Using Pre-NBA Draft Data to Project Success in the NBA

Ryan Edwards
Stanford University
Civil and Environmental Engineering
Email: rwe@stanford.edu

Chris House
Stanford University
Electrical Engineering
Email: chouse12@stanford.edu

Nathan Lord
Stanford University
Management Science and Engineering
Email: nathanl@stanford.edu

Abstract—**The NBA draft poses a unique challenge in that predicting a player's future success in the NBA is incredibly difficult. We seek to use machine learning techniques to quantify the attributes that tend to indicate a college player's playing ability in the NBA. Using historical data from players' college careers in combination with their NBA career data, we have developed a model to predict where a player should be drafted (if at all) in the NBA draft. While predicting the exact number of win shares per season that a player will contribute was difficult, predicting whether a player would be successful, or what discretized level of success that player would attain in the NBA was easier to predict with-Wheeler, Kevin. "Predicting NBA Player Performance," 2012. [Online]. Available: http://cs229.stanford.edu/proj2012/Wheeler-PredictingNBAPlayerPerformance.pdf higher certainty.**

I. Introduction

In June 2015, Forbes reported that the number one draft pick of the 2015 NBA draft, Karl-Anthony Towns, is contractually guaranteed $11,663,760 over the first two years of the contract with the potential to make $25,720,035 over its entirety of four years. Subsequent draft pick values decrease to just above a $2 Million, 2 year contract guarantee at the end of the first round of the draft (30th draft pick). Such non-trivial investments indicate the high value placed on these players by general managers who are ultimately looking to improve the production of their organization. It is thus extremely valuable to an organization to be able to accurately estimate the production of a particular player before making such an investment.

For many years, general managers made these predictions based on scouting reports, college production, and other objective and subjective criteria. The input to our algorithm, unlike with general managers, is purely objective data - a player's college statistics. We then use a SVM to output whether a player will be successful in the NBA and linear regression to predict how successful the player will be. Thus, we aim to turn a subjective process of evaluating a player's NBA prospects into an objective process and therefore able to make better draft choices and investments in NBA players.

II. Related Work

While the NBA draft is an interesting problem in terms of predicting a player's future success, not many attempts have been made at quantitative solutions. A former CS229 student used linear regression on NBA statistics in order to predict how many points a player would score given a specified opponent team [1]. This is somewhat similar to predicting future success, like we are attempting to do, with the main difference being that the feature space in this situation includes past NBA data, which would not be available prior to a player being drafted. Furthermore, predicting a player's point total is not exactly the same as predicting a player's level of success: while a player like Allen Iverson was a prolific scorer, he was notorious for being an incredibly inefficient scorer, ultimately hurting his team's chances at winning in some cases. Therefore, classifying success based on win shares is more accurate.

Another CS229 project involved predicting a player's success from year to year [2]. This group created their own success metric, which was a weighted average of a combination of statistics to create a canonical player success average, and then taking the L2-norm of a player's given statistics for a specific year to determine that player's yearly success. Then, the group classified all players in their dataset into clusters based on their career trajectories. Their methods were interesting, as they came up with a cluster that contained predominantly star players, another cluster with average players, and another cluster with below average players. However, like the above project, their methods predicted future NBA success based on previous NBA data, not previous college data. In this sense, our project is unique for attempting to classify players into different groups based off of their college statistics.

III. Dataset and Features

A. Data Collection

It was difficult to collect the data due to lack of easy availability, innacurate sources, and a lack of organized basketball statistics for both college and NBA players. After searching through many sports websites, we found that basketball-reference.com had all the data we needed and the easiest html code to parse.

We programmed two scripts in Python to download and scrape thousands of NBA player html files from basketball-reference.com. These players only included those that had data recorded for both college and the NBA, so some players before 2006 were not included due to a lack of college data (players like Kobe Bryant or Lebron James went straight from high school to the NBA - a rule change in 2006 prevented this practice). Once downloaded, we parsed through each html file to extract the basic and advanced statistics for each season the player played in the NBA, and stored the data in the respective basic and advanced tab-separated data files. The downloading and parsing scripts combined were about 100 lines of code, and the scraping took about 6 hours total to complete.

B. Win Shares - An NBA Success Metric

In order to quantify how successful or useful a player in the NBA is, basketball-reference developed a metric known as, "win shares." This metric is a measure of how many wins a player is responsible for during their NBA career. For example, during a given game, LeBron James might perform at a level to produce 0.3 win shares while his teammates combine to produce 0.7 win shares (assuming their team wins), for a total of 1 (team) win. Basketball-reference computes win shares based off of several advanced basketball metrics, including modified points (attributing the number of points scored by a player minus his team's contributions to the point total), modified shot attempts, league averages for modified shot attempts, a player's defensive rating (measure of how many points allowed by a player over 100 possessions), and a measure of his team's defensive ability. Essentially, win shares provides information on how effective a player is at increasing his team's point total (offensive win shares) as well as stopping the other team's point total from increasing (defensive win shares). Win shares were our metric for an NBA player's success and our goal was to determine whether there were certain attributes from college players that accurately projected their NBA win share total.

C. College and NBA features

In order to classify or predict a player's success based on win shares, we used data collected from their college careers. These features included games played, minutes played, points scored, field goal percentage, 3 point percentage, free throw percentage, assists, and rebounds, among others. Each player's college data was taken as an average over their career, so certain statistics which were cumulative, like points scored or games played, were somewhat skewed dueto some players playing as few as 1 college season all the way to 4 college seasons. We also had access to each player's NBA data which we used for comparison with our linear regression method on college statistics for predicting NBA win shares.

IV. Methods

A. SVM

Support vector machines are a widely used machine learning technique for classification problems (although the technique can also be extended for regression problems). SVM's work by attempting to find separating hyperplanes between data classes that maximize the geometric margin between the data and separating hyperplane. This gives rise to the following optimization problem:

$$\min_{\gamma,w,b} \frac{1}{2}\|w\|^2$$

$$s.t. y^{(i)}(w^T x^{(i)} + b) \geq \gamma, i = 1, ..., m$$

In the above equations, gamma represents the geometric margin between each training example and the support vector is w. To solve this problem, it is transformed into a convex optimization problem by solving the dual equation [3].

B. PCA

Principal components analysis is an algorithm for reducing feature space by identifying the axis of highest variation for the feature space. Typically, PCA is performed on data that is normalized to have zero mean and unit variance. PCA then identifies the axis of the data with the highest variance (and subsequent axes) by computing the eigenvectors of the data's covariance matrix in order to maximize the quantity $\frac{1}{m}\sum_{i=1}^{m}(x^{(i)^T}u)^2$. Under the condition of $\|u\|_2 = 1$, this amounts to finding the eigenvectors of $\Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}x^{(i)^T}$. Finally, the data is then projected onto a k-dimensional subspace (to reduce the feature space dimensionality) by choosing the projection vectors to be the first k eigenvectors of the data [4].

C. Linear Regression

Linear regression is a method used for generating predictions on continuous data. The goal is to determine the relative weights (theta) of $h(x) = \sum_{i=0}^{m}\Theta_i x_i$ of each feature (x) so as to minimize the given cost function:

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})^2$$

This cost function gives rise to least squares, where the objective is to minimize the mean squared error between the model and data [5].

V. Results and Discussion

A. Baseline Win Share Expectation

In order to determine a baseline expectation for our model's performance, we looked at historical data from the previous 15 NBA draft classes to determine the average win share total by draft position. Using an exponential regression, we determined a baseline prediction for win shares/game for each draft position, summarized in Figure 1.

While this data was useful for understanding the success of general managers at predicting player value, predicting exact win shares was somewhat difficult and thus this baseline was difficult to apply to our algorithm. Therefore, we needed another way to determine how successful GM's were in predicting a player's NBA success.

One way to measure success for the NBA draft is to better understand the point of the draft. Out of all the players in the

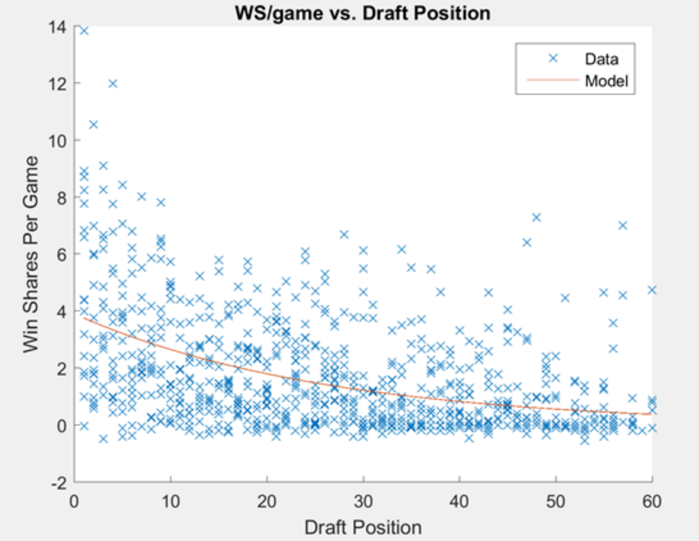

Fig. 1. Exponential regression on win shares based on draft position

NBA currently, approximately 97% that are starters came into the league through the NBA draft [6]. Therefore, one might simply have the goal of drafting a player capable of starting in the NBA. In order to determine players that were "starting level" players, we took all players who averaged greater than 25 minutes played per game (over half of an NBA game) over their career and then compared their win share per game metrics. Using our draft class data, roughly 30-38% of players surpassed the mean/median (use of median in order to control for outliers) win share per game metric, implying that general manger's are succesful at drafting a starter level player only 30-38% of the time. Therefore, in predicting whether a player will be successful based on their college data, we aimed to beat this mark of 30-38% accuracy in classification.

B. Multi-Class Support Vector Machine (SVM)

We implemented a multi-class SVM using libsvm [7] in order to classify players in our data into a different numbers of classes. We tested on three different setups – 2 classes, 3 classes, and 5 classes. The classes are based on win shares per game and are roughly broken down into the following classifications proportional to the number of players in each class in the NBA: 1) MVP candidate players (top 1% of NBA), 2) All star players (25 players each year - 5% of NBA), 3) consistent starters (~26.7%), 4) bench players (~20%), and 5) non-contributing players (~46.7%). We are looking to use features that pertain to their playing abilities before playing in the NBA (i.e. statistics from their careers in college, height, weight, draft position, etc.). The results were encouraging, as we were able to improve significantly over both chance and the baseline of 30-38% success rate for general managers. Table 1shows the results of our multi-class SVM classification attempts.

TABLE I
SVM PREDICTION SUCCESS BY NUMBER OF CLASSES

Number of Classes	Prediction Accuracy	Improvement over Chance	Improvement over GM (38%)
2	70.4%	20.4%	32.4%
3	51.2%	18.2%	13.2%
5	48.4%	28.4%	10.4%

C. Principal Component Analysis (PCA)

After developing this algorithm, PCA was applied to the features to reduce the feature space. After applying PCA, our feature space was reduced down to 6 features: games played, points scored, field goal percentage, 3 point percentage, free throw percentage, and assists. We then attempted classification on the same classes and compared the accuracy to the original classification efforts. The results of this updated model can be seen in Table 2. It can be seen that with fewer classes PCA yielded worse results, but with 5 classes, PCA improved classification accuracy slightly. In all cases, however, it could be seen that classification predictors were far more successful than strict chance as well as current general manager predictions.

TABLE II
PREDICTION SUCCESS AFTER APPLYING PCA

Number of Classes	Prediction Accuracy	Difference from Original
2	68.3%	-2.1%
3	48.9%	-2.3%
5	49.8%	1.5%

D. Linear Regression

General manager's have often admitted to only analyzing college players' basic statistics (e.g. FG%, 3P%, AST) during the scouting process. So performing linear regression on players' basic college statistics and their NBA win shares can create the "idealized GM" predictor.

We were curious to see whether it is able to significantly outperform the "idealized GM" by analyzing college players' advanced statistics as well (e.g. BPM). So we performed linear regression on players' advanced college statistics and their NBA win shares. We then computed the absolute win share errors of the basic and advanced predictors to see if the advanced predictor outperforms the "idealized GM." The results are shown below.

As seen in the boxplots, the advanced predictor produces essentially the same error as the "idealized GM," which indicates that the basic college parameters (FG%, 3P%, AST) are the most valuable ones used when predicting college players' future NBA success.

There are far more advanced statistics available for NBA players than for college players (real plus/minus, value over replacement player, etc.). So for a further study we performed linear regression on basic NBA statistics and NBA win shares, and on advanced NBA statistics and NBA win shares. We then graphed the absolute win share errors below.

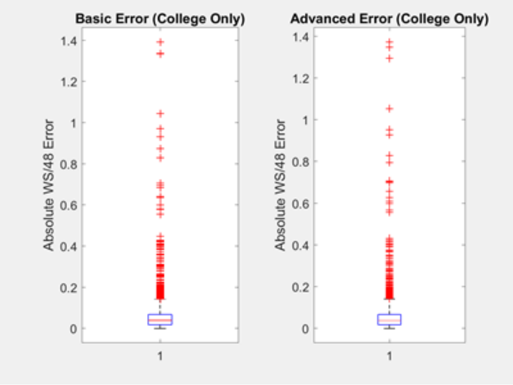

Fig. 2. Box plot displaying error of win share prediction linear regression. The feature space includes only college statistics and the basic plot only has the six most relevant features.

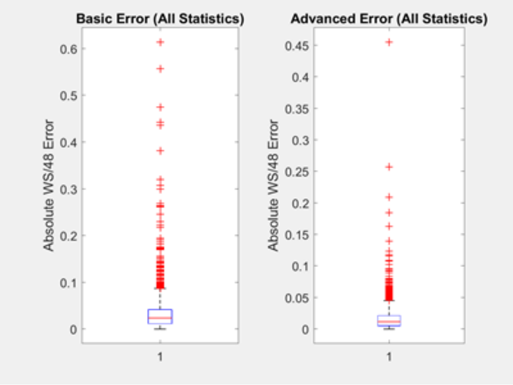

Fig. 3. Box plot displaying error of win share prediction linear regression. The feature space includes all statistics (college and NBA) and the basic plot only has the six most relevant features.

VI. Future Work

While predicting win shares via linear regression was difficult with sufficient accuracy, classifying players by discretized success levels yielded much more success. The information yielded from these models could be used to inform general managers, owners, and coaches when determining whom to draft at what position. These models are useful not only on an individual team's basis, but could also prove to be useful to the league as a whole. Currently, even very low draft picks are paid large sums of money. The fact of the matter is that some players are statistically unlikely to perform successfully. Being able to determine this likelihood will enable the league as a whole to more accurately negotiate contract values for different draft positions.

Future attempts at improving these methods would primarily focus on expanding the feature space to include more relevant features. Our features did not include phyiscal measurables such as height, weight, wing span, etc. An additional feature that would be interesting to study would be the university that each player played for. Some NBA draft "experts" conjecture that better players tend to come from college basketball powerhouses such as Duke and Kentucky, whereas a player from a division II or III school is less likely to be successful in the NBA.

VII. Acknowledgements

Thank you to Professor Ng and the CS229 teaching staff for a great class.

References

[1] Wheeler, Kevin. "Predicting NBA Player Performance," 2012. [Online]. Available: http://cs229.stanford.edu/proj2012/WheelerPredictingNBAPlayerPerformance.pdf

[2] Cousland, Alex, et. al. "Predicting Career Paths of NBA Players," 2012. [Online]. Available: http://cs229.stanford.edu/proj2012/ShahCou

[3] A. Ng, Support Vector Machines, 1st ed. Stanford university, 2015, pp. 1-25.

[4] A. Ng, Principal Components Analysis, 1st ed. Stanford university, 2015, pp. 1-6.

[5] A. Ng, Supervised Learning, 1st ed. Stanford university, 2015, pp. 1-30.

[6] Crabdribbles.com, 'NBA Draft Over the Last 15 Years: A Statistical Overview', 2015. [Online]. Available: http://www.crabdribbles.com/nba-draft-over-thelast-15-years-a-statistical-overview/. [Accessed: 07- Dec-2015].

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm