How Influential Are Certain College and NBA Statistics To Top Ten NBA Draft Pick Rankings?

Written by:

Aiden Kempen

Professor: Aaron Nielsen

STAT 451: Sports Statistics and Analytics II

Colorado State University

May 5, 2024

Abstract

In this research study I will be trying to find the most significant statistics that increase top ten college players' draft pick value, whether significant college statistics can correctly predict when a top ten college player will be drafted, and whether or not NBA statistics can predict draft order more accurately. To address these topics I will be using variable selection and predictive modeling techniques that we learned this semester, with NBA and College data scraped from Sports Reference. The variable selection techniques being used in this research included best subset selection, forward selection, backward selection, and lasso regression. The predictive models used were multiple linear regression, weighted least squares regression, ridge regression, lasso regression, elastic net regression and finally principal component analysis. The results showed that different variables were significant towards a top ten college players draft pick value depending on what metrics were optimized or minimized. It should also be noted that predicting top ten college players draft pick value with college statistics was inherently hard and had underwhelming accuracy. Finally, predicting top ten college pick values with NBA statistics was less accurate then with college statistics.

Introduction:

Each year the NBA draft can be monumental in determining a team's future, and gives teams a chance to pick from the young talent from colleges around the U.S. College athletes have to make the decision of whether or not they are ready to make their leap from collegiate to professional sports by declaring for the draft. This decision comes with a lot of thought and pressure, and I believe that certain attributes or statistical values a player produces during their college career will affect the position they will be drafted.

For this project, I want to explore the relationship between draft value and significant college and NBA statistics. This relationship is extremely relevant for players today because it can help NBA teams decide who to draft and whether to draft them earlier or later in the draft. It can also help collegiate athletes optimize their potential by declaring for the draft at the right time, and could help athletes learn what attributes of their game they need to improve to raise their draft stock while playing college basketball. In 2006, the NBA made a rule that basketball athletes had to finish at least one year of education before being draft eligible. I will be using the data after this rule was made to get a better understanding of what makes teams choose certain collegiate players over others who have had at least one year of college experience.

Data Acquisition:

To acquire the first dataset, I scraped NBA draft data from 2013-2023 through basketball reference. This dataset included 24 different variables including draft pick, a rank value, and statistics from each player's NBA career. I decided to look at only the top ten picks of each draft that had attended college, since acquiring the college data for each player was going to be a little more difficult. To do this, I limited the dataset to only players that had attended college, and then only took the top ten players from each draft. The first dataset had around 2,530 total observations to use for predictive modeling.

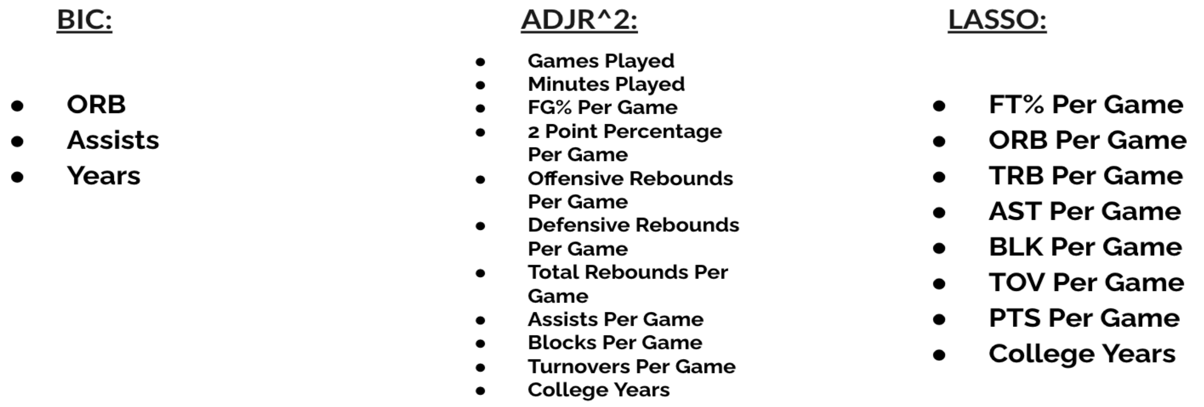
The second dataset which contained each top ten player's college statistics, was more difficult to get since I had to acquire it manually. I went through each player's name in the first data set and used sports reference college basketball to search each player and grab their career college statistics manually. I added each player's statistics to a google document, and when finished exported it to a csv file. The second dataset had around 2,860 observations with 26 variables including years spent in college. Once I had the csv file, I was able to combine the first and second dataset in R creating the final dataset with 50 different variables, and around 5,500 observations.

Statistical Analyses:

Now that the data had been acquired, it was time to start addressing the main topics of this research. To start, I wanted to find the most significant college variables that contributed to the top ten college players' draft pick value. Before I chose the variable selection techniques, I chose what criteria I was going to minimize or maximize for each technique. I ultimately chose to look at what variables were selected from each method with a minimized BIC and maximized ADJR, and then to see if the variables selected from lasso regression differed.

The first method I used for variable selection was best subset selection. With a minimized BIC best subset selection ultimately chose ORB/G, TOV/G, and years played as the three most important variables considering player draft rank. With a maximized ADJR best subset selection ultimately chose games played, minutes played, FG/G, FG%/G, 2P/G, ORB/G, DRB/G, TRB/G, AST/G, BLK/G, TOV/G, and years played as the most important variables considering player draft rank. I continued this analysis for both forward and backward selection techniques and ultimately got similar results for variables chosen with both minimized BIC and maximized ADJR.

Then, I fit a lasso regression model to see which variables the model chose, and whether or not there were major differences between the variables chosen previously. The lasso model chose FT%/G, DRB/G, TRB/G, AST/G, BLK/G, TOV/G, PTS/G, and years played as the most important variables considering player draft rank. FT%/G and PTS/G were not variables chosen in the BIC and ADJR variables selection, but were selected in lasso. This was surprising since I believed PTS/G would be an extremely important variable considering draft rank, but wasn't included in the BIC or ADJR variable selections. Below, I have lists of the most important variables chosen considering player draft rank for BIC, ADJR , and lasso.



BIC:

- ORB
- Assists
- Years

ADJR^2:

- Games Played
- Minutes Played
- FG% Per Game
- 2 Point Percentage Per Game
- Offensive Rebounds Per Game
- Defensive Rebounds Per Game
- Total Rebounds Per Game
- Assists Per Game
- Blocks Per Game
- Turnovers Per Game
- College Years

LASSO:

- FT% Per Game
- ORB Per Game
- TRB Per Game
- AST Per Game
- BLK Per Game
- TOV Per Game
- PTS Per Game
- College Years

Next, using these selected variables it was time to fit predictive models to see how accurately we could predict players' draft pick rank using college data. The first models that were fit were just normal multiple linear regression models for a baseline accuracy for both BIC and ADJR selected variables on draft pick rank. Multiple linear regression is a basic way of predicting a dependent outcome variable using multiple independent variables which is why it was the most basic model fit.

I then went on to fit weighted least squares regression models for each BIC and ADJR variables on draft pick rank to try and account for the unequal variance in the amount of years each player played in college. I decided to use this weighted least squares model because it could have performed better in terms of accuracy compared to the multiple linear regression. The next model I fit was a ridge regression model to see if I could eliminate any high correlation between the variables that were in my dataset, and again hopefully have better accuracy then the models used beforehand.

Lasso regression was the next model used, in which we saw what variables this model selected from earlier in this paper. Lasso regression performs both variable selection and regularization in order to enhance the prediction accuracy of a model by shrinking coefficients that are not needed to zero. Next, was elastic net which is a combination of both lasso and ridge regression. I was hoping if I did not have great results from lasso and ridge, a combination of the two could be helpful in increasing prediction accuracy for players' draft pick rank. Finally the last model I fit was a principal component model to possibly reduce the dimensionality of my dataset and get more accurate predictions with the college data.

Finally, after I had fit all my models using the college data, I wanted to fit these same models with NBA data to see if NBA data was more accurate in predicting draft pick rank then college data. I ended up fitting the lasso regression model, ridge regression model, elastic net regression model, and principal component analysis model with NBA data to see if the prediction accuracy was higher or lower than the models prediction accuracy with college data.

Results:

Test Results For Predictive Models On College Data Using RMSE and R^2

METRIC	MLR BIC	MLR ADJR	WLS BIC	WLS ADJR	Ridge	Lasso	Elastic Net	PCA

Test RMSE	2.348	2.226	2.306	2.239	2.476	2.468	2.438	2.436
Test R^2	0.353	0.434	0.374	0.413	0.342	0.340	0.341	0.305

As we can see in the table above, the model that performed the best in terms of test RMSE and test R^2 was the multiple linear regression model with ADJR maximized with a test RMSE of 2.226 and a test R^2 of 0.434. This was a little disappointing since the multiple linear regression models were the most basic models fit during this project. It was also disappointing to see how low the prediction accuracy was with R^2 values all under 0.44. I was surprised to see that the lasso, ridge, elastic net, and PCA models all had worse test RMSE and test R^2 than both the multiple linear regression models and both the weighted least squares models. Overall, it looks like when using college data we can predict players draft pick rank with a low accuracy.

Test Results For Predictive Models On NBA Data Using RMSE and R^2

METRIC	**Lasso**	**Ridge**	**Elastic Net**	**PCA**
Test RMSE	2.746	2.720	2.680	2.664
Test R^2	0.154	0.174	0.185	0.257

Unlike the predictive models on college data, we can see that PCA was the best performing model out of the four predictive models on NBA data with a test RMSE of 2.664 and a R^2 of 0.257. Elastic net regression in this case did outperform lasso and ridge regression which was interesting and shows how the use of both lasso and ridge penalties could be helpful in increasing model accuracy. Overall, these predictive models had a lot worse accuracies then the models we used with college data. This shows that player performance in the NBA may be less indicative of their draft pick rank then college performance.

Possible Further Research/Conclusions:

Although I did not get the exact results I wanted during this project, I still was able to learn from my results, and think about how I could further this research with more time and models. The first thing I could do to improve this project is obtain more data! Instead of grabbing the college data manually I could scrape the college data by taking the name column from the first data frame and use the list of names to locate urls for each player's college data. By doing this, I wouldn't have to limit myself to only the top ten picks of each draft, and could have a lot more useful observations in my data set.

The second thing I could do to further this research is to use more models that may be better for predicting. I didn't have much time to explore an intense amount of predictive model building, but I could use random forest or gradient boosting models to maybe see better predictions accuracy for both college and NBA data on draft pick rank. I also think I could have improved my project by obtaining NBA draft combine statistics and more physical characteristic statistics that may also influence a team to draft a certain player over another. Statistics like height, weight, vertical, etc… could definitely play a role in what position players are drafted.

Ultimately, we found that different variables were significant to draft pick rank based on different variable selection techniques. Based on the most accurate model we saw that Games Played, Minutes Played, FG% Per Game, 2P% Per Game, ORB Per Game, DRB Per Game, TRB Per Game, AST Per game, BLK Per Game, TOV Per Game, and College Years may be most significant when evaluating draft rank for top ten picks. Based on college data we predicted with a low accuracy where top ten college players will be drafted within the top ten picks of the draft. It was also found that NBA statistics have lower accuracy in predicting a top ten college players draft rank then college statistics, which could mean that a player's performance in the

NBA may not depend on draft rank or how well a player performed in college. Overall, I had a great time creating and executing this project and feel that with more data and more work we could see some more efficient prediction accuracy on top ten college players' draft rank.

Bibliography

Men's and women's College Basketball Statistics and history: College basketball at sports. Reference.com. (n.d.). https://www.sports-reference.com/cbb/

Basketball Statistics & History of every Team & NBA and WNBA players. Basketball. (n.d.). https://www.basketball-reference.com/

Code is on github link below:

https://github.com/Akempen17/STAT-451-Final_Project