# Does Where You Go to School Matter in Basketball?

Brian Liao

May 11, 2020

**Abstract**

What makes the University of Kansas, a world class institution for basketball, different than the University of California, Berkeley that has much fewer players in the NBA? Looking at statistics like average points per game, NBA tenure length and player positions, there does not seem to to be any significant difference between schools like Berkeley and schools like Kansas. In addition, using college basketball statistics, I try to predict the tenure of a player in the NBA. My model is able to predict a player's tenure with a root mean square error of 4.47. Features that help predict a players tenure include the player's field goals per game and field goal percentage, which predicted a longer tenure and if the player go to an less popular college, which predicts a shorter tenure. A video walkthrough is included here. ***Keywords:*** *Basketball, NCAA, NBA*

## 1 Introduction

My dad went to the University of Kansas, which is famous for their basketball program. His freshman year, the Kansas Jayhawks won the NCAA National Championship in 1988. The University of California: Berkeley, by comparison, is not as famously well known for their basketball program. This made me curious, did where a basketball player went to college predict how well they would do in the NBA? To investigate this, I posed four questions to answer:

- **How do schools that send many players to the NBA like the University of Kansas compare to schools like the University of California: Berkeley, which send much fewer players?**
  - **Do their college NCAA basketball statistics, NBA basketball statistics, tenure in the NBA, weight, height, and position differ significantly?**
- **What school produces the most NBA players?**
- **What school produces the longest tenured NBA players?**
- **What player data in college are good predictors of tenure in the NBA?**

A player's tenure in the NBA is a good measure of how good of a player. We assume that good player should be able to continue playing long in their career. If they are not a good player, they will be cut from the team as the NBA is very competitive. However there are factors out of the players control such as injury that affect their tenure in the NBA.

For my analysis, I used the `college.csv` dataset. It contains players in the NBA, and data such as where they went to college, their height, weight, position, and their NCAA and NBA basketball statistics.

## 2 EDA and Data Cleaning

In our dataset, we noticed null values for several fields. The major ones were that there were 300 null colleges, over a thousand nulls for 3 pointer statistics, and around two thousand nulls for NCAA statistics. To see if I could drop these null values without biasing our data, I looked at the distribution of the data.

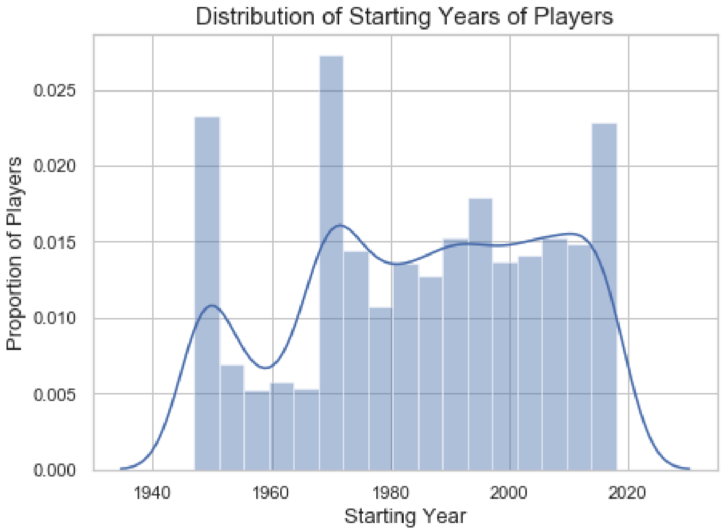

Figure 1: Histogram of Player Starting Years.

In this data, I noticed a spike at 1947/1950, 1968, and 2018. Looking at the data and Googling, the 1968 spike was due to to the introduction of the Seattle SuperSonics and the San Diego Rockets, the 1947 spike was due to the creation of the Basketball Association of America and the 1950 spike was due to the creation of the National Basketball Association (NBA). I could not find an explanation for the 2018 spike but I hypothesize this was due to it being the last year added in our dataset.

Next, I looked at the distribution of players with null colleges, null 3 pointer statistics, and missing NCAA statistics. The differences in distribution seem to only be due to data not being collecting at those times such as for 3 pointers in the 1950s. The statistics of those players should be proportion to those that were in the dataset, so it is ok to drop then. I dropped the players missing the college field and NCAA statistics. For the 3 pointer fields, I dropped the field instead of the players because more NBA players were missing from those statistics and this would require dropping less people.

Another field I had to clean was the `college` field. First, the University of California: Berkeley was not listed and was actually listed as the University of California. To fix this, I did a regex match and replace. Second, players that played at multiple universities were separated by a comma. To address this, I created a row for each university the player played at using Pandas' `explode` function. This left whitespace in the college field and I cleaned that using Pandas' `strip` function.

Two other fields I cleaned were the height field, which gave height as a string in the form of feet-inches and the player position field. I create a function to convert the height field into a float. I also replaced the player positions with the three basic positions, `Forward`, `Guard`, and `Center`.

## 3 Methods and Experiments

The first question I answers was which college sent the most players to the NBA. Doing a `groupby('college').size().sort_values(ascending=False)` listed the colleges and the number of players they had in the NBA. The largest was the **University of California: Los Angeles with 79 players** and second largest was the University of Kentucky with 76 players. In actually, I found by Googlingt that the **University of Kentucky was actually first with 111 players**. It's possible these players were removed in our cleaning process showing some bias in our cleaning method.

Next I compared Berkeley and Kansas. Berkeley sends 20 players to the NBA and Kansas sends 57 players to the NBA. To compare schools like Berkeley and Kansas, I added a field `school_popularity` with classes `High` if the school has greater than 30 players, `Medium` if the school is between 15 and up to 30 players, and `Low` if the school has 15 players or less. Berkeley has a `Medium` school popularity and Kansas has a `High` school popularity.

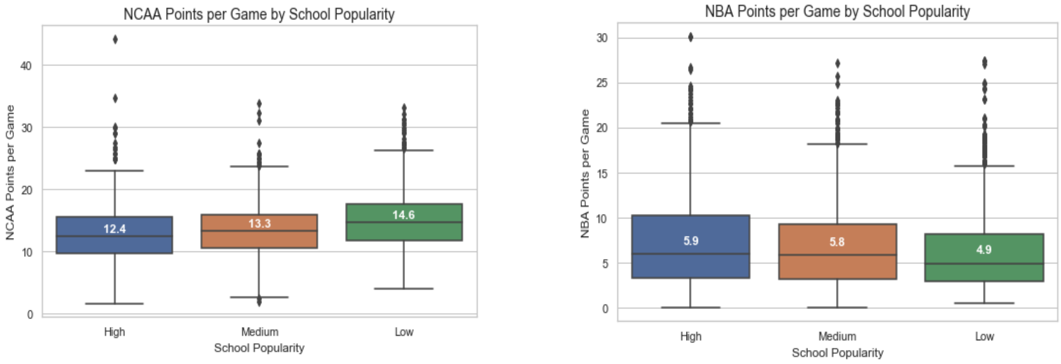

Figure 2: Boxplots of NCAA and NBA points per game. NCAA points are much higher than NBA points.

By plotting the data such as height, position, and basketball statistics between high, medium, and low popularity schools, I did not see any significant difference the classes of school popularity. The only case that did was points per game in both the NBA and NCAA. In the NCAA, low popularity school players scored 14.6 points on average while high popularity school players scored only 12.6 points on average. This may be because the low popularity players were stars on their team and scored more than those in high popularity schools which may have had multiple stars on the team. The opposite relationship is seen in the NBA, players from high popularity schools score 5.9 points on average compared to players from low popularity schools that score 4.9 points on average.

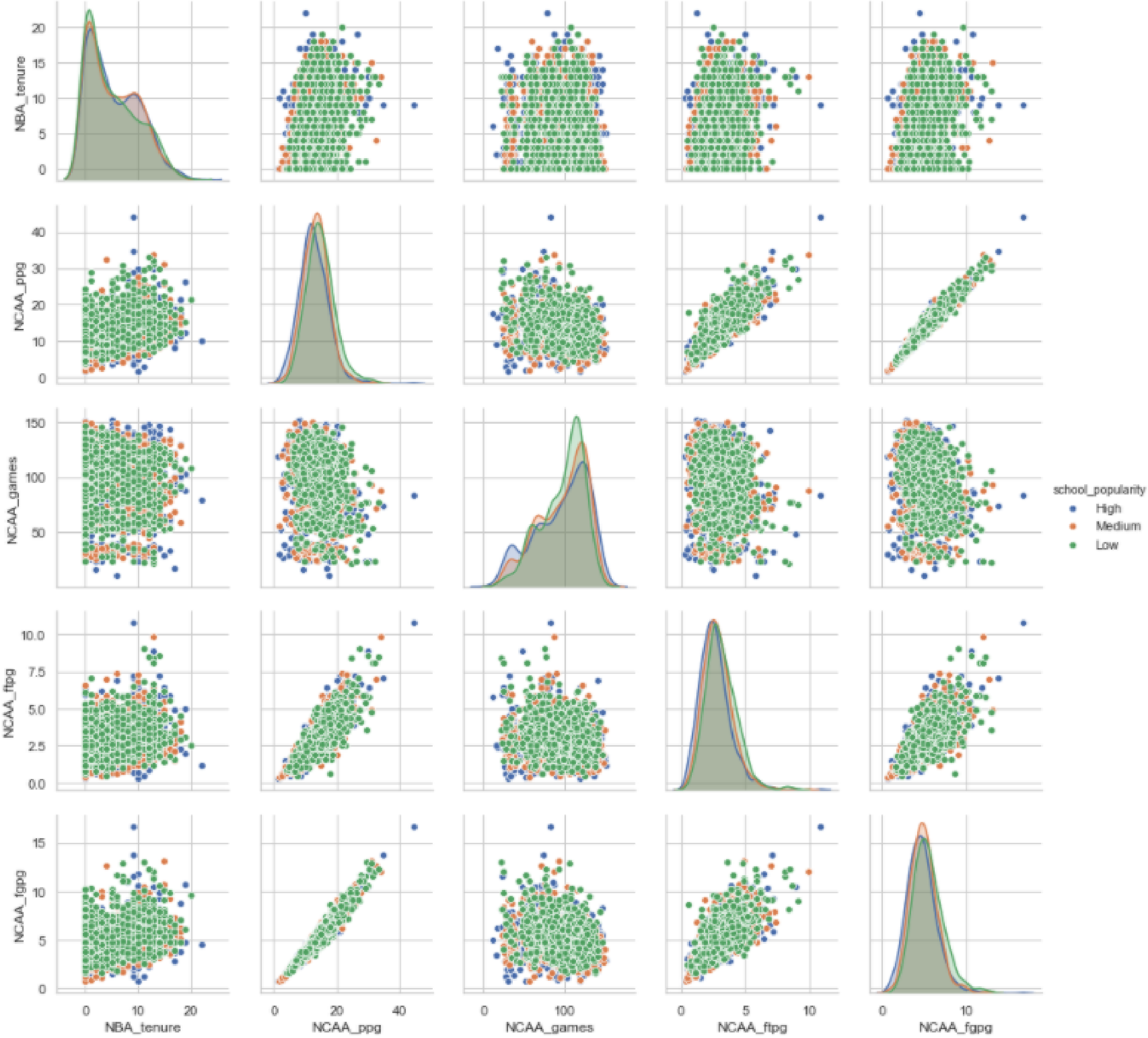

Figure 3: Pairplot of NCAA statistics. The distributions of High, Medium, and Low school popularity does not have any significant deviations. NCAA points per game is highly correlated with field goals per game and moderately correlated with free throws per game.

Using this information, the answer to my first question is that **there isn't a significant difference between players of high, medium, and low popularity schools expect slightly for the number of points they score per game.**

We also investigated which school produced players with the longest tenure. The tenure field, `NBA_tenure` was added by getting the difference of `active_to` and `active_from`. Doing a groupby college and mean of NBA tenure, **Gardner-Webb University had the longest average tenure of 16 years**. This may have been biased because there were several schools that may send one or two players to the NBA. To address this, I also calculated the longest tenure for schools with at least three players, which **Louisiana Tech University has the longest with a mean tenure of 10.6 years**.

Finally, I explored what features from data in college are good predictors of a player's

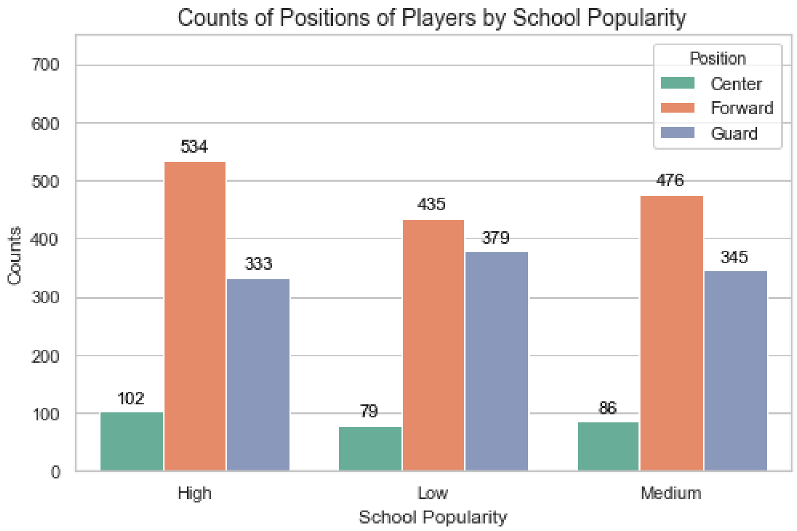

Figure 4: Positions of players from High, Medium, and Low popularity schools. Each school has a majority of Forward positions, then Guard, and Center last.

tenure in the NBA. Using these features, I would attempt to create a regression model to predict their NBA tenure. Some features I hypothesized would be useful were school popularity, points per game in the NCAA, the number of games a player played in the NCAA and their height. Graphing this in comparison to tenure did not show any strong relationships but a player from low popularity universities was slightly more likely to have shorter tenures and players that scored more points in the NCAA had a weak positive correlation to longer NBA tenures.

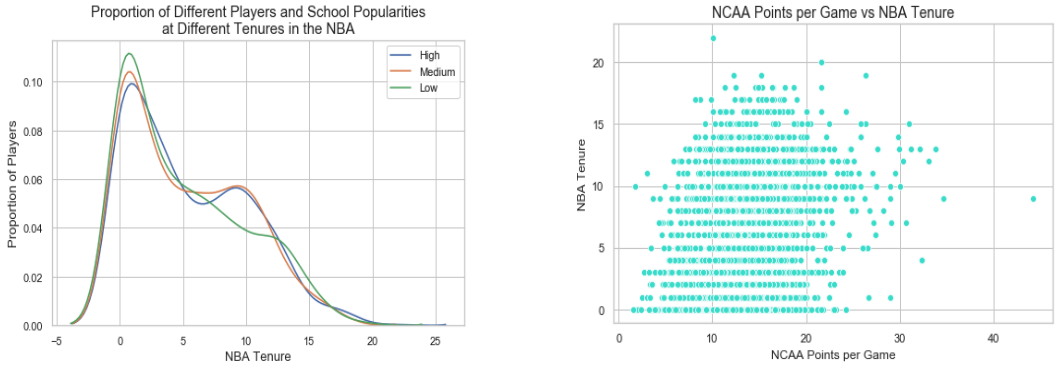

Figure 5: One the Left: Distribution of NBA tenure from players of different school popularity. The distributions are approximately the same, but players from low popularity schools are more likely to have shorter tenures. One the Right: NCAA points per game vs NBA Tenure. There is a weak positive correlation between points scored and a players tenure

To prepare our data for prediction, I normalized all of the quantitative variables and one-hot encoded the categorical variables `position` and `school_popularity`. I split the data into an 80/20 training/test set. To set a baseline model, I created a model that predicted the tenure of each player in the test set as the mean of the training set. This had a root mean square error (rmse) of 4.76.

To investigate features, I trained a Linear Regression model, a Ridge Regression model, and a Lasso Regression model with each individual feature. For each model, I took the average of a 5-fold cross validation rmse. This showed that NCAA field goals per game, field goals percentage, points per game, and free throws attempted were good features. These features are all correlated and related to the number of points scored in the game.

Using this data, I created a final model using Ridge Regression and college features I found that helped predict NBA tenure. Answering my final question, the features I choose were NCAA field goals per game, field goals attempted per game, height, and school popularity. Field goals measure how well a player could score, field goal percentage measured how accurate a player was, and height would be beneficial to a player throughout their NBA career. School popularity had been shown to have a slight impact as players from low popularity schools tend to have shorter tenures in the NBA. Using these features, my final model has a test rmse of 4.47, an improvement of 0.29.

## 4 Analysis and Conclusion

The most interesting features for my particular question was the NCAA points per game because they were two times higher than in the NBA and players from low popularity schools had the highest average points per game. The next most interesting was school popularity which showed players from low and medium popularity schools were able to compete at the NBA level against high popularity school players. One feature I thought would be useful was player position, but it seemed to be ineffective as all basketball teams need each position so no position was more preferred than any other. One challenge with the data was the amount of null values. We tried to deal with them fairly, in a way that removing them would not bias our data, but it did impact some analysis. An example is finding the school with the most number of players, as our data said it was UCLA when in reality it was the University of Kentucky.

A limitation of my analysis is the weak correlation between college statistics and tenure in the NBA. While we were able to improve our predictive model, player tenure is more affected by other factors such as lack of injury and performance and consistency in the NBA. One assumption I made was that player's with null values would have statistics distributed proportionally to our dataset population data. An example is that the players missing NCAA statistics are more common before 1980. If this assumption was incorrect, there may have different statistics like points per game, which would have skewed our analysis.

An ethical dilemma I faced in the data was that uncommon schools may be underrepresented. We saw that the University of California: Berkeley was actually listed at the University of California. Many other schools may actually be underrepresented because we did not know them and could not have done the data cleaning required to accurately represent them.

An additional piece of data that would help our analysis would be the tenure of the players in college. In the NBA, it is common to be "one and done," which means the player plays in college for an year then goes to the NBA. These players are usually the best of the best, and may have helped us predict their tenure in the NBA.

A ethical concern I might encounter in studying this problem was creating selective bias for players from popular schools. Our analysis showed that less popular schools tend to have shorter tenures in the NBA, but only slightly. Somebody might use this analysis to start only selecting players from top colleges, leaving good players from medium and low popularity schools unselected, even though their NBA and tenure do not seem to be statistically different.

Based on my investigation, I do not think there is a strong relationship between a players college performance, where the player went to school and the player's tenure in the NBA. While we did improve on the baseline model, we were only able to capture weak correlations. Other factors have more control over tenure such as player injuries and player performance in the NBA. Therefore, our methods are limited for the task we are trying to solve.

The most surprising thing I found was that players score more points in the NCAA than in the NBA, around 13 points compared to 5 points per game. My hypothesis for this is that the players that get to the NBA are superstars and have a disproportionate amount of scoring compared to the rest of their teams. Evidence for this is also supported by another surprising thing I found. Low popularity school players score more on average, at 14.6 points, than high popularity schools, at 12.4 points, in the NCAA. The opposite is true in the NBA with high popularity schools scoring an average of 5.9 and low popularity schools scoring an average of 4.9. In the NCAA, the players from low popularity schools that make it to the NBA would have to be a superstar and would be able to score more than if they were at a high popularity school with multiple superstars.

Overall, I was able to investigate the relationships between NBA player college popularity and the tenure of players in the NBA. Future work could explore finding undervalued players from uncommon universities, as we have shown that there are great, high scoring players with long careers in the NBA from uncommon schools. Good players can come from anywhere and even if a school is less common like Berkeley compared to Kansas, there are still super star players waiting to be discovered.