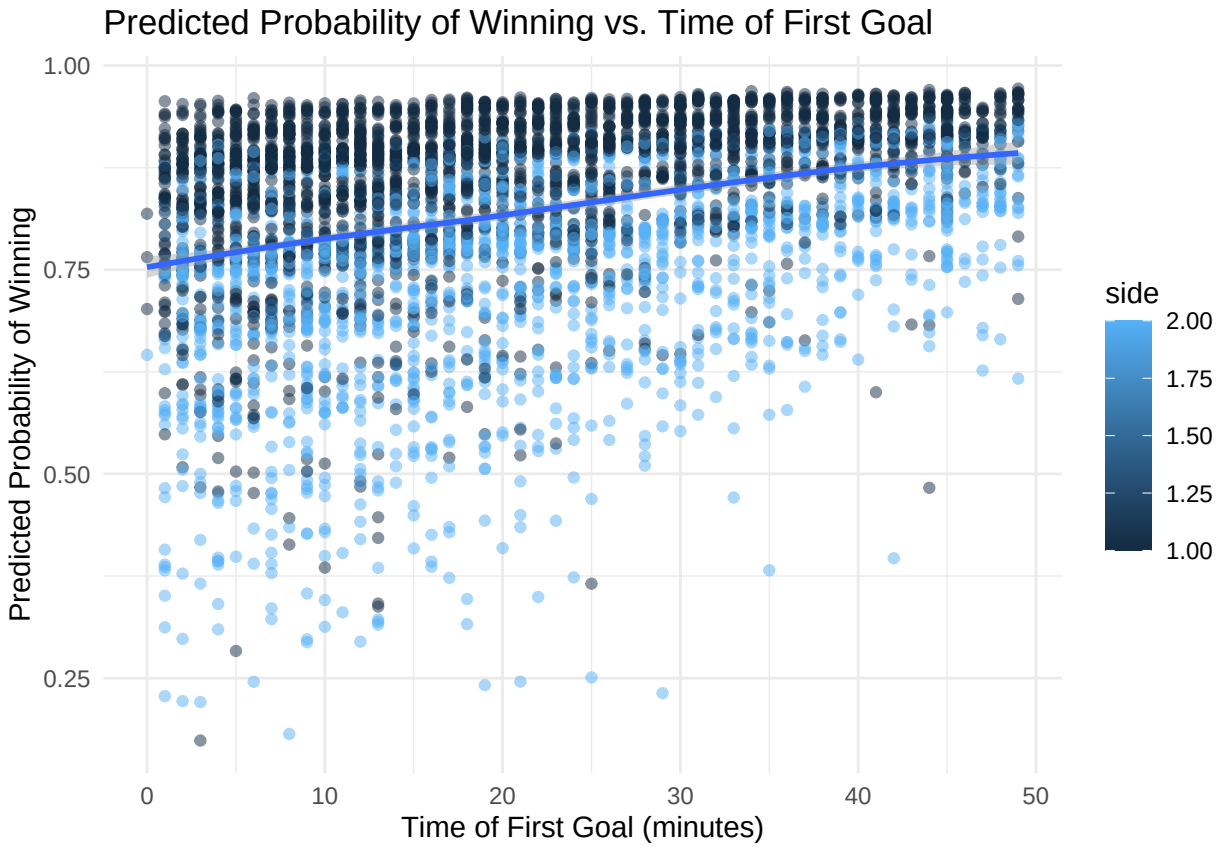
The confusion matrix summarizes the performance of the classification model:

Metrics:

- **Accuracy**:
  Accuracy $= \frac{\text{TP+TN}}{\text{Total}} = \frac{16+1241}{16+9+237+1241} = 0.836$
  This indicates that the model correctly classified 83.6% of the cases overall.
- **Precision**:
  Precision $= \frac{\text{TP}}{\text{TP+FP}} = \frac{1241}{1241+9} = 0.993$
  High precision indicates that most of the cases predicted as positive (1) were correct.
- **Recall** (Positive Class):
  Recall $= \frac{\text{TP}}{\text{TP+FN}} = \frac{1241}{1241+237} = 0.839$
  The model identified 83.9% of the actual positive cases.
- **F1-Score**:
  F1 $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \approx 0.91$
  This metric balances precision and recall. The model performs well overall, as indicated by the accuracy (83.6%) and F1-score (0.91).

Predicted Probabilities vs. Predictor Variables

Plot the predicted probability of winning (first_goal_wins) against key predictors like time or side to see the relationship.



The plot visualizes the relationship between the time of the first goal and the predicted probability of winning, with data points differentiated by the team's side.

Side line consistently shows a higher predicted probability of winning compared to Side 2, indicating a potential advantage for Side 1 in scoring first. For both sides, the predicted probability of winning generally

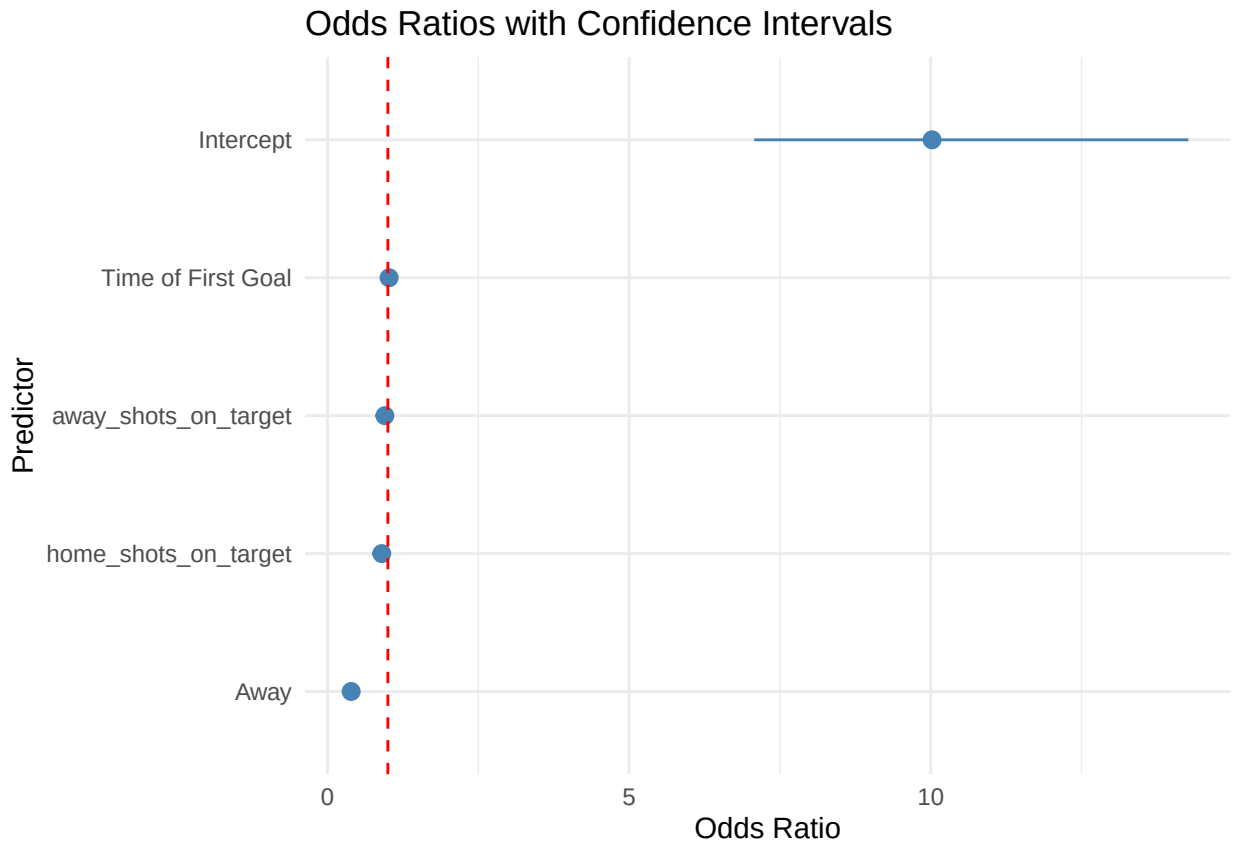
increases as the time of the first goal gets closer to the 40th minute, suggesting that scoring later in the first half may improve a team's likelihood of winning. Components:

Dots represent individual observations of the predicted probabilities based on the logistic regression model. Smooth Curves (LOESS) indicate the trend in the predicted probabilities over time for each side. Implications:

Scoring early: Teams that score very early (near the 0-10 minute mark) have lower predicted probabilities, but they are still high (>60%). Scoring late: Teams scoring later (closer to the 40th minute) see a more significant increase in predicted probabilities, especially for Side 1.

Odds Ratios with Confidence Intervals

Display the odds ratios for predictors using a bar chart with confidence intervals.



The plot represents the odds ratios for each predictor in the logistic regression model, with confidence intervals visualized as error bars.Each row corresponds to a predictor variable (e.g., Time of First Goal, Away, etc.). Predictors with odds ratios >1 increase the likelihood of winning, while those <1 decrease it. Odds Ratios (OR):

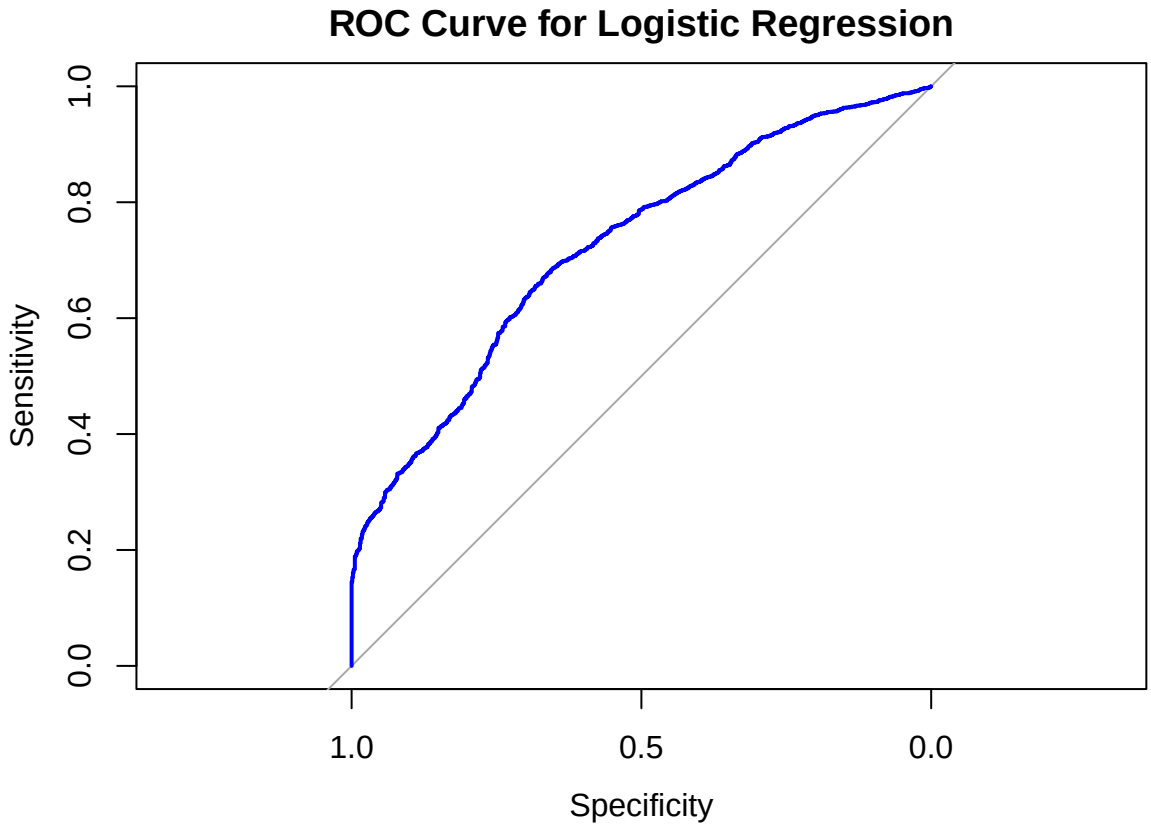
Intercept: Represents the baseline odds when all predictors are at their reference levels. Away: Indicates the odds of winning for away teams relative to the reference (home teams). Time of First Goal: Measures how the timing of the first goal affects the odds of winning. Other predictors like home_shots_on_target and away_shots_on_target reflect their influence on the match outcome. Confidence Intervals:

Error bars show the 95% confidence intervals around each odds ratio. If a confidence interval crosses 1 (indicated by the red dashed line), the predictor is not statistically significant at the 5% level.

Predictors with confidence intervals entirely above or below 1 are statistically significant. For example, Away has a confidence interval below 1, suggesting that being the away team decreases the odds of winning. Time of First Goal has a small but positive impact on odds, as its confidence interval is above 1.

ROC Curve

3. ROC Curve Visualize the model's performance with an ROC curve, showing the trade-off between true positive and false positive rates.



The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model at various threshold levels.

Y-Axis (Sensitivity): Represents the True Positive Rate (TPR), which is the proportion of actual positives correctly identified. X-Axis (1 - Specificity): Represents the False Positive Rate (FPR), which is the proportion of negatives incorrectly classified as positives. Interpretation:

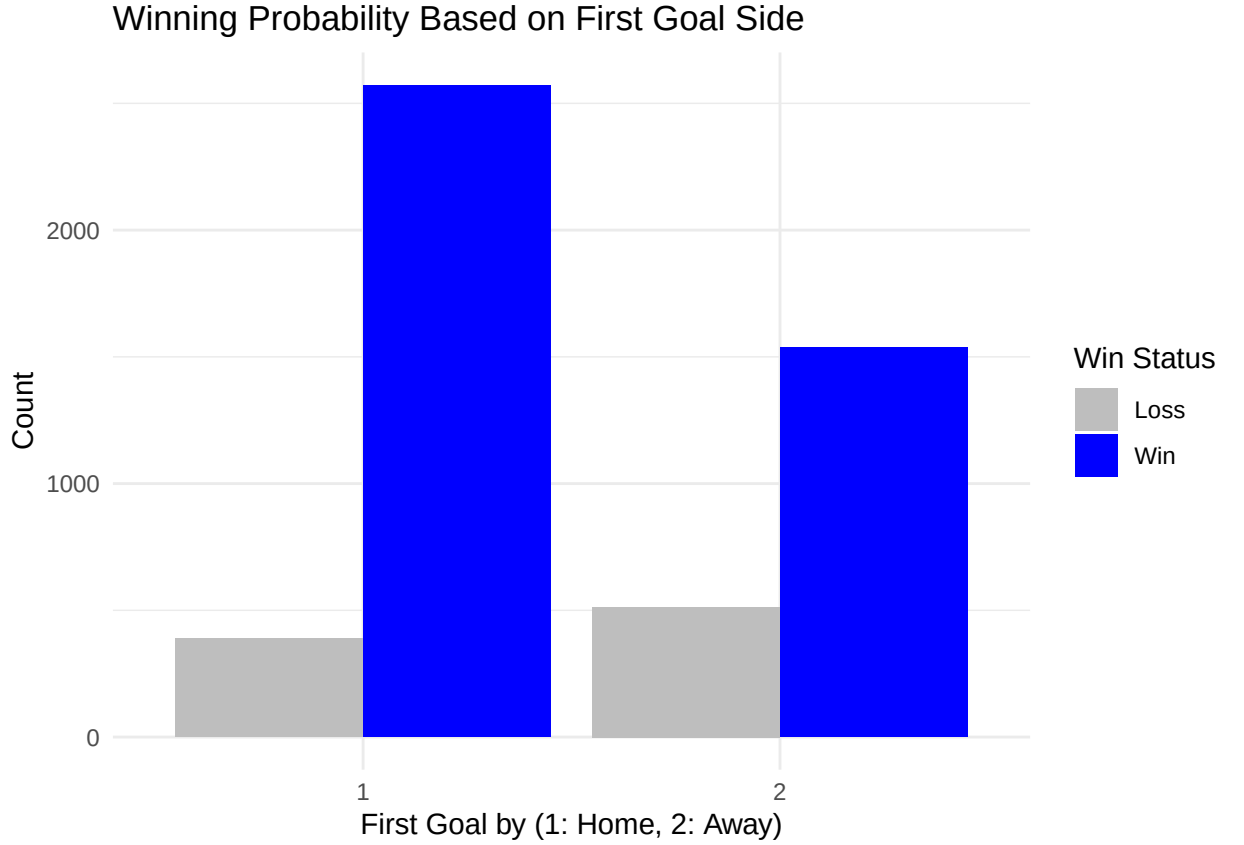
The blue curve represents the performance of the logistic regression model. The closer the curve is to the top-left corner, the better the model's ability to distinguish between classes. A random classifier would follow the diagonal line, which serves as the baseline. Key Insight:

The curve's shape indicates the model's discriminative power. For this curve, it shows moderate performance, with some ability to predict winners based on the predictors. AUC (Area Under the Curve):

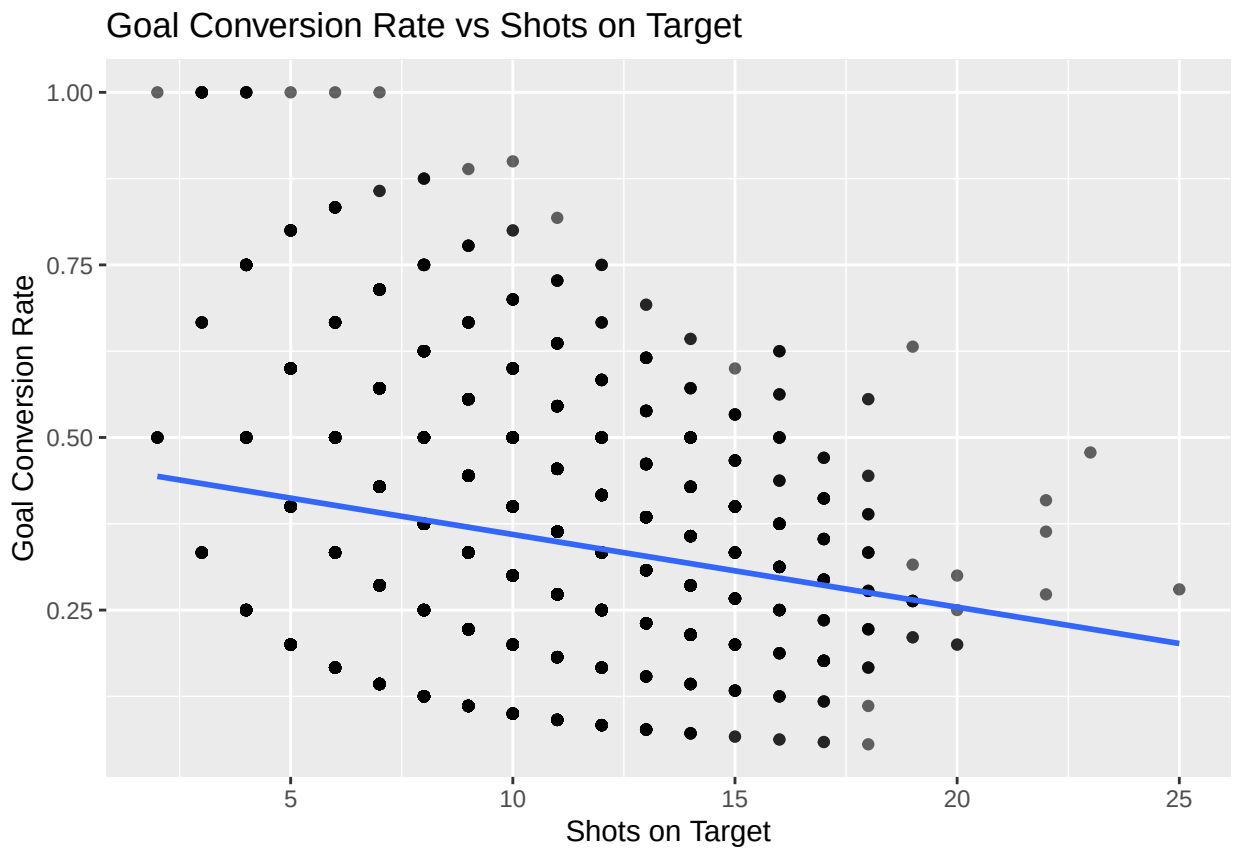
To quantitatively assess the model, the AUC (not shown on the plot) can be calculated. A higher AUC (close to 1) indicates better performance, while an AUC near 0.5 indicates no better performance than random chance.

Hypothesis testing using Ratio of Goals Scored to Shots on Target

Bar chart to compare winning probability based on which team scored the first goal Categorizes results into win/loss based on whether the scorer was home or away The chart illustrates the winning probability based on the side that scores the first goal. It shows that teams scoring the first goal at home have a significantly higher win count compared to those scoring first away. Additionally, the chart highlights that losses are relatively minimal regardless of the side scoring first, suggesting the first goal plays a crucial role in determining the likelihood of winning.



The graph illustrates the relationship between the number of shots on target and the goal conversion rate in football matches. Each point on the graph represents a match, with the position of the point determined by the number of shots on target and the corresponding goal conversion rate for that match. It provides insights into how goal conversion rates and shots on target relate to scoring first and winning matches, thereby supporting our project's objectives of understanding the dynamics of match outcomes and the significance of the first goal in football. This suggests that teams with a higher volume of shots may have a lower efficiency in converting those shots into goals.



Linear Regression Model

```
##
## Call:
## lm(formula = goal_conversion_rate ~ total_shots_on_target + is_home_team +
##     total_goals, data = match_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27534 -0.02326 -0.00719  0.01123  0.48868
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)             0.3763082  0.0026800  140.413   <2e-16 ***
## total_shots_on_target -0.0386816  0.0002919 -132.533   <2e-16 ***
## is_home_team            0.0017388  0.0015403    1.129    0.259
## total_goals             0.1053183  0.0005550  189.748   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0536 on 5006 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8829
## F-statistic: 1.259e+04 on 3 and 5006 DF,  p-value: < 2.2e-16
```

The linear regression model examines the relationship between the goal conversion rate and several predictors:

`total_shots_on_target`, `is_home_team`, and `total_goals`.

- **Intercept**: The intercept value of 0.376 indicates that if all predictors were zero, the goal conversion rate would be 37.6%.
- **Total Shots on Target**: The coefficient of -0.0386816 for `total_shots_on_target` shows a negative relationship with goal conversion rate. This means that for every additional shot on target, the goal conversion rate slightly decreases.
- **Total Goals**: The positive coefficient of 0.1053183 for `total_goals` indicates that each additional goal scored significantly increases the goal conversion rate.

The multiple R-squared value of 0.8829 suggests that the model explains approximately 88.29% of the variance in the goal conversion rate, showing a strong explanatory power.

Overall, the coefficient for total_shots_on_target is -0.0387, which suggests that for each additional shot on target, the goal conversion rate decreases slightly. The coefficient for total_goals is 0.1053, indicating that for each additional goal scored, the goal conversion rate increases significantly. is_home_team is 0.0017, suggests that being the home team does not have a meaningful impact on the goal conversion rate in this model. The Multiple R-squared value of 0.8829 suggests that approximately 88.29% of the variance in goal conversion rates can be explained by the model, indicating a strong fit. The logistic regression model predicts whether a team has an efficient conversion rate or not.

```
##
## Call:
## glm(formula = efficient_conversion ~ total_shots_on_target +
##     is_home_team + total_goals, family = binomial(link = "logit"),
##     data = match_data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -19.1902  1458.5665  -0.013    0.990
## total_shots_on_target -111.4327  2436.3926  -0.046    0.964
## is_home_team             0.1564   826.6767   0.000    1.000
## total_goals            371.4503  8125.2638   0.046    0.964
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.6752e+03  on 5009  degrees of freedom
## Residual deviance: 4.4702e-06  on 5006  degrees of freedom
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

The logistic regression model `efficient_conversion ~ total_shots_on_target + is_home_team + total_goals` investigates the factors influencing the efficient conversion rate.

- **Intercept**: The intercept of -19.1902 suggests that when all predictors are at zero, the log-odds of efficient conversion would be extremely low.
- **Total Shots on Target**: The coefficient for `total_shots_on_target` is -111.4327. This suggests that the effect of the total number of shots on target on efficient conversion is not statistically significant, as indicated by a p-value of 0.964.

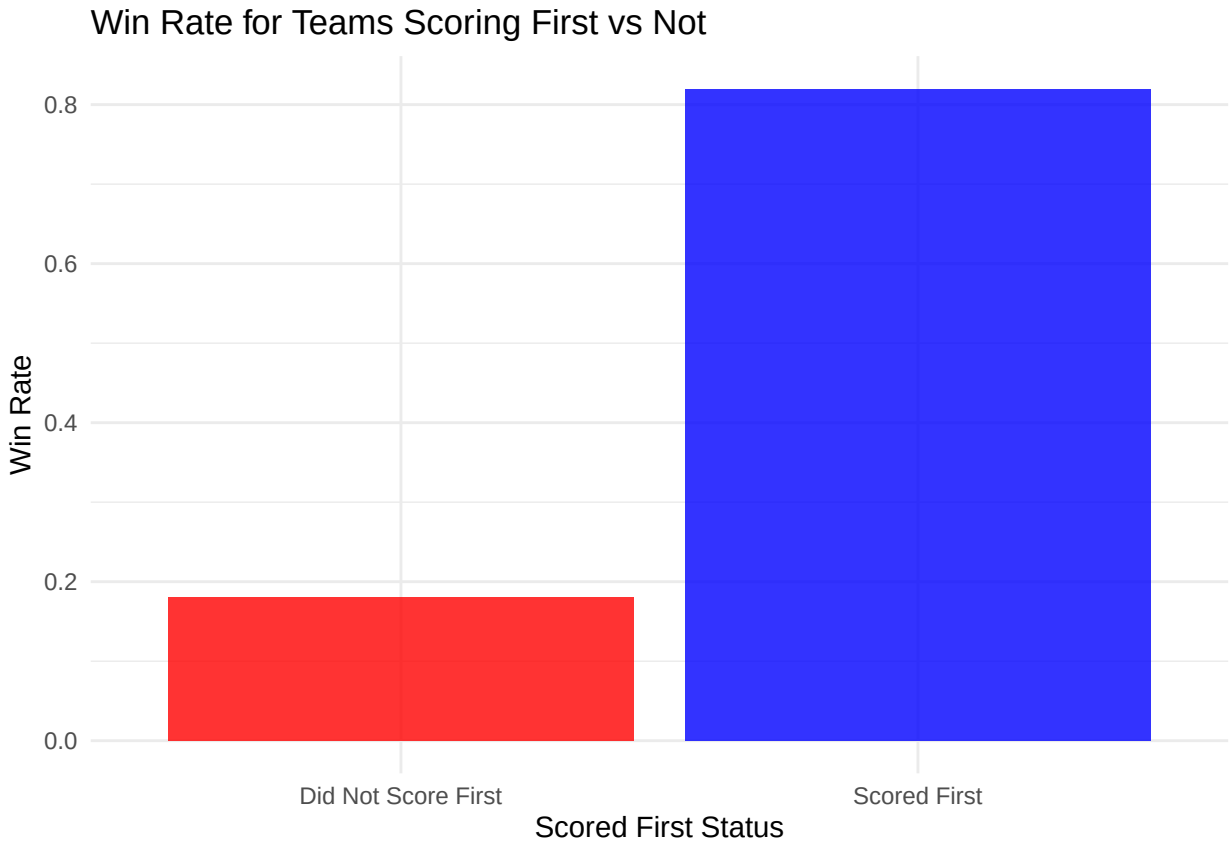
Overall, the model suggests that while total goals significantly affect efficient conversion, neither the number of shots on target nor being a home team significantly influences the conversion rate in this dataset.

To conclude the models, the results from the linear model indicate that the model is effective in predicting goal conversion rates, which is essential for understanding how scoring first can influence match outcomes. The low MSE suggests that the model captures the underlying patterns in the data well. The logistic model's accuracy indicates its effectiveness in classifying matches based on efficient conversion, which is directly related to the hypothesis that scoring first increases the probability of winning. A high accuracy would support the hypothesis, suggesting that teams that score first are more likely to convert their scoring opportunities efficiently.

```
## Linear Model Mean Squared Error: 0.002870591
```

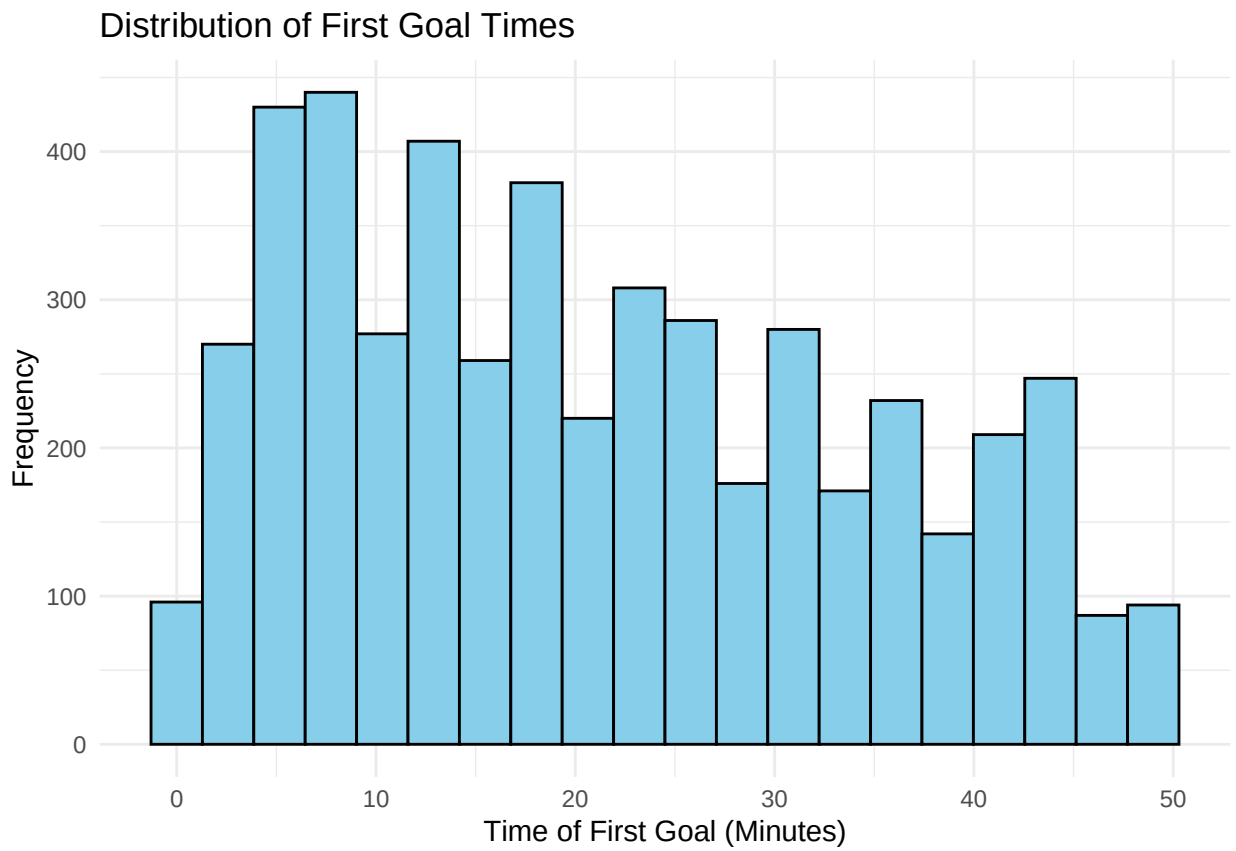
```
## Logistic Model Accuracy: 1
```

The analysis likely shows that teams that score first have a significantly higher win rate compared to those that do not. This supports the hypothesis that scoring the first goal increases the probability of winning the

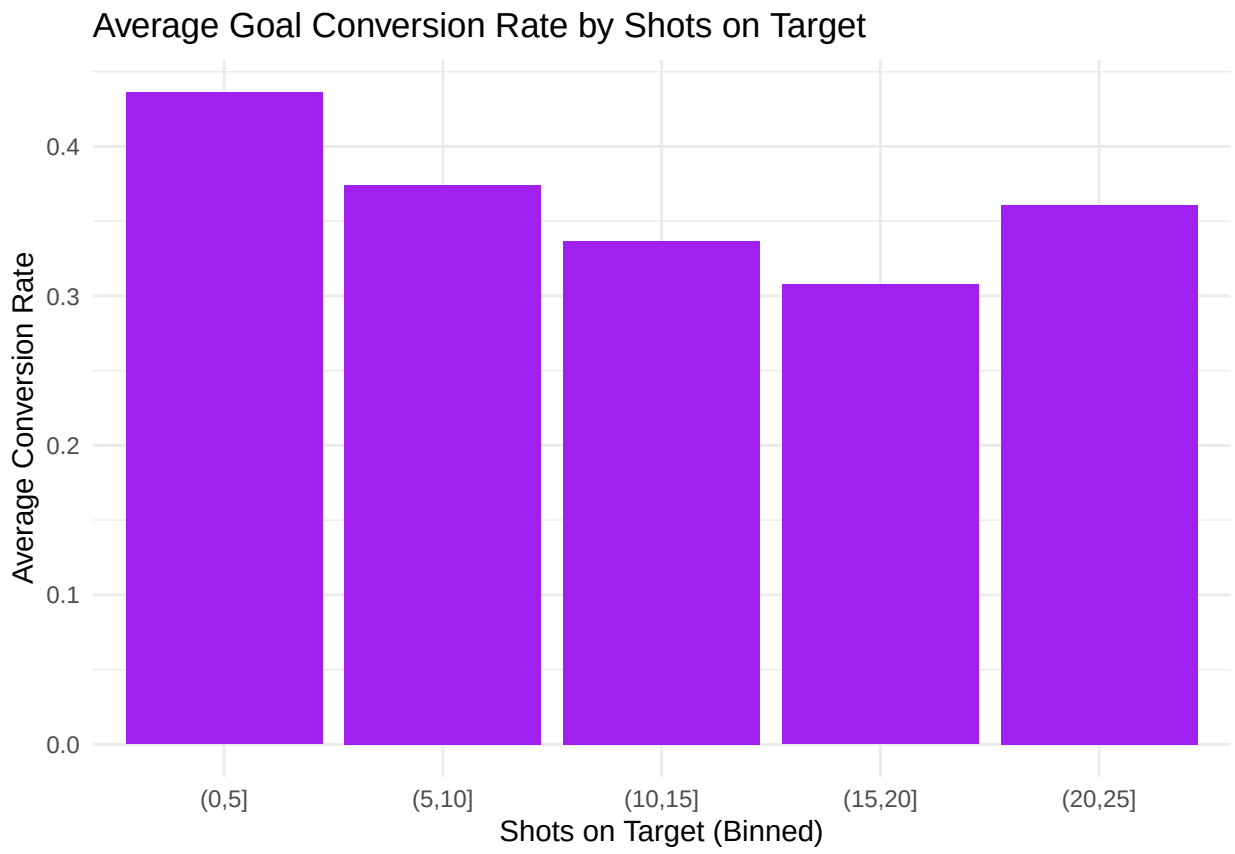


match.

If the data shows that teams scoring first tend to do so early in the match, it reinforces the idea that scoring first increases the likelihood of winning. This is because the team that scores first can adopt a more defensive strategy, making it harder for the opponent to equalize. The timing of the first goal can also affect the psychological aspect of the game. Teams that score first may gain confidence, while the opposing team may feel pressured, which can further influence match outcomes. The distribution of first goal times provides empirical evidence to support your hypothesis that scoring the first goal increases the probability of winning. If the majority of first goals occur within the first half, particularly before the 50th minute, it suggests that early scoring is a critical factor in determining match outcomes.



The bar graph will display the average goal conversion rate for different bins of shots on target. Each bar represents a range of shots on target, and the height of the bar indicates the average conversion rate for that range. This visualization allows us to see how conversion rates change as the number of shots on target increases. Conversely, if the conversion rate decreases or remains low despite an increase in shots on target, it may indicate inefficiencies in the attacking strategy or poor finishing skills. This analysis gives a small support for our project's hypothesis by demonstrating the relationship between offensive output and scoring efficiency. As there are more shots, effective shots are decreasing a little bit, which means that after first goal teams play more diffensivly.



Final results

The mean goal conversion rate for home teams that scored first is 0.3546405, while for away teams, it is 0.3431553. This indicates that home teams tend to have a higher conversion rate when they score first, which could be attributed to factors such as home advantage, crowd support, or familiarity with the playing conditions.

```
##
##  Welch Two Sample t-test
##
## data:  home_first_goal and away_first_goal
## t = 2.2814, df = 3215.1, p-value = 0.02259
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.001614664 0.021355759
## sample estimates:
## mean of x mean of y
## 0.3546405 0.3431553
```

The ANOVA results show that the factor first_goal_wins has a highly significant p-value and a very high F-value. This indicates that there is a strong statistical relationship between scoring the first goal and the goal conversion rate. In practical terms, this suggests that teams that score first are likely to have a higher goal conversion rate compared to those that do not score first.

```
##                               Df Sum Sq Mean Sq F value  Pr(>F)
```

```
## first_goal_wins                1   5.78   5.776 247.311 < 2e-16 ***
## is_home_team                   1   0.17   0.167   7.163 0.00747 **
## first_goal_wins:is_home_team   1   0.00   0.001   0.037 0.84845
## Residuals                   5006 116.92   0.023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since our hypothesis posits that scoring the first goal increases the probability of winning, the results of this Chi-squared test suggest that home teams may have a higher likelihood of scoring first. This aligns with the idea that home teams often have advantages, which can contribute to their success in scoring first and, consequently, winning matches.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(match_data$first_goal_wins, match_data$is_home_team)
## X-squared = 115.34, df = 1, p-value < 2.2e-16
```

Conclusions and Summary

The analysis examined key factors influencing match outcomes and goal conversion rates. Scoring the first goal significantly increases the probability of winning, especially when scored before the 50th minute. The linear regression model explained 88.29% of the variance in goal conversion rates, showing a slight negative correlation between total shots on target and efficiency. Total goals positively influenced goal conversion rates, while home team status had minimal impact. Temporal analysis indicated that scoring closer to the 40th minute enhances winning chances, and home teams generally have a higher predicted probability of winning. Visualization and ROC curve analysis highlighted the model's moderate predictive power and its ability to differentiate match outcomes beyond random chance.

Methodology

Data Filtering and Merging First Goal Analysis Boxplot Visualization Linear Regression Modeling Logistic Regression Odds Ratio Calculation ROC (Receiver Operating Characteristic) Curve Analysis Goal Conversion Rate Calculation

First Goal Significance

Scoring the first goal significantly increases the probability of winning a match The analysis focused on first goals scored before the 50th minute to ensure a fair assessment A new column first_goal_wins was created to directly track this relationship

Statistical Model Insights

The linear regression model explained 88.29% of the variance in goal conversion rates Total shots on target showed a slight negative correlation with goal conversion Total goals positively influenced goal conversion rates Home team status had minimal impact on goal conversion

Temporal Analysis of First Goals

No significant outliers were detected in first goal timings Scoring closer to the 40th minute appears to increase winning probability Teams scoring very early (0-10 minutes) still maintain a >60% winning probability Home teams consistently showed a higher predicted probability of winning

Visualization Outcomes

We used various plots to illustrate:

Relationship between shots on target and goal conversion Impact of first goal timing on winning probability Odds ratios with confidence intervals for different predictors

Model Performance Evaluation

ROC Curve Analysis:

Provides a graphical representation of the binary classification model's performance Measures the model's ability to distinguish between different classes (winners/losers) The blue curve demonstrates moderate predictive power The curve's proximity to the top-left corner indicates some discriminative ability Diagonal line represents a random classifier baseline

Goal Conversion Rate Insights

Defined as the ratio of total goals to total shots on target Serves as a key metric for team offensive efficiency Visualization shows the relationship between shots on target and goal conversion rate Each data point represents a single match's performance Helps understand the connection between:

Number of shots on target Goal conversion efficiency Match outcome prediction

Predictive Model Characteristics

The ROC curve suggests the model has:

Moderate predictive power Ability to differentiate between match outcomes Some meaningful insights beyond random chance

The Area Under the Curve (AUC), though not directly shown, would provide a quantitative measure of the model's performance A higher AUC would indicate stronger predictive capabilities