

Projecting Three-Point Percentages for the NBA Draft

Hilary Sun

hsun3@stanford.edu

Jerold Yu

jeroldyu@stanford.edu

Roland Centeno

rcenteno@stanford.edu

December 16, 2017

1 Introduction

As NBA teams have begun to embrace data analytics, more emphasis has been placed on shooting and spacing the floor. The three-point shot, in particular, is becoming increasingly prevalent in the NBA. Players and teams have been taking more threes than ever, and recruiters are increasingly looking to recruit members who will be proficient three-point shooters in the NBA [3]. The goal of this project is to project the best three-point shooters among top NCAA prospects.

Using a player's statistics from the NCAA and the NBA as inputs, along with team statistics on both leagues, our model outputted a three-point shooting percentage for that player.

2 Related Work

Previous work has been done into building models for recruitment from the NCAA to the NBA, but most of these focus on predicting overall NBA success from the draft itself or trying to predict the draft itself. Harris and Berri research what factors affect the draft for the WNBA, modeling the draft using a Poisson Distribution model and a Negative Binomial model and incorporating the three-point percentage into their model[5]. Using data from division-one players, Bishop and Gajewski used principal compo-

nent analysis, logistic regression, and cross validation to predict a players potential in the draft to the NBA[4].

There has been some work evaluating relationships between pre-NBA statistics and NBA career success. Coates and Oguntimein analyzed correlations and regressions from data collected about players drafted in 1987 to 1989 to see if it is possible to use college statistics to predict success in the NBA, using a variety of features such as points per game, minutes per game, and throw percentages. They concluded that there was a strong relationship between a players performance in their college career and their professional career[2]. Another paper does the same, analyzing data from 1988-2002 instead using multiple regression tests as well for various positions. It concluded that there is a relationship between pre-NBA statistics and a players career longevity for guards and forwards, but not for centers[1]. However, these studies focus on a players performance holistically, while we hope to specifically focus on three-point percentages.

Although there are not a lot of published research focusing in on threes, there exists some fan-made statistical analyses of three-point percentages, performing linear regression on selected features. One fans conclusion was that there is significant information conveyed about a players later three-point percentages from before their college statistics, but also

that this was a difficult problem[6]. We hope to extend the research attempted on this problem.

3 Dataset and Features

3.1 Data Collection

Our dataset was scraped from basketball-reference.com and sports-reference.com/cbb. We selected 149 NCAA prospects from nbadraft.net’s Top 100 Big Board from 2009-2016 that met the following criteria:

- Player must have taken at least 30 threes in the NCAA.
- If the player was drafted before the 2015 NBA Draft, he must have taken 300 threes in the NBA.
- If the player was drafted in 2015 or 2016, he must have taken 100 threes in the NBA.

We chose players from this list because we wanted our model to be relevant to NBA scouts who would want to learn more about the best prospects in the country. Additionally, our filtering method allowed us to narrow down our data to only apply to players who are considered regular 3-point shooters.

3.2 Feature Selection

We included the following features for each player:

- Player’s total number of 3-point field goal attempts in the NCAA.
- Player’s NCAA 3-point shooting percentage.
- Player’s NCAA free throw shooting percentage.
- The strength of schedule of the player’s NCAA team.

Principal Component	Variance Explained
1	0.348
2	0.180
3	0.158
4	0.136

- The average number of threes per game the player’s NCAA team took.
- The average offensive rating of the NBA team(s) the player has played for.
- The average number of threes the player’s NBA team(s) took, relative to the league average.
- The player’s NBA 3-point shooting percentage. This is our output.

In selecting features, our hunch was to not only include player statistics that would intuitively have some correlation with the specified output, but to also include the environment surrounding the player. For example, we thought that if the NBA team the player played for had a good offense (i.e. a high offensive rating), then he is more likely to take higher quality shots and thus have a higher shooting percentage.

We decided to break up our dataset into 80/20 split between training and testing. Because a lot of our features are closely related, we guessed that there were probably correlations between variables and extraneous information. We decided to run principal component analysis on our data and reduce the dimensionality of our data. After running PCA from `sklearn` library on our training data, we saw that the variance explained was as such:

As we can see, the first four principal components already explain 82.273% of the variance of the data, we decided that reducing the dimensionality of our data down to four features was enough, so we pro-

jected our training and test datasets into our new feature spaces.

After reducing the dimensionality of the data, we ran our dataset through k-fold cross validation to decide which model best suited the data.

4 Methods

4.1 Linear Regression

We used linear regression as a baseline. This allowed us to assign coefficients θ to each feature depending on its effect on the labels, according to the following equation below. It does so by minimizing the residual sum of squares between the labels and the values predicted by the model.

$$h(x) = \theta^T x \quad x \in \mathbb{R}^{Nx^4}, \theta \in \mathbb{R}^4$$

We used the `LinearRegression` function with the `sklearn` library to run this function.

4.2 Weighted Linear Regression

We used weighted linear regression to place more emphasis on points with less variance. We therefore defined weights as the inverse of the variance squared so that examples with smaller error variance will be assigned more weight since they should provide relatively more information about the model. We therefore minimized the following function:

$$\sum_i w^{(i)} (y^{(i)} - \theta^T x)$$
$$x \in \mathbb{R}^{Nx^4}, \theta \in \mathbb{R}^4, w \in \mathbb{R}^4, y \in \mathbb{R}^4$$

The WLS function from within the `statsmodels` library was used to run this regression.

4.3 Random Forests Regression

We wanted to include models to see if our problem was a non-parametric regression, where the data is

not directly based off of the features, but information that is derived from the features. One way to do this was through decision trees and combining a number of "weak" models to create a strong end model. Random forests chooses random bootstrap samples of the data to construct n trees. Each of these nodes then branch off based on the best split using the best predictor among a randomly-chosen subset of predictors at that node. The algorithm then averages the predictions from each of these trees are then averaged to get the final prediction [7]. This also helps to prevent any overfitting. To run random forests, we used Scikit-learn's machine learning Python library, which includes a random forest regressor: `sklearn.ensemble.RandomForestRegressor` [9]. We used minimizing mean squared error as the criterion that would determine each node split.

$$MSE = \frac{1}{n} \sum_i (y - \hat{y})^2$$

4.4 Gradient Boosted Regression

Another decision tree regression was a gradient-boosted one. It assumes that at each level of the algorithm is a "weak" model that can be improved by fitting a stronger model using it [8]. We used Scikit-learn's machine learning Python library for this as well, using `sklearn.ensemble.GradientBoostingRegressor` [9]. For each stage, we minimized the least squares function and fit a regression tree based on it.

$$LS = \frac{1}{2} \sum_i (y - \hat{y})^2$$

This, and random forests, allows us to create complex regression models.

Model	Training MSE	Dev Test MSE
Linear Regression	9.141	9.983
Weighted Linear Regression	9.202	9.947
Random Forest	5.276	11.792
Gradient Boosting	10.061	13.056

5 Experiments, Results, and Discussion

5.1 k -fold Cross-Validation

Since we were not sure if our data would work better with a simple or a complex regression, we decided to run k -fold cross validation on all of the methods above. We used $k = 10$.

In order to choose our hyper-parameters for random forest and gradient boosting, we first tuned them on the whole training set and then used those parameters in the cross-validation. For random forest, we tuned the maximum depth of the tree between a range of 4-6 using `sklearn`'s `GridSearchCV`, which gave us the optimal parameters of 4.

For gradient boosting, we tuned the learning rate and the maximum depth of the tree. For the learning rate, we tuned between a range of 0.000001-1 on a log scale and the maximum depth between a range of 4-6 using `GridSearchCV` as well. The best parameters were a learning rate of 0.0001 and a maximum depth of 5.

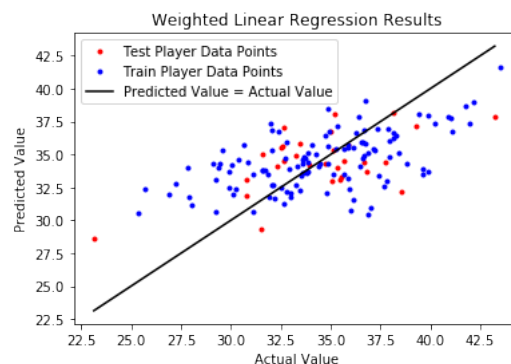
For each of the models, we calculate the average mean-squared error as our metric, which is as follows:

$$MSE = \frac{1}{n} \sum_i (y - \hat{y})^2$$

For each model, we found the average mean-squared error for the training set and the development test sets that was held out using k -fold. The following table summarizes the results from the cross validation.

Parameter	Standardized Coefficients
ncaa fg3	4.54E-04
ncaa fg3a	4.54E-04
ncaa fg3 pct	4.32E-04
ncaa ft pct	9.06E-04
ncaa ft sos	4.63E-04
ncaa team fg3a avg	4.24E-03
nba avg team ortg	2.34E-03
nba relative team fg3a	3.31E-04

As the weighted linear regression performed best on the development test sets in k -fold, we decided to proceed forward with that model. We therefore trained our model on on the complete training set using a weighted linear regression model, and used that model to predict NBA 3-point percentages for our test set. Below is a graph of actual 3-point percentages of the players in our training and test set plotted vs the 3-point percentages predicted by our model.



One can see that while many points do cluster around the predicted value = actual value line, some outliers do exist that our model failed to accurately predict.

In addition, in order to see which parameters were most significant in influencing our predicted nba 3-point percentage, we normalized the theta values by the average of that parameter. The standardized thetas are in the table above. We can observe from the relative sizes of the standardized coefficients that

the player's college team volume of shooting as well as the offensive rating of the nba team they joined ended up having the most impact in our model. Therefore, in our model, much of the player's projected 3-point skill is dependent on their team's ability rather than their individual statistics

6 Conclusion/Future Work

Though our model was able to accurately capture the shooting percentages of players who were average three-point shooters, we found that there were too many outliers (e.g. Tony Wroten) for the model to be used reliably. It's unclear where the problem lies in catching and accounting for outliers. Perhaps there is a more effective method to capture the nonlinearities of predicting three-point shooting percentages. Additionally, we could have missed key features that are highly correlated with a player's three-point shooting percentage.

With more time and data, we would first refine the model to be more accurate. From there, it would be interesting to expand the model to include international players. As the NBA expands its market globally, players like Kristaps Porzingis and Nikola Jokic have been drafted from overseas and have performed at all-star levels. In general, NBA fans know less about the abilities of international prospects compared to their NCAA counterparts, so building models that generated more information on them could be very insightful.

Additionally, we could expand the model to account for players who only began to shoot threes in the NBA. For instance, since NBA teams began to covet the 3-point shot more in recent years, a higher value has been placed on bigs that can shoot the ball. Consequently, players such as Marc Gasol and Brook Lopez have expanded their range beyond the arc, and have become proficient 3-point shoot-

ers. Including more features that could be indicative of a player's 3-point shooting capabilities (e.g. mid-range proficiency) could help predict which players might add the 3-point shot to their game.

7 Contributions

Jerold wrote the web-scraping code and collected the data, and wrote the gradient boosting and k-fold cross validation code.

Hilary wrote the code to split the data into its training and test set, and wrote the principal component analysis and random forest code.

Roland wrote the linear regression and weighted linear regression code.

All team members collaborated on debugging the code, analyzing the results, and creating the write-ups.

References

- [1] W. Abrams, J. Barnes, and A. Clement, "Relationship of Selected Pre-NBA Career Variables to NBA Players Career Longevity," *The Sports Journal*, April 2, 2008. [Online]. Available: <http://thesportjournal.org/article/relationship-of-selected-pre-nba-career-variables-to-nba-players-career-longevity/>. [Accessed November 10, 2017].
- [2] D. Coates and B. Oguntimein, "The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes?" 2008. [Online]. Available: http://web.holycross.edu/RePEc/spe/CoatesOguntimein_NBA.pdf. [Accessed November 18, 2017].
- [3] C. Gaines. Nearly 30% of all shots in the NBA now come from behind the 3-point

- line. *Business Insider*, March 8, 2016. <http://www.businessinsider.com/nba-three-point-shooting-2016-3>. [Accessed November 1, 2017].
- [4] B. Gajewski and T. Bishop. "Drafting a Career in Sports: Determining Underclassmen College Players Stock in the NBA Draft," 2004. [Online]. Available: https://works.bepress.com/byron_gajewski/10/download/. [Accessed November 20, 2017].
- [5] J. Harris and D. Berri. "Predicting the WNBA Draft: What Matters Most from College Performance?" 2005. [Online]. Available: <http://daveberri.weebly.com/uploads/6/1/3/8/61387427/2015harrisberrijsf.pdf>. [Accessed November 20, 2017].
- [6] A. Johnson, Predictions Are Hard, Especially About Three Point Shooting, *Counting the Baskets*, Sept. 2014. <http://counting-the-baskets.typepad.com/my-blog/2014/09/prediction-are-hard-especially-about-three-point-shooting.html>. [Accessed November 18, 2017].
- [7] A. Liaw and M. Wiener. "Classification and Regression by randomForest," *R News*, vol. 2/3, 2002. [Online]. Available: <http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf>. [Accessed Dec. 8, 2017].
- [8] A. Natekin and A. Knoll, Alois, "Gradient boosting machines, a tutorial," 2013. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/>. [Accessed Dec. 8, 2017].
- [9] Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR 12*, pp. 2825-2830, 2011. [Accessed Dec. 8, 2017].