NBA 2014/2015 Season Shot Analysis

Hrdya Bhaskaran

09 February 2021

Abstract

The purpose of this study is to analyse the influence of various features like Shooting Distance, Nearest Defender Distance, Dribbling, Touch time and the time remaining in the Shot Clock on a shot attempted in the basketball game. This study also investigates the advantages of Home Team over Away Team in the NBA basketball games. Various statistical tests are performed on the NBA 2014-2015 season data set to investigate if there is any statistical evidence to show that these features play a significant role in the shot outcome. Finally, a shot prediction model is built using logistic regression to explore the explanatory power of these features.

Contents

1 Introduction

The National Basketball Association (NBA) is the premier men's professional basketball league in the world. The league is composed of 30 teams and the regular season runs from October to April. The official website of the NBA - https://www.nba.com provides latest NBA basketball news, scores, stats and many other interesting information.

The data set analyzed in this report consists of 12,8609 observations of the shots taken by the players for NBA 2014-2015 season. The main objective of this report is to perform statistical analysis on the variables (listed below) that are crucial for scoring maximum points in a basketball game. The analysis provides information about whether these variables are significantly different for the observations belonging to 'Shots Made' group and the 'Shots Missed' group. Finally a shot prediction model is built using logistic regression to check if the results predicted using these variables are better than random guessing.

Because of the complexity of the game, the list of variables that determine the success of a shot is not exhaustive. As a result, the logistic regression model built using these specific features cannot achieve high accuracy. But the statistical analysis performed in this report help us to understand the significance of these features in the basketball game outcome.

- Shooting Distance
- Nearest Defender Distance
- Shot Clock
- Touch Time
- Location

2 Background

Data preprocessing is done before performing statistical analysis. Among 12,8609 observations, 312 records had negative values for `Touch Time` and 5,567 rows had missing values for `Shot Clock`. Negative `Touch Time` observations were removed and missing values for `Shot Clock` were imputed with the mean value.

In section: 3, we will explore various features in the data set by performing univariate and bivariate analysis to find interesting hidden insights. In section: 4, we will perform statistical tests to investigate if the means of these features are significantly different for the 'Shots Made' and the 'Shots Missed' observations.We will also perform analysis of variance (ANOVA) to check if there is any significant difference between the means of the `Closest Defender Distance` in the 'Two Points Scored', 'Two Points Missed', 'Three Points Scored' and 'Three points missed' groups. To analyze the correlation between the remaining time in the shot clock and the shooting efficiency, we will perform Spearman's rank correlation test.

In section: 5, we will conduct proportion test to investigate if the location can influence the game outcome and finally in section: 6, logistic regression model is built to assess the explanatory power of these variables on the outcome of a shot.

Since the parametric tests assume that the means of the various samples are normally distributed, sample sizes of all groups are checked to ensure that the normal approximation of the means are realistic. Histograms and Q-Q plots are also plotted to know more about the data distribution.

3 Inital Dataset Analysis

In this section, we will analyze the relation between the successful shot percentage and various variables (shot distance, defender distance, time remaining in the shot clock).

Figure 1 indicates a positive linear relation between the defender distance and the successful shot percentage. When the defender is very close, scoring three points seems to be more challenging than scoring two points. A probable reasoning for this could be that, for two pointers, even when the defender is nearby, they have the advantage of being nearer to the basket. The linear relation is relatively strong until around 8ft defender distance. At greater defender distances, all the attempted two points were successful. The Figure: 2 (Right) also illustrates that in the 2014/2015 season, more than three quarters of the points scored were two points.

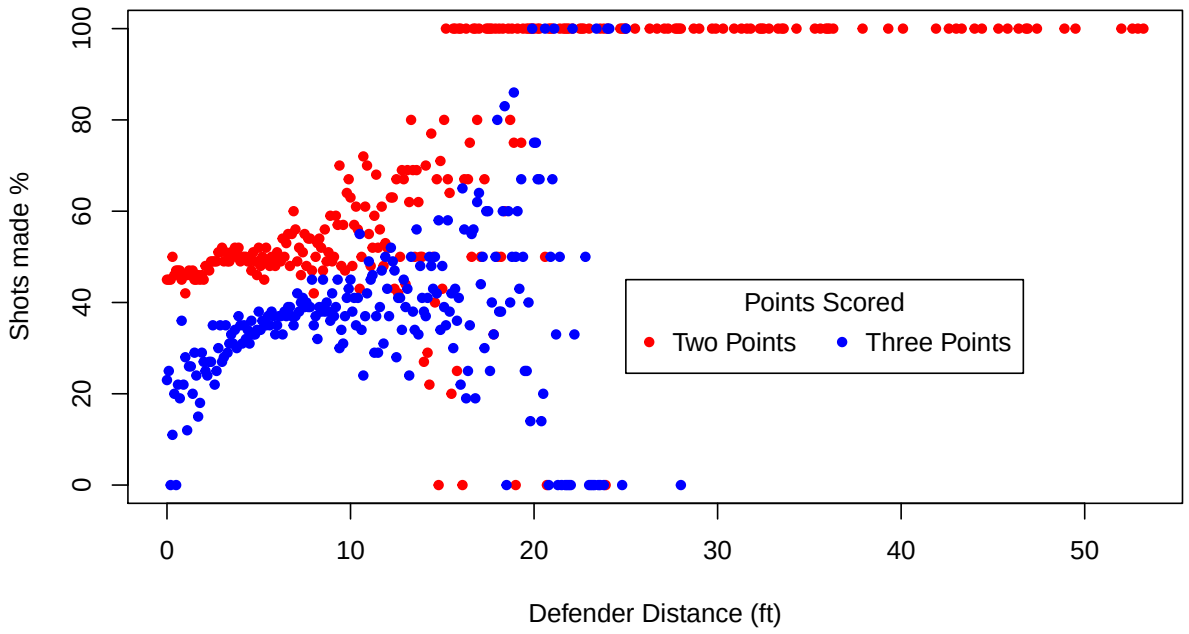

Figure 1: Relation between Defender Distance and Shot Made Percentage

As expected, the Figure: 2 (Left) illustrates negative linear relation between the shooting distance and the successful shots made percentage. A remarkable shot was taken by the player **Derrick Rose** from 47.2ft shooting distance. The nearest defender distance when this shot was taken was 6.2ft which is more than the average defender distance in the data set.

Figure: 2 (Right) illustrates the influence of remaining time on the shot clock over the shooting efficiency. It appears that, as the time remaining reduces, shooting efficiency also decreases. In section: 4, we will perform statistical test to investigate more about this.

As shown in the Figure: 3 (Left), majority of games were won by the `Home Team`. To investigate if the proportion of games won by the `Home Team` is significantly greater, we will perform proportion test in the section: 5.

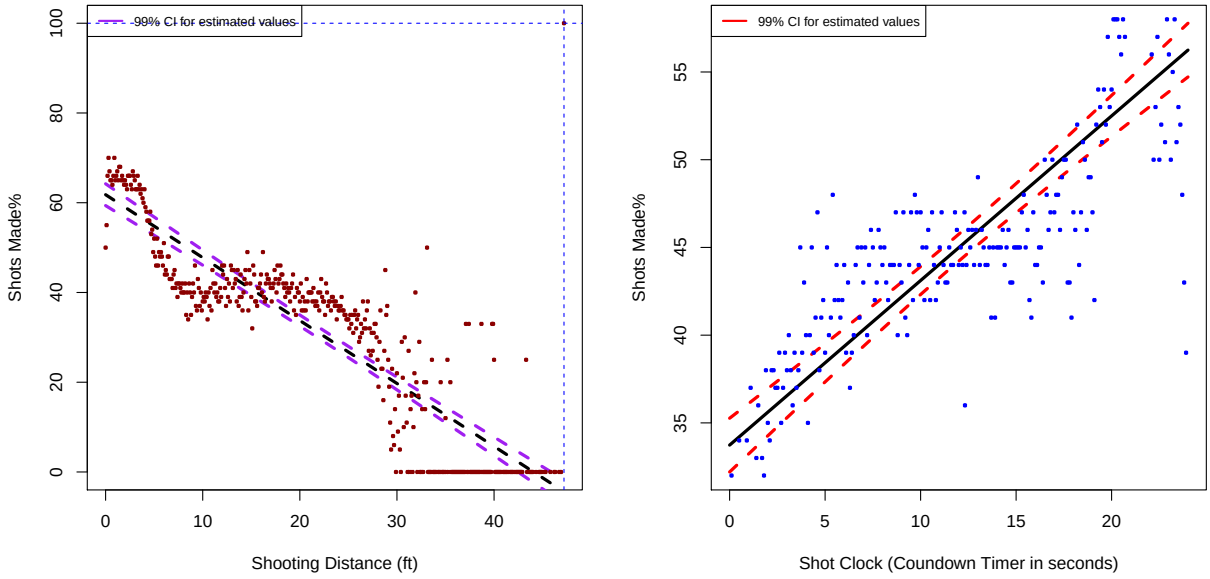

Figure 2: Graphical representation of the relation between the Shooting Distance and the Shots Made Percentage (Left) and the relation between the time remaining in the Shot Clock and the Shots Made Percentage (Right) are shown here, the regression lines (dashed black line) are added along with the corresponding 99% confidence intervals for both the graphs.

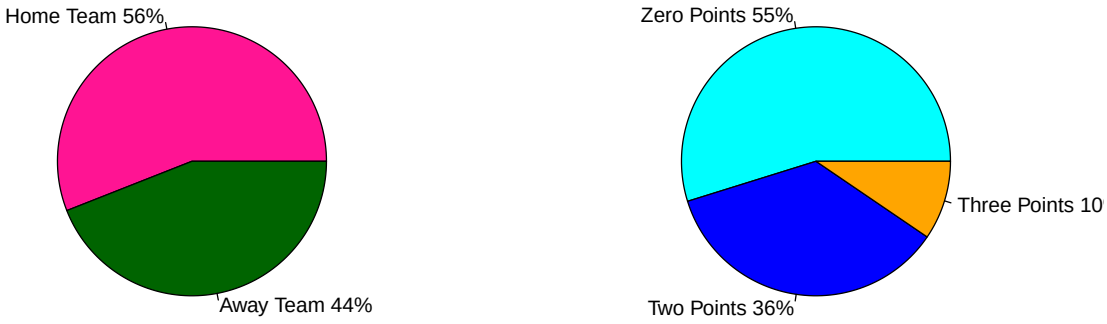

Figure 3: Pie chart representation of the proportion of games won by Home Team and Away Team (Left) and the proportion of various points scored in the 2014/2015 season.(Right)

4 Significance of various features

In this section, we will perform various statistical tests on some of the features to understand their role in the shot outcome.

4.1 Influence of shooting distance on the shot outcome

Figure 4 indicates that the mean shooting distance might be less for `Shots Made` observations. To investigate this, we will perform t-test to assess if the mean shooting distance is significantly less for the `Shots Made` observations than the `Shots Missed` observations. The histograms and Q-Q plots suggested that the samples are not normally distributed, but the sample sizes are very large ($56{,}355$ observations for `Shots Made` and $68{,}356$ observations for `Shots Missed`) to assume that the normal distribution is realistic for sample means.

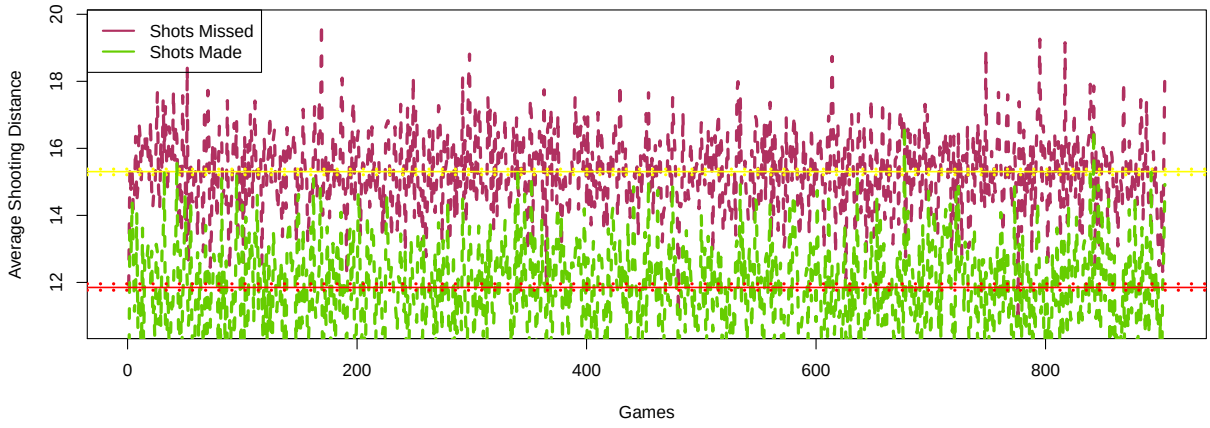

Figure 4: Mean shooting distance for shots made and shots missed observations calculated for every games. Corresponding means and their 99% confidence intervals are also shown (yellow and red dashed lines).

To check if the variance of the shooting distance for 'Shots Made' and 'Shots Missed' observations are homogeneous, we will perform *Two Sample Test for Variance* at the significance level of $\alpha = 0.5$.

```
## [1] F-test (compararison of variances); p-value=9.6197e-08
```

Since the p-value obtained is very small ($9.62e-08$), we will perform *Two Sample Test For Means* assuming unequal variances at the significance level of $\alpha = 0.5$.

- H_0 : The mean shooting distance is equal for 'Shots Made' and 'Shots Missed' observations.
- H_1 : The mean shooting distance is less for observations belonging to 'Shots Made' group.

```
## [1] t-test (compararison of means); p-value= < 2.2e-16
```

The obtained p value ($< 2.2e-16$) is very small, suggesting that the mean shooting distance might be less for the successful shots than the missed ones. Thus it appears that the shooting distance might influence the shot outcome significantly.

Table 1: Sample size in each group.

Group	Sample_Size
Two Points Missed	46631
Two Points Made	44498
Three Points Missed	21725
Three Points Made	11857

4.2 Influence of closest defender distance on the shot outcome

In the initial data exploration (section: 3), we have seen that the closest defender distance is different for the `Two Pointers` and the `Three Pointers`. In this section, we will perform Analysis of Variance (ANOVA) to analyze the differences in means of the closest defender distance in the following groups - 'Two Points Missed', 'Three Points Missed', 'Two Pointers Made' and 'Three Pointers Made'. The Figure: 5 gives a visual representation of the closest defender distance belonging to each group.

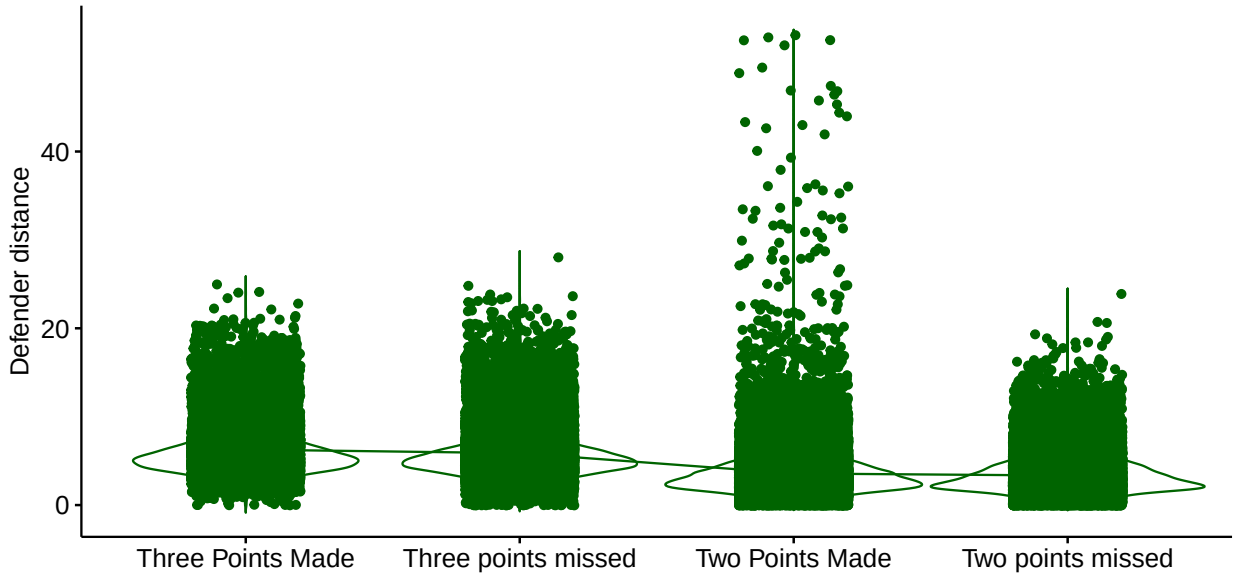

Figure 5: Closest Defender Distance belonging to each group.

All the three tests for homogeneity of variances (see Table: 2) returned very small p-values indicating that the variances are not homogeneous. Note that, as shown in the Table 1, the groups have unequal sample sizes. But the sample size in each group is large, hence we will perform *Welch's one-way test* (which does not assume equality of variance).

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  defenderdist and pointsscored
## F = 7351.6, num df = 3, denom df = 39735, p-value < 2.2e-16
```

Table 2: Tests for homogeneity of variance for the nearest defender distance belonging to each group

Test	p.value
Bartlett	< 0.0001
Fligner-Killeen	< 0.0001
Levene	< 0.0001

```
## [1] Pair wise t-test :
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  defenderdist and pointsscored
##
##                      Three Points Made Three points missed Two Points Made
## Three points missed <2e-16            -                   -
## Two Points Made     <2e-16            <2e-16              -
## Two points missed   <2e-16            <2e-16              <2e-16
##
## P value adjustment method: bonferroni
```

The p-values obtained from *Welch one-way test* and *Pairwise t- test with Bonferroni correction* are very small, suggesting that there is statistical evidence to show that the mean defender distance is significantly different between every group. `ANOVA.lm` function was also checked and it returned the same result. It thus appears that the variable 'Closest Defender Distance' might be able to influence the outcome of the shot.

4.3 Effect of shot clock on shooting efficiency

In the initial data analysis (section: 3), we have seen that the shot clock has linear relation with the shooting efficiency. Here we will investigate the correlation between the time remaining in the shot clock and the successful shots made percentage. To find this, the percentage of shots made was calculated for all the 241 unique values in the variable `Shot clock`.

```
## [1] Shapiro Wilk Test for Normality; p-value=0.0003
```

```
## [1] Spearman's rank correlation test; p-value=1.6689e-63
```

Since the p-value (0.00029) returned from the Shapiro-Wilk test suggest that the multivariate-normality assumption is not realistic, we will perform test using the method *Spearman's rank correlation* at the significance level of $\alpha = 0.01$.

As shown in the figure 6, the p-value obtained ($< 2e - 16$) is very small suggesting that the time remaining in the shot clock and the shots made percentage might be correlated. It appears that the players might be under pressure as the time in the shot clock decreases, which might be affecting the shooting efficiency.

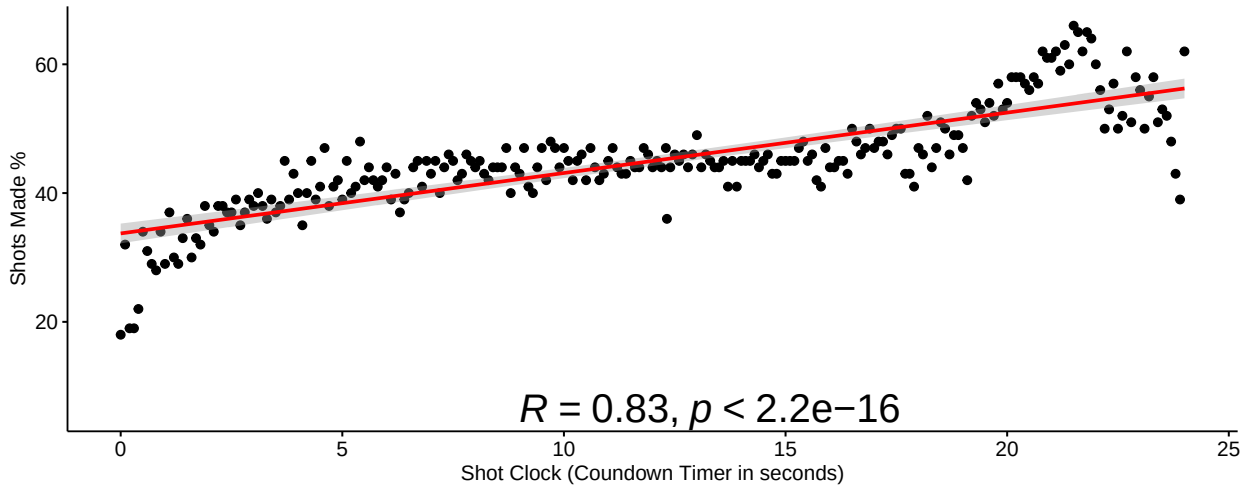

Figure 6: Relation between time remaining in the shot clock and shooting efficiency.

4.4 Influence of touch time on the shot outcome

Figure: 7 illustrates the mean touch time for `Shots Made` and `Shots Missed` observations calculated for all the 904 games. It suggest that the mean touch time for `Shots Missed` observations are less than the `Shots Made` observations. Here, we will perform t-test to investigate if the difference in the mean touch time is significant. Note that the histograms and Q-Q plots, indicated that the samples are not normally distributed, however the sample sizes are large enough to assume that the normal distribution of the sample means are realistic.

- H_0 : The mean touch time is equal for 'Shots Made' and 'Shots Missed' observations.
- H_1 : The mean touch time is less for 'Shots Made' observations than 'Shots Missed' observations.

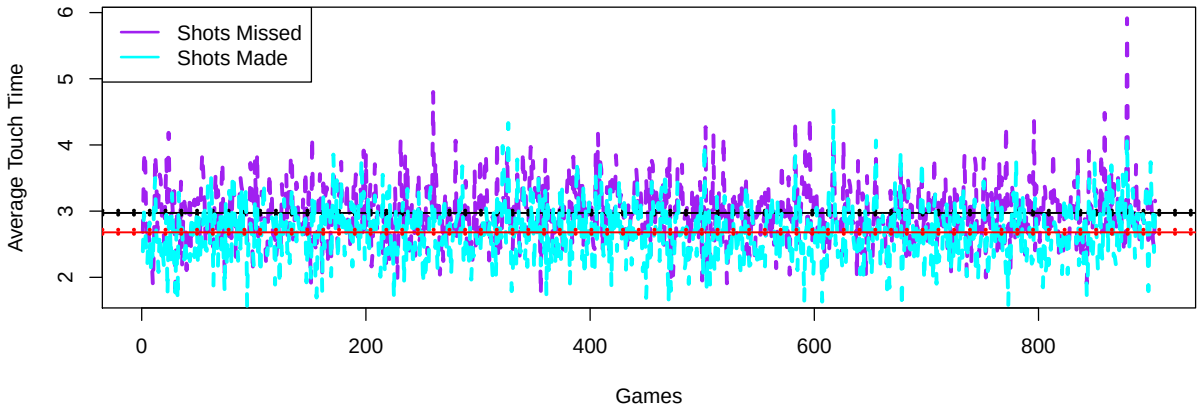

Figure 7: Mean touch time for 'Shots Missed' and 'Shots Made' observations calculated for all the games. Corresponding means and their 99% confidence intervals are also shown (black and red dashed lines).

First we will test for equality of variance at the significance level of $\alpha = 0.5$.

```
## [1] F-test (compararison of variances); p-value=1.3462e-67
```

Since the p-value obtained is very small ($P < 2.2e-16$), we conduct lower tail t-test at the significance level of $\alpha = 0.5$ assuming that the variances are unequal.

```
## [1] t-test (compararison of means); p-value=7.5412e-66
```

The p-value obtained ($P < 2.2e-16$) from the t-test, suggests that the mean `Touch Time` is less for the `Shots Made` observations than the `Shots Missed` observations. It appears that, it may not be a good strategy for a player to keep the ball for a long time. Making the right decision at the right time, to pass the ball or to shoot, without giving an opportunity for the defending team, appears to be very important. Thus we can conclude with reasonable certainty that the `Touch Time` can influence the `Shot Outcome`.

5 Home Court Advantage

In this section we will investigate whether playing at home can have an impact on the game outcome. Though the underlying reasons behind this advantage is not fully understood, research done by (Entine and Small, 2008) suggest that lack of rest for the `Away Team` might be contributing to the home court advantage in the NBA.

5.1 Proportion of games won by the Home Team and the Away Team

Here we will perform proportion test to check if the proportion of games won by the `Home Team` is significantly greater than the `Away Team` at the significance level of $\alpha = 0.05$. Among the 904 games played during this season, 506 games were won by the `Home Team`. This test is conducted under the assumption that the estimated proportions are empirical proportions and the sample size is sufficiently large.

H_0 : The proportion of games won by the `Home Team` is equal to the proportion of games won by the `Away Team`.
H_1 : The proportion of games won by the `Home Team` is greater than the proportion of the games won by the `Away Team`.

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(home_won_count, away_won_count) out of c(total_games, total_games)
## X-squared = 25.33, df = 1, p-value = 2.416e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.07995618 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5597345 0.4402655
```

The obtained p-value ($2.416e-07$) is very small suggesting that there is statistical evidence indicating that the `Home Team` has some advantage over the `Away Team`. Support of the crowd and the comfort of being at home rather than traveling might be helping the `Home Team` to perform better. It thus appears that the feature `Location` may have some predictive power on the `Shot Outcome`.

6 Logistic Regression to predict a shot

In this section, we will build a logistic regression model to evaluate the explanatory power of the features we discussed so far. The model tries to predict whether a shot will be made or missed based on the values of selected features.

The stepwise selection using `stepAIC` function in MASS package returned the full model itself indicating that all the features are relevant. Note that the model was built using only the features that we discussed in this report.

```
## [1] Coefficients of step.model:

##    (Intercept) CLOSE_DEF_DIST      SHOT_DIST     TOUCH_TIME     SHOT_CLOCK
##    -0.00951249     0.10254133    -0.06551953    -0.07326547     0.01573870

##   PTS_TYPE  LOCATIONH
## 0.09054312 0.02764795

## [1] p-values for the Z-values of step.model:

##    (Intercept) CLOSE_DEF_DIST      SHOT_DIST     TOUCH_TIME     SHOT_CLOCK
##   8.200961e-01  2.706253e-299   0.000000e+00   2.200530e-39   7.783422e-46

##     PTS_TYPE    LOCATIONH
## 8.957620e-06 1.832907e-02
```

The negative coefficients of shooting distance and touch time indicate that the shots are likely to be made when their values are less. The t-tests performed in section: 4 also suggested the same. The variable `LOCATIONH` shown in the coefficients of step model (see above) is the home team which has positive coefficient. However based on the p-value for the Z values, the location appears to be less significant than the other features. But removing this feature gave similar results.

The Area under the ROC curve (Figure: 8) that represents the capability of the model to distinguish between `Shots Made` and `Shots Missed` observations is 63.22%.

```
## [1] Confusion matrix for step.model with 44.6% decision treshold:

##       FALSE  TRUE class.error
## FALSE 39809 23010   0.4176224
## TRUE  28547 33345   0.4083045
```

The decision threshold was set at 44.6% to obtain balanced percentage of classification errors. With this threshold value, the model could correctly classify 58% of `Shots Missed` observations and 59% of `Shots Made` observations. Though the score is not very impressive, the model is performing better than random guessing.

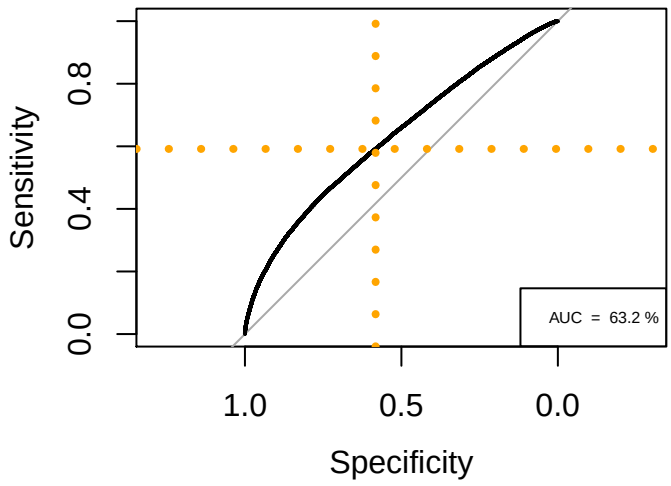

Figure 8: ROC curve for the logistic regression model; the orange dashed lines correspond to the sensitivity and specificty of the binary classifer at the 44.6% decision threshold.

7 Conclusion

In this report, we analyzed `NBA` data set to explore the importance of some of the features that can influence a shot taken. Various statistical tools, both parametric and non parametric methods were used for analysis. The assumptions were checked to ensure that the results we get from these methods are valid.

The main observations can be summarized as follows:

- Time remaining in the shot clock appear to be correlated with the shooting efficiency.
- Making a decision to shoot or pass without keeping the ball for long, thus reducing the touch time appear to be a better strategy.
- Home Team might have some advantage over Away Team, but their influence on the shot outcome is less significant than other features.
- Distance from the nearest defender appear to be an important variable in terms of the points scored (Two points or Three points). Players tend to score maximum total points by taking more number of two pointers than three pointers.
- Shooting distance appear to play a significant role in determining the success of a shot taken.

Considering the complexity of the game, only a subset of important features are analyzed in this report. The observations discussed in this report should however be further analyzed by adding features like free throws and by including more recent NBA data.

References

Entine, O. and Small, D. S. (2008). The role of rest in the nba home-court advantage. *Journal of Quantitative Analysis in Sports*, 4.