Final Results

Curtis Higa

February 6, 2018

Introduction

In 1979, the National Basketball Association introduced a new component to their game that would take how points are scored away from the rim, the three point shot. Since then we've seen a plethora of players score a majority of their points from behind the arc like Reggie Miller and Ray Allen. But does this shift change the way the game is played? Do you have a better chance of winning a game if you outscore your opponent in three-point shots? What about for the season? With teams like the Golden State Warriors and Houston Rockets, it may seem that your teams ability to convert three-point shots would equate to more wins but to verify this, data was taken from the 2016-2017 NBA regular season.

Data Wrangling

This data was taken off a website called MySportsFeeds where they house countless sports statistics for baseball, hockey, football, and basketball. From their data set for the 2016-2017 NBA season, I've removed columns such as rebounds and assists along with adding columns for three point percentage, total points, and percentage of points contributed by three-pointers.

The total points column was particularly important to identifying who was the winner of each game and indicating it within the data set. It was done by mutating a column that took the highest score for the game and compared it to the total points of each team. Depending if the total points equaled the new column, a "1" or "0" was added to signify if the team won with the following code.

```
# Find max points for each game
with_win_col = with_opp_tpp %>%
  group_by(game.id) %>%
  mutate(win_c = max(total_pts.points))
# add 1 for max row
with_win_col$win_team = ifelse(with_win_col$total_pts.points == with_win_col$win_c,1,0)
# drop win_c column
with_win_col$win_c = NULL
```

Once those statistics were added to the data set, I appended columns that coincided with their opponents statistics for each of their games and the differences between their statistics to their opponents.

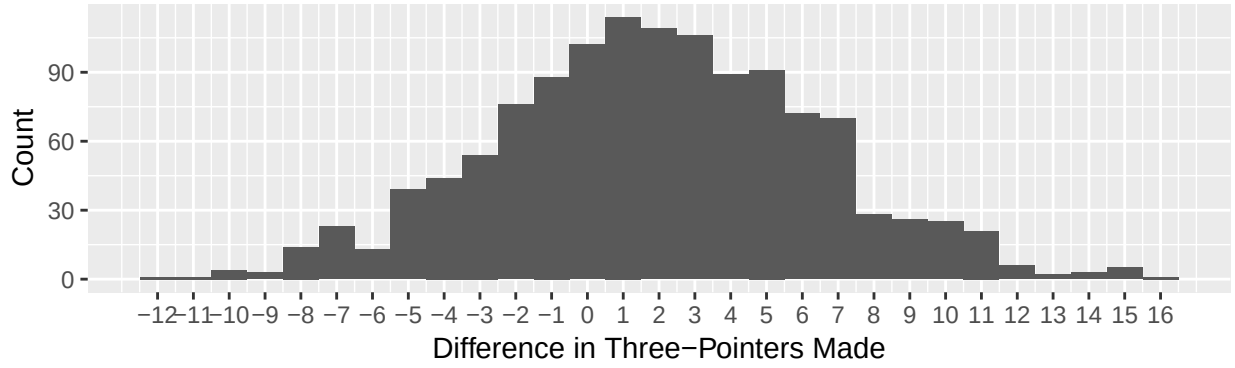
Statistical Analysis

Over the 1,230 games played throughout the season, the difference between numerous three-point statistics of the winning and losing teams were plotted to visualize any patterns. They were as follows:

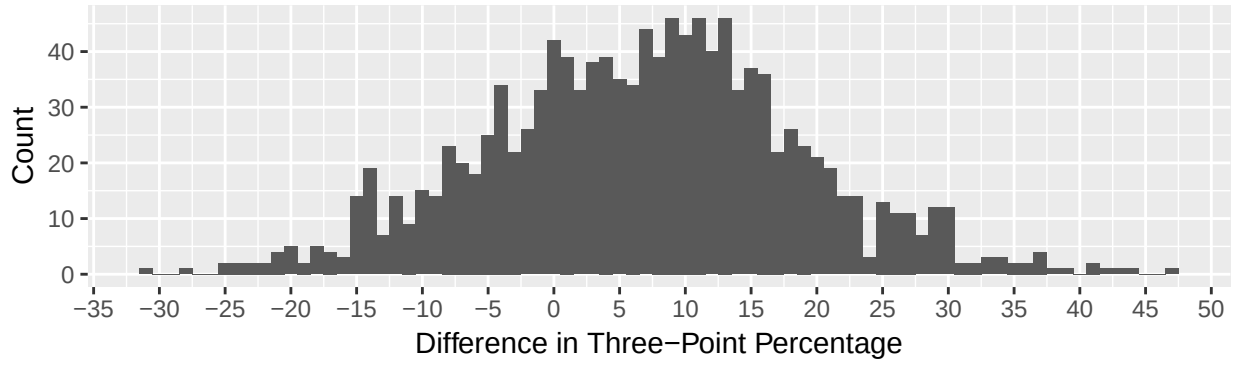
```
game_filter <- complete_gamelogs %>% filter(win_team == 1)
geom_hist <- geom_histogram(aes(y = ..count..), position = position_dodge(), binwidth = 1)

game_filter %>%
  ggplot(aes(x = diff_tpm)) +
  geom_hist +
```

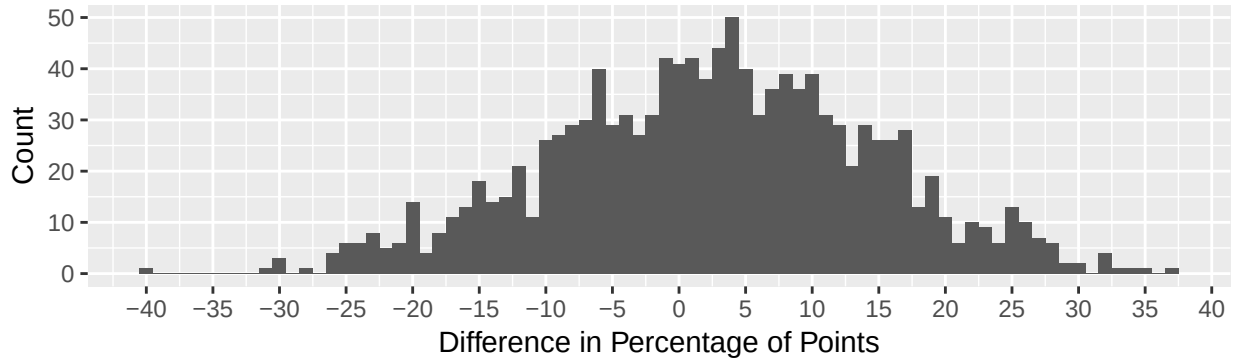
```
  scale_x_continuous(breaks = round(seq(-12,16, by = 1),1)) +
  labs(x = "Difference in Three-Pointers Made", y = "Count")
```




```
game_filter %>%
  ggplot(aes(x = diff_tpp)) +
  geom_hist +
  scale_x_continuous(breaks = round(seq(-70,70, by = 5),1)) +
  labs(x = "Difference in Three-Point Percentage", y = "Count")
```




```
game_filter %>%
  ggplot(aes(x = diff_pop)) +
  geom_hist +
  scale_x_continuous(breaks = round(seq(-40,40, by = 5),1)) +
  labs(x = "Difference in Percentage of Points", y = "Count")
```



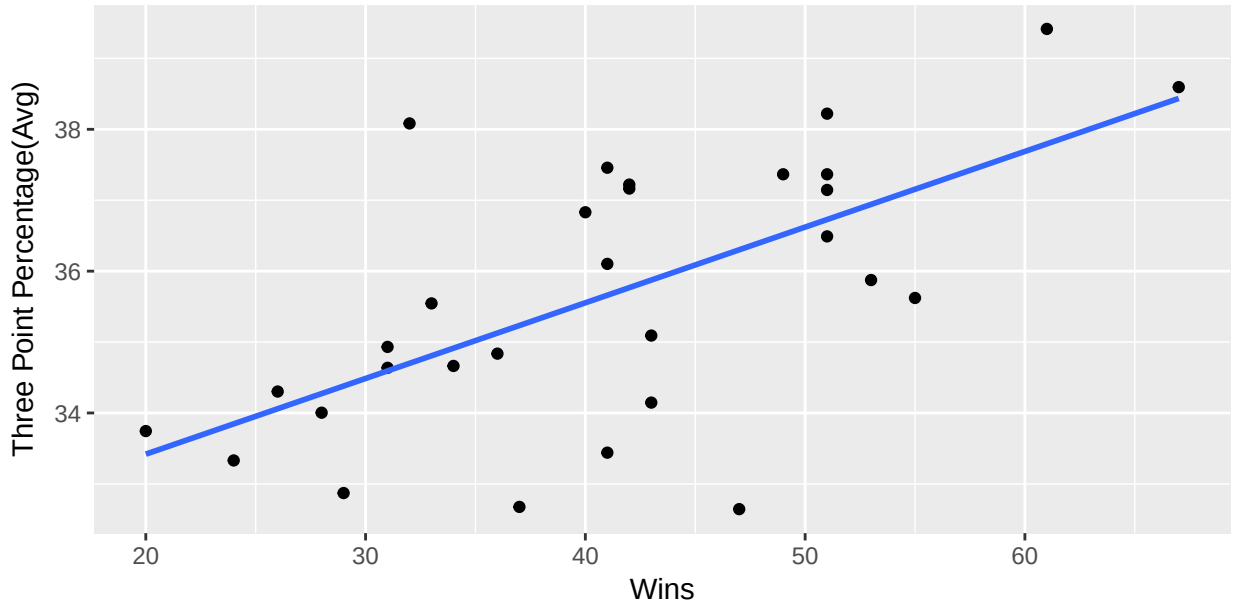
As the histograms illustrate, the winning team typically out performs their opponents in all the three-point

statistics studied. On average, the winning team makes 2 more three's, shoots threes better by around 7%, and has roughly 2.5% more of their points coming from threes. But looking at the statistics in the table below, there's no statistical significance to out-producing your opponents in the three-point statistics looked at in this study with 95% confidence.

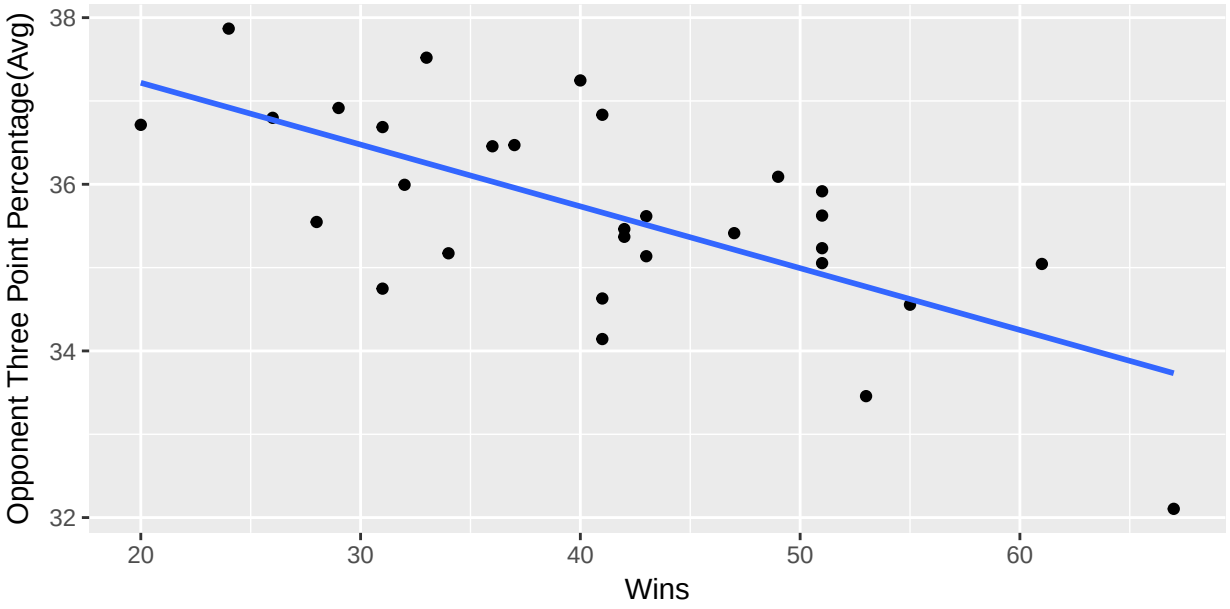
diff_tpm_mean	diff_tpp_mean	diff_pop_mean	diff_tpm_sd	diff_tpp_sd	diff_pop_sd
1.902	6.809	2.476	4.496	11.96	12.09

However, if a team is able to excel at a couple three-point statistics, it could correlate to an above average winning percentage throughout the season. Splitting up the same data of the 2016-2017 NBA regular season into teams, a different pattern emerges when we plot these statistics with respect to the number of wins/win percentage. Using the *corrplot* in the *corrplot* library, we're able to visually see the correlation between different statistics.

```
team_statistics %>% ggplot(aes(x = win_team_sum, y = three_pt_Percentage_mean)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Wins", y = "Three Point Percentage(Avg)")
```

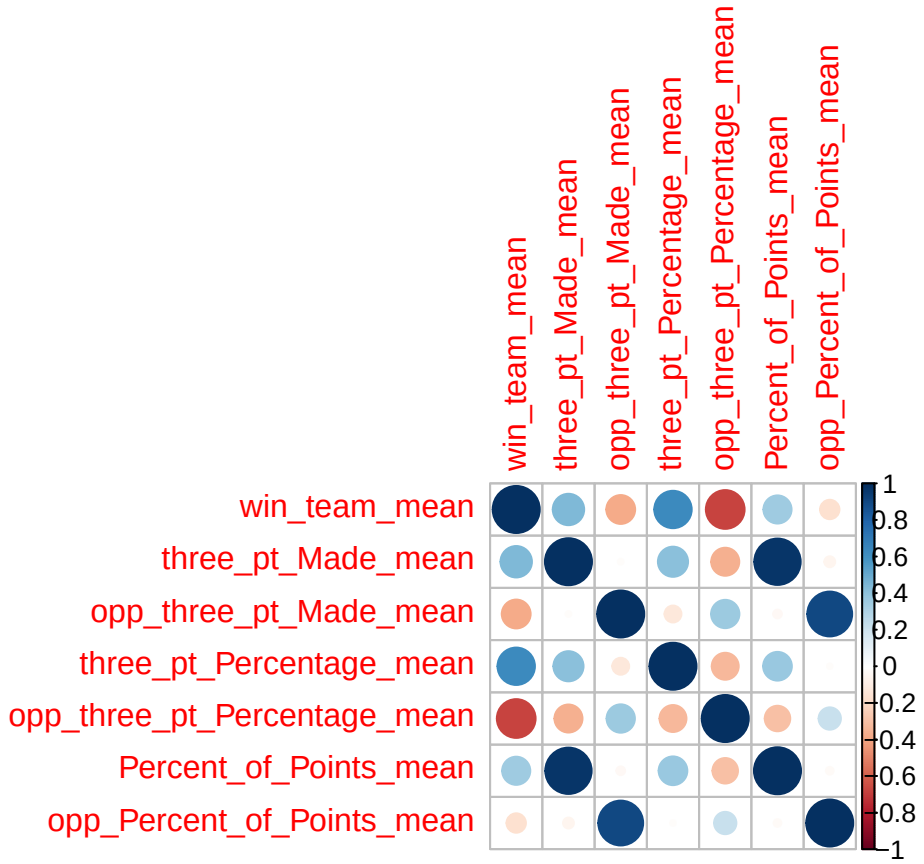



```
team_statistics %>% ggplot(aes(x = win_team_sum, y = opp_three_pt_Percentage_mean)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Wins", y = "Opponent Three Point Percentage(Avg)")
```




```
cor <- cor(team_statistics[,c("win_team_mean", "three_pt_Made_mean",
                              "opp_three_pt_Made_mean", "three_pt_Percentage_mean",
                              "opp_three_pt_Percentage_mean", "Percent_of_Points_mean",
                              "opp_Percent_of_Points_mean")])

corrplot(cor, method="circle")
```



As shown in these graphs, there seems to be a positive correlation to the number of wins and a teams three-point percentage in addition to being a negative correlation to their opponents three point percentage. The last figure confirms that of the statistics, three point percentage and opponents three point percentage hold the strongest linear relationship to your teams win percentage by 0.632 and -0.678 respectively. The next step was to develop a model to predict a teams winning percentage based on their overall three point percentage and opponents three points percentage throughout the entire season.

Machine Learning

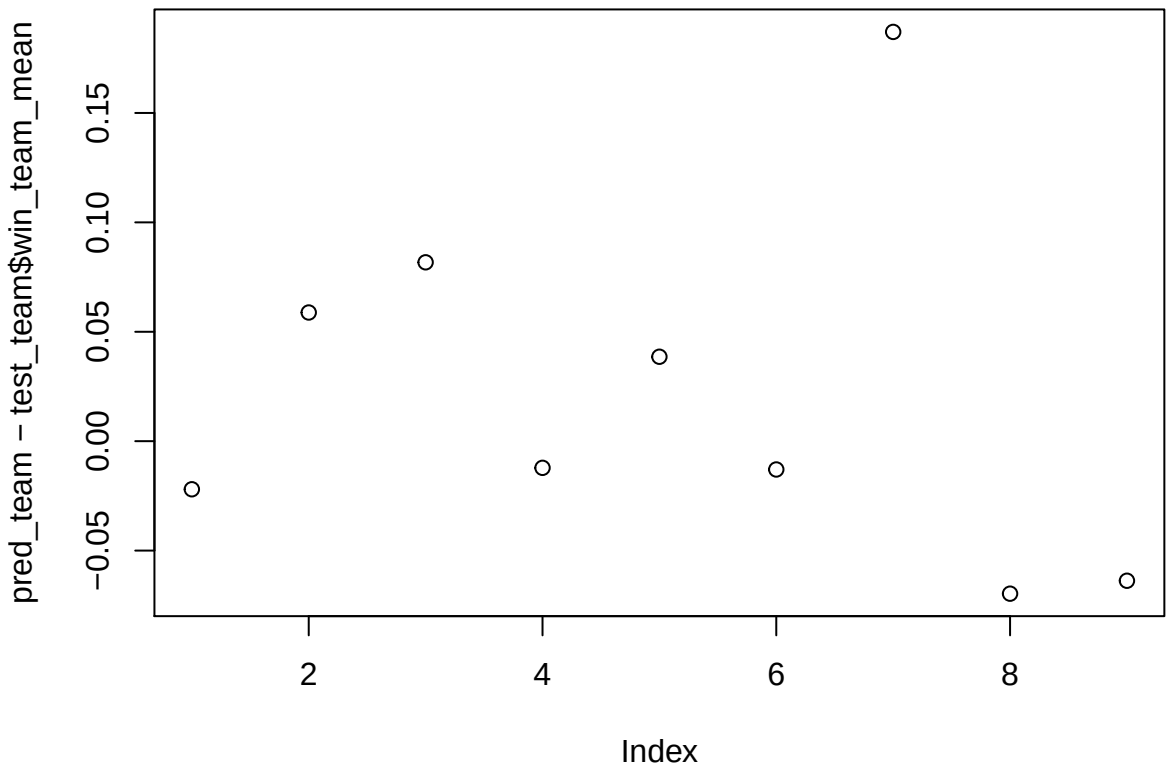
When trying to predict each teams winning percentage, a linear regression model was used using the independent variables *three_pt_Percentage_mean* and *opp_three_pt_Percentage_mean* which are the three point percentage and allowed three point percentage respectively for the team over the course of the season. The model is shown below as *reg_mode*.

```
reg_mode <- lm(win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
               data = team_statistics)
summary(reg_mode)
```

```
##
## Call:
## lm(formula = win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
##     data = team_statistics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17033 -0.05472  0.01265  0.06035  0.15848
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.415593   0.653277   2.167 0.039232 *
## three_pt_Percentage_mean      0.033095   0.008733   3.790 0.000770 ***
## opp_three_pt_Percentage_mean -0.058770   0.013488  -4.357 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08401 on 27 degrees of freedom
## Multiple R-squared:  0.647,  Adjusted R-squared:  0.6209
## F-statistic: 24.75 on 2 and 27 DF,  p-value: 7.848e-07
```

As shown in the summary, the R^2 term is *0.647* which indicates it's a fairly decent model and both independent variables are indicated with "***" signifying they are significant. After splitting the data set into a training and test set 70-30 respectively and creating a new model using the training data, the following prediction model was generated.

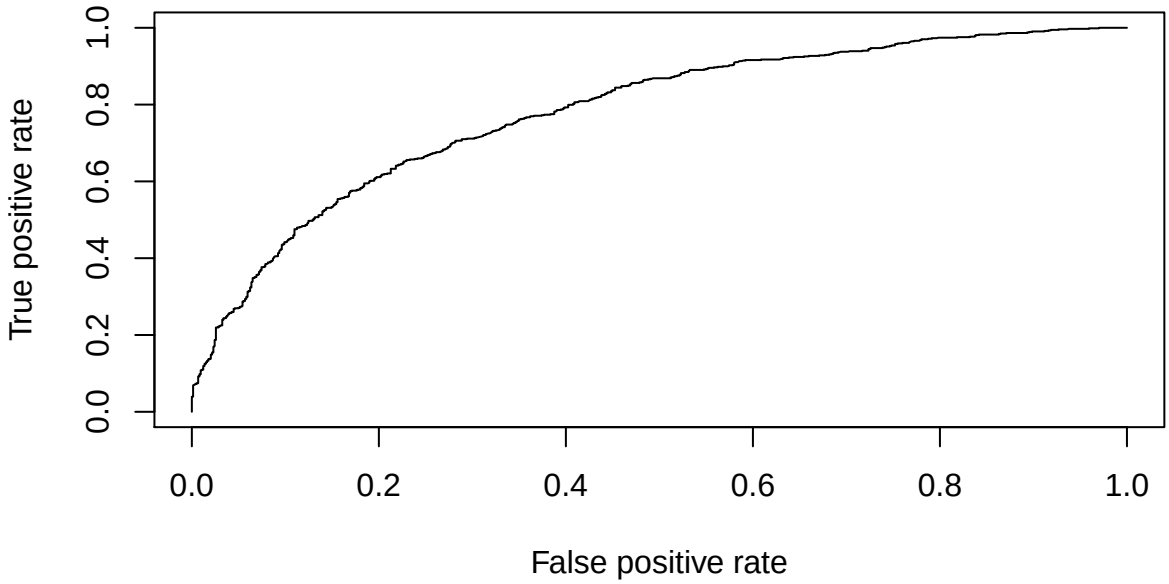
```
mod_team <- lm(win_team_mean ~ three_pt_Percentage_mean + opp_three_pt_Percentage_mean,
               data = train_team)
pred_team <- predict(mod_team, type = "response", newdata = test_team)
plot(pred_team - test_team$win_team_mean)
```



The above plot shows how the residuals are distributed among the test set and it shows the residuals are randomly plotted. This means the model produces no bias towards one result and is therefore valid. However, the performance of the model may not be great due to a couple reasons. One of which may be because of the sample size. The team statistics are for all the teams that played in the 2016-2017 NBA regular season which is 30 teams. Another reason may be because there are other factors of basketball that may contribute more to a higher winning percentage than just focusing on three point shots.

Models predicting game-by-game outcomes were also investigated and considered as a classification problem due to there being exactly one winner and one loser per game. On a game-by-game basis, we're trying to build a model to predict a winner based on the various final three point statistics and the two three point percentage statistics were most relevant resulting in the following logistic regression model, where *train* is the training data.

```
mod1 <- glm(win_team ~ three_pt_Percentage + opp_three_pt_Percentage,
            data = train,
            family = "binomial")
ROCRperf <- performance(ROCRpred, "tpr","fpr")
plot(ROCRperf)
```




```
table(test$win_team, predictTest >= 0.5)
```

```
##
##     FALSE TRUE
##   0   375  117
##   1   138  354
```

With a threshold value of 0.5, this model correctly predicted 74% of the test data indicating it's a good model shown in the confusion matrix above. However, this model takes into account middle and end game statistics to determine a winnner which is good for which team has a high probability of winning at a specific moment of the game. It doesn't include predetermining factors similar to venue or active players to predict a winner before the game occurs. If we want to be able to predict the winner of a game before the game occurs, we'd need to take those factors as independent variables instead.

Summary

With that said, of the three-point statistics analyzed here, being able to defend and shoot behind the three-point line efficiently is a key factor if you want to win basketball games. However, this doesn't take into consideration your offensive and defensive philosophies and isn't the only factor that matters in basketball. On offense, being able to run a system of plays that get players open looks and easy shots will help your efficiency and outscore your opponents. The opposite is true for defense. Being able to disrupt and guard shots effectively will lower your opponents efficiency so they can't score as much. This doesn't just go for three point offense and defense but to all of basketball, which is why it makes sense that overall offensive and defensive efficiency factors into how often you win over the course of an entire season.

In the future, more factors and data entries could be added to improve this model. Other basketball statistics could be added like turnovers, assists, and rebounds. These are all factors that represent how effciently your offense runs and how disruptive your defense is. You can even account for individual player statistics to

determine your teams performance and which lineups work best. More seasons could also be added that would add more team statistics and could improve the model.

The moral of this study is that efficient offense and defense from the three point line could help in winning more games. Coaches could utilize a scheme to best utilize their men to optimize their efficiency. In recent years, we've seen teams evolve their lineups to be more agile and lengthy to some success. Organizations could also work with the coaches and use this information to develop/sign more players who excel at both aspects of the game, known as 3-and-D specialists. It should be noted, that the model used wasn't the best at predicting a teams winning percentage, so an organization shouldn't invest all it's resources to specialize in one aspect of the game. Coaches will still need to mix up their plays and defensive schemes to keep their opposition guessing as to what to expect. Players will also need to keep improving and changing up their game to become more unpredictable.