Analyzing the Relationship Between College Stats, Draft Performance, and Career Longevity

Ethan Senatore

May 16, 2024

Contents

List of Figures

List of Tables

1 Introduction

Every year men's college basketball players across the country start the season with one dream-like goal in mind; play well enough to get drafted by an NBA team. Each player has to meet certain standards to be considered elite enough to warrant an NBA roster spot and faces near insurmountable odds to do so. In my report, I first investigate the relationship between a player's draft position and their performance metrics in his final year of collegiate basketball.

Assuming a player does indeed get drafted, he now faces even tougher odds to remain within the NBA for more than four years. It is notoriously difficult to make it past your 'rookie contract' in the NBA as there is a never-ending influx of incredibly talented players into the league. In the second half of my report, I seek to investigate if a relationship exists between a player's stats in his final year of collegiate basketball and whether they receive a second contract in the NBA.

1.1 Research questions

- Regression

How do performance metrics from a player's final year in collegiate basketball influence their draft position?

- Classification

Do performance metrics from a player's final year in collegiate basketball predict their potential to "bust"?

- Important Variables

In my report I uncover different variables that play a larger role in predicting draft performance and career longevity. Pts (Points per game), Ast (Assists per game), $Conf$ (Conference), and Yr (Academic year) are some of the variables that I highlighted as potential major players in achieving accurate predictions. However, as I will note in my report, variables that may appear to be important at face-value do not play as much of a role as I had anticipated.

1.2 Data set desription

I sourced my data from Kaggle and found two sources that I thought were easily interpretable and detailed enough for my analysis. I first found a data-set containing contract information

of NBA players from 1990 - 2017. There was little to no information regarding how this data was collected, but I can assume the author used publicly available information regarding NBA player's salaries. While I don't use most of the variables, this data-set gave us the necessary information in cataloging all the players that played in the NBA for more than 4 years.

This data-set can be found using this link:

https://www.kaggle.com/datasets/whitefero/nba-player-salary-19902017

My second data-set is a collection of individual basketball statistics of all collegiate players from 2009 - 2021. The author, Bart Torvik, chose to compile this data-set as a project that combined two of his passions - data science and basketball. This is an incredibly complex data-set with season average statistics on ~25,000 players. I focused on just 670 as they are the only player's that were drafted during this period. All of the relevant performance metrics along with the pick in which they were chosen are held in this data-set and proved to the backbone of my statistical analysis.

This data-set can be found using this link:

https://www.kaggle.com/datasets/adityak2003/college-basketball-players-20092021?select=CollegeBasketballPlayers2009-2021.csv

2 Statistical Methods

2.1 Regression

Before performing any analysis, I cleaned up the data to remove any variables that were either correlated with other variables or contained a substantial amount NA values within their column.

Collinearity, or redudancy at the very least, between variables was expected as this data-set contained complex basketball statistics that were based off formulas that incorporated other variables included in the data set. For example, usg (Usage Percentage), is a stat that measures the percentage of a team's possessions a player uses while on the court. It is calculated using this equation:

$$usg = 100 * (\frac{(FGA + 0.44FTA + TO) * (TMP/5)}{(MP * (TFGA + 0.44TFTA + TTO))})$$

Variables such as FGA, FTA, and TO are all variables that appear in my data-set resulting in collinearity with usg. This was just one of the many examples collinearity or redundancy resulting in us omitting a variable from my analysis. In the end, I elected to include the variables in Table

1 in my regression analysis as I believed they would be the most relevant predictors for draft performance:

Table 1: Variables Used in Analysis

Variable	Description
conf	Conference
yr	Academic Year
ftr	Free-Throw Rating
GP	Games Played
eFG	Effective Field Goal Percentage
TP_per	Three-Point Percentage
blk_per	Block Percentage
stl_per	Steal Percentage
pfr	Personal Foul Rating
asttov	Assist-Turnover Ratio
pick	Pick Drafted
drtg	Defensive Rating
oreb	Offensive Rebounds
dreb	Defensive Rebounds
ast	Assists per Game
stl	Steals per Game
blk	Blocks per Game
pts	Points per Game

My final adjustment to my data-set was to exclude any conferences that had less than 20 players drafted in them. My reasoning behind this decision lies in the fact that the few who make it out of the smaller name conferences are outliers compared to their peers. Take Damian Lillard for example, he is now a star player in the NBA yet he attended Weber St. University that plays out of Big Sky Conference. His presence in my data-set would have skewed my analysis resulting in a less accurate model.

As I previously noted, I suspected that the conference in which a player competes in has a major impact on their draft potential. The logic is that the stronger the competition a player plays against, the more it would translate into the professional game. See Figure 1 for a clearer view into my data-set and how each conference stacks up against each other.

I gathered that if a player competes in the ACC, SEC, Pac-12, Big 10, Big 12, and Big East they face better odds to get drafted. Moreover, below in Figure 2 I highlight the number of lottery

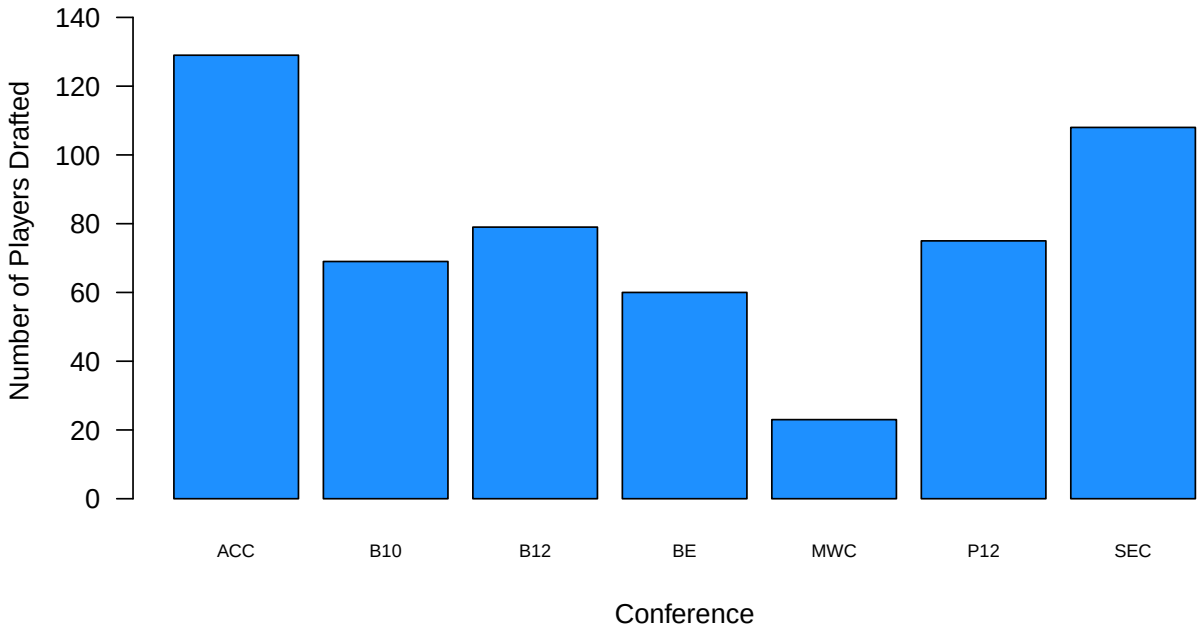

Figure 1: Number of Players Drafted per Conference

draft picks (picked 1-14 in the draft) for each conference. The figure suggests that not only does a player have a better chance of being drafted if they compete in the aforementioned conferences, but they are also poised to get drafted ahead of their peers in other conferences.

A relationship between the conference in which a player competes in and their draft performance clearly exists. However, when inspecting the relationships between a player's performance metrics and the position in which they were drafted it becomes clear that there is certain level of nuance that hides the relationship from the plot. Take a player's points per game average compared with their draft position. Below in Figure 3, no clear relationship can be identified resulting in us needing to make use of statistical techniques to unveil them.

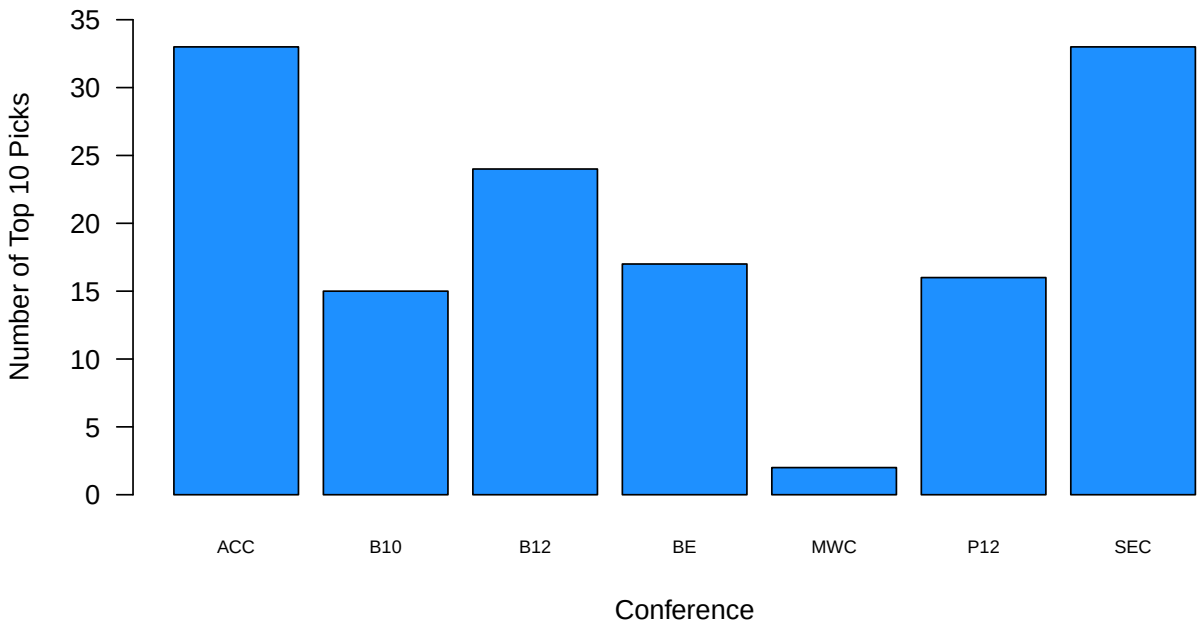

Figure 2: Number of Lottery Draft Picks per Conference

The nuance behind a player's draft position and their performance metrics is precisely the reason why I chose to investigate it. In my initial attempt to answer my regression question I chose to use a linear regression model including all the variables mentioned in Table 1. I elected to use this approach as it was widely covered in my course this semester, and offers a fair amount of insight into how each variable impacts the response variable. Below is the model equation I elected to use:

$$\begin{aligned}\hat{pick} = \beta_0 + \beta_1 I_{B10}(conf) + \beta_2 I_{B12}(conf) + \\ \beta_3 I_{BE}(conf) + \beta_4 I_{MWC}(conf) + \beta_5 I_{P12}(conf) + \\ \beta_6 I_{SEC}(conf)\beta_{12} I_{JR}(yr) + \beta_{13} I_{So}(Yr) + \\ \beta_{14} I_{SR}(yr) + \beta_{15} ftr + \beta_{16} GP + \beta_{17} eFG + \\ \beta_{18} TPper + \beta_{19} blkper + \beta_{20} stlper + \beta_{21} pfr + \\ \beta_{22} ast.tov + \beta_{23} drtg + \beta_{24} oreb + \beta_{25} dreb + \\ \beta_{26} ast + \beta_{27} stl + \beta_{28} blk + \beta_{29} pts\end{aligned}$$

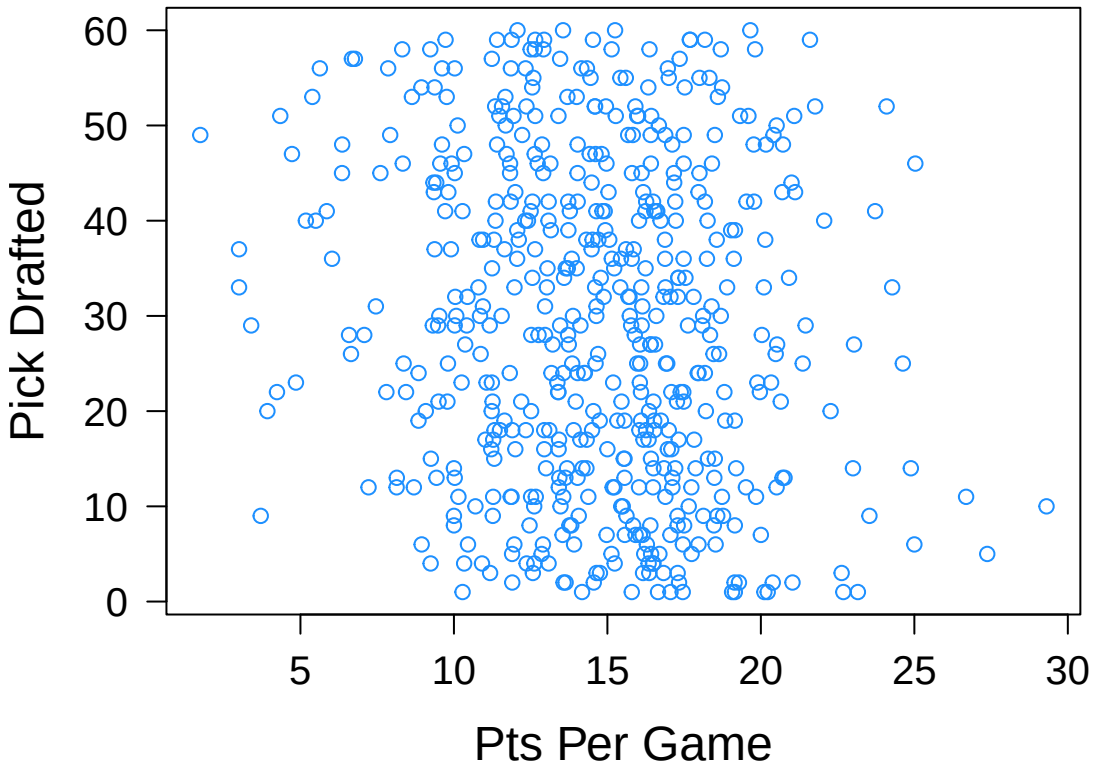

Figure 3: Draft pick versus points per game in final year of college career

For my second technique, I chose to used a cross-validated ridge regression model in order to identify the most meaningful variables within my data. I inputted the exact same coefficients as the ones used for the linear model. This process allowed us to highlight which aspects of a player's performance metrics had the greatest impact on his draft performance.

2.2 Classification

The first of my classification techniques used is a basic logistic regression model. I use the exact same variables listed in Table 1 with the one exception being that I include a variable called *scndContract*. This variable is either a 0 (did not receive a second contract) or a 1 (did receive a second contract) for collegiate basketball players that were drafted in the 2009 - 2012 seasons. Additionally, I are now only including players from conferences who had more than 10 players drafted due to my smaller sample size. For my logistic regression model, I did not use any cross-validation and included all variables in my model.

For similar reasons why I elected to use a linear regression model, I chose a logistic regression model due to my familiarity with it. I had covered it in my course and it proved to be a handy model that is simple to use and interpret.

The amount of player's that received a second contract compared to the amount who didn't is fairly close. In Figure 4, it can be seen that it is essentially an even split amongst the players who were drafted in my range of seasons.

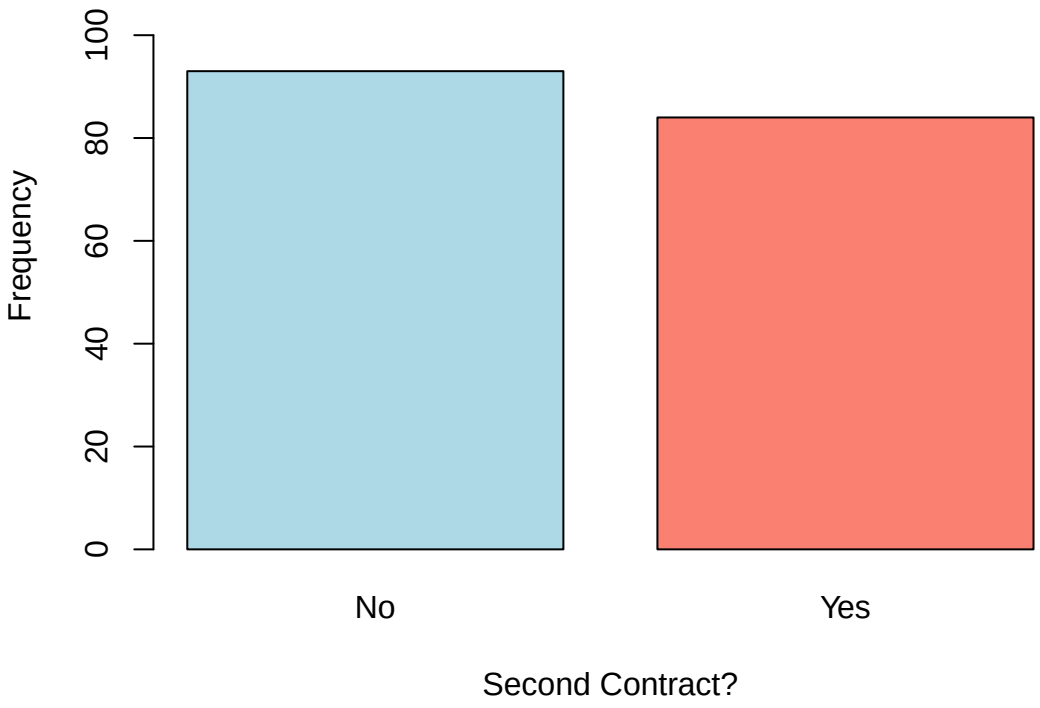

Figure 4: Number of players who received or did not receive their second contract (2009 - 2012)

Once again, I suspected that the conference in which a player competes in is a major factor in how long their career lasts. Typically, if a player comes out of an elite conference they are exposed to the professional lifestyle and regiment fairly early on. The "student" in student-athlete is far from the top priority for many of these players who have professional aspirations. They dive into a professional regiment early on and therefore are better equipped to last in the NBA. Below in Figure 5, I highlight the percentage of players drafted who did not receive their second contract per conference. Interestingly, the Pacific-10 conference has differentiated itself as a pro-producing conference alongside the ACC.

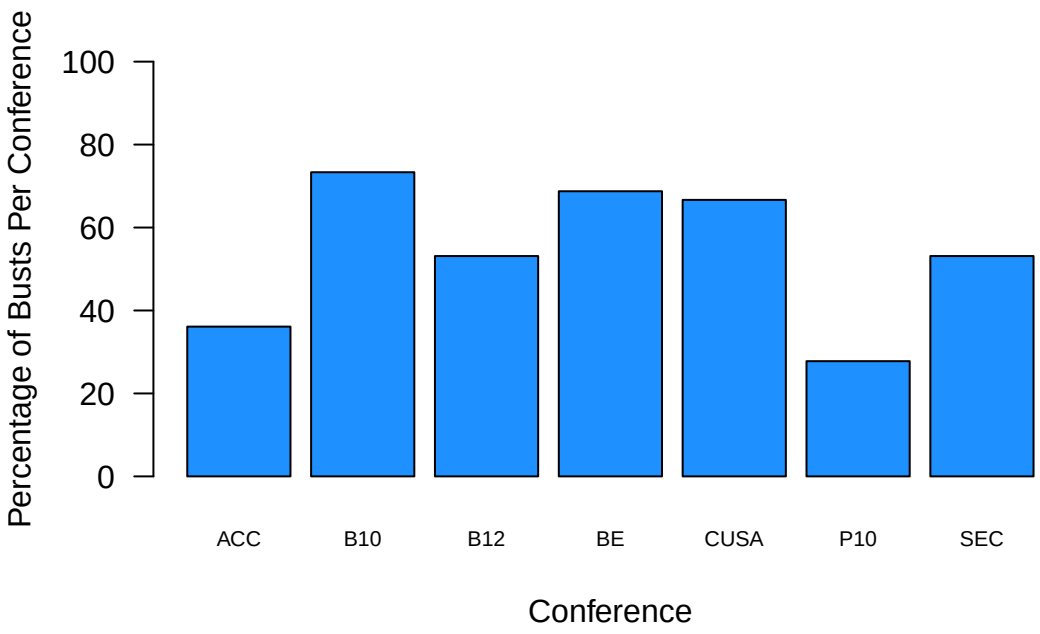

Figure 5: Number of players who did not receive their second contract per conference (2009 - 2012)

Another telling sign of a how well a college player translates into the professional game is how good of a scorer they are in college. Below in Figure 6, I highlight the spread of points per game for the category of players who receive and don't receive their second contract. It can be seen that a player who receives his second contract averages more points per game in college than his peer who doesn't.

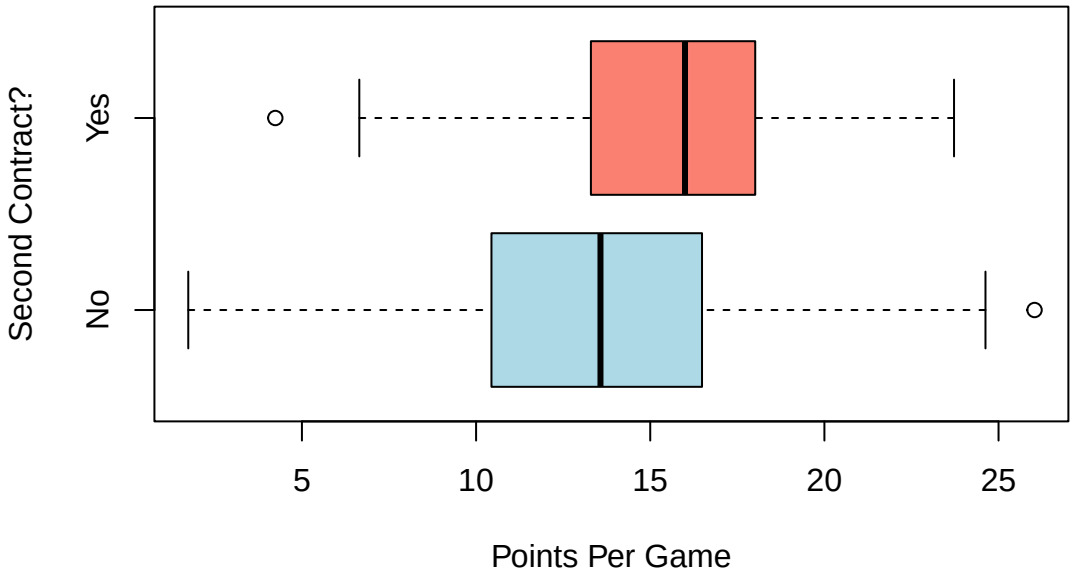

Figure 6: Boxplot of players points per game versus whether they receive a second contract or not (2009 - 2012)

My second technique used was trees and random forests. This approach allowed us to present an easily interpretable model based off of my data-set. The draw backs of using a tree was that it was non-robust and does not have the same level of predictive accuracy as the other models.

3 Results

3.1 Regression

The estimated final model for my linear model was:

$$\begin{aligned}\hat{pick} = 41.33 + 4.90I_{B10}(conf) + 1.41I_{B12}(conf) + \\ 0.27I_{BE}(conf) + 5.27I_{MWC}(conf) + 1.61I_{P12}(conf) + \\ 0.99I_{SEC}(conf) + 17.68I_{JR}(yr) + 11.24I_{So}(Yr) + \\ 25.80I_{SR}(yr) - 0.06ftr - 0.36GP - 0.73eFG \\ -2.48TPper + 2.24blkper - 2.71stlper + 0.35pfr + \\ 2.96ast.tov + 0.50drtg + -0.76oreb + 0.71dreb + \\ -1.80ast + 0.47stl + -9.29blk + -1.07pts\end{aligned}$$

My model did not perform well scoring a mean squared error of 177.66. Essentially, on average, I were off by 13 picks for my predictions. This could be for a variety of reasons, but my analysis determined a few key components that factored into producing a flawed model. First and foremost, I included 24 variables in my linear model which may have resulted in us overfitting to my data. Additionally, when reviewing the summary of my linear model only 10 variables would have been able to pass a p-test suggesting that only a handful of variables really matter. A few examples in particular that caught my attention were *yr* (academic year drafted in), *blk* (blocks per game), and *drtg* (defensive rating). The *yr* variable stood out as the single most important variable in my model as it indicated that the younger that you get drafted the better you will perform in the draft. Being a freshman in the ACC, you would, on average, be drafted 13 picks higher than a sophomore. You would be drafted 18 picks higher than a junior and a startling 27 picks higher than a senior. This made sense as, more often than not, a young freshman that is drafted is typically a highly touted prospect that treats collegiate basketball solely as a buffer year until they go pro.

The variables *blk* and *drtg* also proved to be insightful variables as they shed light on what is sought after in the NBA Draft. For context, *drtg* is calculated by estimating how many points a player would allow per 100 possessions when they're on the court, and a lower rating is better. Both *drtg* and *blk* had p-values < 0.05 and based off of the model I believe that this suggests a player's defensive prowess is highly sought after amongst NBA scouts. Consider that if a player averages 2 blocks per game they are, on average, going to be drafted a full 18 picks higher than their peer who averages none. Now, this is more than likely my data-set that inflated this value, and I suspect that if I took many more years of college basketball into consideration, blocks per game would have less of an impact on draft performance.

In my second model I chose to use a cross-validation approach to pick my best ridge regression model for my data. Figure 7, shows the various λ values and their respective MSE. In my case, the best λ value I found was 0.998.

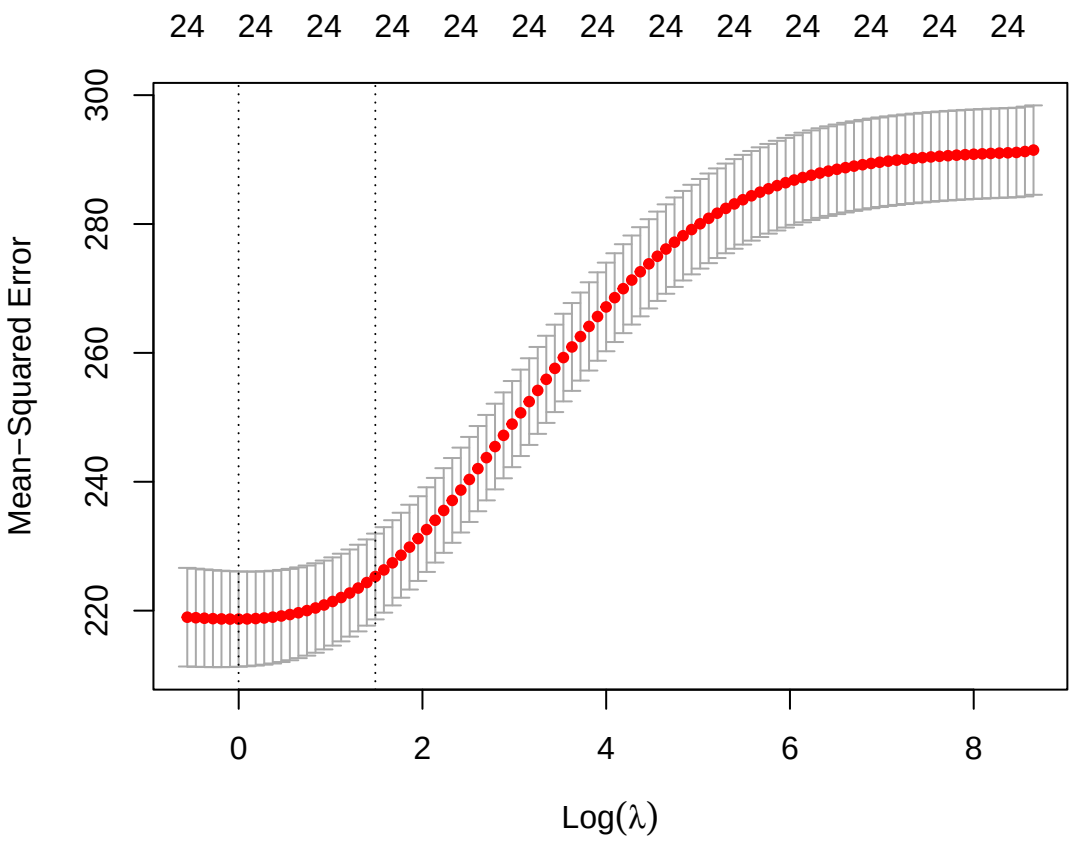

Figure 7: Cross-Validation insight onto log(lambda) versus Mean-Squared Error

Below is my estimated model using the cross-validated ridge regression approach:

$$\begin{aligned}\hat{pick} = 72.12 + 3.84I_{B10}(conf) + 0.63I_{B12}(conf) - \\ 0.20I_{BE}(conf) + 4.00I_{MWC}(conf) + 2.29I_{P12}(conf) - \\ 1.46I_{SEC}(conf) + 14.39I_{JR}(yr) + 7.47I_{So}(Yr) + \\ 20.51I_{SR}(yr) - 0.03ftr - 0.27GP - 0.58eFG - \\ 4.49TPper + 0.33blkper - 1.37stlper - 0.20pfr + \\ 1.38ast.tov + 0.14drtg + 0.15oreb - 1.16dreb + \\ -0.45ast - 2.45stl - 1.86blk - 0.92pts\end{aligned}$$

This model tells a very similar story as my previous one despite performing worse with an MSE of 222.31. It places significant importance on yr which is what I had anticipated. Interestingly, TP_Per (Three Point Percentage) is far more important in this model than the previous one.

Overall, I came to the conclusion that the academic year in which you declare for the draft is the most significant of factors when predicting draft performance. This is not to say that I believe if you are a freshman you are automatically more likely to get drafted, but I believe that if you are a freshman producing a performance output on the same level as older players you are far more likely to get drafted ahead of them. Another conclusion I reached was that when inspecting a data-set of solely drafted players, the performance metrics of each player will look fairly similar. This results in the models not putting much importance on statistics that I believed would play a pivotal role. Perhaps if I had included all players from college basketball certain performance metrics would have separated themselves as the tell-tale signs of an elite draft prospect.

3.2 Classification

Below is my estimated logistic regression model:

$$
\begin{aligned}
ln(\frac{p_{scndContract}}{1-p_{scndContract}}) = \\
-9.58 - 2.05I_{B10}(conf) - 1.06I_{B12}(conf) - 1.65I_{CUSA}(conf) + \\
0.46I_{P10}(conf) - 1.20I_{SEC}(conf) - 1.39I_{JR}(yr) - \\
1.76I_{So}(Yr) - 0.01ftr + 1.21I_{SR}(yr) - 0.03ftr + \\
0.10GP + 0.01eFG + 1.01TPper + 0.08blkper + \\
1.41stlper - 1.41pfr + 1.33ast.tov - 0.04pick + \\
0.03drtg + 0.83oreb - 0.03dreb - 0.14ast - \\
1.85stl - 0.05blk + 0.15pts
\end{aligned}
$$

My model performed better than I had anticipated. My confusion matrix indicated that I had achieved a test error rate of roughly 35%. For context, the probability prediction produced by the model was set to either 0 or 1 depending on if it was greater than or equal to 0.5. My confusion matrix can be seen below in Table 2.

Table 2: Confusion Matrix of Logistic Regression Model

	0	1
0	13	9
1	3	10

The coefficients in my model did not surprise us as many of my predictions on which would have the most impact came to light. Variables such as *yr* and *conf* had a tremendous impact on your chances of reaching a second contract. As I previously noted, the conference in which you play could have a massive developmental impact therefore impacting your career longevity. Additionally, as I saw in my regression analysis, the younger draft prospects seemed to fair better than their older peers. This could be attributed to the fact that a younger draft prospect has more upside to them resulting in teams betting on their long term development.

My second model, using trees and random forests proved to be a difficult undertaking. I first used a random forest model that highlighted the relevant variables to make predictions. According to the Figure 8, there's a considerable drop off after the first six, so I fit a new tree using only those six.

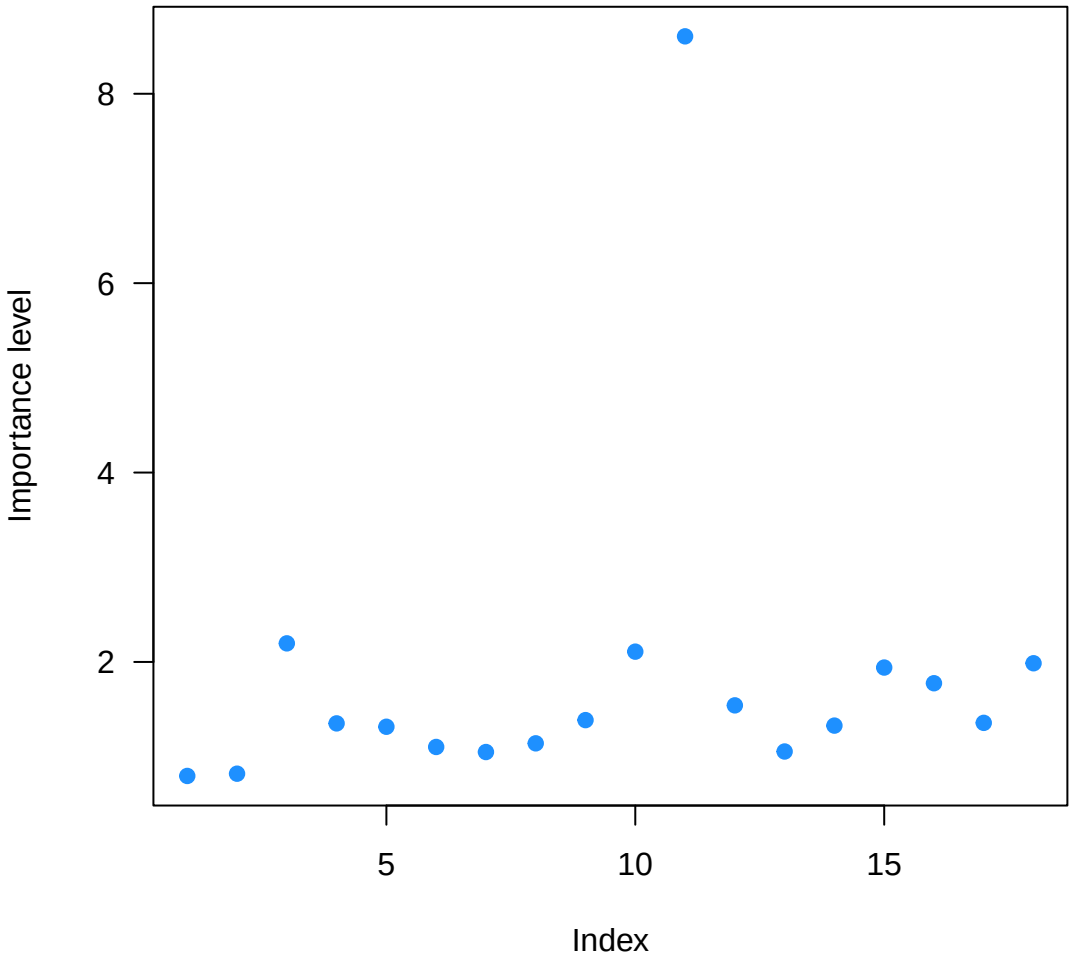

Figure 8: Importance Plot

The tree I fit can be seen in Figure 9

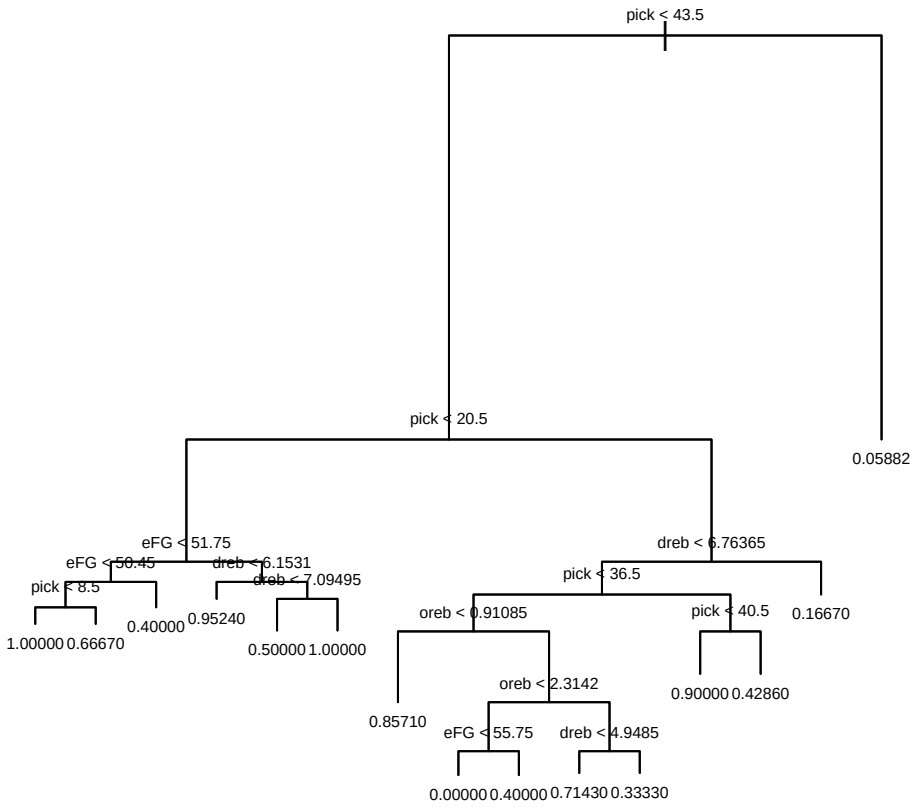

Figure 9: Final Tree

Table 3: Confusion Matrix of Tree Model

	0	1
0	17	6
1	5	7

According to Table 3 my test error rate was roughly 32%. This was marginally better than my previous model despite being a far more interpretable model. Moreover, it was interesting to

note which variables my random forests process selected to be the relevant predictors. This was an aspect that differed from all other models fitted in my report.

4 Conclusions

My analysis on college basketball proved to be a difficult task that demanded well-fitted models to handle my data. Unfortunately, my regression models proved to be less than useful as they could not handle the amount of variables present. The insight gained on each variable was by far the most useful portion of my regression analysis. As I noted in my report, a more interesting approach I could have taken was to analyze all collegiate basketball players. This would have made the 670 players who were drafted look far more special than how they appear to be when all piled into one data-set. Additionally, this analysis might have shed more light onto the performance metrics that seemingly appeared to be a non-factor in my analysis on draft performance.

However, my classification analysis on career longevity proved to be useful. A test error rate of 35% and 32%, while being trained on an incredibly small data-set was a metric I were proud to achieve. I believe that if I had included a larger data-set I could have achieved even better. Moreover, it would have been fascinating to perform an analysis that included a player's NBA statistics in his first few years. Finding a balance to be able to weigh collegiate and NBA statistics appropriately would have proved to have been a fun challenge.

My report highlights the nuance of player scouting and the difficulty of capturing a player's ability solely by statistical analysis. Some believe in the "eye test" and others believe in data. I believe that if a balance is struck then calculated decisions can be made to highlight players that will have a lasting impact in the NBA. As time passes the data-set at my disposal only grows larger and it presents an incredible opportunity to be able to digest and present this data in a manner that is useful to others.