Introduction

Analytics utilization is in increasing in the sports industry. As an avid sports fan and interested in this field, this is a perfect opportunity to work on a project with two things that I enjoy, analytics and basketball. Of all sports, I am a huge fan of basketball and chose to explore basketball data for the project. Having a keen interest in basketball and knowledgeable about teams and statistics, it shows the sport has embraced the analytics avenue the most. In addition, basketball has changed the most because of analytics as compared to the other sports. If you look at basketball now and how it was played ten years ago it is a completely different game and the reason is because of the emphasis placed into the utilization of analytics.

DataSet

In searching for an appropriate data set, one was found on Kaggle. The data set chosen was an NBA dataset with Teams statistics from the past four years. This data set was interesting as there wasn't much to clean and it had no null values for any of the columns. The set consisted of Teams and included all thirty NBA (National Basketball Association) teams, the dates for every game from the past four years, win/loss column to note if the team won or lost the game, the opponent on who the team played, the game number played for the year and also a column to note if the team was home or away. These are all the categorical data. The statistics columns includes the following information; TeamPoints for how many points the team scored, field goals made, field goals attempted, the percentage of field goals, 3 pointers and free throws for a specific game night. In addition, they had offensive and defensive rebounds, assists, steals, turnovers, fouls and blocks. The same statistics was utilized for the opponent team.

Statement

There were two questions chosen for this dataset. One, how steals affected other categories on the offensive side of the ball and if steals are indicator to winning or is it just

another statistic like rebounding. The reason for choosing these two questions is in today's NBA games, the offensive side of the ball is most important and desired that the defense has gone by the way side. The reason for its importance was placed into the 3 pointer as shown from all the analytic output that has driven the NBA in the past 10 years. The data set was broken down by only using two seasons, 2015-2016 and 2016-2017. The two seasons produced about 162 games per team and it was found to be a big enough sample to obtain respectable results. Offensive side statistics was used other than steals and the opponent's statistics was not used. Before I get into the data and what was completed, I would like to discuss what the analysis would say before it was run. It is believed that steals would be highly correlated to team points, field goals attempted and three pointers attempted. The reason for this thought process is that when you get a steal you usually are out on a fast break for a lay-up. For Steals being an indicator for winning, it is believed to be a major factor for the reasons mentioned before.

The R Packages utilized were dpylr , dbpyler, plyr, ggplot2 and MLR ' Machine Learner in R. Dpyler was used to select certain columns from the main data set and also to filter through certain items in the matrix.

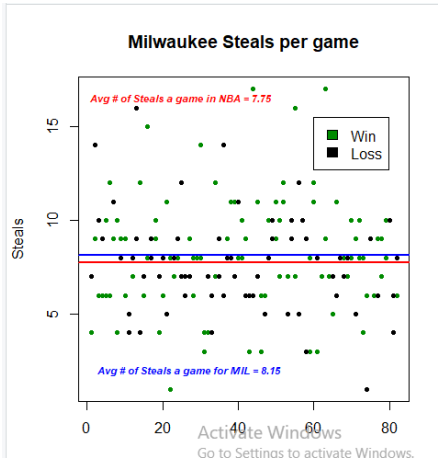
Analyses

The first test used steals the y variable and also a binary variable. A new table was created pulling all the columns and data chosen from the main table and also filtering it down to the correct seasons. Next, steals were made a binary y variable using the ifelse statement and also using the mean to divide it which came to be 7.75 steals per game, making anything greater than 7.75 a one and anything below a zero.

```
mydata <- select(NBA, Season, Game, Date, Team, WinLoss, TeamPoints, Steals, Assists, FreeThrows, FieldGoalsAt
mydata$StealsB <- ifelse(mydata$Steals >= median(mydata$Steals), 1, 0)
mydata <- filter(mydata, Season != '2017-2018')
mydata <- filter(mydata, Season != '2014-2015')
mydata <- createDummyFeatures(mydata, cols = "WinLoss")
```

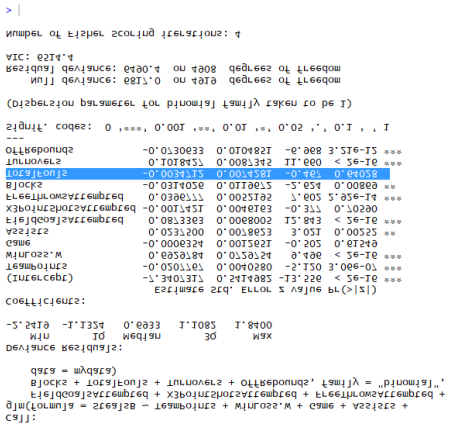
In reporting the actual analysis, it was decided to go with a Logistical regression using the General Liner Model ('glm'). It is believed this was the best way to attack the question around how much an impact does steals have on other statistics in the game of basketball. The first model started with 11 variables as I wasn't sure if this would be too many or too little. I also kept going back and forth on which variables to use, thus it was decided to run just two regressions. In

the first regression, the following were chosen; field goals attempted, assists, turnovers, blocks, 3 points attempted, team points, fouls, free throws attempted, off rebounds, win/loss and game. It was run about 8 times losing a variable along the way until I got down to the final 4. The next step, the coef and predict values were run as well as the odds ratio rounding out with then ROCR graph. For the second regression, the same process was followed except I swapped free throws attempted, field goals attempted, X3points attempted for Field goals made, 3 Points made and Free throw made. For the second analysis I ran some functions plot on 5 random teams to see if the outcome for steals was significant on wins or losses. This analysis is more a visual to demonstrate how teams fall with their median and the NBA's median.



Results

The first run was General Liner Logistic with the attempted free throws, field goals, 3 pointers and the results were very surprising. This is what was received when I ran my first model.



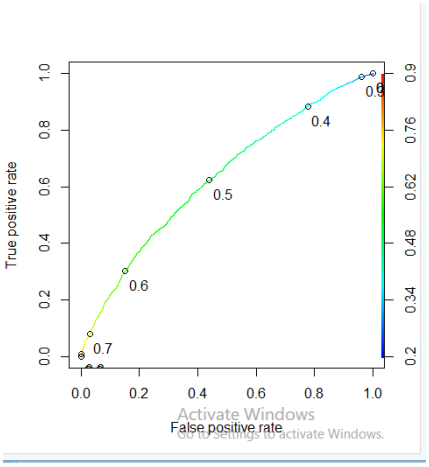
It was very surprising to see the 3 pointer attempted was not correlated with steals whereas free throws attempted and field goals attempted were. Everything else was as expected. From the beginning it looks like steals have an effect on a many other statistics. Another surprise was turnovers, however, it can depend on how you look at it because a steal is forcing a turn over. Afterward I ran 8 Models and was left with turnovers, offensive rebounds, win and field goals attempted and they all had Pvalues significantly close to zero. Below are the coefficients.

```
Number of Fisher Scoring iterations: 4

> exp(coef(Logistic.model8))
       (Intercept)           WinLoss.W FieldGoalsAttempted           Turnovers          OffRebounds
       0.001990692        1.743772304        1.063205700        1.097034490        0.955501268
>
```

As noted for every field goal attempted means it is 6 percent more likely that it would have been a steal. Every turnover that happens you have a 9% percent chance that you will be over the Median of steals. Next is the odd ratio.

```
      FALSE TRUE
FALSE  1450  944
TRUE   1051 1475
```

Sensitivity = 58 percent
Specificity = 60 percent
Below is the ROCR curve it is sort of flat but it does have a Threshold of .5, 0,0 has a threshold of 7.9 8.0 and 1,.9 has a threshold of .6

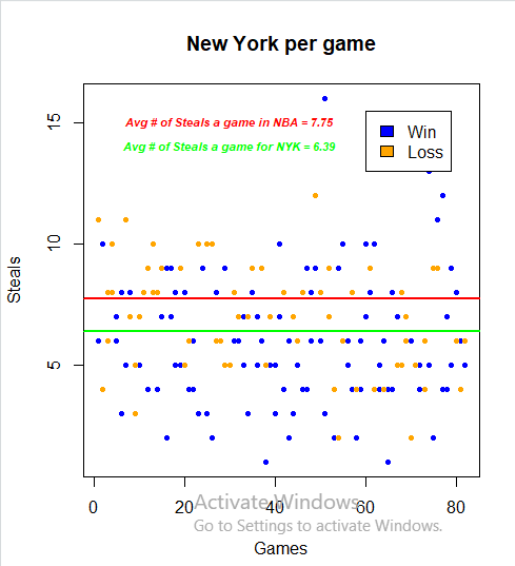


I decide to run another model but instead of using shots attempted, I used shots made such as field goals, three pointers and free throws. Steals had the same effect on most of the

variables but the two that stuck out the most to me was assists and three pointers. Which assists makes sense because assists only count when a basket is made and with three pointers, I find it odd that steals have more of an effect on three pointers then field goals. My only thought for this is that a lot more three pointer are being made than any other type of shot. Below is my last run GLM for my second model

```
Call:
glm(formula = StealsB ~ WinLoss.W + Assists + X3PointShots +
    Turnovers, family = "binomial", data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6880  -1.1675   0.8613   1.1325   1.5804

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.330947   0.171784  -7.748 9.35e-15 ***
WinLoss.W    0.463207   0.062035   7.467 8.21e-14 ***
Assists      0.031240   0.006427   4.860 1.17e-06 ***
X3PointShots -0.039438   0.009180  -4.296 1.74e-05 ***
Turnovers    0.059734   0.007589   7.872 3.50e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6817.0  on 4919  degrees of freedom
Residual deviance: 6667.5  on 4915  degrees of freedom
AIC: 6677.5

Number of Fisher Scoring iterations: 4
```

Last Analysis I ran was a plot chart to show all the steals that happened in the 2 seasons I used. In the chart shows all the games steals and the ones that they won, show the ones they lost as the example below. Also in the graph shows the average steals they had over the two seasons and then the NBA's overall average. For this analysis I used five random teams not knowing there record or overall steals. My results were mix some teams won more games where they had over the mean of steals and some didn't as shown below the Knicks.



Conclusion

In my conclusion I can say that steals do have a major impact in a lot of the offensive parts on the game in the NBA, but to say that steals alone move the Wins or Loss needle, you can't. The evidence does say thought that if steals have major impact on things that can move that needle. Steals did say that your team points are higher when having more steals, but that is more common sense then it is analytical finding. Overall I really enjoyed this project I learned a lot about R and force me to find ways to build on what I learned in Class. If there is a next step, I would say it would be collecting game to game data or second to second data to really see how steals impact from a on the court view. The pitfalls I found in the data were that a lot of it was predicting on a base number. The biases in the data is that you really don't know what the end result of a steal was as in there could have been a game with 11 steals but all of them lead to turnovers themselves.

Final NBA sets.R