THE EFFECT OF FIRST GOAL ON THE FOOTBALL MATCH OUTCOME

SPORTS ANALYTICS

PRESENTED BY SILVA YEGHIAZARYAN, LEONID SARKISYAN, TIGRAN KOSTANYAN, ARAM GRIGORYAN, SURAM BAGRATYAN

TABLE OF CONTENTS

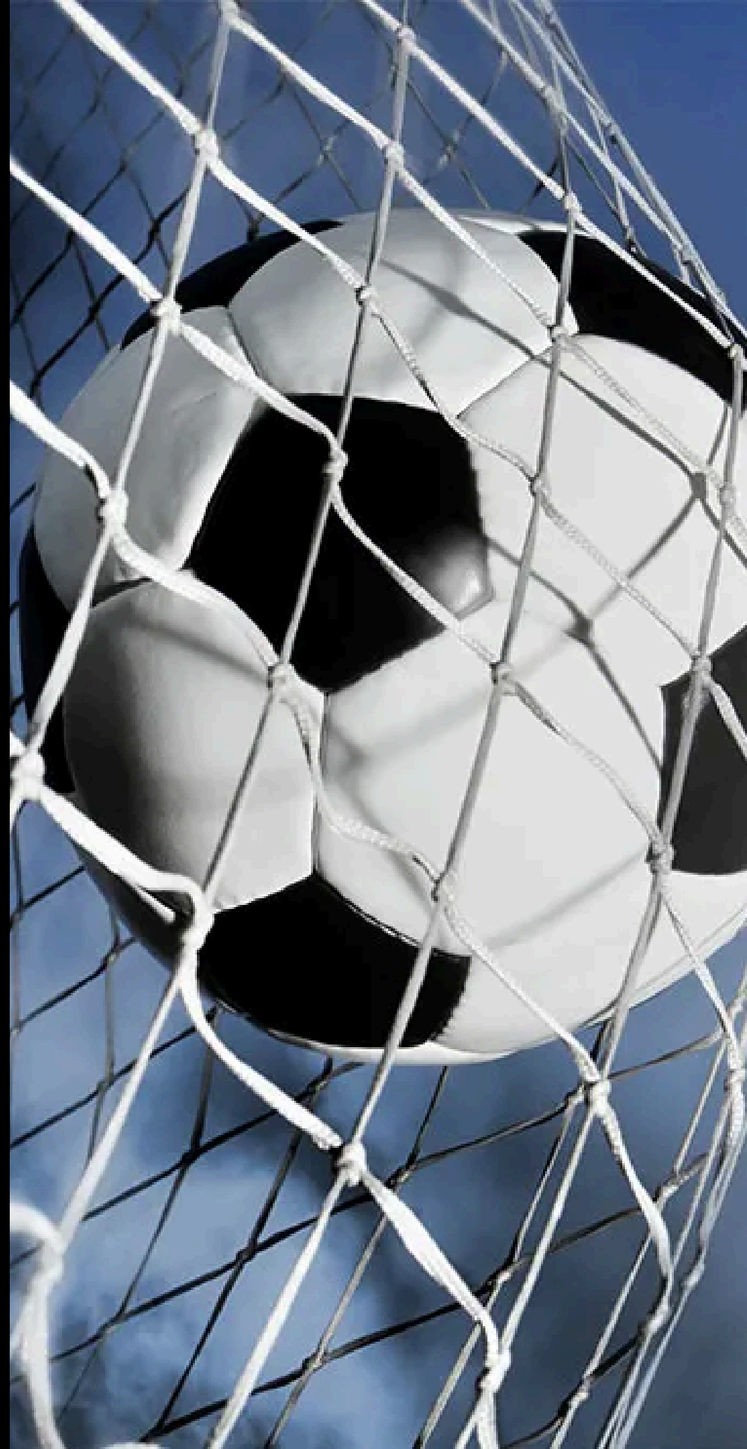
HYPOTHESIS

Scoring the first goal increases the probability of winning the match.

THE DATASET

Information about events from the biggest 5 European football leagues: England, Spain, Germany, Italy, France from 2011/2012 season to 2016/2017 season as of 25.01.2017.

Includes information about more than 900000 events which are not only goals but yellow cards, red cards, offside, attempt, foul, corner, substitution etc.



DATA PREPARATION

1. Select Relevant Columns and Filter The Data

Select the events which are either goal or shots on target

2. Calculate The Game Results

Calculate the match results and shots on target during the match for home and away teams

3. Identify The First Goal and Scored Side

Identify the first goal, the scored team and the winner (home or away)
Add Goal Conversion Rate for later analysis

DATA PREPARATION

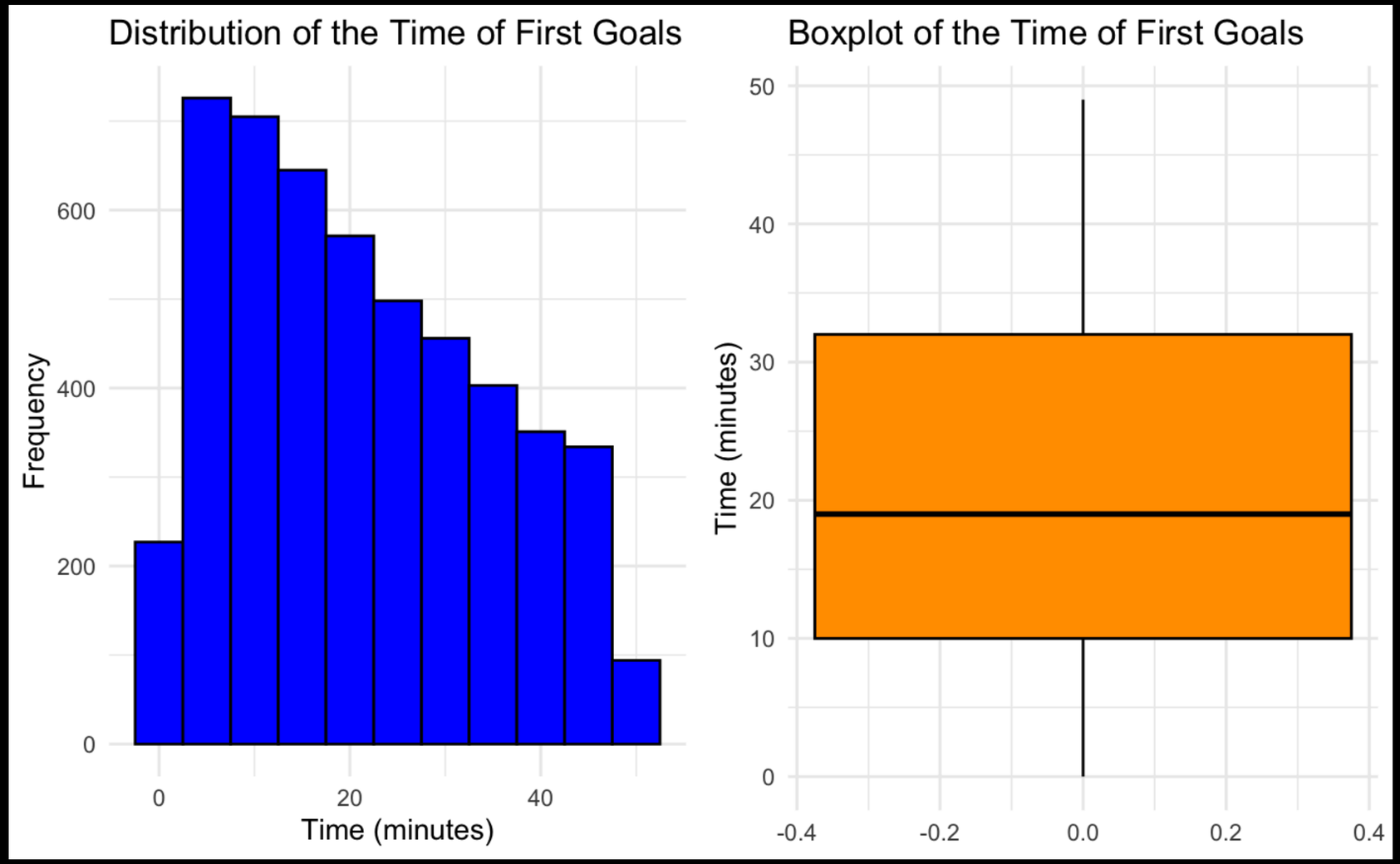
4. Keep The Cases When The First Goal is Scored After 50th Minute

For example the first goal is scored in the 90th minute of the game, the likelihood of the team winning increases simply because there is little time remained for the opponent team to score. Our threshold is the 50th minute.

After data preparation, the dataset has 5010 rows

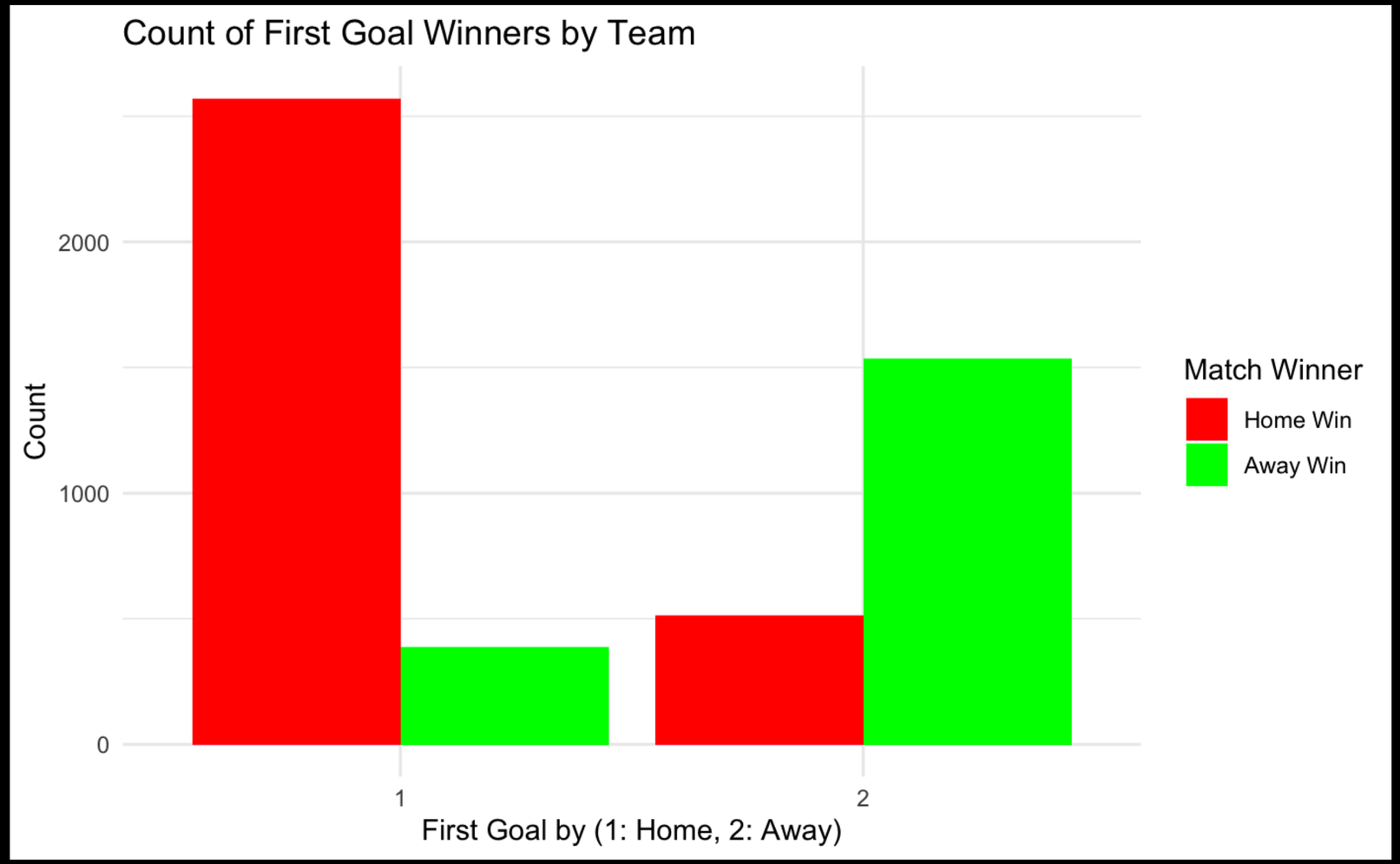
EXPLORATORY DATA ANALYSIS

First goals are typically scored early in football matches, with the majority occurring within the first 20 minutes.



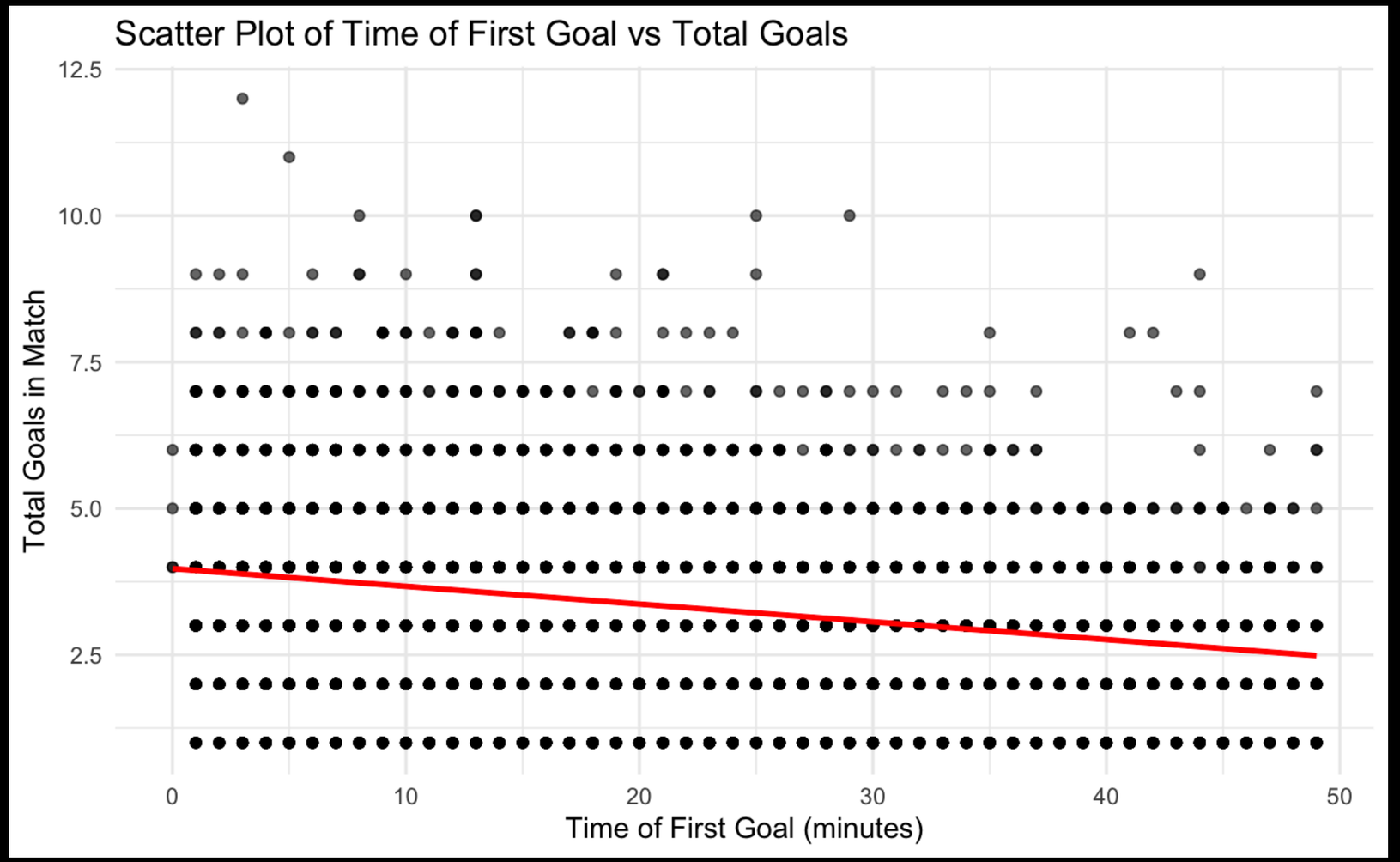
EXPLORATORY DATA ANALYSIS

When the home team scores first, they predominantly win. Similarly, when the away team scores first, they often secure victory but there is notable home advantage.



EXPLORATORY DATA ANALYSIS

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.



HYPOTHESIS TESTING

Chi-Square Test of Independence

The p-value of 0.001529 indicates a statistically significant relationship between scoring the first goal and the match outcome. 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
```

HYPOTHESIS TESTING

Logistic Model

```
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
```

HYPOTHESIS TESTING

Logistic Regression

Intercept (3.81): The positive intercept indicates a strong likelihood of winning for the first-goal scorer.

Side (-0.87): The negative coefficient suggests that when the away team scores the first goal, their chances of winning are significantly lower. 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.

HYPOTHESIS TESTING

Predictions with Logistic Model

```
          Actual
Predicted    0    1
        0   16    9
        1  237 1241
[1] 0.8363273
```

1. **Accuracy**: $\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{Total}}$

   Accuracy = $\frac{16+1241}{16+9+237+1241} = 0.836$

   This means the model correctly classified 83.6% of the cases overall.

2. **Precision** (for positive class):

   $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{1241}{1241+9} = 0.993$

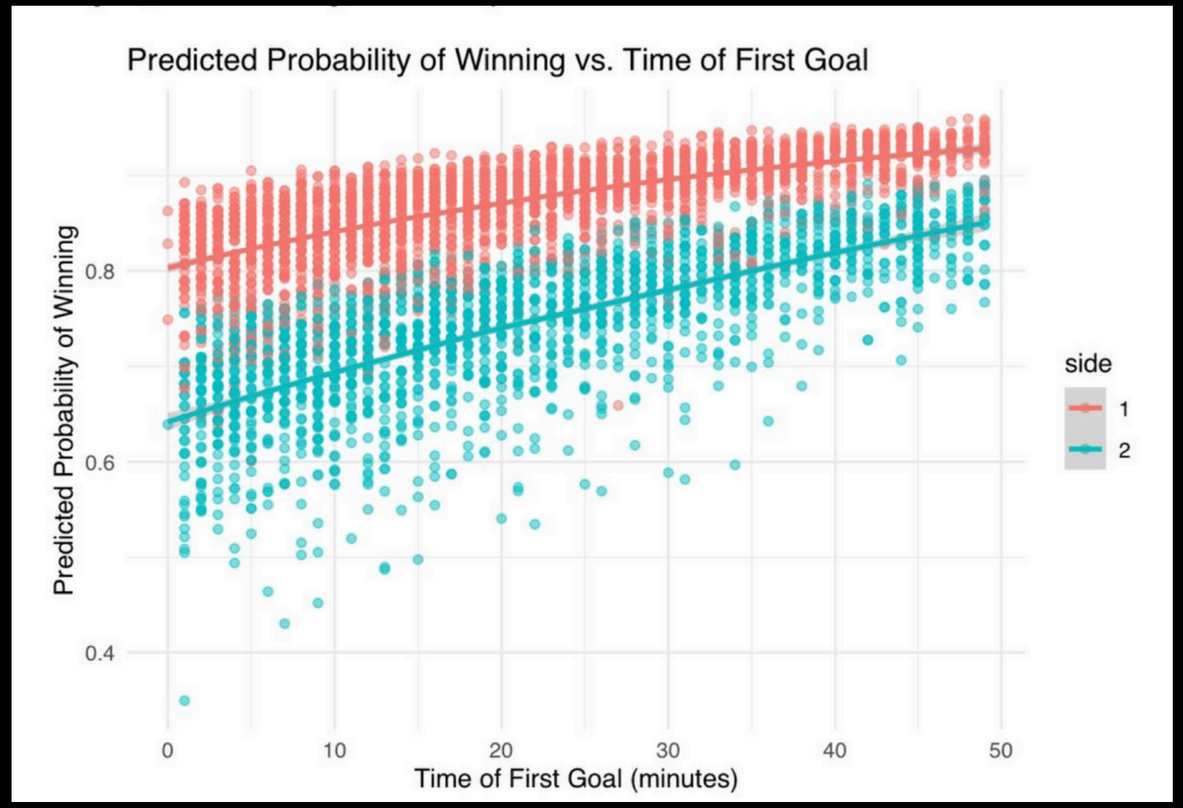
   High precision indicates that most of the cases predicted as 1 were correct.

HYPOTHESIS TESTING

Predictions with Logistic Model

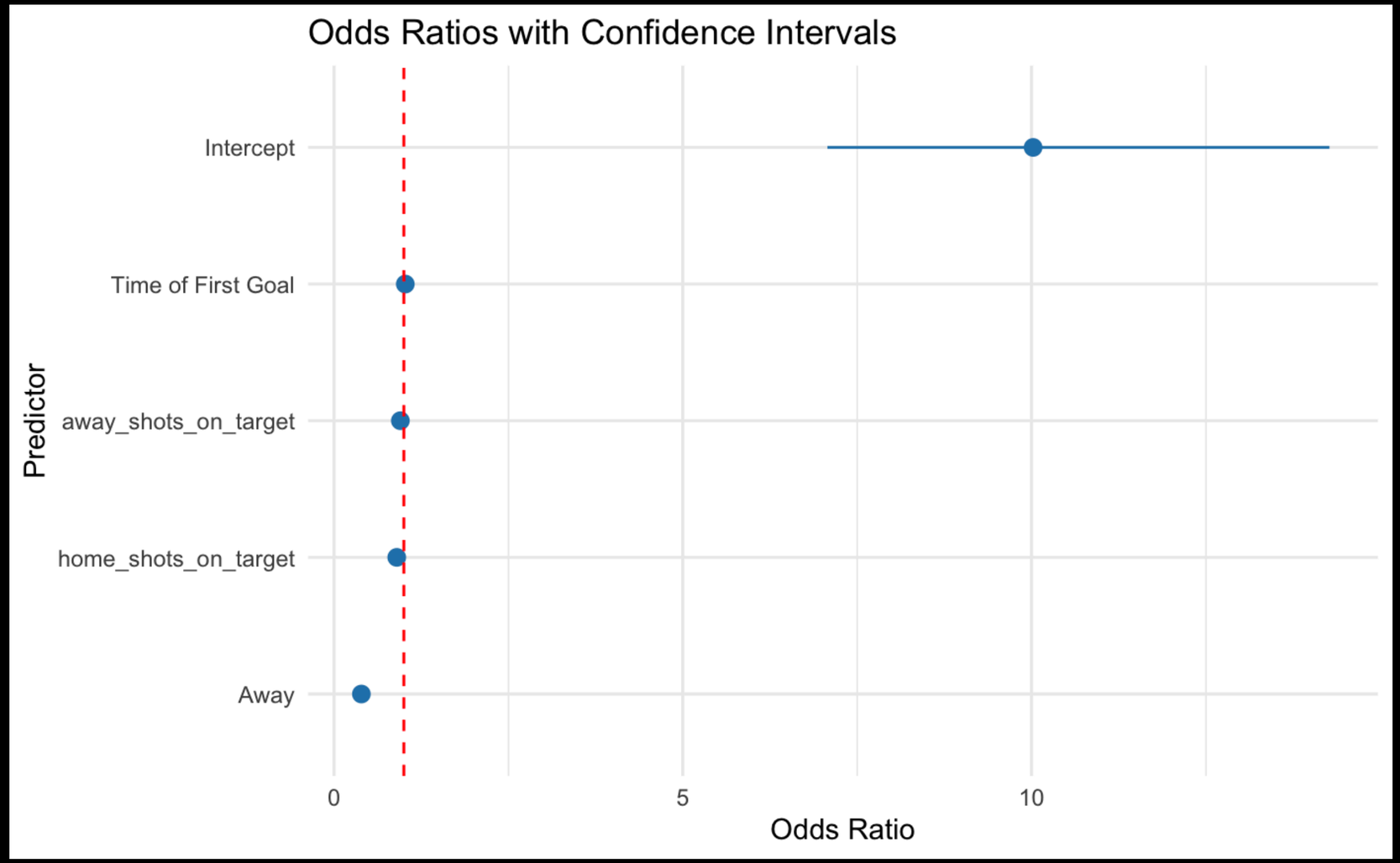
Scoring early: Teams that score very early (near the 0-10 minute) have lower predicted probabilities.
Scoring late: Teams scoring later (closer to the 40th minute) see a more significant increase in predicted probabilities.

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



HYPOTHESIS TESTING

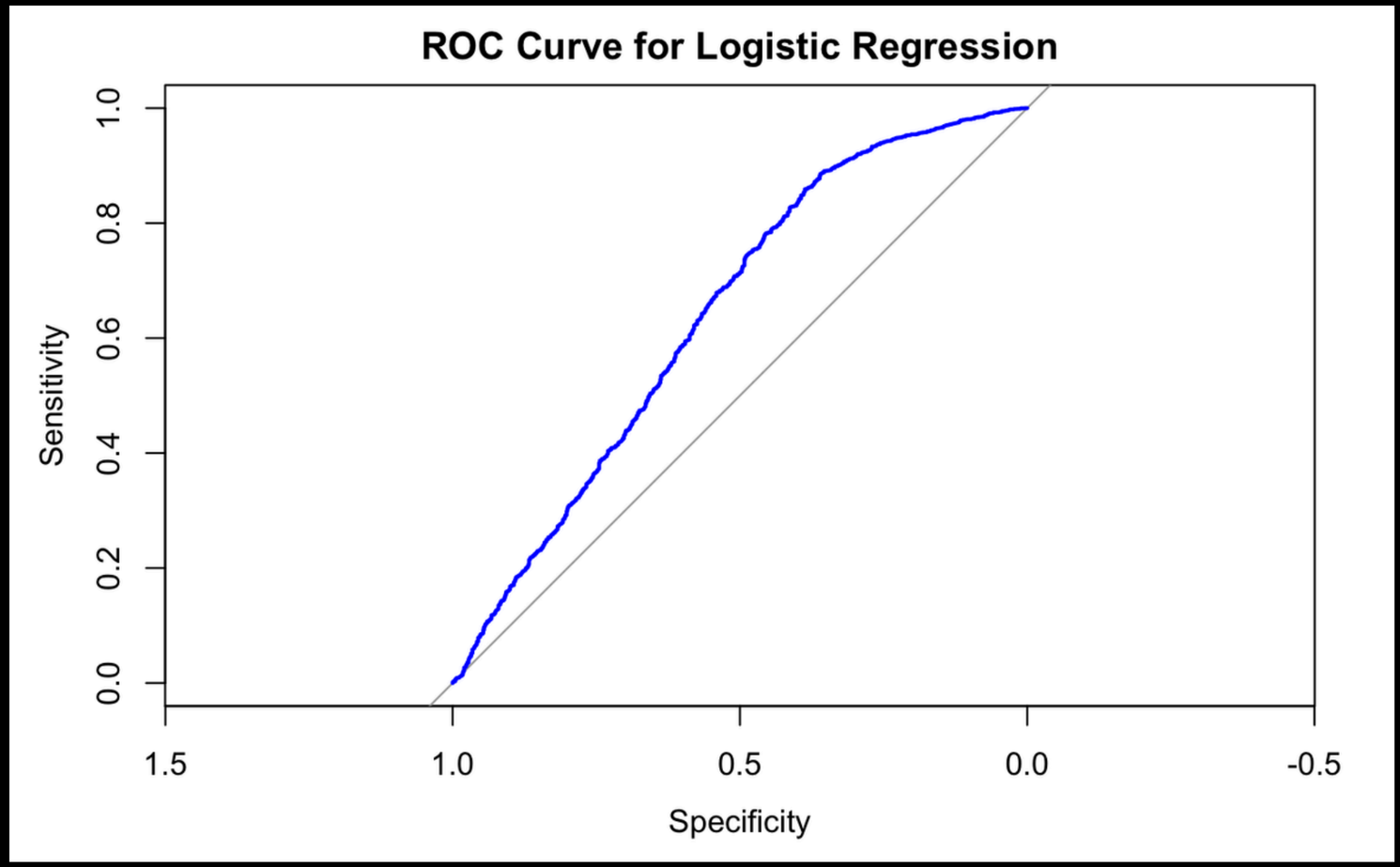
Odds Ratios with Confidence Intervals



HYPOTHESIS TESTING

ROC Curve

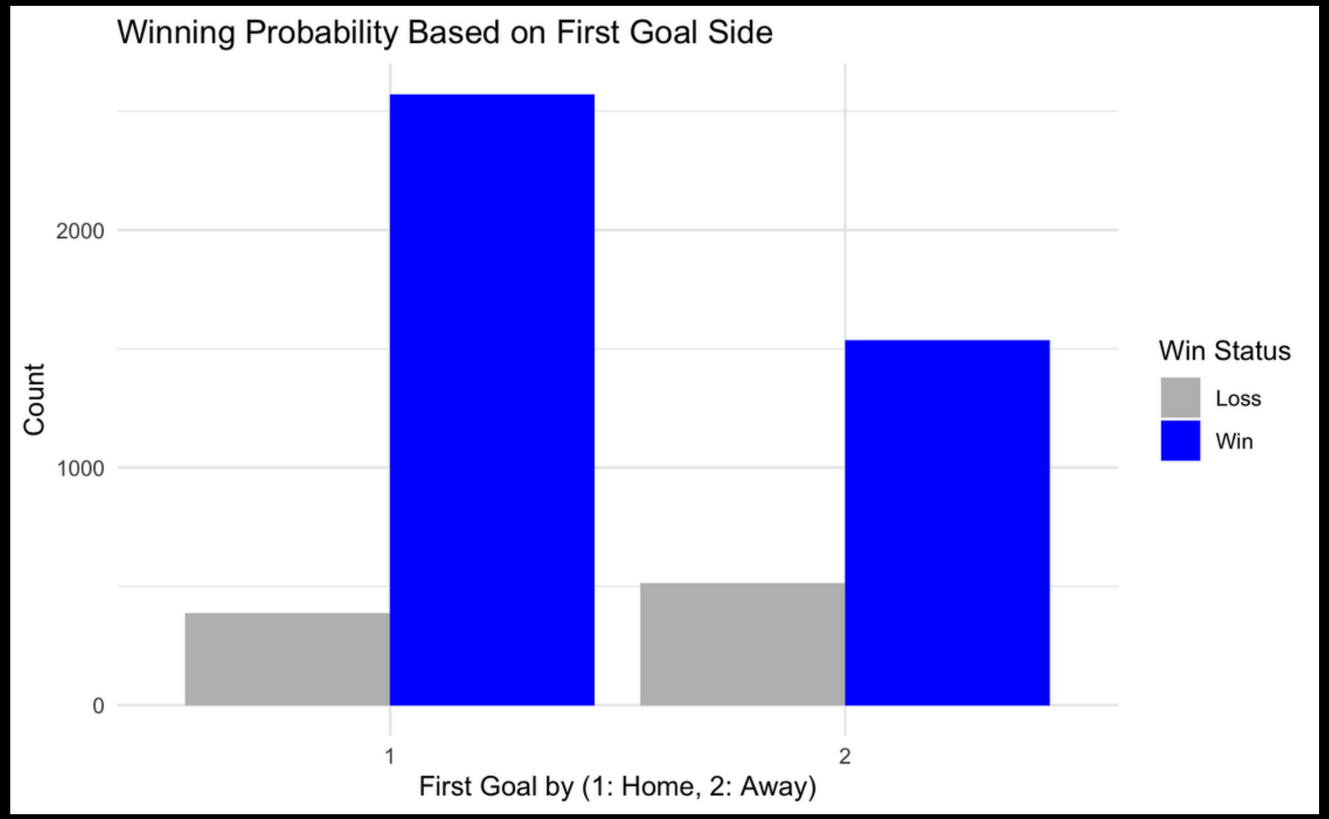
For this curve, it shows moderate performance, with some ability to predict winners based on the predictors.



HYPOTHESIS TESTING

Model Using the Ratio of Goals Scored to Shots on Target

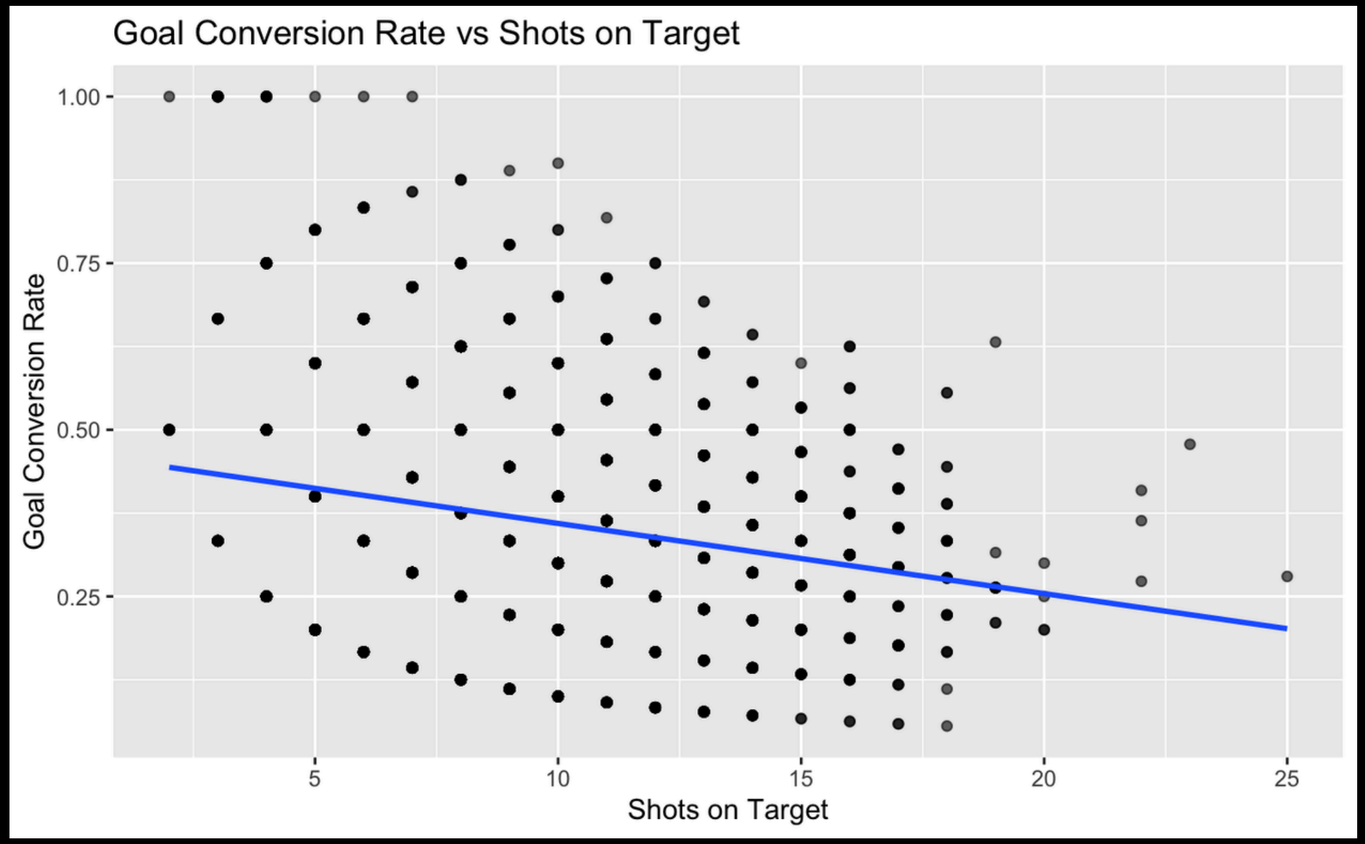
Teams scoring the first goal at home have a significantly higher win count compared to those scoring first away.



HYPOTHESIS TESTING

Model Using the Ratio of Goals Scored to Shots on Target

Teams with a higher volume of shots may have a lower efficiency in converting those shots into goals



HYPOTHESIS TESTING

Linear 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 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.

HYPOTHESIS TESTING

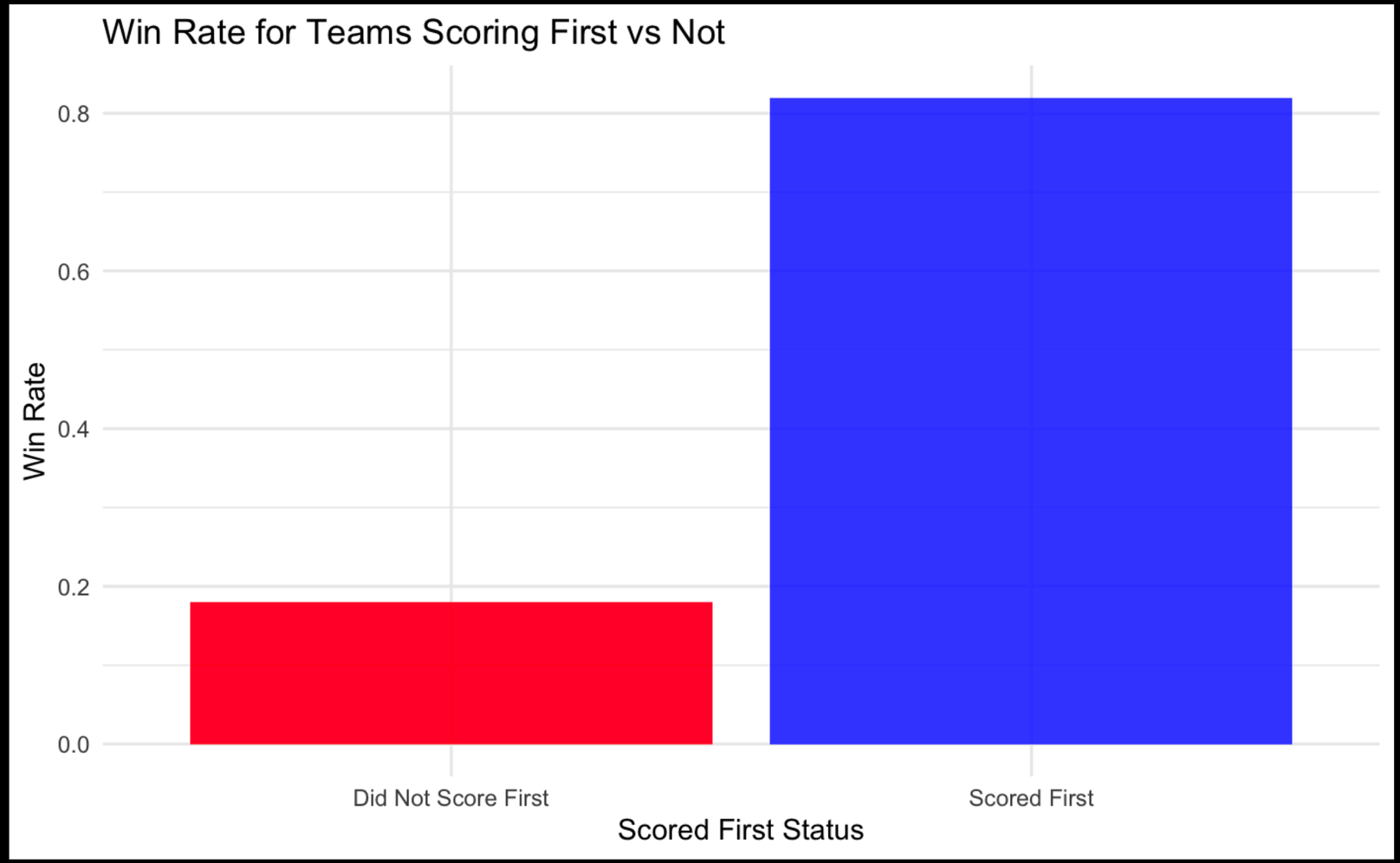
Logistic Model

```
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 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.

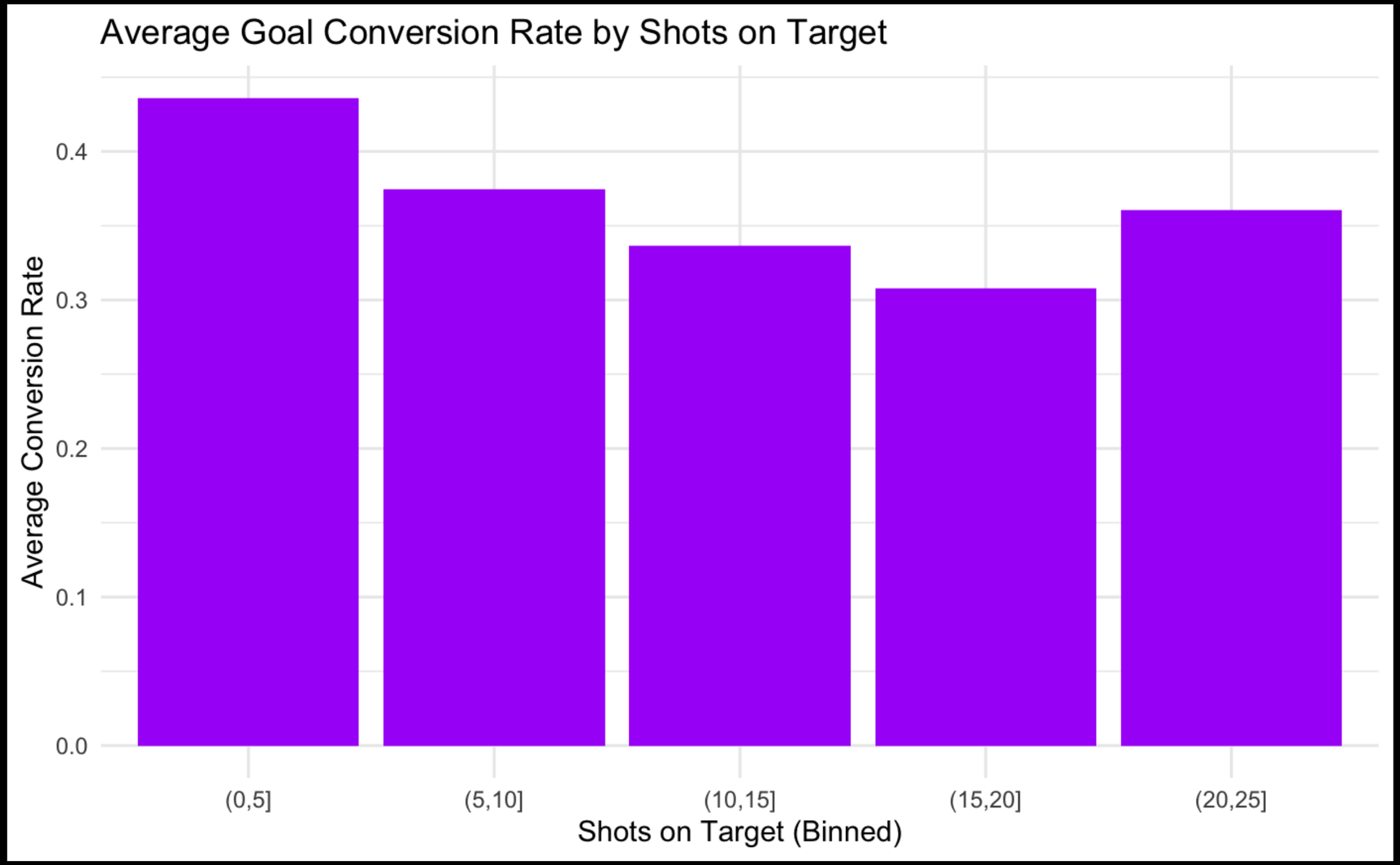
HYPOTHESIS TESTING

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.



HYPOTHESIS TESTING

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



HYPOTHESIS TESTING

```
                                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
```

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.

RESULTS AND CONCLUSION

- 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
- The linear regression model explained 88.29% of the variance in goal conversion rates
- Home teams consistently showed a higher predicted probability of winning
- Scoring closer to the 40th minute appears to increase winning probability

THANK YOU FOR YOU ATTENTION!