Predicting the Draft Slot of NBA Prospects in the Modern Era

Kevin Hillyard & Brent George
CS 472, Winter 2021
Department of Computer Science
Brigham Young University

Abstract

Every year experts attempt to determine which basketball players are the most valuable and when those players should be picked in the NBA draft. In this paper, we record our efforts to accurately predict where players should be drafted based on their stats using different machine learning models. We first describe by what means our data was gathered and what type of data we gathered. We then review the results we had in predicting a player's draft position with our different models, namely KNN, decision tree, multi-layer perceptron, and linear regression. We then review our findings, especially from linear regression, in summarizing which statistics were important, and which were not as important. Lastly our ideas for future work and improvements to our research are made known.

1 Introduction

NBA scouts are always looking for NBA potential players. When determining how good a player is they are almost always compared to past players. Most people are able to predict the top 3 to 5 NBA picks from the available players but it is nigh impossible to predict each draft pick accurately.

In our research we attempted to use machine learning models to predict where a player would be drafted. These models use the past 11 years of NBA draft results, along with the players' associated statistics, to determine where the player should be drafted if they are to be drafted at all.

There was some difficulty in predicting a player's draft position as it is a human perception of that player's capability and is often swayed by what each NBA teams' current needs are. We were also limited to a certain number of players as the era of basketball has changed, encouraging more long range field goals compared to years prior. This drastically affects the typical stats for players making comparisons between newer and older players less viable.

Despite those difficulties we were able to predict a player's draft position to within 10 picks. In a future study if we were to account for the aforementioned issues the accuracy of our results could be increased.

2 Methods

2.1 Data Source

There are various online providers of basketball statistics. Official sources include NBA.com and the official site of NCAA basketball. However, official sources often do not include more than the most basic statistics.

Several sites exist that provide extensive college statistics. Many of these have a paywall (kenpom.com) or do not include statistics for international prospects. Other sites request not to be scraped by bots.

After further exploration, we found realgm.com, a site that consented to limited scraping that also included historical data for past statistics, data on international prospects, and a wide array of advanced statistics. We wrote a web scraper in Python to traverse the element tree of the web page and lift each of the players selected in the past 11 NBA drafts and every available statistics associated with those 660 players. This formed our initial dataset, which was stored in a CSV file. We also pulled the records of nearly 736 prospects that went undrafted during that same time period.

2.2 Featurization of Data

Due to the nature of basketball statistics, the data scraped from the website was already largely in the form we needed. We wrote a Python script to perform a few sanitizing steps and convert the CSV file to an ARFF file.

The main sanitization we needed to perform was the consolidation of various leagues that prospects played in. Prospects played in over 35 leagues and countries. While there was consistency for prospects that played in the NCAA in the United States, there was extreme fragmentation among the teams and leagues that international prospects played for. In order to help the model, we grouped all international teams under one label, leaving us with three total league labels: NCAA, G-League, and International.

While drafted players were labeled according to their draft selection, the undrafted players that we pulled were unlabeled. We didn't want to use a clustering algorithm to label these instances - a clustering algorithm would attempt to label these players with a pick number from 1 to 60, whereas we wanted to be able to predict if a player went undrafted. We decided to label any undrafted player as being picked 61st, with the plan of interpreting our future model's output under this assumption. This led to over half of our dataset consisting of players "selecting 61st in the draft". This would prove to have some positive effects, such as giving the model many examples of poor prospects on the higher end of the draft. It would also have negative effects, such as leading the models to be biased towards higher pick numbers when it was unsure.

2.3 Models

We used a variety of models in this project, to be described in further detail in results sections. All implementations came from scikit learn v1.0.1.

3 Initial Results

Initially, we trained Decision Tree, Multi-Layer Perceptron, KNN-Neighbor, and Linear Regression models. Each of these had a grid of hyperparameters that was exhaustively searched to find the best results, as recorded in Table 1.

4 Feature and Model Improvements

4.1 Improvements

To improve our results we reviewed the most important features and least important features as determined by the weights from our linear regression model. The features that had the lowest weights included the following features: Usage %, Assist %, Field Goals Made, Field Goals Attempted, Free Throws Made, Free Throws Attempted, 3-pointers Made, and 3-pointers Attempted. We deemed these features safe to remove as they had low weights and the essence of their statistic was captured in a different feature. For example, Field Goals Made and Field Goals Attempted is captured in the feature Field Goal % which has a high weight. In removing these somewhat redundant features we were able to slightly improve our predictions.

5 Final Results

6 Discussion and Conclusion

7 Future Work

There are some minor changes we would make to our research to improve the results of our predictions. These changes include further research about the data, how the error of our models was determined, and some additional features.

7.1 Further Draft Research

As previously mentioned in this report there were some contributing factors to a player's draft stock that we did not take into account when gathering our data. When gathering our data and when a player was drafted it would be worthwhile to research the team's needs that drafted that player. Knowing this would allow us to determine if the player was drafted at that position for their skills as a player or more for their fit with that team.

7.2 Error Scoring

There was an issue with the error scoring with our models as we were using the scikit models with their own error scoring function. Because undrafted players don't have a draft position associated with them we edited our data so that all undrafted players were drafted 61st. The issue appears when our models predict a draft position greater than 61. Our data has a max of 61 while our models are able to predict draft positions higher than this. The error scoring considers this difference as an error when in reality if the player has a draft position of 61 any placement greater than or equal to 61 should be considered correct.

7.3 Additional Features/Sources

As a final change to our research we would gather data from one or more additional sites in order to gather all advanced statistics possible. The site we pulled our data from had a large number of advanced statistics but not all of them.

Dataset	Algorithm	Centroid #	SSE	Size	Silhouette Score
Debug	Single Link HAC	Total	54.4392	200	0.3453
Debug	Single Link HAC	1	54.3917	195	
Debug	Single Link HAC	2	0.0000	1	
Debug	Single Link HAC	3	0.0475	2	
Debug	Single Link HAC	4	0.0000	1	
Debug	Single Link HAC	5	0.0000	1	

Table 1: Test table