Dapo Adegbile
IDS 702: Modeling and Representation of Data

Determining Early NBA Success

SUMMARY

The purpose of this report is to infer how NCAA Division I basketball player's stats translate to their value as NBA players in the first four years of their career. I chose to evaluate their performance over the first four years of their career because that is the life of their rookie contracts, and this might help determine whether or not a player (or type of player) will be worth a significant contract upon the completion of their rookie contract. To conduct this analysis a Multiple Linear Regression was used, where the predictor variables the stats of NCAA D1 players that were drafted between 2009 and 2015. The response variable used was the average NBA Box Plus Minus (BPM) of the first four years of each drafted player's career. After some data manipulation my final model was chosen using Backwards AIC Selection, which left me with the most statistically significant predictors. The results of the regression showed that the best indicators of early NBA Success were a college player's Offensive Rating, Assist Percentage, Defensive Rebound percentage, Steal percentage, if they were a Sophomore, if they played in the Pac12, if they played in the SEC, Block percentage, Height in inches, the interaction between Assist Percentage and Sophomores, and finally, the interaction between assist percentage and Seniors.

INTRODUCTION

For NBA franchises the easiest and most consistent way to acquire talent is through the NBA draft. The player a franchise selects could propel them into championship contention, or just as easily leave them floundering at the bottom of the NBA. Through this project I hope to shed some light on the draft analysis process, and potentially identify some key trends regarding successful draft picks. To evaluate this, I wanted to study more recent NBA draft picks (considering the ever-changing basketball landscape). Due to ease of access and the fact that currently, 84.7% of players in the NBA played college basketball, I decided to analyze players who played in college basketball immediately prior to entering the NBA. To emphasize the importance of drafting well, it's important to provide some historical context. Since the NBA started giving out Finals MVP awards in 1969, the recipient of the Finals MVP award received it while playing for the team that drafted him roughly 73% of the time. This is likely to decrease given the amount of player movement in this current player empowerment era, but nonetheless indicative of the importance of drafting the correct players.

Additionally, through this analysis, I'm seeking evaluate the validity of some common tropes associated with NBA draft prospects. Some typical tropes are "the best college scorers make for the NBA best players" , "Seniors (older players) aren't going to be as effective as younger players (Freshman) in the NBA", or "Players in some conferences fare better than players in other conferences in the NBA".

DATA

There is no good, single metric to determine the value of an NBA player, but if there were one, it'd arguably be Box Plus Minus (BPM). Basketball Reference defines BPM as "**a basketball box score-based metric** that estimates a basketball player's contribution to the team when that player is on the court. It is based only on the information in the traditional basketball box score." Additionally, "BPM uses a player's box score information, position, and the team's overall performance to estimate the player's contribution in **points above league average per 100 possessions played**. BPM does not take into account playing time -- **it is purely a rate stat!"** BPM being a rate statistic is key in my analysis because it can help identify quality players that may not have played a large number of minutes. To get a sense of the BPM scale:

- +10.0 is an all-time season (think peak Jordan or LeBron)
- +8.0 is an MVP season (think peak Dirk or peak Shaq)
- +6.0 is an all-NBA season
- +4.0 is in all-star consideration
- +2.0 is a good starter
- +0.0 is a decent starter or solid 6th man
- -2.0 is a bench player (this is also defined as "replacement level")
- Below -2.0 are many end-of-bench players

Continuing with the theme of rate statistics, I wanted to use rate statistics (i.e., Assist Percentage) as my predictor variables because they normalize for minutes played and help remove volume bias from my data. The variables I collected for my analysis are as followed:

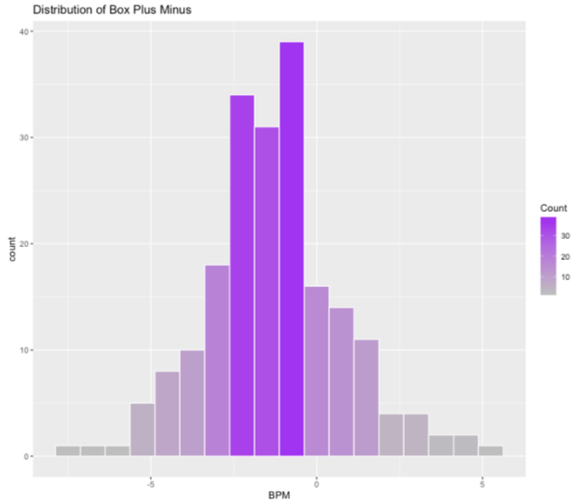
- Class – class
- Conference – conference
- Offensive Rating – ORTG
- Usage Rate -- USG
- Effective Field Goal Percentage – EFG
- True Shooting Percentage – TS
- Offensive Rebound Rate -- OR
- Defensive Rebound Rate -- DR
- Assist Percentage -- AST
- Turnover Percentage -- TO
- Block Percentage -- BLK
- Steal Percentage -- STL
- Free Throw Rate -- FTR
- Free Throw Percentage – FT%
- 2 Point Percentage -- 2P%
- 3 Point Rate – 3PR
- 3 Point Percentage – 3P%
- Height in Inches – height

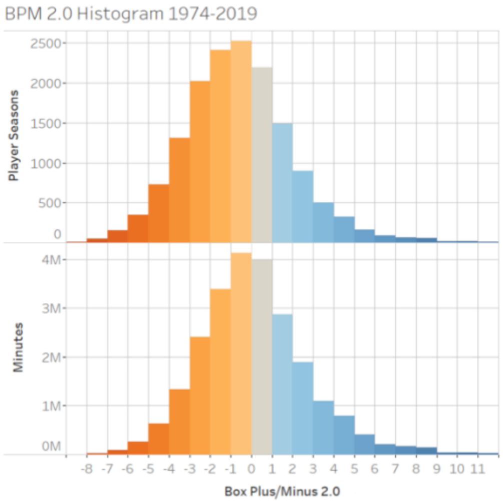
I obtained this data by taking a dataset of all drafted players from 2009 - 2015. I then pulled all Division I players' stats (barttovik.com) from 2009 – 2015 and merged each data set to its corresponding year by player name to get all drafted players' college statistics. To get each player's BPM for the first year of their career I web scraped yearly stat sheets from basketball reference. I concatenated every 4-year window of stats, grouped by player and year and took the average of their BPM for those 4 years. Lastly I merged these datasets to the draft data set corresponding with the first year of the four year window. This data collection left me with a dataset consisting of 205 observations of 19 variables.

Exploratory Data Analysis

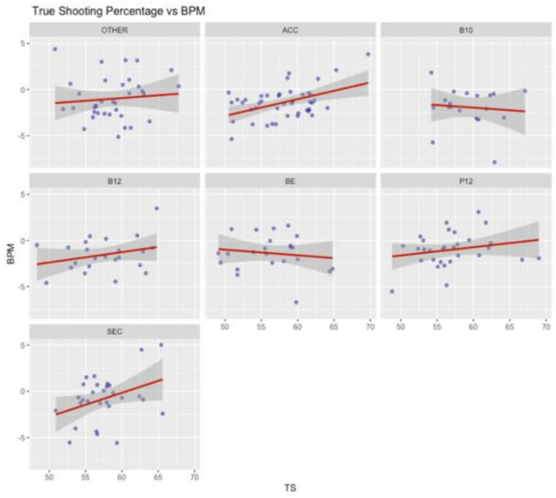
When exploring the data, I decided to plot a histogram of BPM in my dataset and found what I thought were outliers. Upon further investigation, these apparent outliers were in fact not good observations because they had unnaturally high or low BPM's as a result of playing little minutes. To correct this issue, I removed them from my data and was presented with a more natural distribution. My distribution of BPM from 2009 -2015 is similar to distribution the of BPM 1974(when it could first be calculated) to 2019.

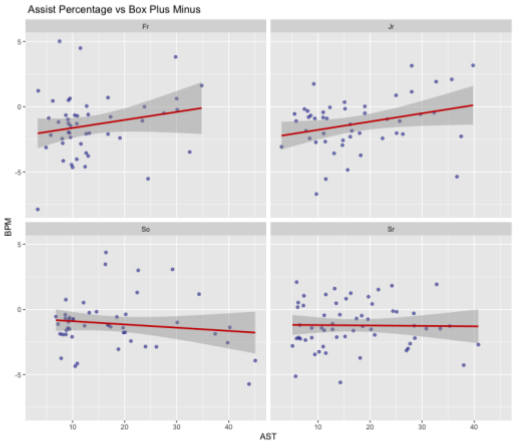
Additionally, when looking at the number of observations in each conference, some conferences have very few observations. To address this issue, I re-leveled the categorical variable "Conference" into each of the Power 5 Conferences + the Big East, and relabeled all other conferences to "Other".



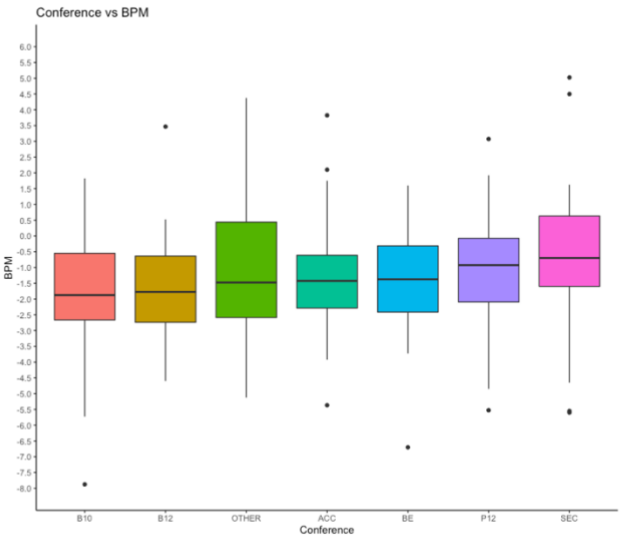


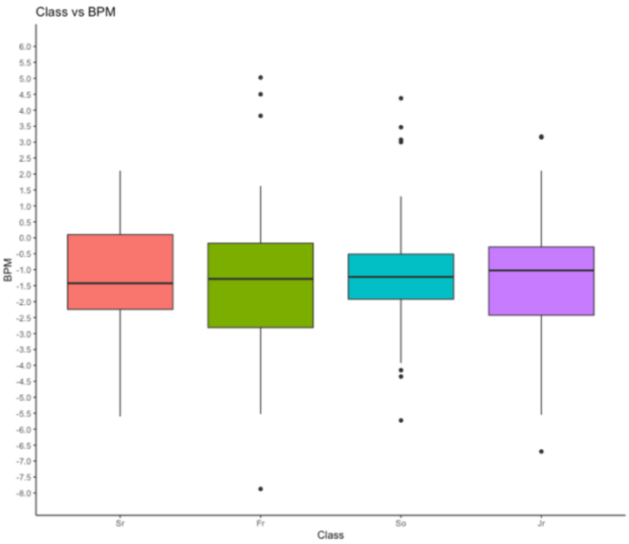
I wanted to start exploring the relationships between BPM and each predictor variable and include the relevant findings below. For starters, more than any other metric, a high True Shooting Percentage is the mark of a good scorer because it is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws. We can see the importance of scoring by conference below, and see that there is a strong positive relationship between college True Shooting Percentage and NBA BPM in the ACC, B12, P12 and SEC. Another interesting relationship was the relationship between Assist Percentage and BPM by class. This plot seems to reinforce the notion that talented freshman tasked with generating the lions share of their team's offense tend to fare well in the NBA.





Furthermore, I wanted to understand the BPM vs Conference relationship, and the BPM vs Class relationship. From the plots below we can see that Pac 12 and SEC tend to produce better NBA players, while each class tends to produce similar levels of talent, with Freshman appearing to be the highest risk and reward players.





MODEL

The full model is as followed.

$$y_i = \beta_0 + \beta_1 conference + \beta_2 class + \beta_3 ORTG + \beta_4 USG + \beta_5 TS + \beta_6 TO + \beta_7 DR + \beta_8 OR + \beta_9 STL + \beta_{10} FTR + \beta_{11} 3PR + \beta_{12} height + \beta_{13} AST + \beta_{14} BLK$$

MODEL SELECTION

After verifying regression assumptions, it is reasonable to conclude the linear model is appropriate for the data.

To reduce complexity and potentially reduce our standard errors, an AIC backwards model selection and an AIC forwards selection model are performed on the initial model. The p value from the AIC backwards selection is the largest, therefore we fail to reject the null hypothesis that these coefficients are zero. I'm left with a final model of :

$$y_i = \beta_0 + \beta_1 conference + \beta_2 class + \beta_3 ORTG + \beta_4 AST + \beta_5 DR + \beta_6 FTR + \beta_7 DR + \beta_8 BLK + \beta_9 height + \beta_9 (AST : CLASS)$$

The results of the final model are as followed:

		BPM	
Predictors	***Estimates***	***CI***	***p-value***
(Intercept)	-5.71	-18.07 – 6.65	0.363
ORTG	**0.09**	**0.05 – 0.13**	**<0.001**
AST	**0.1**	**0.02 – 0.17**	**0.009**
DR	**0.12**	**0.06 – 0.18**	**<0.001**
STL	**0.35**	**0.02 – 0.69**	**0.04**
FTR	-0.02	-0.03 – 0.00	0.124
CLASS [Jr]	-0.74	-2.21 – 0.72	0.319
CLASS [So]	**1.7**	**0.23 – 3.17**	**0.024**
CLASS [Sr]	0.88	-0.56 – 2.32	0.23
CONF [ACC]	-0.05	-0.88 – 0.77	0.897
CONF [B10]	-0.71	-1.74 – 0.33	0.179
CONF [B12]	-0.02	-1.01 – 0.97	0.969

CONF [BE]	0.14	-0.84 – 1.13	0.772
CONF [P12]	0.81	-0.11 – 1.73	0.086
CONF [SEC]	**1.03**	**0.06 – 2.00**	**0.037**
BLK	**0.24**	**0.12 – 0.37**	**<0.001**
HEIGHT_IN	-0.13	-0.27 – 0.01	0.074
AST * CLASS [Jr]	0.01	-0.07 – 0.09	0.765
AST * CLASS [So]	**-0.1**	**-0.18 – -0.02**	**0.02**
AST * CLASS [Sr]	-0.08	-0.16 – 0.00	0.064

From my model, offensive rating, assist percentage, defensive rebound percentage, steal percentage, sophomores, SEC Conference, PAC 12 conference, block percentage and the interaction between assist percentage and sophomores are the statistically significant variables at a 5% level. The results of the most scientifically significant variables are as followed:

- If the player played in the SEC, their NBA BPM is would increase by 1.03.
- For every unit increase in steal percentage, we can expect the BPM to increase by 0.35
- If the player is a sophomore, we can expect the BPM to increase by 1.7.

The final model assumptions are assessed by plotting the residuals and the Q-Q plots. It can be seen from the residual plots that the points are scattered roughly evenly around zero, and there is no discernible trend. Additionally, most points on the Q-Q plot fall on the diagonal with a few points tapering off at the beginning and the end of the plot. Since the VIF scores for all of the continuous variables are low multicollinearity doesn't appear to be an issue.

CONCLUSION

Using the final model specified above, I found that the most important college statistics when predicting early NBA success are the Offensive Rating, Assist Percentage, Defensive Rebound Percentage, Steal Percentage, Sophomore Status, playing in the SEC, Block Percentage, and the interaction between Assist Percentage and Sophomore status. More specifically, for a college Freshman not in any of the 6 major conferences, we can expect his average NBA BPM in the first four years of his career to increase by 1.03 if he played in the SEC. Holding all other variables constant, we can expect his average BPM to increase by 0.35 for every unit increase in steal percentage. Lastly, of the scientifically significant variables, holding all other variables constant, if the player is a sophomore we can expect his average BPM to increase by 1.7. Additionally we are 95% confident that holding all other variables the same, the interaction between sophomores and assist percentage will cause a decrease in the range of 0.02 to 0.18. It is interesting to note that even though separately, sophomore status and assist percentage have positive relationships with BPM, the interaction between the two results in a decrease of BPM.

Also, a potential limitation of this analysis is the high variance for each variable. Each player is very unique and trying to find consistent trends using a smaller sample size is difficult. One way to improve this would be to take player data for more years, increasing the sample size. Another stat worth exploring as a response variable could be the (current) holy grail of analytics, Adjusted Plus Minus.

Other stats that may be worth exploring as a response variable could be PIPM, RPM, and RAPTOR.

One important thing to note is that player analysis should start from what is taking place on film. Conducting statistical analysis helps identify trends that may not be apparent, but it is always good to make sure that our analysis matches what we're seeing on the court.

I also used my model to predict how players from the 2020 draft class would fair in the NBA, but the prediction interval was too large to draw any meaningful results.