Predicting the Draft and Career Success of NBA Players from College Statistics and Draft Timing

Mingxian Gao

A PRACTICUM

In

Data Science

Presented to the Faculties of the University of Pennsylvania in Fulfillment of the Requirements for the Degree of Master of Science in Engineering

Acknowledgement

I would like to express my great appreciation to my advisor Professor Shane Jenson for his valuable suggestions and patient guidance throughout this independent study.

Abstract

National Basketball Association teams are spending an increasing amount of time on scouting draft prospects from college as the rising NBA salary gap helps players earn lucrative contracts while pushes team to spend more. As the stakes of NBA draft are getting higher, teams have complex drafting strategies based on college statistics and positional fits that are intended to predict career success of players in the NBA. In this paper, we focus on the timing when college players declare the draft along with their college statistics. We create prediction models for career success, draft rank and draft success with different predictors and outcome variables by conducting multiple linear regression, random forest and stochastic gradient boosting. With these modeling approaches, we find that the variables that are most predictive of career success in NBA are not necessarily predictive of draft rank. The finding suggest that we can predict career success and the ideal draft rank with Win Shares per 48 minutes and use classifiers to measure the draft success of players.

Key Words: NBA draft; basketball; draft prediction; career success; first round draft; rank

Contents

1 Introduction **1**
1.1 Project Goal 2

2 Background **4**
2.1 Draft History. 4
2.2 Draft Timing 4
2.3 Draft Rank and Career Success 6
2.4 Draft Predictions 7

3 Literature Review **8**

4 Data Retrieval, Creation and Analysis **9**
4.1 Data 9
4.1.1 Source. 9
4.1.2 Data Description 9
4.2 Exploration Data Analysis 12
4.2.1 Univariate Analysis 12
4.2.2 Bivariate Analysis 14

5 Modeling on Career Success and Draft with Regression **17**
5.1 Modeling on Career Success 17
5.2 Modeling on Draft Rank 21
5.3 Model Predictions 22
5.3.1 Model Comparison on Predicting Career Success 23
5.3.2 Model Comparion on Ideal Draft Rank 25

6 Classifying Draft Success **27**
6.1 Building First Round Draft Classifier with Random Forest 27
6.2 Predictions on First Round Draft 31
6.3 Stochastic Gradient Boosting 33
6.3.1 Model Building 33
6.3.2 Model Predictions 34

7 Summary and Conclusion **36**

List of Figures

1 Draft distribution by class year and college location. 12
2 Distribution of WS, WS/48, BPM and VORP 14
3 Correlation Heatmap of all variables 15
4 Q-Q plot of residuals of the WS model 17
5 Summary Statistics of the WS/48 model. 19
6 Summary Statistics of the BPM model 21
7 Summary Statistics of the Draft Rank model 21
8 Variance important Plots for the First Round and the Second Round outcome 28
9 Partial Dependence Plots on Class Year, SOS and Position. 29
10 Summary Statistics on the First Round classifier 31
11 OOB Bernoulli Deviance Change v. s. Iteration plot 33
12 Relative Influence of variables 34

List of Tables

1 Model Accuracy on WS/48 based on four metrics 18
2 Prediction Accuracy for selected players in the 2013 draft 23
3 Prediction Error of the draft rank model and career success models 23
4 Prediction Error of draft rank and career success models by Class Year. 24
5 Prediction Errors on ideal draft rank by model. 25
6 Prediction Error on the ideal draft rank of Al Horford 26
7 Process of feature selection in random forest 30
8 Prediction error of First Round Classifier on the test set 32
9 Prediction error on players drafted from No.21 to 40 and those outside this range 32
10 Prediction error of boosted classifier on the test set 34
11 Prediction error on players drafted from No.21 to 40 and those outside this range 35

Chapter 1

Introduction

Given the rising competitiveness and the hunger for winning championships in recent years, NBA teams have been actively following the blueprints of the 2010-2014 edition of Miami Heat with superstar trio Lebron James, Dwayne Wade and Chris Bosh and the 2014-2017 edition of Golden State Warriors with Stephen Curry, Klay Thompson and Draymond Green by teaming up at least two superstars during free agency to compete for a championship. However, the unprecedented $24-million leap in salary cap, the maximum amount of money spent on player salaries per season without paying luxury tax, has risen from 70 million to 94 million in 2016.[1] followed by another $25 million to $109 million for the current 2019-2020 season.

This trend has enabled relatively mediocre players to obtain big contracts after a few free-agent superstars like Stephen Curry (5-year, $201 million contract) and James Harden (4-year, $171 million contract), known as “the big fish”, have already signed huge contracts with championship-contending teams once free agency started every summer over recent years. Two of the worst contracts signed during the 2016 free agency came from former Memphis Grizzlies forward Chandler Parsons and current Charlotte Hornets forward Nicholas Batum. The former became injury prone and only appeared in 95 games over three seasons with the Grizzlies under a 4-year, $94.4 million contract, shooting at an abysmal rate of around 31 percent from the three-point line at a league that embraces floor spacing and outside shooting.[2] The latter signed a 5-year, $120 million contract and never lived up to expectations after becoming a passive player who attempted only 8.6 goals per 36 minutes.[2]

As a matter of fact, teams now start to understand the high costs of initiating trades to add quality players to the roster because these players can often times be overpaid. Sports economist

[1] Favale, 2018
[2] Favale, 2019

David Berri invented a methodology that uses the estimated number of wins in a team and the expected win share of the player in a team to calculate the salary that he actually deserves: for instance, Sacramento Kings forward Harrison Barnes only deserved a $3.5 million salary instead of $24.7 million paid out based on his estimated win share.[3] The stark $21 million difference can be used to acquire a lottery draft pick from other teams or saved for cap space to acquire superstars in future seasons.

Scouting talent from NBA draft has been an increasingly important approach for teams to achieve success in professional sports, especially for small-market teams that find it difficult to attract superstar talent like the two teams mentioned above or teams that operate above the salary cap and have to pay luxury taxes. In the second case, Golden State Warriors drafted Jordan Bell with the 38th pick in 2017 and immediately added him to the roster rotation for his defensive prowess and ability to rebound and block shots. While draft picks can be of crucial assets to teams, their value is only guaranteed when the right player is taken since there are countless instances of booms and busts from draft prospects. One of the most undervalued draft prospects in the history of professional sports is former New English Patriots quarterback Tom Brady: drafted 199th overall in the sixth round of the 200 NFL draft, Brady has become the most accomplished football star with six super-bowl championships.

1.1 Project Goal

The goal of this project is to apply multiple linear regressions to predict the draft rank and the career success of college players based on their draft timing and college statistics and also apply supervised machine learning techniques including random forest and stochastic gradient boosting to predict whether players will be drafted in the first round (No.1 – 30) of the draft as an indicator of their draft success.

[3] Knight, 2019

An accurate prediction of being drafted in the lottery round or the first round can help players better position themselves in the NBA draft for salary negotiations and future plans after getting a sense of where they will get drafted. The prediction of career success, depending on what metrics are used, can not only help players gain insights about how good their careers will turn into and how long they will be able to play in the league but also help teams gain information beyond traditional scouting reports and draft rank predictions in order to better identify which players they should select in the draft. Since college statistics contain an important portion of the features in our data, this project does not take into account players who were selected in the draft right after their senior year of high school.

Chapter 2

Background

2.1 Draft History

The NBA draft is an annual event dating back to 1947 for NBA teams to scout for young talent who wish to join the league. There were many rounds of draft until 1989, when the number of draft rounds was reduced to two.[4] There are 30 picks in both the first and the second round. As there are 30 teams in the league, each team is given a first and a second round pick. Historically, the draft was run in a lottery style where teams with the worst record during the previous season have the best chance of receiving the first pick. Because the top 16 teams enter playoffs every season, the rest 14 teams that are out of playoffs are guaranteed to receive picks from No.1 to No.14 and these picks are thus regarded as the lottery picks. Starting in 2019, the three teams with the worst record over the previous season are each given a 14% chance of winning the first pick in the lottery, with the rest 11 teams assigned chances based on the reverse order of their record during the previous season.[5]

2.2 Draft Timing

Draft timing plays a huge role of player success on the court. It is obvious to notice that players who declare for NBA draft after freshmen and sophomores tend to be selected with higher picks early in the first round of the draft, while juniors and seniors tend to be selected at lower picks late in the second round of the draft. This trend can be explained by the intention of teams to develop talent at a young age and allows players to player for a longer period of time at a professional level. This not only helps young players to grow quickly into stars and make a name for themselves but also positions teams at a winning situation. From a financial

[4] Wayback Machine
[5] NBA.com Staff, 2019

perspective, young players are less injury-prone and teams can exploit these players with long rookie contracts that last four or five years at a low cost before they enter their prime in their late 20s and demand an expensive contract.

However, some players who are predicted to be drafted after finishing off their college freshman year choose to stay for another year of college instead in order to have a better chance of being selected at a higher pick in the following year. This can be a huge risk because while the draft value of some players rises over the year, others are predicted to be drafted with lower picks when they did not improve dramatically through another year in college. A stark comparison between the destiny of two players that were both projected to be selected in the lottery round in 2016 can speak the truth. University of Oklahoma guard Buddy Hield was expected to enter the draft in 2015 after winning the Big-12 Conference Player of the Year but chose to return for his senior year after he made the announcement that "You can't go in there and take a gamble and bury yourself."[6] The decision paid off in a year later when he was drafted by New Orleans Pelicans with the No.6 pick and later became a star in the Sacramento Kings. Meanwhile, California Bears forward Ivan Rabb was originally projected as a No.14 lottery pick in 2016 but chose to return for his sophomore year of college.[7] When he decided to go professional a year later, Rabb was the No.35 pick selected by Memphis Grizzlies; after two unimpressive seasons with the team and a short-stint with New York Knicks, he is now playing for the Westchester Knicks in the G-league, the development league of NBA.[8] The stakes are high when it comes to making a decision to enter the NBA draft and stay in college because the difference could be millions of dollars for these players that chase their basketball dreams.

[6] Riedel, 2015
[7] O'Donnell, 2016
[8] Rotowire Staff, 2020

2.3 Draft Rank and Career Success

The career success of many players can be highly related to their draft ranks. Players who are selected early in the first round tend to have higher starting salaries and play a higher number of minutes in their starting season because teams with bad records that drafted these players give them opportunities to develop themselves on the court and make immediate contributions on the team. Some researchers discovered that the high playing time of players drafted with higher picks, including the times when they are less efficient than their teammates, is due to the sunk-cost effect,[9] which indirectly justifies the importance of selecting the right player in the draft. Two prominent examples of success go to the former Houston Rockets center Hakeem Olajuwon and the former San Antonio Spurs forwards Tim Duncan were both selected as the No.1 pick in 1984 and 1997 respectively and both had incredible hall-of-fame careers after leading their teams to multiple championships. However, it is inconclusive that high draft picks can lead to phenomenal careers and change the destiny of teams due to the presence of some huge busts coming out of lottery picks. Former No.1 picks Kwame Brown and Anthony Bennett were both touted as explosive scorers but ended up being notoriously known as draft busts: Bennett was given a chance of "redemption" by signing a non-guaranteed deal with the Houston Rockets in 2019 and was soon waived after a knee injury several months later.[10]

Thus, draft rank alone cannot be a good predictor of the future success of players and teams. In fact, teams started to use advanced metrics such as Player Efficiency Rating (PER) and Win Shares (WS) to holistically analyze player performance. While the majority of superstars came from the first round of draft, a number of them come from the second round of the draft: former San Antonio Spurs hall-of-famer guard, Manu Ginobili, selected as the 57th overall pick in 1999

[9] Staw, 1995
[10] Dubose, 2019

and known as one of the best players who came from the bench, finished his career with a WS of 106.4, placed fifth in the Spurs team history[11] and 71th in the NBA history.[12]

2.4 Draft Predictions

To translate draft rank into team success, sports analysts and team managers have been actively analyzing game statistics, the skills and physical attributes of players along with the positional fits of players in specific teams to predict the draft rank before the actual draft. While Bleacher Report analyst Jake Rill, like numerous other sources, projected University of Memphis center James Wiseman as a top-3 pick of the upcoming 2020 NBA draft,[13] Kevin O'Connor from The Ringers projected him at the 7th spot in his scouting report instead.[14] The dissentions among analysts are due to the analytical focus from different angles such as statistics and breakdowns of strengths and weaknesses. Thus, it is intriguing to dive into the prediction of NBA draft to see what metrics make players like LaMelo Ball and James Wiseman high on the draft board.

[11] Urbina, 2018
[12] Basketball Reference
[13] Rill, 2020
[14] O'Connor, 2020

Chapter 3

Literature Review

Multiple scholars have investigated on the prediction of the draft rank and career success across different leagues. Ryan Edwards and two other researchers from Stanford used a combination of players' college career data and their NBA career data to develop a model to predict where a player should be drafted (if at all) in the draft through classifying players by different levels of career success with principal component analysis, linear regression and support vector machines.[15] Meanwhile, Alexander Greene used a regression-heavy approach to build models based on player positions on the court (e.g. point guard) and found that the college-year field goal percentage, the number of blocks per game and the 1st team All-American honor are important predictors of player success in NBA and found a higher career WS by 14.12 for every player named as 1st team All-American.[16]

Outside basketball, researchers in football also actively explored the prediction of the draft and the career success. Former New York Jets analyst Jason Mullholland and Professor Shane Jenson created prediction models for the NFL draft and NFL career performance with a focus on the tight end position and found that the predictors of these two models are different. In addition to college statistics, they included the NFL combines and physical measures in their models.[17]

[15] Edwards, 2015
[16] Greene, 2015
[17] Mullholland and Jenson, 2014

Chapter 4

Data Retrieval, Creation and Analysis

4.1 Data

4.1.1 Source

The data comes from Basketball Reference and Sports Reference, two linked sources that exclusively record the historical statistics of basketball players from college to NBA as well as numbers from some international leagues in Europe and the Middle East. For NBA players who went through colleges in NCAA, there are statistics per game broken down by each college year. However, for players who competed in international leagues before joining NBA, their statistics from international leagues were unrecorded due to two reasons. First, some of these statistics are available on a related source called Sports Reference and others are completely missing. Second, the variety of basketball leagues outside the USA made it difficult to compare the prior-NBA statistics of these players with those of players that competed in NCAA during college.

4.1.2 Data Description

The dataset includes all drafted players from 2006 to 2014 in both the first and the second round. Each row of data consists of the rank of the player in the draft and their corresponding NBA statistics and college statistics.

In addition to their draft ranks and statistics from these two stages, the position that they played on the court is recorded in the “Position” variable and the college year when they declared for NBA draft is recorded in an indicator variable called “Class Year”. For Position, there are 5 levels composed of the five on-court positions on the court, namely PG, SG, SF, PF and C. If a player can play multiple positions such as PG and SG due to his versatility, the final position recorded is based on the highest number of games he played under that position during his NBA career. For Class Year, there are 6 levels including freshman, sophomore, junior, senior,

super-senior (players who stayed five years in college) and IN. “IN” are assigned to international players who played internationally before entering NBA to demonstrate their prior international experience. The NBA statistics of these players are aggregated over a player’s entire career as of November 2019. The following list of variables can be used as candidates of the outcome variable when modeling on career success:

1. **FGP:** the career field goal percentage
2. **Points:** average points per game over the career
3. **Rebounds:** average rebounds per game over the career
4. **Assists:** average assists per game over the career
5. **WS:** career win share
6. **WS/48:** win share per 48 minutes
7. **BPM:** Box plus and minus
8. **VORP:** Value Over Replacement Player

The college statistics of these players are broken into the statistics per game from the most recent year and the average statistics per game of the previous years combined excluding the most recent college year. For instance, if a player was drafted after junior year, we take his statistics per game during his junior year as the most recent year data and his statistics per game during his freshman and sophomore year combined as the data for the previous years combined. If a player was drafted after freshman year, we take his freshman year statistics as the most recent year data and leave the data of his previous years as blank. The following list of variables are the candidates of predictors to model both career success and draft rank:
(**Note**: we use the variables listed below for the most recent year and added the suffix “.1” to each variable for the previous years combined.)

1. **FG:** field goal percentage
2. **FGA:** field goals attempted per game
3. **FG.PG:** field goal percentage

4. **P2M:** 2-point goals made per game
5. **P2MA:** 2-point goals attempted per game
6. **P2.PG:** 2-point goal percentage
7. **P3M:** 3-point goals made per game
8. **P3MA:** 3-point goals attempted
9. **P3.PG:** 3-point goal percentage
10. **FT:** free throws made per game
11. **FTA:** free throws attempted per game
12. **ORB:** offensive rebounds per game
13. **DRB:** defensive rebounds per game
14. **TRB:** total rebounds per game
15. **AST:** assists per game
16. **STL:** steals per game
17. **BLK:** blocks per game
18. **TOV:** turnovers per game
19. **PF:** personal fouls per game
20. **PTS:** points per game
21. **SOS:** strength of schedule (college players play in different leagues and SOS can differ.)

4.2 Exploratory Data Analysis

The data was split into the first 7 years (2006-2012) as the training set and the latter 2 years (2013-2014) as the test set. The EDA will be performed on the training set and the model predictions will be conducted on the test set. The reason for this test-training split is to preserve a complete draft round in the test set in order to see how well the model predicts for different draft ranks in the same round. Moreover, the 2013 and 2014 draft are chosen because players will have at least 5 full years of career experiences in NBA and the career statistics in the dataset can be a better reflection of their career performance compared with players in the recent 2018 and 2019 draft.

4.2.1 Univariate Analysis

Draft Distribution

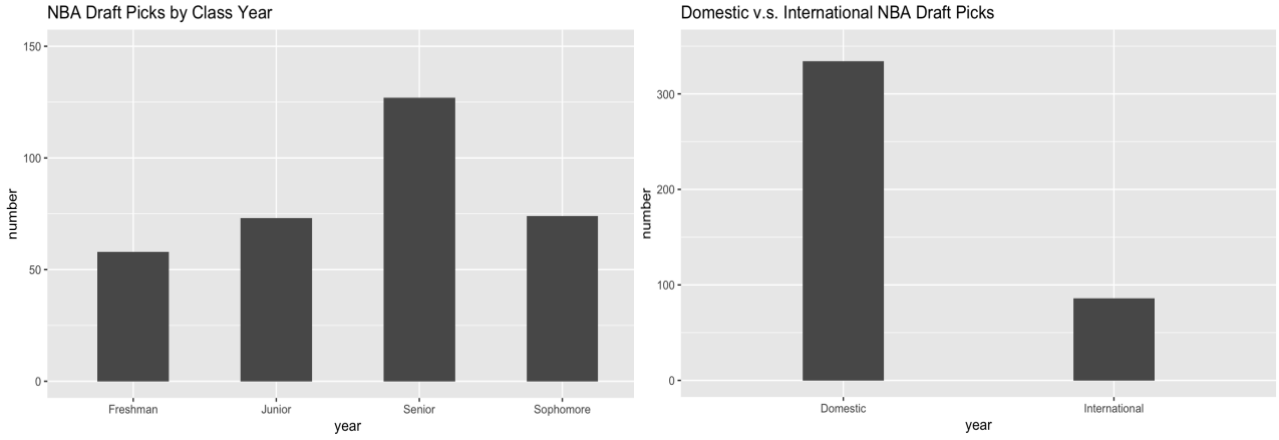

Figure 1. Draft distribution by class year and college location

From the distribution of NBA draft picks on the left histogram of Figure 1, we notice that seniors have the highest number of drafted players at 127 over 7 years of draft. While there were only 2 super-seniors drafted over this time period, there is an even spread of freshmen, sophomores and juniors from 60 to 80. The right histogram of Figure 1 shows that about 80% of the drafted players played domestically in NCAA while 20% played internationally. Because international

players count as an un-negligible number of players in the draft, this distribution justifies including these players as part of the analysis instead of only analyzing domestic players.

Outcome Variable Selection for Career Success

The NBA career statistics of players in the dataset are aggregated over the entire player career as of November 2019. These statistics are used instead of game statistics based on individual seasons because the career length of players varies widely, and the fact that players hit their stride or decline at different times of their career would create difficulty to compare statistics from individual seasons between players. Among all the outcome variables for career success, the first four candidates listed above in the Data section, namely FGP, Points, Rebounds and Assists, are basic career statistics that only capture one aspect of a player's game and thus are not good indicators of career success. These metrics along with other basic game statistics such as turnovers and skills can be converted altogether into an advanced stat called John Hollinger's PER (Player Efficiency Rating) through a formula shown in Appendix I. Meanwhile, the latter four candidates for career success, namely WS, WS/48, BPM and VORP are all advanced stats that already factored multiple basic statistics in their calculations.

Among these five advanced stats, PER has a two major weaknesses: first, it is a metric based on player performance per minute and cannot truly reflect the performance of players with limited minutes; second, it does not take into account defensive contributions such as blocks and steals.[18] Meanwhile, it tailors exclusively to individual performance while other advanced stats such as WS and BPM are team-centric metrics computed based on player's contributions to the team and are useful to both teams and players concurrently; an accurate prediction model of career success with metrics like WS thus could help players sell themselves and earn more lucrative contracts.

[18] "Player Efficiency Rating", Wikipedia

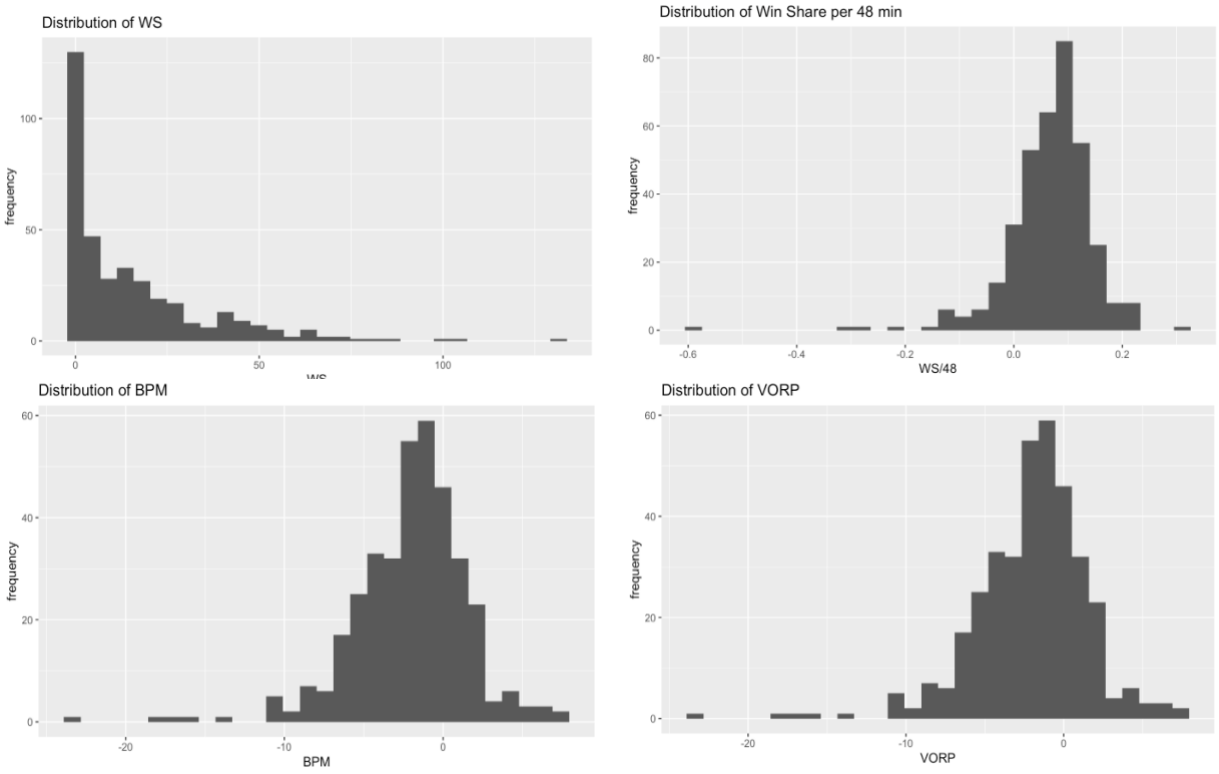

Figure 2. Distribution of WS, WS/48, BPM and VORP

From the four histograms in Figure 2, we observe that WS/48, BPM and VORP might be slightly negatively skewed but overall are normally distributed. WS does not follow a normal distribution and this can be explained by the cumulative nature of WS where the longer players stay in the NBA, the higher the WS will be. Since many players stay for 2 to 3 years in the league, their WS will mostly be around 0 with positive WS outliers being a small number of superstars who thrive in NBA for more than 8 to 10 years.

4.2.2 Bivariate Analysis

Variable Correlation

The correlation analysis focused on removing variables with high linear correlations and was conducted based on the domestic player data due to the high number of null values of

international players who had no college experience in NCAA. From the correlation heatmap in Figure 3, we observed a strong correlation between several pairs of variables, namely FGA, FG.PG and FG per game. It becomes intuitive to acknowledge the high correlation coefficient between FG and FGA at 0.91 because players score more with a higher number of shot attempts. Thus, we removed FGA, the variable more strongly correlated with FG.PG at a coefficient -0.46 compared with FG.

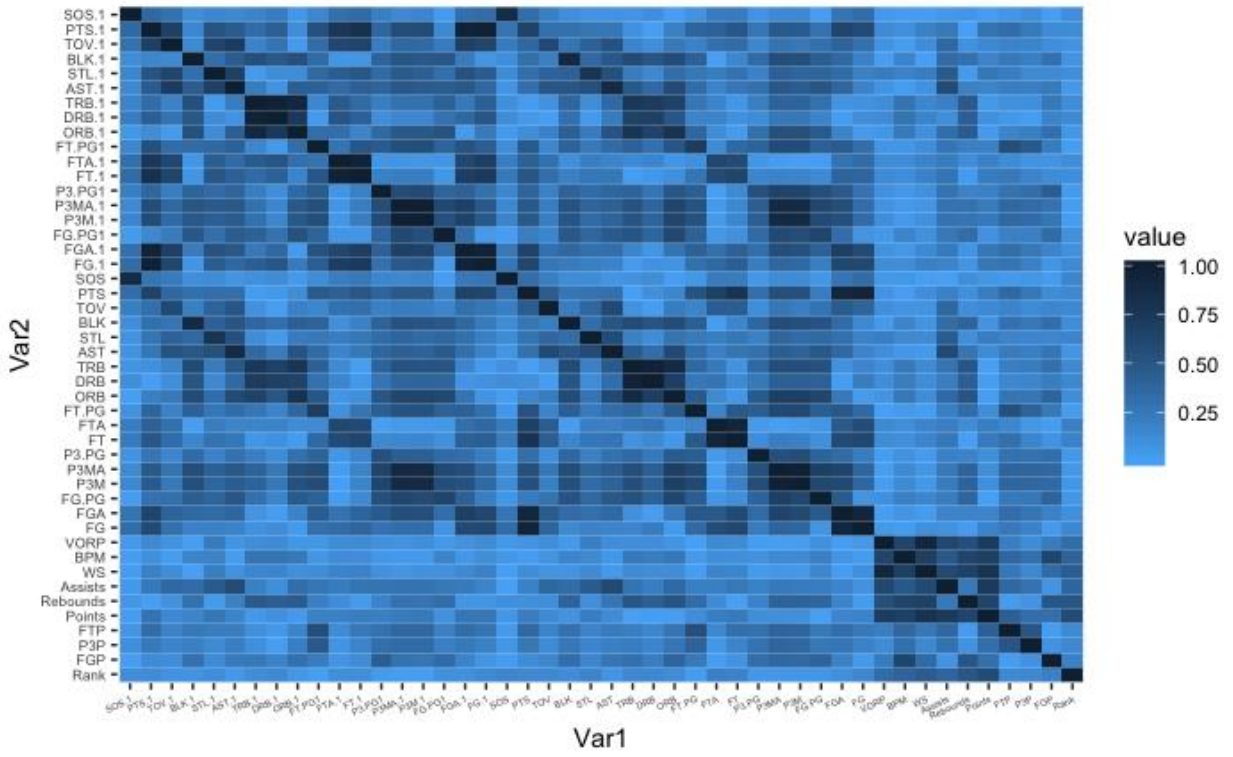

Figure 3. Correlation Heatmap of All Variables

The same rule was applied to other pairs of highly correlated variables related to free throws and 3-point shots. Owing to the high correlation between FT and FTA per game at 0.945, we removed FT as it has a stronger correlation with FT.PG at 0.47. All the correlation coefficients are displayed as a table in Appendix II. The choice of keeping FTA instead of FT can also be justified by the fact that FTA serves as an important metric of individual and team performance by drawing fouls from opponents to achieve points. However, similar to the FGA case, we removed P3MA because P3MA and P3M shared a similar correlation with P3.PG and P3M is a

direct indicator of game performance especially during the modern small-ball era of professional basketball world where the ability to knock down three-point shots and create space for teammates is strongly emphasized and traditional big-men who are not able to shoot beyond the free throw lines are gradually abandoned. The most recent example that exemplified this trend came from a trade completed on February 4th, 2020 by NBA team Houston Rockets, who paid a premium on the sharp three-point shooter and wing defender Robert Covington by trading their starting center Client Capela, a solid rim protector who can post double-double figures on a nightly basis.

Following the same approach, we performed correlation analysis on the statistics from the years combined prior to the final college year and removed FGA.1, FT.1 and P3A.1. Last but not least, because PTS (points) and FG both indicate how much players score, we decided to remove FG and FG.1 because FG does not discern the value between a 2-point goal and a 3-point goal.

Chapter 5

Modeling on Career Success and Draft Rank with Regression

5.1 Modeling on Career Success

We first turn our attention to predicting the career success, using Class Year, Position and college statistics with multiple linear regression. For outcome variable, we used WS, WS/48, BPM and VORP and checked the validity of each model. All four variables are team-centric metrics of player career success because they account the contribution of each player to team's wins in WS and WS/48, and to team's scores in BPM and VORP. VORP is derived from BPM and takes into account the playing time of players.[19] When we attempted to use WS as the outcome variable, based on the Q-Q plot in Figure 4, we found that the residuals of the model follows a non-normal distribution due to a skew in both tails, and this corresponds to the non-normal distribution of WS itself. Due to the difficulty to conduct a log transformation on WS since there are many values of 0 and the conversion would lead to values of negative infinity, we do not believe that WS is a good outcome variable to model career success.

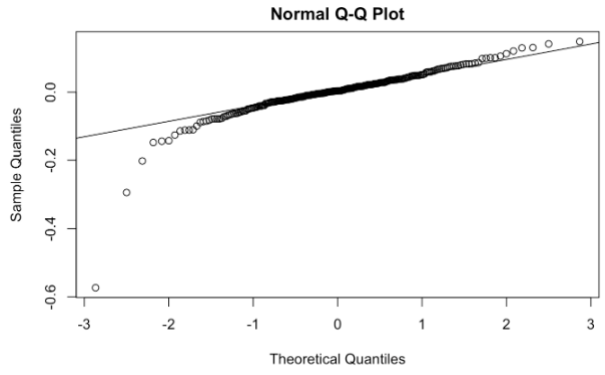

Figure 4. Q-Q plot of residuals of the WS model

[19] "About Box Plus/Minus (BPM)."

We then focus on the remaining three variables and found that the residuals of the three models under each variable follow a normal distribution. First, we model career success with WS/48 as the outcome variable and use the best subset regression to conduct feature selection and compare model accuracies based on four model performance metrics: adjusted R-squared, RMSE, Mallow's Cp and BIC. Because the selected predictors of each model are different due to the differences between these 4 metrics, we measure the accuracies of the four models with the highest R-squared value, the lowest RMSE, the lowest Cp and the lowest BIC and then compare the adjusted R-squared values of these four different models. RMSE here is measured by the formula: $$RMSE = \sqrt{MSE} = \sqrt{\frac{RSS}{n-p-1}}$$

The parameter n stands for the total number of observations and p stands for the number of predictors in the model. Table 1 shows that among all four models, the two models based on the lowest RMSE and the highest adjusted R-squared have the same set of predictors and the highest adjusted R-Squared value at 0.2094, which is significant at the 0.01% level, indicating that the selected predictors have significant power for WS/48. Because RMSE measures the average prediction error, we consider RMSE as a better metric to determine how good the model is and will apply this model performance metric for model selection when BPM and VORP are used as outcome variables later on.

Model	Adjusted R^2 Value of Model
Highest adjusted R-squared	0.2094
Lowest RMSE	0.2094
Lowest Cp	0.205
Lowest BIC	0.2016

Table 1. Model Accuracy on WS/48 based on 4 metrics

In Figure 5 below, the selected predictor variables of WS/48 include FG.PG, TRB, STL, TOV, Class.YearSenior, Class.YearSophomore, PositionPG, FG.PG1 and FT.PG1. As only 2 out of 6 levels of Class Year and 1 out of 5 levels of Position are significant, we perform two F-tests by removing each of these two predictors in a new model to check their significance.

			WS/48
	Coefficient estimate	Standard error	p-Value
Intercept	-0.300	0.08	0.000
Class.YearSenior	-0.017	0.01	0.134
Class.YearSophomore	0.019	0.01	0.115
Class.YearSuperSenior	-0.008	0.05	0.882
PositionPF	-0.006	0.02	0.724
PositionPG	0.022	0.02	0.410
PositionSF	-0.010	0.02	0.602
PositionSG	-0.014	0.02	0.545
FG.PG	0.171	0.12	0.147
FG.PG1	0.226	0.10	0.021
FT.PG1	0.183	0.05	0.001
TOV	-0.019	0.01	0.023
TRB	0.010	0.00	0.000
STL	0.017	0.01	0.130

Figure 5. Summary Statistics of the WS/48 model

The first F-test shows that adding Class Year did lead to a significantly improved fit over the model without Class Year based on a p-value of 0.017. While the second F-test shows that adding Position does not lead to a significantly improved fit over the model without this variable based on a p-value of 0.19, Position can impact career success because the NBA is gradually evolving from a big-man oriented league in the early 2000s to a guard-oriented league in the 2010s. Meanwhile, the removal of Position variable will decrease the adjusted R-squared model to 0.202. Ultimately, we decided to keep both Class Year and Position in the model and keep these two categorical variables. For Class Year, we recoded the Freshman, Junior and IN levels as a level called “Other” because these three levels are insignificant and thus did not appear in the summary statistics. We chose to recode this variable instead of keeping the original interpretation because the “Other” category can combine the non-selected levels and serve as a

reference category. Moreover, leaving levels that miss statistical significance can create trouble when predicting the career success of players whose Class Year are among the insignificant levels because that level is not a parameter in the model.

After recoding, the summary statistics of the final model remain the same. FG.PG1 and FT.PG1, which measure the shooting accuracy of field goals and free throws respectively over the previous years combined, are the most significant predictors of WS/48 according to Figure 5. The reference level are players who were drafted after freshman year and play the Center position due to the presence of two categorical variables. The coefficient estimates can be interpreted as the partial effect of that variable when other variables are held constant. They show that declaring draft after sophomore year of college might boost WS/48 by a little while waiting till senior year to be drafted can have a negative impact on WS/48.Moreover, in terms of player position, PG seems to have the highest WS/48 on average among all five positions, and this can be interpreted by the guard-oriented trend of the league starting in late 2000s and early 2010s.

The same approach was repeated by using BPM and VORP as the outcome variable to estimate career success and conducting model selection by choosing the predictors that minimize RMSE and recoding insignificant levels. The VORP model has a very low adjusted R-squared value at 0.1105 while the BPM model has the best adjusted R-squared value among all three at 0.2135. The summary of the final model on BPM is shown in Figure 6 on the next page. Similar to the WS/48 model, FG.PG and FT.PG1 are important predictor variables. Here we also observe that players drafted after the sophomore year of college have a slightly higher BPM given all other factors staying constant. As the reference level of the Position variable is C, both PG and SG demonstrate a slightly advantage on BPM over the other three positions, namely SF, PF and C. Overall, the WS/48 and BPM model have a similar set of predictors.

	BPM		
	Coefficient estimate	Standard error	p-Value
Intercept	-14.6	3.38	2.26e-5
Class.YearSenior	-1.27	0.52	0.134
Class.YearSophomore	0.42	0.56	0.454
Class.YearSuperSenior	-2.26	2.35	0.338
PositionPF	-0.86	0.74	0.242
PositionPG	1.90	1.17	0.104
PositionSF	-0.01	0.83	0.986
PositionSG	0.18	0.93	0.850
FG.PG	10.25	4.85	0.036
FT.PG1	6.01	2.45	0.017
TRB	0.28	0.17	0.092
TRB.1	0.51	0.20	0.010
STL.1	1.83	0.61	0.003
TOV.1	-1.60	0.47	0.001

Figure 6. Summary Statistics of the BPM model

5.2 Modeling on Draft Rank

We model the draft rank with the same candidates of predictors and conduct feature selection with the best subset regression by minimizing RMSE and then recoding the insignificant levels of categorical variables as “Other”. The draft rank follows a uniform distribution because there are exactly 7 players drafted with each pick from No.1 to No.60 during the 7 rounds of draft. The summary statistics of the model below in Figure 7 demonstrate that the predictors of draft rank are different from the predictors of career success.

	Draft Rank		
	Coefficient estimate	Standard error	p-Value
Intercept	94.2	11.43	8.94e-15
Class.YearSenior	9.9	2.15	6.11e-6
Class.YearSophomore	-6.0	2.34	0.011
Class.YearSuperSenior	16.7	9.93	0.094
FG.PG	-50.0	17.71	0.005
P3.PG	-14.2	6.28	0.025
STL	-9.7	2.48	0.000
BLK	-4.4	1.27	0.001
PTS	-1.3	0.25	6.82e-7
SOS	-0.9	0.30	0.003
STL.1	5.7	2.83	0.046

Figure 7. Summary Statistics of the Draft Rank model

The Position variable is no longer in this model and the reference level is the "Other" category in the Class Year variable. Negative coefficients indicate that higher values of that variable can predict a lower (i.e. better draft rank) and vice versa. The positive value of estimate for Class.YearSenior and the negative value of estimate for Class.YearSenior show that declaring draft early during sophomore year would mean a better draft rank on average given all other factors constant. Meanwhile, FG.PG and P3.PG are the two most important predictors and this shows that the team emphasize on finding three-point shooters to space the floor in the modern NBA era. While the models on career success have STL and TRB as predictors with low coefficient estimates, the model on draft rank places a heavier emphasis on defensive statistics by including STL, BLK and STL.1. An improvement by one block and one steal game in the most recent year of college could mean a rise in the draft board by 4.4 and 9.7 positions respectively. Moreover, the model also factors into SOS because the statistics in different college leagues have meanings of different magnitude.

5.3 Model Predictions

For prediction, we used the final models for WS/48, BPM, VORP and Draft Rank and applied them to the test dataset, which includes 120 players in total from the 2013 and 2014 draft. We used RMSE as an estimation for the prediction error by taking the standard deviation of the residuals as shown in the formula below:

$$RMSE = [\sum_{i=1}^{n}(a-p)^2/n]^{1/2},$$

where n stands for the number of observations, a stands for the actual value and p stands for the predicted value from the regression model. Because all the models contain at least one stat from the previous years combined among their predictors, the model predictions are only limited to sophomores, juniors, seniors and super-seniors while freshmen and international players are left because they either only had statistics from the most recent year or no statistics at all. The Table 2 on the following page shows a part of the prediction from the 2013 draft. The complete

prediction results for the 2013 and 2014 draft are shown in Appendix III. The table already omits freshmen and international players, thus the "Other" category of the Class Year variable essentially means juniors. (Note: The prediction results for the VORP model are not shown due to limited column space and high prediction error.)

Player	Class	rank	fitted_Rank	Rank (a-p)^2	WS48	fitted_WS48	WS48 (a-p)^2	BPM	fitted_BPM	BPM (a-p)^2
Al Horford	Other	3	31	771	0.16	0.12	0.001	3.2	0.8	6.0
Jeff Green	Other	5	29	553	0.07	0.05	0.001	-1.4	-2.4	1.0
Corey Brewer	Other	7	31	556	0.06	0.05	0.000	-1	-3.0	4.1
Joakim Noah	Other	9	13	17	0.16	0.12	0.001	4.1	0.0	17.0
Acie Law	Senior	11	42	968	0.03	0.01	0.000	-5.4	-3.5	3.7
Julian Wright	Sophomore	13	22	78	0.06	0.09	0.001	-1.3	-2.1	0.7
Al Thronton	Senior	14	22	71	0.04	0.04	0.000	-3.5	-3.7	0.1
Rodney Stuckey	Sophomore	15	21	34	0.08	0.06	0.000	-1.4	-2.8	1.9
Nick Young	Other	16	29	175	0.05	0.05	0.000	-3.4	-1.7	2.7
Jason Smith	Other	20	31	129	0.07	0.09	0.000	-2.7	-2.0	0.5

Table 2. Prediction Accuracy for selected players in the 2013 draft

5.3.1 Model Comparison on Predicting Career Success

The total number of observations in the predictions is 65, with 30 from the 2013 draft and 35 from the 2014 draft. By computing the RMSE for each model, we noticed that WS/48 is the best model to predict career success because it only has an error of 0.06 measured by RMSE in the as shown in Table 3 below. A deviation of 0.06 in WS/48 is not as big as the deviation measured by RMSE in other indicators of career success. For instance, according to the NBA career leaderboard in WS/48, current Portland Trail Blazers forward Carmelo Anthony has a career WS/48 of 0.1232 and former Minnesota Timberwolves forward Kevin Garnett has a career WS/48 of 0.1822.[20] Garnett will be inducted to the hall of fame while Anthony has the caliber of a future hall-of-famer based on his career performances thus far, and the 0.06 difference between their career WS/48 does not show a huge difference.

Model	RMSE on Predictions
Draft Rank	15.70
WS/48	0.06
BPM	2.96
VORP	8.91

Table 3. Prediction Error of the draft rank model and career success models

[20] "NBA & ABA Career Leaders and Records for Win Shares Per 48 Minutes."

The BPM has a very high RMSE value at 2.96 because an increment in BPM by a value of 2.0 would mean that the caliber of the players improves to a new level according to Basketball Reference.[19] For instance, the 2.0 increase in BPM is equivalent to the improvement from a good starter to a player that receives all-star consideration. Thus, a prediction error of greater than 2 means that the BPM model is not a good predictor of career success. Meanwhile, the VORP model does a worse job than BPM due to its high prediction error at 8.91. VORP is a metric that converts the BPM rate into an estimate of player's overall contribution to the team, and a value of -2.0 is established as the replacement level for the NBA.[19] According to Basketball Refence The reigning MVP Giannis Antetokounmpo had a high VORP of 7.4 during the 2018-2019 season. The difference between this value and the replacement level is a stunning 9.4, which is very close to the prediction of the VORP model. This large spread proves that the VORP model cannot accurately predict the career success of players on average.

To see how well each model predicts the outcome by Class Year, we compute the RMSE of each model for predictions under a specific class year in Table 4.

Class Year	Rank	WS/48	BPM
Super-Senior	16.0	0.021	1.28
Sophomore	10.0	0.058	2.71
Senior	16.1	0.066	3.48
Junior (Other)	17.3	0.047	2.34

Table 4. Prediction Error of draft rank and career success models by Class Year

The draft model has the best predictions on sophomores based on the lowest RMSE among all four class years at 10.0. While WS/48 has the lowest prediction error for super-seniors, this observation is inconclusive because there is only one observation that is a super-senior in the test set. From the WS/48 and BPM model, we notice that both models are better at predicting sophomores and juniors compared with seniors due to a lower prediction error.

5.3.2 Model Comparison on Ideal Draft Rank

The draft rank model has a very high RMSE at 15.70, which means that a prediction of No.1 pick could have an error by more than 15 picks down to No.16, a position outside the lottery round (No.1-14). The huge difference between actual draft rank and predicted draft rank can be attributed to having college statistics as predictors. In reality, teams make decisions based on not only college statistics but also things outside statistics such as the performance of these players in individual workouts and positional fit. Some teams also use have the wrong judgments and waste draft picks on busts, which are detailed in countless stories on social media. Thus, we decided to rank the career success of players to determine the ideal draft rank of players since more accomplished players deserve to be drafted earlier.

To do this, we first take the 2013 draft and rescale the rank from 1 to 30 since there are only 30 out of 60 players remaining in the prediction after players who declared for draft after freshmen and international players are removed from the prediction. Then we took the rank of WS/48 as the ideal draft rank because win share is a more widely used metric of career success compared with BPM. To compare how well each model predict the ideal draft rank, we use the MAE (mean absolute error) to measure prediction accuracy instead of RMSE because both the predicted and actual draft ranks are in whole numbers and MAE is easier for interpretation. MAE uses the average absolute difference between the prediction and the actual value over all observations as demonstrated in the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| a - p|,$$

where n stands for the number of observations, a stands for the actual value and p stands for the predicted value from the regression model. The prediction results are demonstrated in Table 5.

Prediction Based On	MAE compared with Ideal Draft Rank (based on actual WS/48)
Actual Rank	8.67
Fitted Rank	8.33
Fitted WS/48	**5.33**
Fitted BPM	7.47

Table 5. Prediction Errors on ideal draft rank by model

For the Fitted Rank, Fitted WS/48 and Fitted BPM cases, we get the predicted ideal draft ranks by ranking the results of these variables and then compare with the respective ideal draft rank based on the actual WS/48. The entire process of derivation is attached in Appendix IV.

Player	Ideal Draft Rank	Actual Draft Rank	Rank (Fitted WS/48)	Rank (Fitted BPM)
Al Horford	**2**	12	**2**	11

Table 6. Prediction Error on the ideal draft rank of Al Horford

Taking player Al Horford as an example in Table 6. Because his draft rank based on the fitted values of the WS/48 model is identical to its ideal draft rank No.2, the value of MAE is 0 and the draft rank based on the fitted values of the WS/48 model best predicts the ideal draft rank of Horford. We found that the actual draft rank has the largest deviation from the ideal draft rank at 8.67 positions. The best approach to predict the ideal draft rank is to rank the fitted values of the WS/48 model as the predicted draft ranks because it has the lowest MAE at 5.33 positions while the second-best approach is to rank the fitted values of the BPM model. However, this set of results can cause bias because the ideal draft rank is based on the actual WS/48 and the fitted values of WS/48 best predict the WS/48. To avoid this bias, we provide a sanity check for the high model accuracy of WS/48 on predicting ideal draft rank by using the ranks of the actual BPM values as the ideal draft rank instead of the actual WS/48 values instead. Here we directly compare the MAE when the draft ranks based on the fitted values of WS/48 and BPM are respectively measured against the ideal draft rank. The entire process of derivation is attached in Appendix V. Surprisingly, even though the fitted values of BPM better predict BPM itself, the MAE based on ranking the fitted values of WS/48 is 6.53, which is lower than 8.93, the value of MAE based on ranking the fitted values of BPM. Thus, it is safe to say that the draft ranks based on ranking the fitted values of the WS/48 best predict the ideal draft rank.

Chapter 6

Classifying Draft Success

6.1 Building First Round Draft Classifier with Random Forest

The destiny between first round and second round draft picks can be different. First round picks tend to be drafted by teams with worse records. Not only do they get more playing time to develop themselves but also do almost all of them get guaranteed contracts at a larger dollar value. Second round draft picks do not get the same treatment: seven college players taken in the second round of the 2017 NBA draft did not get guaranteed contracts.[21] Thus, it becomes increasingly important to help players predict whether they will be drafted in the first round so that players can better position themselves and have better expectations of what contracts they will sign with teams.

Since the model on draft rank poses a high prediction error and thus is not good for predicting draft rank, we can use random forest to classify whether players will be drafted in the first round or the second round, a good indicator of draft success. To use random forest classifier, we construct an indicator variable called "firstRound" with True for players whose draft ranks are between No.1 and No.30 and False for players whose draft ranks are between No.31 and No.60. Because there is a certain amount of missing values in variables that show the college statistics from the previous years combined due to the presence of freshmen and international players in the Class Year variable, we perform imputation on missing values with two different methods and compare the model accuracies after these two methods of imputation. We first used the default random forest imputation package in R to perform data imputation, which uses a proximity matrix to impute missing values with the weighted average of the non-missing observations for continuous variables and the category with the largest average proximity according to the R documentation. This is equivalent to the k-nearest-neighbors algorithm by

[21] Dauster

uses feature similarity to predict the values for new data points. Second, we used the Multivariate Imputation by Chained Equation (MICE) package in R because the chained equation approach is flexible and can handle both continuous and categorical variables as well as complexities such as bounds or survey skip patterns.[22]

When running random forest on the imputed data of these two methods with the same seed, we found that the out-of-bag (OOB) estimate error rate of the first approach is 35.95%, higher than 32.86%, the value of the OOB estimate error of the second approach. Thus, we proceed with the random forest classifier of the MICE approach and conduct feature selection through eliminating unimportant features from the variable important plot in Figure 8.

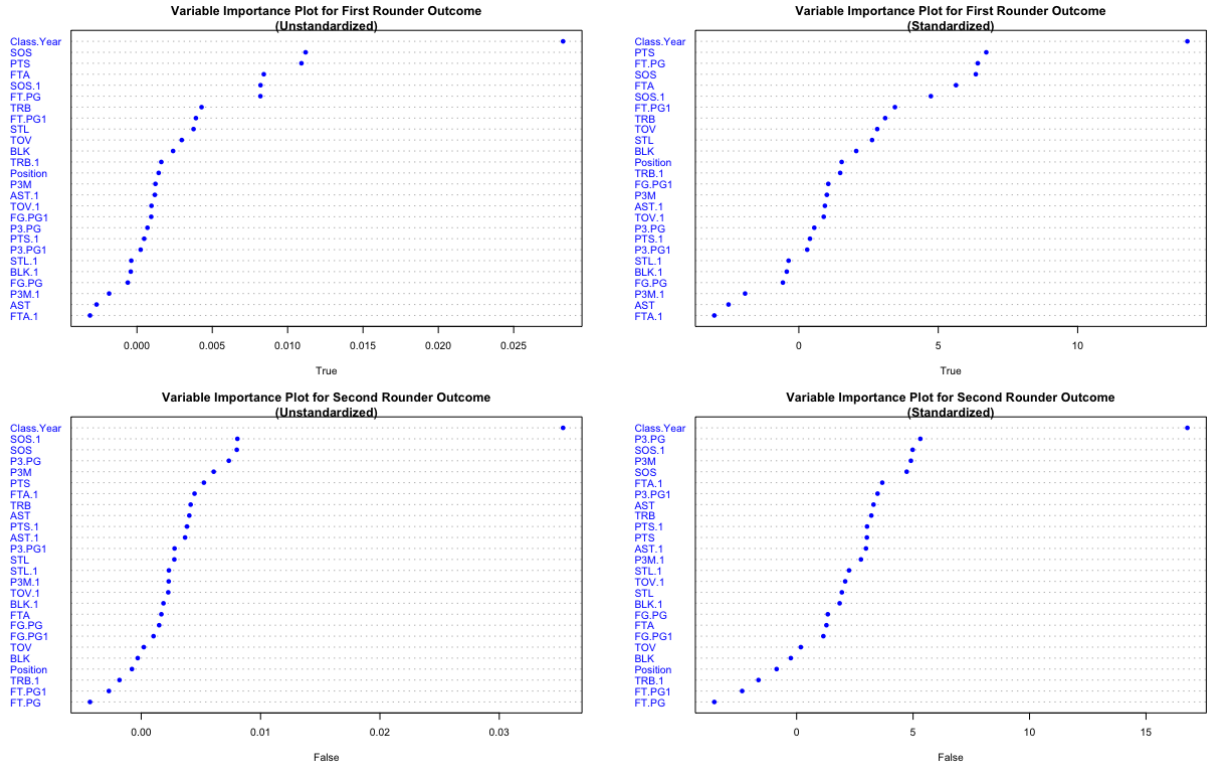

Figure 8. Variance important Plots for the First Round and the Second Round outcome

[22] Badr, 2019

The values on the x-axis show the decrease in the mean accuracy of the model if the corresponding predictor is removed from the model. For the First Rounder outcome, it seems that the expected important variables with a basketball mind hold more weight as the outcome seems to be most affected by Class Year, SOS, PTS and FTA. However, for the Second Rounder outcome, the plot shows that in addition to Class Year, SOS and PTS, the two variables P3.PG and P3M are also among the most important variables. This indicates that the draft value of players can be significantly dropped from the first round to the second round if players are not good at shooting three-pointers to space the floor in the modern NBA era. Moreover, the Position variable is a more important variable for the First Rounder outcome than the Second Rounder outcome. We can infer from this that the first round of the draft takes account the positional fit of player on a team more than the second round.

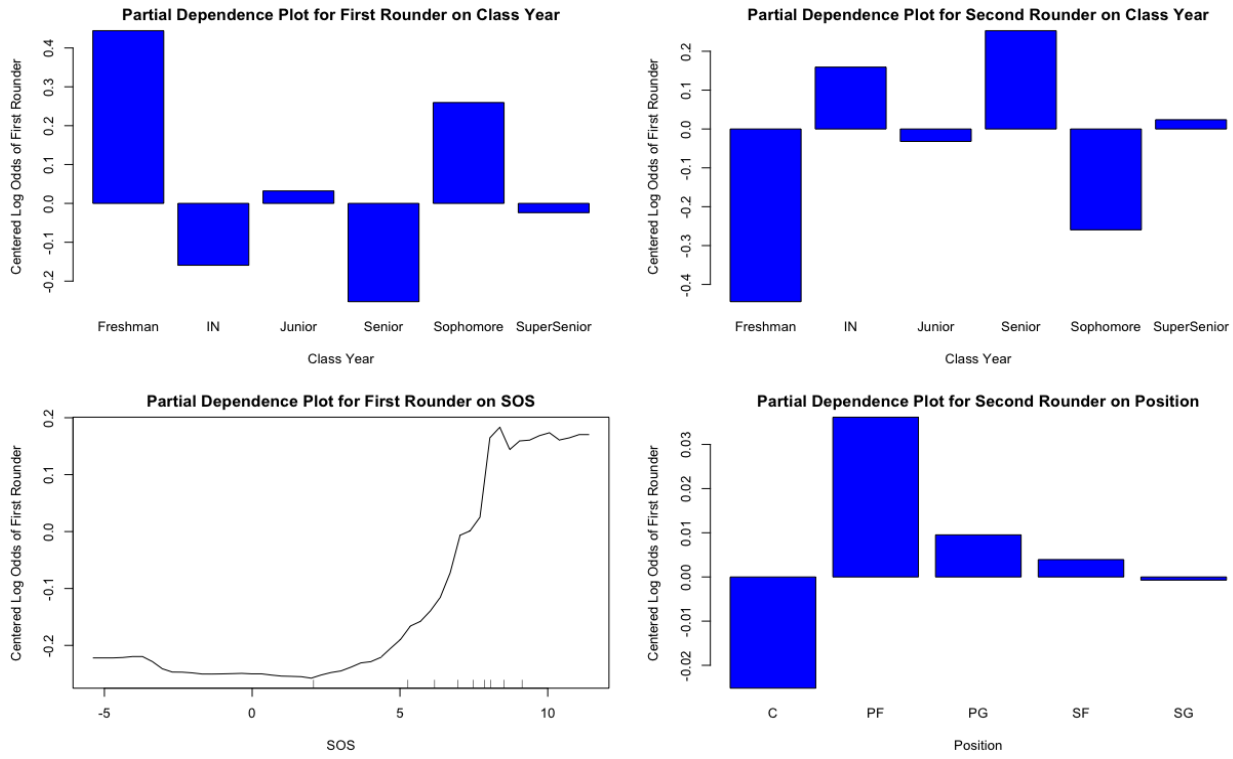

Figure 9. Partial Dependence Plots on Class Year, SOS and Position

To further explore the relationships between the predictors and the outcome, we dive deeper into the contributions of each predictor to explaining the outcome while holding all other predictors of the model constant in Figure 9 above. To conserve space, we picked Class Year, SOS and position to explain the effects of each predictor on the outcome. The top left figure shows that freshmen and sophomores are more likely to be selected in the first round and the top right figure shows that seniors are more likely to be drafted in the second round. The bottom left figure shows that the higher the strength of schedule in college leagues, the more likely teams will value the players and select them in the first round of the draft based on the increasing log odds value of being selected in the first round at a higher SOS. Last but not least, the bottom right figure demonstrates that players who play the PF position are disfavored because they are more likely to be drafted in the second round due to the highest log odds value of being drafted in the second round.

To select the best features, we take out the variable with the largest negative variable importance from the variance importance plot for the First Rounder to see if the OOB estimate of error rate has improved. We then re-run random forest on the remaining predictors, re-compute the variance importance plot and repeat the same procedure until there are no predictors with negative values of variable importance.

Step	Removed Variable	New OOB Estimate of Error Rate
1	FTA.1	29.52%
2	P3M.1	28.33%
3	FG.PG1	30.48%
4	P3.PG1	29.52%
5	FG.PG	31.67%
6	AST	28.81%

Table 7. Process of feature selection in random forest

From Table 7 above, we noticed that random forest classifier improved after removing FTA.1, P3M.1, FG.PG1, P3.PG1, FG.PG and AST. After this process, we set the target cost ratio of false positive to false negative as 2:1 and then perform parameter tuning on the random forest classifier to make sure that the actual cost ratio approximates the target ratio. The false positive

is when second rounders are misclassified as first rounders while the false negative is when first rounders are misclassified as first rounders. We set a target cost ratio of 2:1 because false positive is clearly more costly because misclassifying second rounders as first rounders overestimates a player's ability and gives players an overly optimistic expectation coming into the draft. This will be detrimental to the career of players as they might prepare less for the draft and work less hard to be what they aspire to be in their careers. Because the number of observations for the first round outcome is 210, we use the two-thirds rule to sample 140 first rounders and adjust the sample size of second rounders as a tuning parameter to approximate the actual cost ratio to the target ratio. By using a sample size of 101 second rounders from observations and 5 variables tried at each split of the tree, we are able to achieve the following results as shown in table 10 below. The target cost ratio 82:43 is very close to the target 2:1 ratio. The false positive rate is 20% while the false negative rate is 39%.

	First Round Classifier		
		No. of Variables tried at each split: 5 OOB estimate of error rate: 29.76%	
	True	False	Classification Error
True	167	43	0.204
False	82	128	0.390

Figure 10. Summary Statistics on the First Round classifier

6.2 Predictions on First Round Draft

To use the random forest classifier above to predict whether players will be drafted in the first round, we first add the indicator variable "First Round" in the test dataset that contains the 2013 and the 2014 draft and then impute the missing values with the MICE package as mentioned above. Because the predictions of random forest are the probabilities of whether a player will be drafted in the first round, we take 0.5 as the threshold and classify players as first

rounders when the probability value is greater than 0.5 and second rounders when the value is less than 0.5. The predictions are demonstrated as a confusion matrix as shown in Table [].

	False	True	Classification Error
False	40	20	33.3%
True	28	32	53.3%

Table 8. Prediction error of First Round Classifier on the test set

We noticed that both the false positive and the false negative rate are relatively higher on the test set compared with the training set. The overall misclassification error is 40%, which is very high. To dive deeper, we assess whether this is due to the high misclassification error for players who are drafted between No.21 and No.40, essentially the last 10 draft picks of the first round and the first 10 picks of the second round because players with draft ranks of this range can be more sensitive to the prediction outcome. Thus, we broke down the confusion matrix for players drafted between No.21 and No.40 and players drafted outside this range in Table 9.

	False	True	Classification Error
False	11	9	45.0%
True	13	7	65.0%

	False	True	Classification Error
False	29	11	27.5%
True	15	25	37.5%

Table 9. Prediction error on players drafted from No.21 to 40 and those outside this range

The false positive and false negative rate for players drafted between N0.21 and 40 are significantly higher than the respective errors on the training data. However, these two rates for players drafted outside this range are both lower compared with the respective errors of the training data. Thus, the model is better at classifying players who are not projected to be selected in the middle of the first round and the second round of draft.

6.3 Stochastic Gradient Boosting

6.3.1 Model Building

To improve the prediction accuracy of the random forest classifier, we use stochastic gradient boosting as the trees in this model grow sequentially based on information previously grown trees to minimize the error of the previous models.[23] We implement boosting through the Generalized Boosted Regression Model (GBM) package in R by using boosted regression on the same set of predictors and the outcome variable in the random forest model. We use the bernoulli distribution because the "First Round" categorical variable is binary. We conducted parameter tuning by starting with a high number of iterations at 5000 to capture the optimum number of iterations later in the Deviance Change v. s. Iteration plot and used a default interaction level of 1, bag.fraction of 0.5, and n.minobsinnode of 1 while minimizing the shrinkage to 0.001 to reduce the learning rate. From the OOB Bernoulli Deviance Change v. s. Iteration plot below, we determine that the optimum number of iterations is 2733.

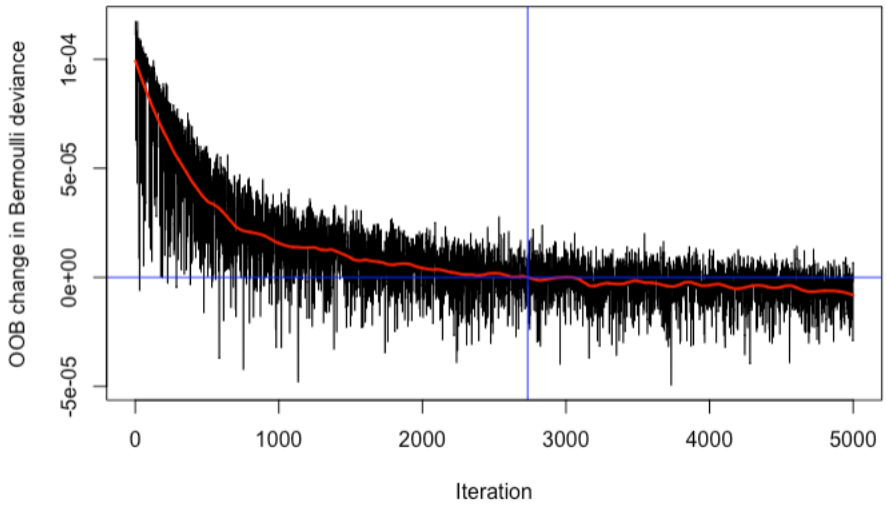

Figure 11. OOB Bernoulli Deviance Change v. s. Iteration plot

[23] James, 2014

By applying this number of iterations to visualize the summary statistics of the boosted regression in standardized units, we are able to generate a relative influence plot in Figure 12, which is equivalent to the Variable Importance Plot in random forest on the next page.

Variable	Relative Influence
Class Year	60.2
SOS	13.2
FTA	5.8
SOS.1	5.4
PTS	4.2

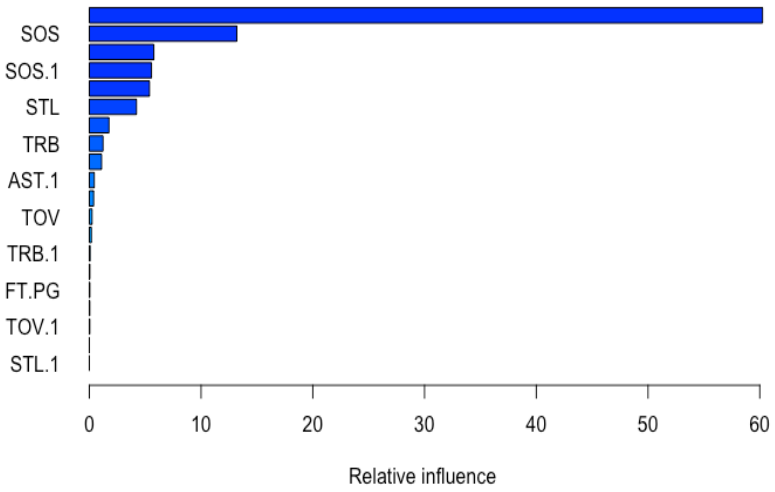

Figure 12. Relative Influence of variables

Because the variable names on the y-axis are only shown partially, the table portion of Figure 12 shows the values of relative influence, with Class Year having a domineering effect due to a value of 60.2, way higher than its relative importance in the variable important plot earlier. This is because boosting uses previously grown trees as information to grow new trees, which minimizes the errors of past models and places emphasis on important variables on early trees.

6.3.2 Model Predictions

After getting the optimum number of iterations, we re-run the boosting algorithm by changing the number of iterations while keeping all other parameters identical to the last model. The predicted results on the test set are shown below in Table 10.

	False	True	Classification Error
False	43	17	28.3%
True	22	38	36.7%

Table 10. Prediction error of boosted classifier on the test set

The new model reduces the false positive and false negative rate by 5% and 17% respectively. The false positive rate 28.3% is now slightly lower than that of the random forest classifier on the training set. Meanwhile, the overall classification error has been reduced from 40% to 34.2%. Breaking down the improve in prediction into players drafted among ranks No.21-40 and players outside this range in Table 11, we find that the false positive rate for both cases go down by 5% and the false negative rate for both cases go down by 10%. The misclassification error for players drafted between No.21 and No.40 thus is still high while the prediction for players drafted outside this range is more accurate due to a low misclassification error.

	False	True	Classification Error
False	12	8	40.0%
True	11	9	55.0%

	False	True	Classification Error
False	29	11	22.5%
True	15	25	27.5%

Table 11. Prediction error on players drafted from No.21 to 40 and those outside this range

Chapter 7
Summary and Conclusion

In this research, we have examined the extent to which the NBA draft rank and the career success for NBA can be predicted from pre-draft statistics. The pre-draft statistics consist of college statistics, player position and the class year that indicates the timing of the draft. The college statistics are broken down into two parts: the statistics of the most recent college year and the average statistics of the previous college years combined excluding the most recent college year.

To predict draft rank, we employed the multiple linear regression model, random forest and stochastic gradient boosting. To predict career success, we employed the linear regression model by using WS/48, BPM and VORP as outcome variables. With these different modeling approaches, we find that the pre-draft statistics that are most predictive of NBA draft rank are not necessarily the most predictive measure of the career success in NBA in Table 12 below, where "X" means that the corresponding variable is included in the selected model. For instance, SOS, BLK and PTS were selected as predictive of draft rank but were not important in predicting career success in any of the three models with different outcome variables. STL was selected as predictive of WS/48, but not of BPM and VORP, the two other indicators of career success. Class Year and FG.PG are the only two variables predictive of both draft rank and all three indicators of career success. On a higher level, the models show that statistics from the last year of college are more important than the statistics from the previous college years combined because more predictors in the models come from the most recent college year.

Variable	Draft Rank	WS/48	BPM	VORP
Class Year	X	X	X	X
Position		X	X	X
FG.PG	X	X	X	
FG.PG1		X		X
FT.PG1		X	X	
TOV		X		
TRB		X	X	X
STL	X	X		
TRB.1			X	
STL.1			X	
TOV.1			X	
P3.PG	X			X
AST				X
P3M.1				X
AST.1				X
STL.1	X			X
BLK	X			
PTS	X			
SOS	X			

Table 12. Selected Predictors of draft rank, WS/48, BPM and VORP model

These findings suggest that players could benefit from drafting after sophomore year compared with delaying till the end of junior and senior year based on a better draft rank and better career success as shown in all four models. The PG position is also favored over the other four positions in both draft rank and career success in the modern NBA era and versatile players could consider switching positions to play the point guard position. For players who would like to have a more successful career, they should work hard to improve their field goal percentage and rebounding ability especially.

In terms of prediction, the models on career success has a higher prediction accuracy than the models on draft rank. Among the career success models, WS/48 not only best predicts the career success but also best predicts the ideal draft rank, regardless of whether this ideal draft rank is ordered by the actual WS/48 or actual BPM values. For the model on draft rank, it accurately predicts neither the actual rank nor the ideal draft rank. Stochastic gradient boosting does a better job than random forest when classifying whether players will be drafted in the first

round. However, the classification of both random forest and boosting is visibly more accurate for players drafted outside the range between No.21 and No.40, the middle of the draft.

Although the linear regression models predict career success relatively well, they cannot predict the draft and career success of freshmen and international players because they either do not have statistics from the previous college years combined after playing in college for one year, or do not have any college statistics in NCAA due to an international background. Meanwhile, recoding the Class Year variable is an imperfect approach since the changes in reference levels could adjust the values of coefficient estimates and affect predictions. By recoding juniors, freshman and international players in the "Other" category could lead to bias in the prediction of juniors because the prediction on juniors also take into account the statistics of freshmen and international players.

We should mention that there are several factors that are not taken into account by our prediction models. First, because players can perform very differently when transitioning from college to NBA, adding statistics from the first few NBA years as predictors may take into the performance differences between college and NBA and improve the prediction on career success. Second, the predictions are conducted upon the 2013 and 2014 draft. The small sample size of prediction can mean that the accuracy of predictions can vary across different draft years.

Due to a higher difficulty to predict draft success than career success, we should consider adding NBA draft combine data such as height, BMI and vertical jump as predictors to better predict the draft rank. These predictors along with the college statistics from the most previous year can help predict the draft rank and the career success of players who declare draft after freshman year of college, a class year that cannot be predicted by the linear models in the paper. Moreover, given the relatively high classification error of random forest and boosting on whether players will be drafted in the first round, we can also collect statistics of undrafted players. Because there will be many more undrafted players compared with those who got drafted, the amount of data can help build a better classifier on whether players will be drafted

by tuning parameters such as sample size. The caveat is that there is no single source that includes all the information about undrafted plyers such as college statistics especially because some players announce their declaration of NBA draft and then later back out of the draft.

The reason for breaking down college statistics into the most recent year and the previous years combined was the increasing scrutiny of scouts on potential draftees and the higher stake of NBA draft. As player contracts are getting increasingly expensive over the past seasons, teams and scouts are paying higher attention on college players to see if they have improved their game from previous years after certain players chose to stay for extra years of college instead of entering NBA draft. Both draft timing and college statistics are crucial to the draft success and career success of players who aspire to leave an imprint in the NBA history. Though we only analyzed NBA players in this research, the general methodology could be applied to other professional sports such as hockey and football.

Appendices

Appendix I.
Formula of PER from Basketball Reference

```
uPER = (1 / MP) *
    [ 3P
    + (2/3) * AST
    + (2 - factor * (team_AST / team_FG)) * FG
    + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG)))
    - VOP * TOV
    - VOP * DRB% * (FGA - FG)
    - VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT)
    + VOP * (1 - DRB%) * (TRB - ORB)
    + VOP * DRB% * ORB
    + VOP * STL
    + VOP * DRB% * BLK
    - PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP) ]

factor = (2 / 3) - (0.5 * (lg_AST / lg_FG)) / (2 * (lg_FG / lg_FT))
VOP    = lg_PTS / (lg_FGA - lg_ORB + lg_TOV + 0.44 * lg_FTA)
DRB%   = (lg_TRB - lg_ORB) / lg_TRB
```

Appendix II.
Correlation Analysis of Key Variables

	FG	FGA	FG.PG
FG	1	0.909	-0.071
FGA	0.909	1	-0.463
FG.PG	-0.071	-0.463	1

	FG.1	FGA.1	FG.PG1
FG.1	1	0.945	-0.033
FGA.1	0.945	1	-0.3178
FG.PG1	-0.033	-0.318	1

	FT	FTA	FT.PG
FT	1	0.936	0.471
FTA	0.936	1	0.158
FT.PG	0.471	0.158	1

	P3M	P3MA	P3.PG
P3M	1	0.985	0.700
P3MA	0.985	1	0.604
P3.PG	0.700	0.604	1

Appendix III. Model Prediction on Draft Rank, WS/48 and BPM for 2013 and 2014 Draft

Player	ClassYear	Rank	fitted_Rank	Rank (a-p)^2	actual_WS48	fitted_WS48	WS48 (a-p)^2	actual_BPM	fitted_BPM	BPM (a-p)^2
Al Horford	other	3	31	771	0.16	0.12	0.001	3.2	0.8	6.0
Jeff Green	other	5	29	553	0.07	0.05	0.001	-1.4	-2.4	1.0
Corey Brewe	other	7	31	556	0.06	0.05	0.000	-1	-3.0	4.1
Joakim Noah	other	9	13	17	0.16	0.12	0.001	4.1	0.0	17.0
Acie Law	Senior	11	42	968	0.03	0.01	0.000	-5.4	-3.5	3.7
Julian Wrigh	Sophomore	13	22	78	0.06	0.09	0.001	-1.3	-2.1	0.7
Al Thornton	Senior	14	22	71	0.04	0.04	0.000	-3.5	-3.7	0.1
Rodney Stucl	Sophomore	15	21	34	0.08	0.06	0.000	-1.4	-2.8	1.9
Nick Young	other	16	29	175	0.05	0.05	0.000	-3.4	-1.7	2.7
Jason Smith	other	20	31	129	0.07	0.09	0.000	-2.7	-2.0	0.5
Jared Dudley	Senior	22	55	1063	0.10	0.02	0.007	1	-3.1	16.4
Wilson Chan	Sophomore	23	27	18	0.07	0.06	0.000	-0.8	-1.7	0.8
Morris Almo	Senior	25	27	3	0.01	0.03	0.001	-6.4	-4.5	3.6
Aaron Brook	Senior	26	38	151	0.07	0.04	0.001	-1.8	-4.0	4.8
Arron Afflalo	other	27	32	26	0.08	0.03	0.002	-1.1	-4.3	10.1
Alando Tucke	SuperSenior	29	45	256	0.05	0.02	0.000	-6.7	-5.4	1.6
Carl Landry	other	31	18	165	0.15	0.10	0.003	-1.1	-2.5	1.9
Gabe Pruitt	other	32	38	36	0.04	0.02	0.001	-3.1	-3.4	0.1
Marcus Willi	Sophomore	33	20	168	-0.08	0.07	0.023	-7.5	-2.4	26.0
Nick Fazekas	Senior	34	32	4	0.15	0.12	0.001	-0.1	0.4	0.2
Glen Davis	other	35	28	56	0.08	0.09	0.000	-2.3	0.0	5.4
Jermareo Da	Senior	36	41	26	0.00	0.05	0.003	-7.6	-4.2	11.9
Josh McRobe	Sophomore	37	18	374	0.11	0.11	0.000	1	-1.2	5.0
Derrick Byars	Senior	42	36	42	0.15	0.02	0.016	0.6	-5.2	33.8
Dominic McC	Senior	47	37	90	0.04	0.00	0.001	-1.6	-5.0	11.4
Aaron Gray	Senior	49	41	67	0.08	0.08	0.000	-1.8	-2.9	1.1
Taurean Gre	other	52	39	175	-0.07	0.01	0.006	-8.2	-4.6	13.1
Demetris Nic	Senior	53	32	441	-0.20	-0.01	0.036	-15.3	-4.8	110.0
Ramon Sessi	other	56	48	67	0.09	0.03	0.003	-1.9	-2.1	0.0
D.J. Strawbe	Senior	59	32	708	-0.03	0.02	0.002	-5.1	-5.1	0.0
Adam Morris	other	3	14	114	-0.02	0.05	0.005	-5.5	-2.8	7.4
Shelden Will	Senior	5	11	38	0.09	0.13	0.002	-2.8	-1.1	2.8
Brandon Roy	Senior	6	29	542	0.16	0.03	0.014	3.2	-4.1	52.6
Randy Foye	Senior	7	37	898	0.06	0.02	0.002	-1.3	-4.8	12.6
Rudy Gay	Sophomore	8	15	55	0.09	0.07	0.000	0.5	-2.4	8.5
J.J. Redick	Senior	11	24	172	0.13	0.00	0.016	0	-4.8	22.6
Hilton Armst	Senior	12	30	326	0.06	0.03	0.001	-2.8	-4.2	1.9
Ronnie Brew	other	14	24	99	0.13	0.04	0.008	1.7	-2.0	13.9
Cedric Simm	Sophomore	15	8	51	0.01	0.07	0.004	-5.7	-3.8	3.5
Rodney Carn	Senior	16	40	595	0.06	0.00	0.004	-2.7	-4.0	1.7
Quincy Doub	other	19	13	36	0.00	0.06	0.004	-5.4	-1.9	12.0
Renaldo Balk	other	20	23	8	0.11	0.06	0.003	0.9	-2.1	9.2
Rajon Rondo	Sophomore	21	33	141	0.11	0.10	0.000	1.4	0.9	0.3
Marcus Willi	other	22	43	448	0.00	0.00	0.000	-4.6	-4.5	0.0
Kyle Lowry	Sophomore	24	20	12	0.16	0.08	0.006	3.8	-1.2	25.0
Shannon Bro	other	25	29	13	0.06	0.03	0.001	-2.3	-2.8	0.3
Jordan Farm	Sophomore	26	35	75	0.07	0.03	0.002	-1.4	-4.4	9.2
Maurice Age	Senior	28	39	130	-0.09	-0.01	0.006	-10.4	-4.8	31.9
Mardy Collin	Senior	29	37	57	-0.03	0.00	0.001	-4.9	-3.2	2.8
James White	Senior	31	37	35	0.06	0.00	0.004	-3.4	-4.7	1.6
Steve Novak	Senior	32	40	72	0.12	0.07	0.003	-0.9	-3.8	8.4
Solomon Jon	Sophomore	33	23	102	0.08	0.07	0.000	-2.9	-3.3	0.2
Paul Davis	Senior	34	36	2	0.04	0.08	0.002	-4.7	-2.0	7.4
P.J. Tucker	other	35	19	269	0.09	0.09	0.000	0.9	0.6	0.1
Craig Smith	Senior	36	39	6	0.11	0.09	0.001	-1.5	-1.6	0.0
Bobby Jones	Senior	37	49	144	0.06	0.03	0.001	-3.5	-2.9	0.4
David Noel	Senior	39	41	2	0.02	0.02	0.000	-3.4	-4.1	0.4
James Augus	Senior	41	31	95	0.10	0.13	0.001	-2.2	-0.4	3.1
Alexander Jo	other	45	31	195	0.06	0.05	0.000	-4.3	-4.7	0.1
Dee Brown	Senior	46	50	14	0.01	0.01	0.000	-4.1	-3.7	0.2
Paul Millsap	other	47	10	1339	0.15	0.16	0.000	3.2	2.1	1.2
Leon Powe	other	49	23	670	0.17	0.08	0.009	-2.3	-0.1	4.9
Hassan Adan	Senior	54	27	740	0.07	0.05	0.001	-4.4	-2.4	3.8
Will Blalock	other	60	28	997	-0.03	0.03	0.004	-5.7	-1.9	14.8

Appendix IV. Model Prediction Errors for Ideal Draft Rank Based on Actual WS/48

Player	Ideal Draft Rank	actual_Rank*	MAE	Rank_FittedRank	MAE	Rank_FittedWS48	MAE	Rank_fittedBPM	MAE
Al Horford	2	1	1	14	12	2	0	11	9
Jeff Green	13	2	11	11	2	15	2	24	11
Corey Brewer	18	3	15	13	5	16	2	18	0
Joakim Noah	1	4	3	1	0	1	0	1	0
Acie Law	24	5	19	27	3	27	3	29	5
Julian Wright	17	6	11	6	11	7	10	13	4
Al Thornton	23	7	16	7	16	18	5	5	18
Rodney Stuckey	10	8	2	5	5	11	1	8	2
Nick Young	19	9	10	12	7	14	5	12	7
Jason Smith	16	10	6	15	1	8	8	30	14
Jared Dudley	7	11	4	30	23	25	18	20	13
Wilson Chandler	14	12	2	9	5	12	2	21	7
Morris Almond	25	13	12	8	17	19	6	27	2
Aaron Brooks	15	14	1	23	8	17	2	23	8
Arron Afflalo	11	15	4	17	6	21	10	17	6
Alando Tucker	20	16	4	28	8	22	2	2	18
Carl Landry	3	17	14	3	0	5	2	16	13
Gabe Pruitt	21	18	3	22	1	24	3	3	18
Marcus Williams	29	19	10	4	25	10	19	9	20
Nick Fazekas	4	20	16	18	14	3	1	4	0
Glen Davis	9	21	12	10	1	6	3	7	2
Jermareo Davidson	26	22	4	26	0	13	13	15	11
Josh McRoberts	6	23	17	2	4	4	2	10	4
Derrick Byars	5	24	19	20	15	23	18	19	14
Dominic McGuire	22	25	3	21	1	29	7	22	0
Aaron Gray	12	26	14	25	13	9	3	6	6
Taurean Green	28	27	1	24	4	28	0	26	2
Demetris Nichols	30	28	2	16	14	30	0	28	2
Ramon Sessions	8	29	21	29	21	20	12	14	6
D.J. Strawberry	27	30	3	19	8	26	1	25	2

Appendix V. Model Prediction Errors for Ideal Draft Rank Based on Actual BPM

Player	Rank_actualBPM	Rank_fittedBPM	MAE	Rank_FittedWS48	MAE
Al Horford	2	11	9	2	0
Jeff Green	12	24	12	15	3
Corey Brewer	8	18	10	16	8
Joakim Noah	1	1	0	1	0
Acie Law	24	29	5	27	3
Julian Wright	11	13	2	7	4
Al Thornton	22	5	17	18	4
Rodney Stuckey	13	8	5	11	2
Nick Young	21	12	9	14	7
Jason Smith	19	30	11	8	11
Jared Dudley	3	20	17	25	22
Wilson Chandler	7	21	14	12	5
Morris Almond	25	27	2	19	6
Aaron Brooks	15	23	8	17	2
Arron Afflalo	9	17	8	21	12
Alando Tucker	26	2	24	22	4
Carl Landry	10	16	6	5	5
Gabe Pruitt	20	3	17	24	4
Marcus Williams	27	9	18	10	17
Nick Fazekas	6	4	2	3	3
Glen Davis	18	7	11	6	12
Jermareo Davidson	28	15	13	13	15
Josh McRoberts	4	10	6	4	0
Derrick Byars	5	19	14	23	18
Dominic McGuire	14	22	8	29	15
Aaron Gray	16	6	10	9	7
Taurean Green	29	26	3	28	1
Demetris Nichols	30	28	2	30	0
Ramon Sessions	17	14	3	20	3
D.J. Strawberry	23	25	2	26	3

Bibliography

"About Box Plus/Minus (BPM)." *Basketball*, www.basketball-reference.com/about/bpm2.html.

Badr, Will. "6 Different Ways to Compensate for Missing Data (Data Imputation with Examples)." *Medium*, Towards Data Science, 12 Jan. 2019, towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

Basketball Reference. "NBA & ABA Career Leaders and Records for Win Shares." *Basketball*, www.basketball-reference.com/leaders/ws_career.html.

Dauster, Rob, and Scott Phillips. "In The Money: Second-Round Picks Cash-in Guaranteed Contracts at Exceedingly High Rates - CollegeBasketballTalk: NBC Sports." *CollegeBasketballTalk | NBC Sports*, 10 May 2018, collegebasketball.nbcsports.com/2018/05/10/in-the-money-second-round-picks-cash-in-guaranteed-contracts-at-exceedingly-high-rates/.

DuBose, Ben. "Former No. 1 Pick Anthony Bennett Waived by Houston Rockets." *USA Today*, Gannett Satellite Information Network, 10 Oct. 2019, rocketswire.usatoday.com/2019/10/09/former-no-1-pick-anthony-bennett-waived-by-houston-rockets/.

Edwards, Ryan, et al. "Using Pre-NBA Draft Data to Project Success in the NBA." 2015, http://cs229.stanford.edu/proj2015/120_report.pdf.

"Evolution of the Draft and Lottery." *Wayback Machine*, web.archive.org/web/20101203184544/www.nba.com/history/draft_evolution.html.

Favale, Dan. "How NBA's Salary-Cap Increases Will Affect 2019 and 2020 Offseasons." *Bleacher Report*, Bleacher Report, 20 Sept. 2018, bleacherreport.com/articles/2796634-how-nbas-salary-cap-increases-will-affect-2019-and-2020-offseasons.

Favale, Dan. "5 Worst Contracts Still on the Books from Ridiculous 2016 Free Agency." *Bleacher Report*, Bleacher Report, 16 June 2019, bleacherreport.com/articles/2840764-5-worst-contracts-still-on-the-books-from-ridiculous-2016-free-agency.

Greene, Alexander C., "The Success of NBA Draft Picks: Can College Careers Predict NBA Winners?" (2015). Culminating Projects in Applied Statistics. 4.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Knight, Brett. "NBA's Most Overpaid Players 2019: Wiggins, Jabari And 8 Others Who Underperformed." *Forbes*, Forbes Magazine, 12 Apr. 2019, www.forbes.com/sites/brettknight/2019/04/12/nba-overpaid-wiggins-jabari/#3e720d9076e3.

Mulholland, Jason, and Shane T. Jensen. "Predicting the Draft and Career Success of Tight Ends in the National Football League." *Journal of Quantitative Analysis in Sports*, vol. 10, no. 4, Jan. 2014, doi:10.1515/jqas-2013-0134.

"NBA & ABA Career Leaders and Records for Win Shares Per 48 Minutes." *Basketball*, www.basketball-reference.com/leaders/ws_per_48_career.html.

O'Connor, Kevin. “The Ringer's 2020 NBA Draft Big Board.” *The Ringer's 2020 NBA Draft Guide*, nbadraft.theringer.com/.

O'Donnell, Ricky. “Projected Lottery Pick Ivan Rabb Skipping NBA Draft to Return to Cal.” *SBNation.com*, SBNation.com, 25 Apr. 2016, www.sbnation.com/college-basketball/2016/4/25/11500318/ivan-rabb-2016-nba-draft-cal-bears.

“Player Efficiency Rating.” *Wikipedia*, Wikimedia Foundation, 16 Apr. 2020, en.wikipedia.org/wiki/Player_efficiency_rating.

Riedel, Charlie. “Big 12 Player of Year Hield Returning to Oklahoma for Senior Season.” *FOX Sports*, 24 Apr. 2015, www.foxsports.com/college-basketball/story/oklahoma-sooners-buddy-hield-returning-senior-season-no-nba-042415.

Rill, Jake. “2020 NBA Mock Draft: Predictions and Analysis for Top Prospects Available.” *Bleacher Report*, Bleacher Report, 20 Apr. 2020, bleacherreport.com/articles/2887471-2020-nba-mock-draft-predictions-and-analysis-for-top-prospects-available.

RotoWire Staff Jan 23. “Ivan Rabb: Added by Westchester.” *CBSSports.com*, 23 Jan. 2020, www.cbssports.com/fantasy/basketball/news/ivan-rabb-added-by-westchester/.

Staw, Barry M., and Ha Hoang. “Sunk Costs in the NBA: Why Draft Order Affects Playing Time and Survival in Professional Basketball.” *Administrative Science Quarterly*, vol. 40, no. 3, 1995, p. 474., doi:10.2307/2393794.

Urbina, Frank. “The Unforgettable Manu Ginobili's Craziest Career Facts and Numbers.” *HoopsHype*, 5 Nov. 2018, hoopshype.com/2018/08/27/manu-ginobili-retirement-career-personal-accolades/.