Mat022 Coursework Report

Hasan Can Uzuner

23 Şubat 2021

Abstract

In this report, we show how to apply inferential and descriptive analysis made on the NBA dataset. We try to extract information from the dataset and interpret them. Some statistical tests will be used to help us to understand the dataset better. We will study how the shot clock affects the score. We will compare performances of teams per game and also in the 2014-2015 NBA season, the final was played between Golden State Warriors and Cleveland. Their best players were Stephen Curry and Lebron James. We will try to analyze their stats to find differences and similarities between Stephen Curry and Lebron James. Then, correlation plots of teams and players will be shown to find correlated features and also the season's best players in different categories will be shown. Finally, we will try to make a prediction about is the team win or lose this game base on statistic per game.

Contents

1 Introduction

The NBA dataset includes shot attempts of players in each game in the NBA 2014-2015 season. Basketball is a game played 5 to 5 and each team has 24 seconds to make a score in a try. There are several features of players given to us in the dataset.

```
##  [1] "GAME_ID"          "DATE"               "HOME_TEAM"
##  [4] "AWAY_TEAM"        "PLAYER_NAME"        "PLAYER_ID"
##  [7] "LOCATION"         "W"                  "FINAL_MARGIN"
## [10] "SHOT_NUMBER"      "PERIOD"             "GAME_CLOCK"
## [13] "SHOT_CLOCK"       "DRIBBLES"           "TOUCH_TIME"
## [16] "SHOT_DIST"        "PTS_TYPE"           "SHOT_RESULT"
## [19] "CLOSEST_DEFENDER" "CLOSEST_DEFENDER_ID" "CLOSE_DEF_DIST"
## [22] "FGM"              "PTS"
```

Each row consists of information about a player who tries to make a score. As we can see above, the dataset is very detailed and lots of different analyses can be made on them. In the dataset, there are 128069 records and 23 features. Also this dataset has records of 281 unique players and 904 unique games.

2 Performances

2.1 Total Point Statistic based on Shot Clock

In this section, we search for are there any effects of the shot clock on points per second. To do that, at first, the NBA dataset has been divided per game. Then the shot clock feature variables have rounded to make data discrete. Figure 1 suggests that, if the shot clock increases, the maximum number of total points in this second also increases. Therefore, it can be seen that there is a positive correlation between shot clock and pts. The same positive correlation is also valid for a mean number of total points per second. These observations are confirmed by Spearman's rank correlation tests.

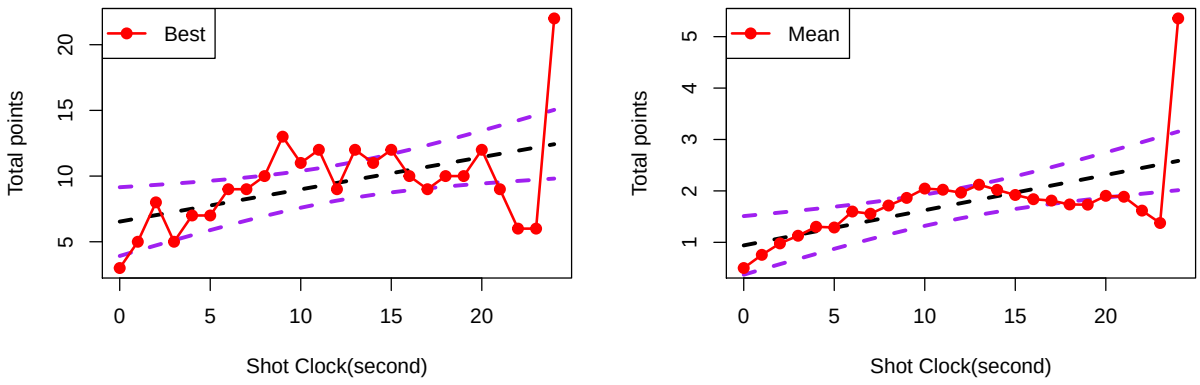

Figure 1: Best and Mean total points per second based on shot clock

In the figure, Several observations have been made. Firstly, at the beginning of the 24-second attack duration, mean and best point numbers are relatively low but it increases over time. Secondly, generally, players make score between 10 and 20 seconds and if they exceed 20 seconds, they tend to use the time completely. This can be seen clearly in the figure. At the end of the 24 seconds, the mean total points are almost 6 and the best total points are almost 23 which is significantly higher than others. Spearman's correlation estimate values of best and mean total points are 0.43 and 0.58 respectively. This is proof of a positive correlation.

performance	alternative	p.value	estimate
Best	greater	0.0160021	0.4298062
Mean	greater	0.0012264	0.5784615

2.2 Comparison of teams points performances per game

Figure 2 gives statistics about the means and variances of teams per game. In this dataset, there are no free throw statistics. Because of that, the exact score cannot find, however, the total of 2 points and 3 points still can be calculated and it will give intuition about teams performances. As we can see in the figure, Minnesota Timberwolves, Philadelphia 76ers, and New York Knicks had the lowest performance. This is not surprising because, in the 2014-2015 season, they were the worst teams in their conferences. Golden State Warriors and Los Angeles Clippers had the highest means. Golden State Warriors were the champion of the 2014-2015 season so this result is also predictable. Besides that, the top scorer player of the season was Stephen Curry which is played for Golden State Warrior.

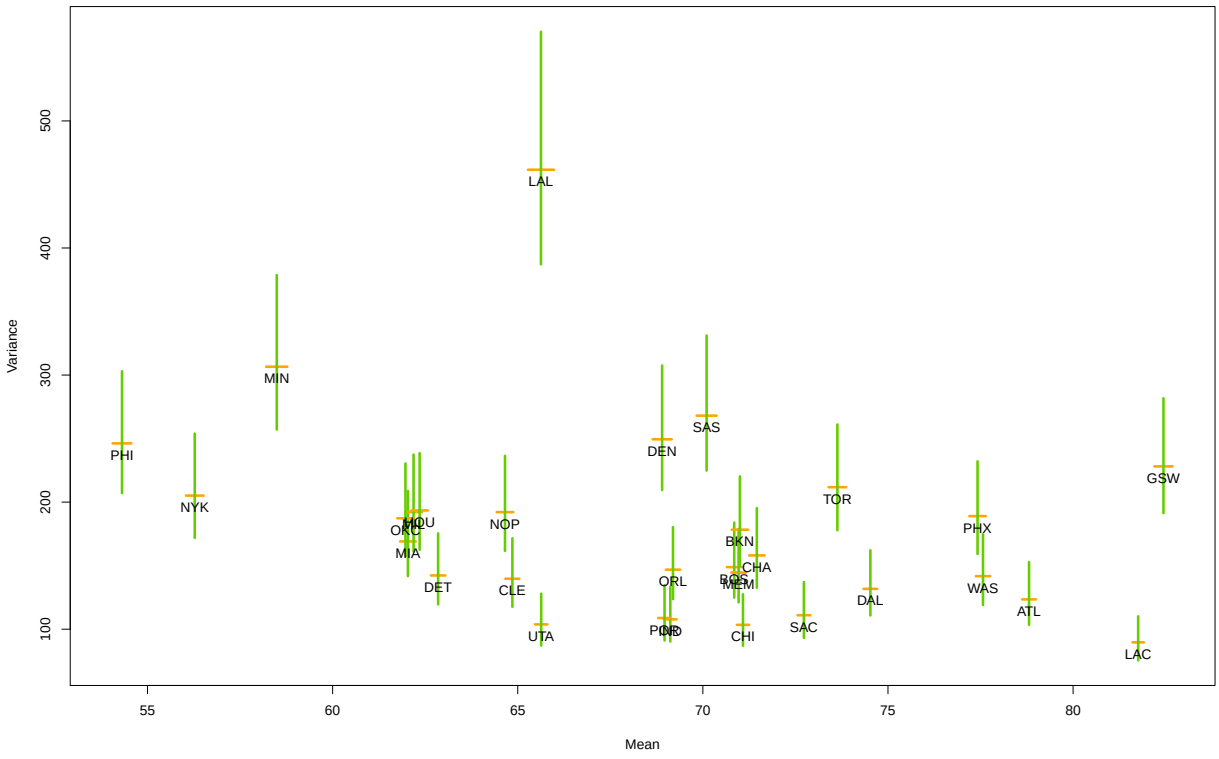

Figure 2: The Figure shows mean and variance of teams per game with confidence intervals. Confidence interval for mean is 0.9 and confidence interval for variance is 0.7

Two homoscedasticity tests were applied to the dataset to find is there any difference between variances of teams. After applied Bartlett and Fligner-Killeen test, it can be seen that the p-values are 2.832030e-13 and 5.507873e-09. This means at significance level=0.05, we can reject the null hypothesis and conclude that there are significant differences of variances among teams.

```
##               Test      p.value
## 1         Bartlett 2.832030e-13
## 2 Fligner-Killeen 5.507873e-09
```

Now, we find that variances of total points of teams per game are different. Then we can check is there any difference in mean points. To do that, a one-way ANOVA test was applied. The p-value of this test is 2e-16 which means there is a significant difference between the mean score of teams per game. In this dataset, each team has between 61-58 games and this means it is balanced and also we have 984 unique games which are relatively large.

Also, we can find lower and higher confidence interval values for the total score of teams both home and away.

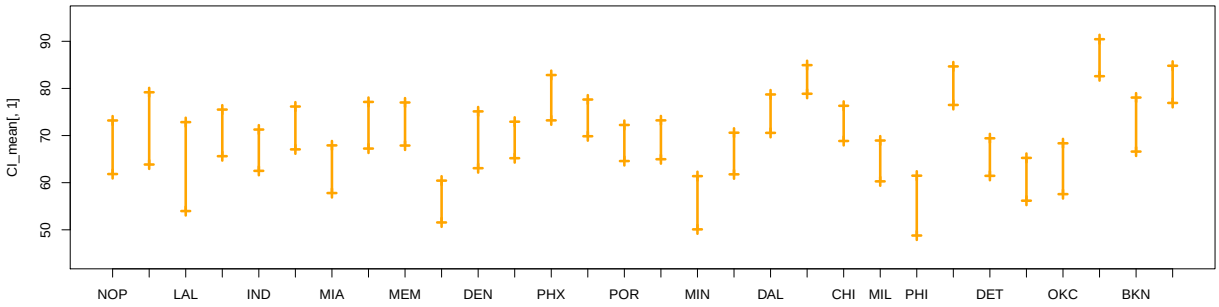

Figure 3: Confidence intervals of totals points of team at home

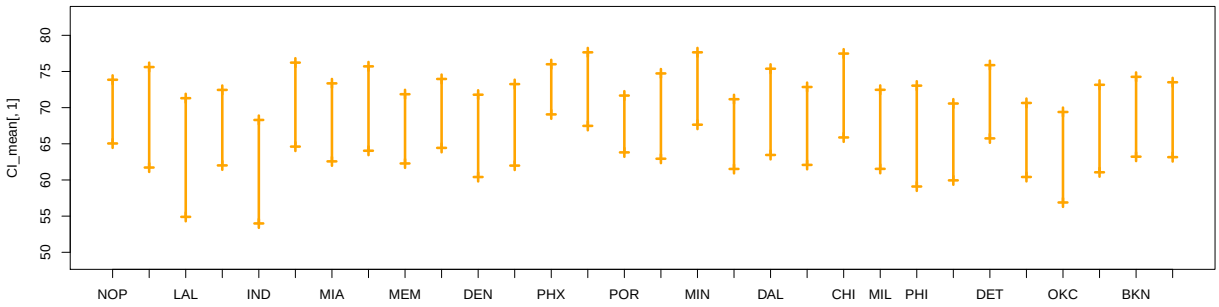

Figure 4: Confidence intervals of totals points of team at away

In Figure 3 and Figure 4, it can be seen that there is a significant difference between teams' home and away

scores. To justify that, Welch Two Sample t-test has been applied and p-value 0.001 which is smaller than 0.05. Therefore, it is justified that, there is a significant difference between home and away points.

2.3 Comparison between Lebron James and Stephen Curry

In the 2014-2015 season, the final game was played between Golden State Warriors and Cleveland Cavaliers. Because of that, we can compare their best players to understand is there any difference between their basketball styles. To do that, at first, important features should be extracted. For instance, average dribbling before shot, shot accuracy, 2 and 3 points, etc. are significant to compare them.

Two Sample t-test and f-test have applied to compare their features. In consider Dribbling, it can be seen that we cannot reject HO and decide any difference between variances in the f test. The p-value is 0.80 which is significantly higher than 0.05. However, in a two-sample t-test, it can be seen that the average dribbling of Lebron James is significantly higher than Stephen Curry. The same situation is also valid for touch time. As a result of that, Lebron James more likely to play with the ball before shooting than Stephen Curry. Also, as a result of the two-sample t-test, we can't conclude that there is a difference between their 2 points accuracy. On the other hand, Stephen Curry has high three-point accuracy. Finally, Curry's average shoot distance is higher than Lebron James.

```
##
##  Two Sample t-test
##
## data:  df_lebron_stats$Dribbling_AVG and df_curry_stats$Dribbling_AVG
## t = 3.3953, df = 107, p-value = 0.9995
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.310518
## sample estimates:
## mean of x mean of y
##  4.305233  3.424912
```

```
##
##  Two Sample t-test
##
## data:  df_lebron_stats$Succ_3_point_rate and df_curry_stats$Succ_3_point_rate
## t = -1.5569, df = 106, p-value = 0.9388
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1365143        Inf
## sample estimates:
## mean of x mean of y
## 0.3381378 0.4042195
```

3 Differences and Bests

3.1 Differences Total Point as periods

In this section, we will analyze total points in periods. At first, we can plot the graph to see total scores better.

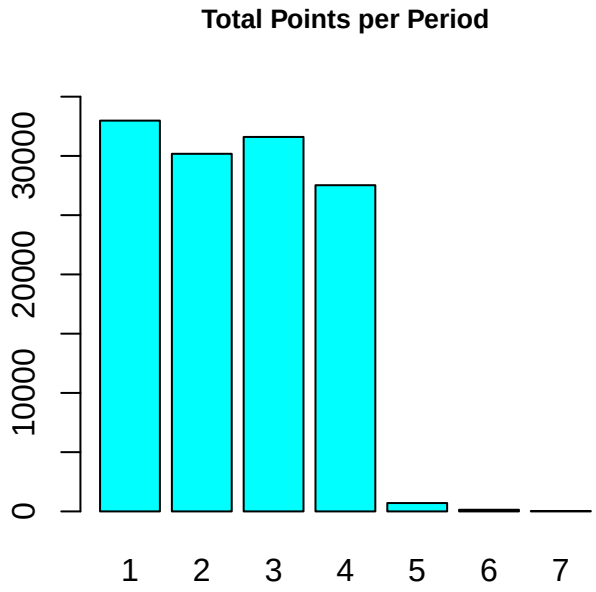

Figure 5: Partial Correlation graph of features of players based on per game

It appears that first four period looks like similar. Other periods are extra time and there is rarely an extension of the game. Therefore it is normal that totals points are lower than the first 4 periods. They understand that are statistically similar or not we make a comparison between them and instead of making it one by one, we will use Tukey Test. In Figure 5, it can be seen that players tend to make more points in periods 1 and 3. This might conclude as, in periods 2 and 4, they are tried. because there is only a long-duration break between period 2 and period 3.

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## PERIOD         6  53721    8953   135.1 <2e-16 ***
## Residuals   3624 240121      66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the Tukey test, a 0.95 confidence interval was used and it can be seen that the total number of points in the first 4 periods are all different. It is said that because, in the comparisons, p values are all almost 0. On the other hand, extension periods means look the same statistically according to the Tukey test.

3.2 Correlations between features of players

In this section, we will try to find both correlation and partial correlation of features of players. To do that, I will use players' stats per game and try to find correlated features.

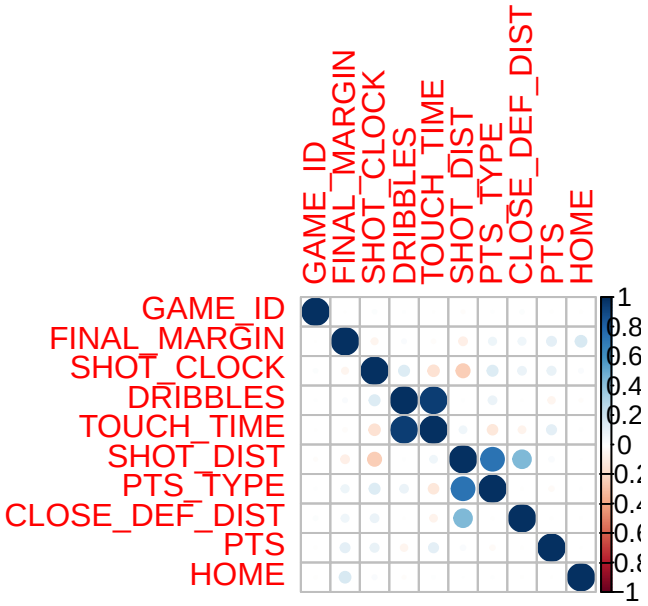

Figure 6: Partial Correlation graph of features of players based on per game

It appears that there are some positive correlations and some negative correlations. if we need to specify;

- There is a positive correlation between HOME and FINAL_MARGIN. This means home teams are more likely to win.
- Points per game slightly positively correlated with FINAL_MARGIN, SHOT_CLOCK, and TOUCH_TIME.
- Shot_Dist and Close_Def_Dist are positively correlated. This means that players prefer to stay away from defensive players when shooting.
- PTS_Type and Shot_Dist positively correlated as expected.
- Shot_Clock and Shot_Dist negatively correlated. This indicated that players use the time to get closer to the basketball hoop.
- Touch_Time is positively correlated with Dribbling. They look strongly correlated.

3.3 Season Best Players

```
## Intersection of Top Scorers and Top 2 point Scorers:  11
```

```
## Intersection of Top Scorers and Top 2 point Scorers:  7
```

```
## Intersection of Top 2 point Scorers and Top 3 point Scorers:  0
```

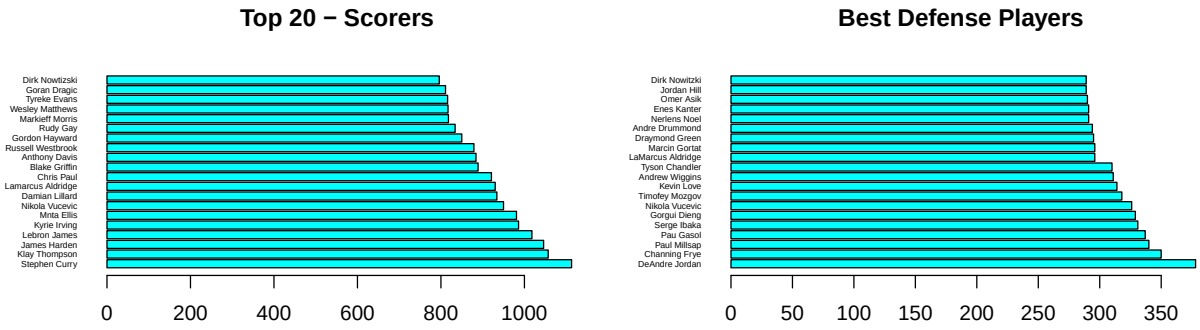

Figure 7: Best Scorer and Defenser Players of Season

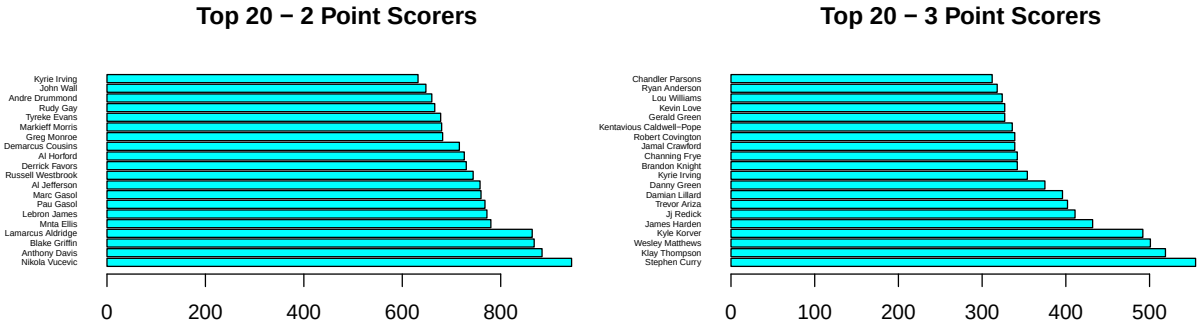

Figure 8: Best Scorer and Defenser Players of Season

According to the intersection of top20 scorers, there are 11 scorers which are both in top20 scorers and 2 point top20 scorers. 7 players are also in top scorers and top 3 points scorers. On the other hand, in top20, there are players that both in them. Besides that, 3 players in Top20 Scorers don't be in top20 2 pointers and top20 3 pointers. This shows that their shooting styles are balanced.

4 Detect Win or Lose

In this section, we will try to guess whether the teams have won or not using team stats. To do this, a new data set containing statistics of the teams per game was created using the data set based on based statistics. This new dataset includes average dribbling, shot distance, touch time, point type, and successful shoot accuracy of players per game, total points per game, also includes whether the team plays at home or away. The new categorical column named "WIN" was created and it takes 1 if the team wins or 0 if the team loses. In the dataset named nbadata, there are 904 unique games and the new dataset also includes all of the unique games but we have 1808 rows because in each game there is one winner and one loser. Therefore, it can be seen that the dataset is balanced with having 50% 0 labels and 50% 1 labels. To detect win or lose, logistic regression which is significantly good to predict binary outputs was used.

```
## Features:  GAME_ID PTS TRY FGM DRIBBLES SHOT_DIST TOUCH_TIME PTS_TYPE WIN LOCATI
```

```
## Number of Games Won:  904
```

```
## Number of Games Lost:  904
```

Feature Selection

Feature selection is very significant to get best results in machine learning models. Multilinearity and unrelevant features can decrease model performance. To prevent that, "stepAIC" function under "MASS" package can be used. It reduce multilinearity and unrelevant features and give values to features based on their relevance.

```
## (Intercept)         PTS         TRY  TOUCH_TIME    LOCATION        TEAM
## -2.15075799  0.11604036 -0.09972387  0.37362865  0.48003535 -0.01604006
```

After feature selection, it can be seen that stepAIC function coefficients can give intuition about important features. Based on these coefficients, if total points and touch time are high, teams more likely to win. Also, location is a very significant feature. As we can see, home teams are more likely to win. On the other hand, surprisingly, there is a negative correlation between a win and a total try. "Team" features are also effective on win or lose. In the coefficient table, it seems negative, however, numerical variables are just used to distinguish them, therefore whether the team values are small or big is meaningless for us.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.773
```

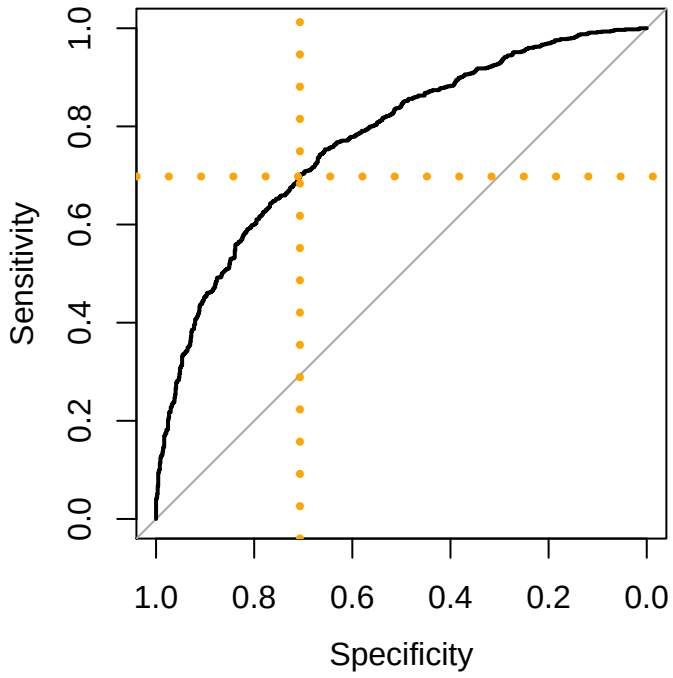

Figure 9: ROC curve for the logistic regression model

Confusion matrix give statistic about true and false labeled. the treshhold determined as 50% percent because this a balanced dataset and if predictor performes higher performance than 50%, it can be said that, it is better than random prediction. In the table below, error rate of both true labels and false labels is very close which is approximately 30%. The logistic model that created can detect win and lose with 70% which is not perfect but better than prediction.

```
## [1] Confusion matrix for logistic model with 50% decision treshold:
```

```
##       FALSE TRUE class.error
## FALSE   639  273   0.2931416
## TRUE    265  631   0.3019912
```

5 Conclusion

In this dataset, we have tried to extract important information from the dataset. To do that, different tests have been applied. We applied regression techniques on the total score based on seconds and find a regression line for both best and mean values per match. Also, to compare teams performances per match f-test, t-test, finger, and bartlett test have applied. Then we compared the two best basketballers of the season and analyze their style using the Shapiro test, var test, and Welch t-test. Also, the seasons of the best basketballers in different areas have been extracted in the dataset. Besides them, to analyze total points in different periods, the Tukey test had applied. In the dataset, at first, there were 128069 records, however, there were some nan

values. These values are especially located in some columns and before using these columns, if na values effect tests, they should have removed from the dataset.

Our main observation on the dataset,

- Players tend to use time completely and make their shot generally last seconds of 24-second duration. Because of that, the Average score increased over 24 seconds time period. (Section 2)
- There is a significant difference between the mean points of teams. On the other hand, according to tests, variances are quite similar. Also, there is a positive correlation between the total win of teams and total scores of teams. (Section 2)
- Total Point on periods is not homogeneous. (Section 3)
- In 20 top scorers, 18 of them based on top 2 pointers or top 3 pointers. Surprisingly, even if the other 2 players not on this top list, they are still on the top scorer list. Therefore, we can say that strong skill on one side is not necessary to be top scorer. (Section 3)
- Winner and Loser can be separated into the dataset. Even if our logistic model doesn't detect winner and loser perfectly. It can be still detected over 70% rate.