Final Paper

STOR 320.02 Group 3

December 04, 2018

Introduction

The Golden State Warriors have broken the NBA. Their offense is so efficient, that they look like they are playing a different game. The Rockets went full blown copycat, assembling a splash brothers duo of their own in Harden and Paul. They even took the Warriors pace and space strategy to newer heights, and nearly succeeded in defeating them in the Western Conference Finals last year. The strategy clearly works when you have personnel like the Rockets or Warriors. But for everyone else who are less fortunate, how should they optimize their game plan: should they run up the court, jack up a three or a layup, and run back? In order to answer the question, "how are the Warriors and Rockets so good?", the modern NBA begs two follow-up questions: Are they choosing more efficient shots than everyone else or are they making a higher percentage of the same shots the rest of the NBA takes?

Data

There were multiple data sets for this project. One of the data sets was from nba savant (http://nbasavant.com/), developed by Daren Willman. The source of this data was stats.nba.com and ESPN. The data ranged from the 2011-12 season to the 2017-18 season. Each row of the data set was a shot in a game with many variables such as the player, team, period, etc. The main variable analyzed in this project was the type of shot. This variable,

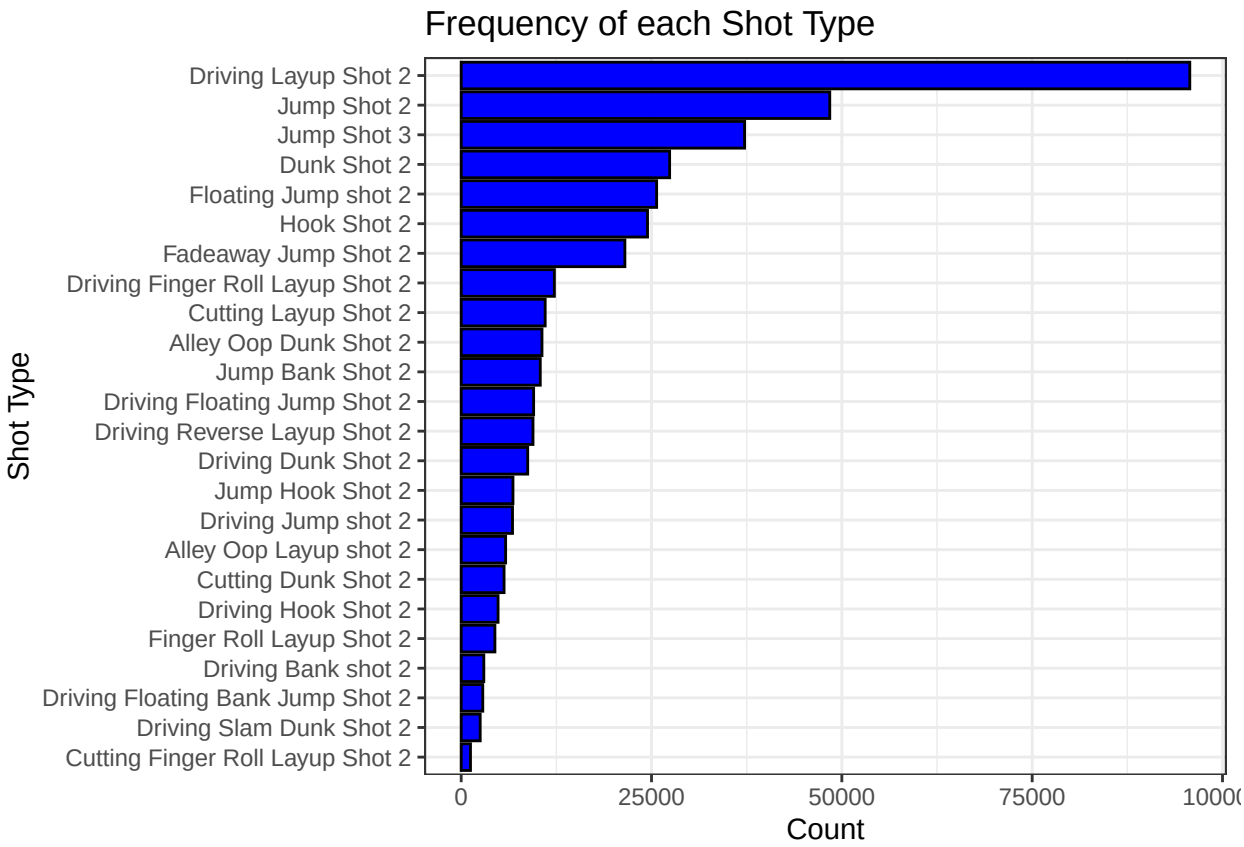
named action_type, had about 25 different factors such as “alley oop dunk shot”, “driving jump shot”, “cutting dunk shot.” This project explored how the amount of shots teams took and allowed of each affected their overall performance.

The other data set used in this analysis came from stats.nba.com with the same season ranges. Each row in this data set corresponded to a team and a season with variables such as offensive rating, pace of play, and net rating (overall rating). The data summarized the team’s performance over a season. By combining these two data sets, the composition of a team’s shots could be compared to their overall performance.

The first data set was manipulated so each row was a team and the columns indicated the percentages the team took and allowed of each type of shot. The data was then merged with the nba.com summary statistics. A snippet of the data is posted below, only the first 4 columns. The remaining columns are about 100 additional shot types and some summary statistics from the stats.nba.com data set.

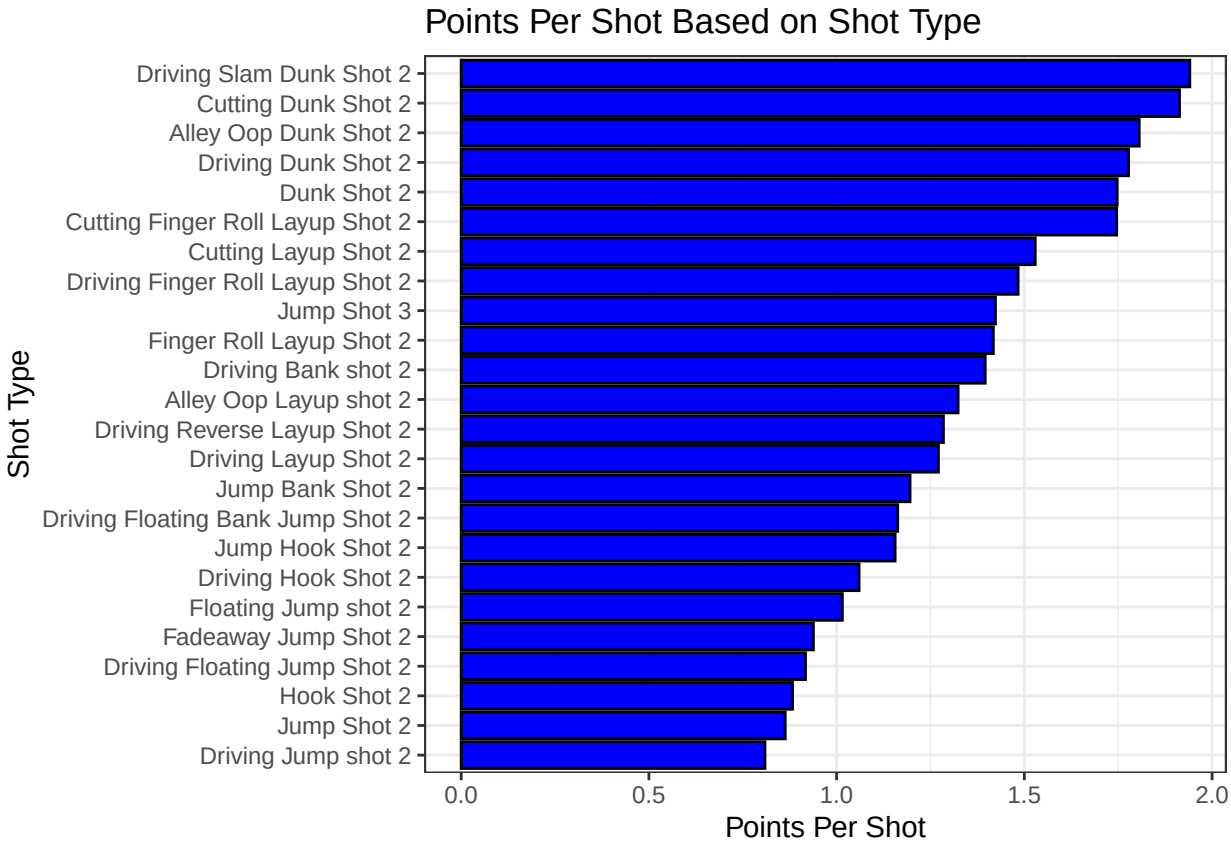
team	year	AlleyOopDunkShot2PTFieldGoaldef	AlleyOopLayupshot2PTFieldGoaldef
Atlanta Hawks	11	1.585728	1.3875124
Atlanta Hawks	12	1.330495	0.6386376
Atlanta Hawks	13	2.232369	1.0147133
Atlanta Hawks	14	2.461539	0.5641026
Atlanta Hawks	15	1.948424	1.3753582
Atlanta Hawks	16	1.962923	1.4721919

Since the primary variable of interest was the type of shot, the league-wide frequency of each shot type was plotted in the figure below to summarize their popularity.



Results

One would expect that teams that take the most efficient shots and force their opponents to take the least efficient shots would have a higher net rating. Here is a bar graph of all shots that occurred more than 1000 times over the past 8 years, with their efficiency on the x axis.



The table shows dunks and layups were, on average, the most efficient types of shots followed by three pointers. The hook shots and mid range jump shots were the least efficient. Therefore, teams that take more dunks and layups on offense, and force hook shots and mid range jump shots on defense, should have a higher rating.

The following models contain step wise regressions, neural networks, elastic nets, Bayes methods, and support vector machine models. These models analyzed a team's offensive rating and net (overall) rating. Future figures and analysis only analyzed offensive ratings as net rating prediction was very similar.

More specifically, the components of these models were the percentages that each team took of that shot type, not made. Only shot types in which the most that a team ever took of that shot type was 10% of their shots in a given season. This choice was arbitrary to help with the chaos that the step wise regression caused due to shots such as "Alley Oop

Dunk shot 3". Additionally, the most efficient shots were combined into a variable called "goodshots" and the least efficient shots into a variable "badshots."

K Fold Cross Validated Stepwise Regression

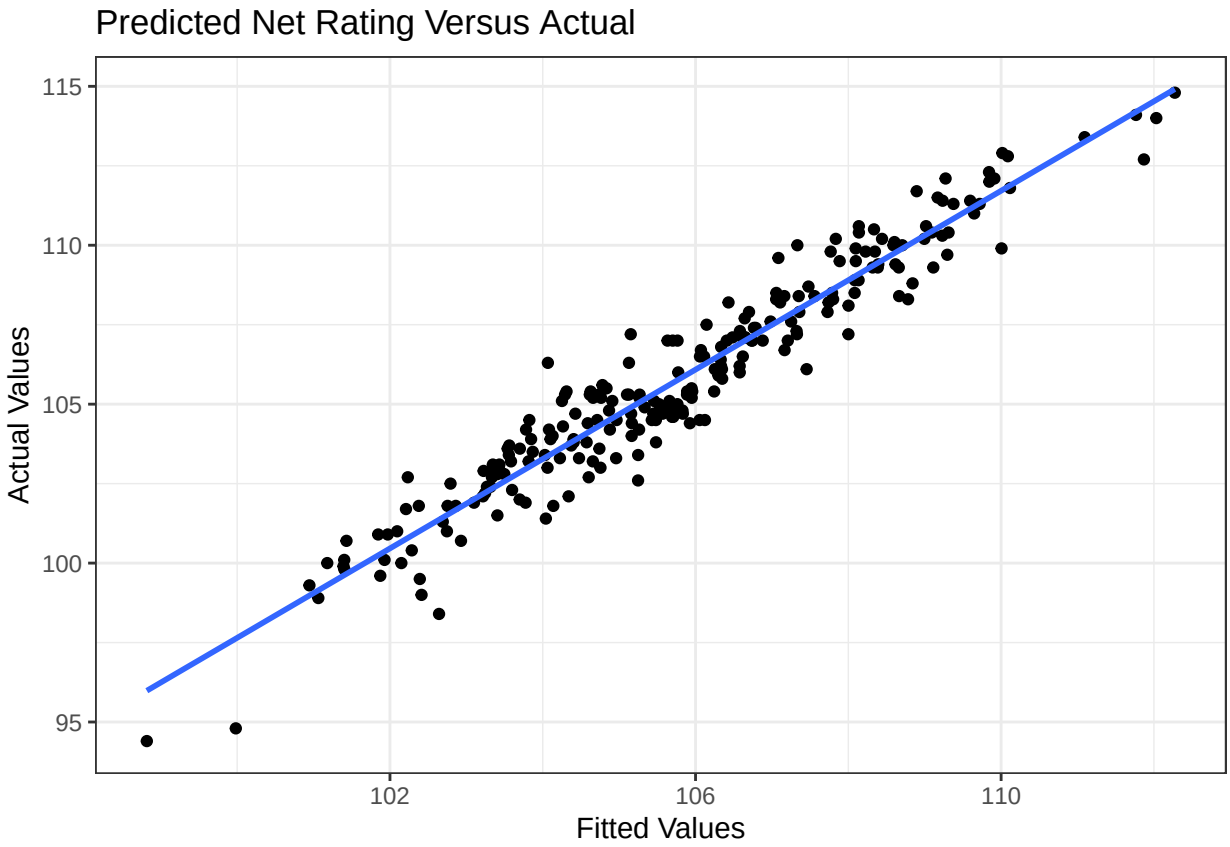
The terms calculated from the bootstrapped step wise regression were used in the base formula for future models. I will print the formula below.

```
## OFFRTG ~ CuttingLayupShot2PTFieldGoal + DrivingFingerRollLayupShot2PTFieldGoal +
##     DrivingFloatingJumpShot2PTFieldGoal + DunkShot2PTFieldGoal +
##     FadeawayJumpShot2PTFieldGoal + FloatingJumpshot2PTFieldGoal +
##     HookShot2PTFieldGoal + JumpShot2PTFieldGoal + cuttingshot +
##     drivingshot + threes + layups + banks + oops + finger_roll +
##     badshots + goodshots
```

Various Bootstrapped Models

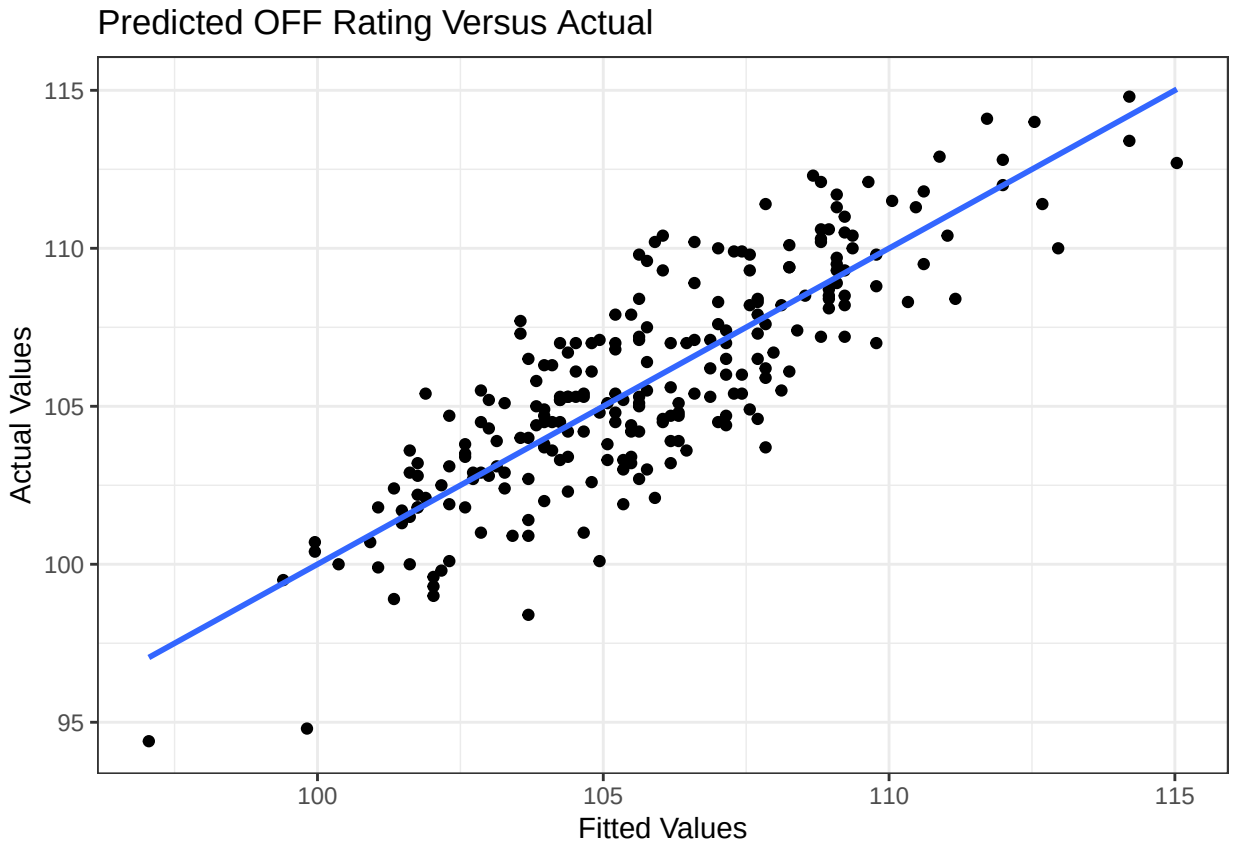
The initial offensive rating predictions made from cross validated step wise regression were less than ideal. While the points did ostensibly follow a linear trend, cross correlation was lower than desired at around .65 with an RMSE around 2.2, which on a 15 unit range, was quite large. The positioning of some of the points in figures also provided evidence that the fit was not ideal as high ratings were consistently underestimated, and low ratings were frequently overestimated. Many other models were tested including support vector machines and Bayes models; however, the results were more or less the same. Our most successful method was using the random forest algorithm. These results will be discussed below.

Bootstrapped Random Forest



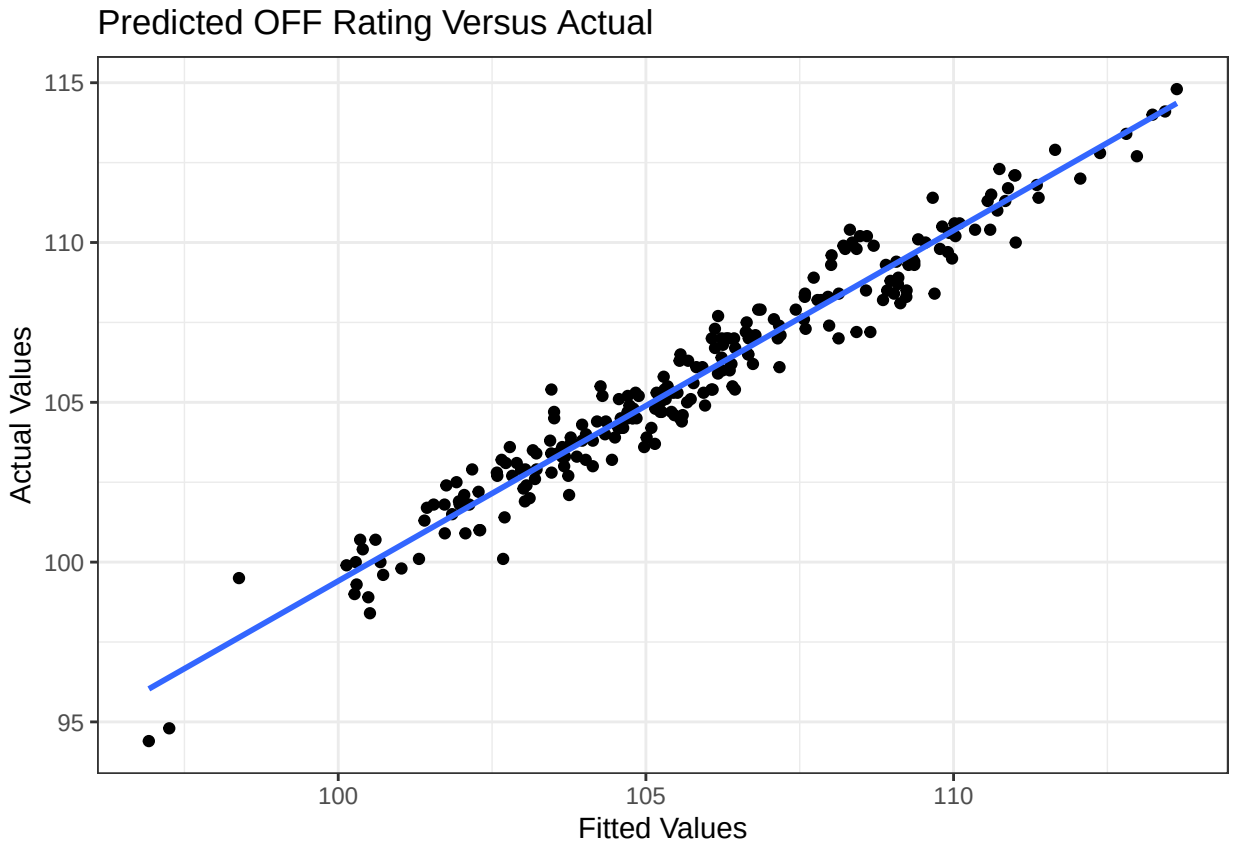
The random forest method produced offensive rating predictions that were much better than those in previous models. Cross correlation improved to a much more acceptable value of around 0.98 with an RMSE between .9 and 1.1. Both underestimation at high ratings and overestimation at low ratings also appeared to be corrected. Therefore this is the best model to compare to the only effective field goal percentage model (EFG%).

Only Effective Field Goal Percentage (EFG%)



Unexpectedly, the random forest model performed better than the EFG%! Offensive rating predictions based only on field goal percentage yielded a lower cross correlation of around 0.85 while also increasing RMSE to about 1.5.

Using Shot Selection Variables and EFG%



Offensive rating predictions made from the addition of EFG% improved the random forest model by retaining the high cross correlation of about 0.98 while also reducing RMSE further to about .63. Overall, the model produced a figure that closely followed the perfect prediction line with few obvious deviations, even at extremes. The two outliers in the bottom left are the 2012 Bobcats, and the 2015 Sixers, who were some of the worst teams in NBA history. Here is a table of the final results for our models.

	Cross correlation	RMSE
K Fold Cross Validated Stepwise Regression	0.6481315	2.1866038
Bootstrapped Regression	0.6481315	2.1866038
Bootstrapped SVM Radial	0.6771317	2.0325304
Bootstrapped Bayes glm	0.6481272	2.1861042
Bootstrapped Random Forest	0.9636609	1.0736776
Bootrstrapped Linear Regression EFG%	0.8458610	1.5082789
Bootstrapped Random Forest: Shot Selection and EFG%	0.9767854	0.6282342

Conclusion

To answer the questions "is shot selection a better predictor of offensive/net rating?" or "is shooting percentage a better predictor of offensive/net rating?", multiple regression methods were implemented to try to predict ratings from these variables. As shown by this analysis, knowing what shots you take indicates a better offensive team shown by the random forest model. In fact, this proved more reliable than purely basing prediction on field goal percentage as seen in the comparisons of cross correlation and RMSE for both models. However, the best option was to combine these two sets of information to further reduce RMSE in the model. It would probably help the model to have information such as how far away is the defender, how fast the shooter is moving, etc. The initial models without EFG% had no way of knowing whether most of a team's threes were Stephen Curry shooting a wide open standing three from the corner with no defender within ten feet, or Josh Smith shooting a contested pull up three pointer with a defender right in his face, yet the generated predictions were still accurate. Once the EFG% was added, the RMSE of around 0.63 was fairly reliable. In the future, using this model (with defensive shot variables) to predict net rating or whether a team can win a championship should be explored.

This analysis certainly does not conclude that your team can simply change strategy and generate better shots on offense. You can have the greatest game plan in the world, know what shots you need to force the opposing team to take, and which shots your players should

seek out; however, if you do not have the ability to create these good shots, knock down these shots, and force poor shots on defense, the odds are not in your favor.

This next draft has some intriguing players, the Duke trio of Reddish, Barrett, and Williamson, and the UNC duo of Little and White. All of them seem to have the potential to get to the cup and convert, create and knock down the three, and generate these shots for their teammates. But who will be able at the next level, regardless of their teammates? Who will drive your teams EFG% up and generate the shots this model sees as efficient? Now that is a question worth exploring down every avenue possible. Whether you have to get shot physics involved or do a comprehensive psychological test, being able to increase your odds of being right on your selection in this next draft can chart the course of your franchise. Ask the Knicks if they are happy they drafted Johnny Flynn over Stephen Curry.