Gabriel Huang

MAS352

Dec. 2023

**Shooting from the NCAA to NBA: A Statistical and Machine Learning Analysis**

In the modern NBA, shooting is the most important skill a player can possess; in almost all scenarios, good shooting will keep you in a rotation, and bad shooting will force you out of it. Consequently, the draft (after the first five or so picks) has become a scramble to identify effective shooters at all five positions. Unfortunately, basketball is an inexact science, and thus the correlation between proficient college shooters and proficient NBA shooters is inexact as well. This effect is bidirectional: bad college shooters can grow into adequate NBA shooters, and good college shooters can regress once they reach the NBA. All of this begs the question: is there a way to predict how well an NCAA prospect's shooting will translate to his NBA rookie year? To answer this, I found patterns using simple statistical analysis, and then created four machine learning models (using K-Nearest Neighbors, Multiple Linear Regression, Decision Trees, and Random Forests) that predict NBA shooting proficiency using college data. By using these models, NBA scouts can more accurately draft the best prospect fit for their team.

My training dataset is a set of 215 players which was created using parameters (1) player played at least 60% of minutes in the season before he got drafted at a D1 college program, (2) player was drafted in the past ten years, and (3) player played at least 500 minutes in his NBA rookie year (a larger dataset would have been preferable, but I deemed pre-2013 data unusable because of the rapidly changing landscape of the NBA). By integrating the CBBReference,

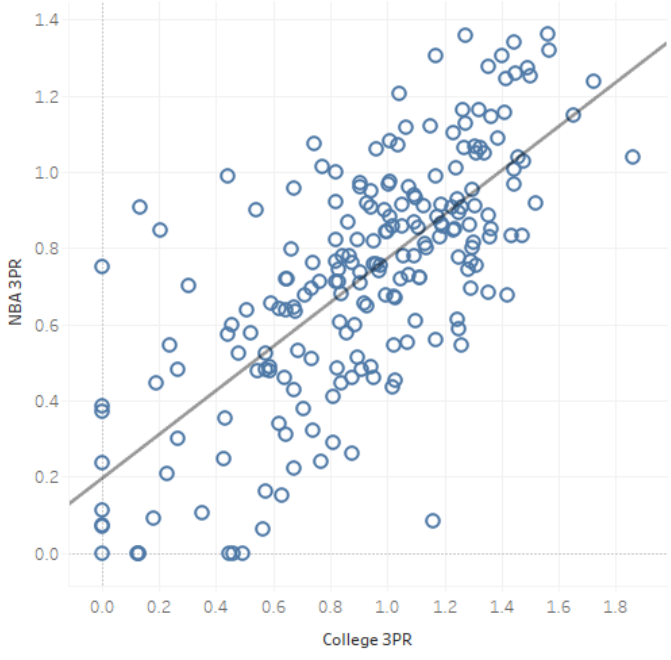
EvanMiya, and Stathead datasets of these players, it created a holistic view of over 200 features from their college careers(which I then parsed down to the most relevant 81).

In order to simplify player analysis, I introduced two new variables into the dataset: ncaa3PR/nba3PR, and ncaaServiceableShooter/nbaServiceableShooter. *nba3PR* acts as the continuous dependent variable for all of my analysis, considering that three point shooting can not be fully analyzed using only 3P% or 3P volume. This is calculated by:

$$3PR = 3P\% * \sqrt{\frac{3PA}{MP} * 100}$$

*ncaaServiceableShooter* and *nbaServiceableShooter* are boolean variables that categorize whether a player can be considered a serviceable shooter at that level. The parameters for being a ServiceableShooter are a 3P% > 0.3 and a 3PA/100 > 4, which are both ~ 33rd percentile in the population.

A simple regression analysis quickly emphasizes that the single most predictive variable of NBA 3PR is, of course, college 3PR. A p-value of <0.0001 details a model that is significantly more strongly correlated than any other single descriptive feature in the dataset. However, an r-squared of only 0.510 confirms that there is much more variation and noise to sift through.




In order to identify the cause of these variations, I focused on the two most important

populations in this study: players who weren't good shooters in college and became good shooters in the NBA, and vice versa.

My training dataset identified 72 players who had a *collegeServicableShooter* value of false, of which 19 players also had an *nbaServicableShooter* value of true. This describes the population of players who's shooting improved from their last collegiate year to their NBA rookie year. The first attribute I found with a strong predictive value for NBA three point shooting is college FT%. I analyzed this variable with simple linear regression.

FT% of Non-Serviceable College Shooters (Continuous)

| Correlation | P-value | R-squared |
|---|---|---|
| Positive | 0.004 | 0.11 |

FT% of Non-Serviceable College Shooters (Bins)

| | NBA Poor Shooter | NBA Serviceable Shooter |
|---|---|---|
| FT% $> 0.67$ | 30 (62.5%) | 18 (37.5%) |
| FT% $\leq 0.67$ | 23 (95.8%) | 1 (4.2%) |

While good free throw shooting in college does not guarantee an improvement in three point shooting in the NBA, the table identifies free throw shooting as a good indicator of three point shooting *potential*. As in, players who are good free throw shooters can grow into adequate NBA shooters (37.5% of players), but bad free throw shooters very rarely do (4.2%).

Another variable that has a similar predictive value of a player's three point shooting *potential* when using simple regression is a player's age.

Age of Non-Serviceable College Shooters (Continuous)

| Correlation | P-value | R-squared |
|---|---|---|
| Negative | 0.011 | 0.09 |

Age of Non-Serviceable College Shooters (Bins)

| | Age: 18 | Age: 19 | Age: 20 | Age: 21 | Age: 22 | Age: 23 |
|---|---|---|---|---|---|---|
| NBA Poor Shooter | 11 | 17 | 10 | 9 | 5 | 1 |
| NBA Serviceable Shooter | 5 | 12 | 1 | 1 | 0 | 0 |

An improvement in three point shooting is almost 5x more likely in younger players than older players; while 37.8% of teenagers who were poor college shooters improved as shooters in their rookie year, only 8% of older players showed the same improvement. If a player's age is not accessible, the same trend can be seen using a player's class.

My training dataset also identified 143 players who had a *collegeServicableShooter* value of true, of which 39 players also had an *nbaServicableShooter* value of false. This describes the population of players who were good college shooters, but then regressed once they got to the NBA. In general, this population did not have too many distinct patterns, but the one predictor I found for sustaining serviceable college shooting is college USG%. USG% estimates how often a player creates a play for their team during their minutes on the floor. A high USG% denotes a ball-dominant player, and a low USG% denotes an off-ball role player.

USG% of Serviceable College Shooters (Continuous)

| Correlation | P-value | R-squared |
|---|---|---|
| Negative | 0.02 | 0.05 |

USG% of Serviceable College Shooters (Bins)

| | NBA Poor Shooter | NBA Serviceable Shooter |
|---|---|---|
| USG% < 21 | 2 (7.7%) | 24 (92.3%) |
| USG% < 21 | 37 (31.6%) | 80 (68.4%) |

In the bins table especially, the correlation is aggressively negative; players with a lower usage percentage in college are more likely to become serviceable NBA shooters. This theory may seem counterintuitive, considering that high-usage players will drop their usage in the NBA, and the general basketball analytics consensus is that when usage goes down, efficiency goes up. However, this finding presents a counterpoint: college players with low usage are already adapted to the low-usage role that they will inherit in the NBA, and are thus better equipped to succeed off the ball as a rookie. On the other hand, high-usage players may struggle to adapt to their low usage role.

Instead of creating a model using just these observed parameters in order to predict NBA shooting success, I decided to create more sophisticated machine learning models that take into account all 81 parameters in the training dataset. These include detailed shooting stats from all over the floor (dunks, close range, mid range, three point range), historical data (shooting data from a player's previous college years), and advanced offensive statistics (bayesian ratings, offensive plus-minus, adjusted efficiency rating), among others. I preprocessed the data by converting categorical data to numerical data using one-hot encoding, and then I scaled each of the numerical variables so that they were each weighted equally.

Each model is trained to predict one dependent variable: NBA3PR. The models I chose to create were a K-Nearest Neighbors regression model, a Multiple Linear Regression model, a Decision Tree regression model, and a Random Forest regression model. For an overview: K-Nearest Neighbors (KNN) Regression models operate by predicting the value of a new data point by averaging the values of the 'K' nearest points in the data set. Multiple Linear Regression models operate by predicting an outcome based on the linear combination of multiple predictor

variables. Decision Tree Regression models use a tree-like model of decisions, where each root represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf represents a value prediction. Finally, Random Forest Regression models operate by constructing a bunch of decision trees for the same data set and outputting the average prediction of the individual trees. Each of these models have their drawbacks (such as K-Nearest Neighbors becoming less effective the more dimensions you add or Decision Tree models being prone to overfitting), so I decided to use all four in order to test which one can predict this year's rookies with the most accuracy and least variance.

I used the 'sci-kit learn' library in Python in order to train these models to predict NBA3PR on the training data set, and then passed a test dataset through the models with the exact same parameters. The test dataset contained the college data from this year's 45 NBA rookies who were drafted out of college. The purpose of all of these models is to predict the NBA3PR for these 45 players in their rookie years.

Because the NBA season is only halfway done, there is no conclusive analysis on how accurate the models were, but here is a sample of the top five shooters (average) in the resulting predictive dataframe:

| Player | KNeighbors NBA3PR | Linear NBA3PR | DecisionTree NBA3PR | RandomForest NBA3PR |
|---|---|---|---|---|
| Julian Strawther | 1.010441 | 1.085915 | 1.319823 | 1.016107 |
| Jordan Hawkins | 0.986616 | 1.053345 | 1.236451 | 1.096316 |
| Brandon Miller | 1.041724 | 1.00849 | 1.236451 | 1.008864 |
| Gradey Dick | 0.925121 | 1.020919 | 1.144745 | 0.99796 |
| Keyonte George | 0.999711 | 0.91362 | 1.144745 | 0.991899 |

Once the NBA season concludes on April 14, 2024, I will compare the actual NBA3PR with each of the models' predictions in order to determine which model, if any, is the most accurate. I will analyze by using standard statistical values such as mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

This analysis reveals significant correlations between college statistics and NBA shooting performance. While college free throw percentages, player age, and usage percentage emerged to me as key predictors, further research could explore additional variables. The insights gained from this study could possibly be used immediately in NBA scouting, offering a new possibility for player shooting evaluation.

Evaluation of college prospects in sports will always be a mix of visual (eye-test) scouting and statistics, but while visual scouting will always be volatile, teams can refine and optimize statistical models. This research shows the potential of statistics and machine learning for NBA scouting, but the concept can be applied to any sport that is willing to accept it.

*Full code and dataset can be found at: github.com/gyhhuang/nbaShootingMLModels