Generating Relative Draft Value in the NBA Draft and Predicting Success from College Basketball

A Major Qualifying Project Report:

submitted to the faculty of the

**WORCESTER POLYTECHNIC INSTITUTE**

in partial fulfillment of the requirements for the

degree of Bachelor of Science

by

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Michael Krebs*

­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Jake Scheide*

Date:

Approved: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Professor Craig Wills, Major Advisor

# Abstract

# Acknowledgements

# Table of Contents

# Table of Figures

# Executive Summary

# 1. Introduction

This project’s goals are threefold: firstly, to evaluate the accuracy of current and developed player performance metrics for basketball players in the National Basketball Association (NBA). Secondly, to identify the relative draft value for the sixty picks in the NBA Draft. Finally, we present a model which predicts the draft position and performance for an individual based on their National Collegiate Athletic Association (NCAA) Division I statistics.

To achieve our first goal, we first select four of the most popular player performance metrics in the NBA, and analyze what sort of players and careers they favor. Additionally, we create two custom metrics which provide contrast to the established statistics and generate additional insight into why certain metrics are used.

Taking these insights into account, we then incorporate draft data to approximate how valuable each selection in the NBA Draft is, both relative to a player and another draft pick. We create a National Football League (NFL)-style draft value chart, and plot each of our six metrics and observe how talent changes with draft position comparably to salary.

For our third goal, we use ­\_\_\_\_\_\_\_\_\_ to construct a model which predicts the expected draft position for a player in a draft year, as well as their projected success in the NBA.

Through this process, we will be able to contextualize the plethora of player performance statistics, examine the value teams assign to draft picks compared to our statistical analysis, and

# 2. Background

## Existing Metrics in the NBA

Basketball on Paper – Dean Oliver

Ranking the Greatest Nba Players: A Sport Metrics Analysis - Mertz, Jeremy ; Hoover, L. Donald ; Burke, Jean Marie ; Bellar, David ; Jones, M. Lani ; Leitzelar, Briana ; Judge, W. Lawrence

## Assessing Draft Value in other sports

Evaluating Talent Acquisition in the NFL – Prof. Wills

## Assessing Draft Value in the NBA

Assessing the Relative Value of Draft Position in the NBA Draft - By Aaron Barzilai, Ph.D.

Sunk Costs in the NBA: Why Draft Order Affects Playing Time and Survival in Professional Basketball - Staw, Barry M & Hoang, Ha

## Predicting NBA success based on college performance

Drafting a Career in Sports: Determining Underclassmen College Players' Stock in the NBA Draft - Bishop, Todd ; Gajewski, Byron J.

The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes? - Coates, Dennis ; Oguntimein, Babatunde

# 3. Methods

To ensure we were extracting a valid performance metric for NBA players, we needed to look at more than just one number. Fortunately, many statisticians have tried to represent the worth of a player in a single figure.

### Metric 1: Player Efficiency Rating

The most common metric used when comparing NBA players is John Hollinger’s Player Efficiency Rating (PER), which is where we began our research. PER is calculated primarily using offensive box score statistics, a common problem identified by Dean Oliver, widely regarded as the pioneer of advanced basketball analytics. PER is normalized to both minutes played and pace, which is an important distinction we will analyze further later.

### Metric 2: Win Shares

A statistic which does not normalize to minutes played, but instead contribution to wins, is Win Shares (WS). Win Shares in basketball is derived from Bill James’ work in baseball, and follows a similar format. It attempts to reflect how many wins a particular player contributed to his team for a season, whilst taking into account pace of play. When evaluating these metrics’ ‘true’ accuracy, we will investigate how these normalizations warp the rankings.

### Metric 3: Value Over Replacement Player

Value Over Replacement Player (VORP) was our third metric analyzed in our project. VORP estimates overall contribution to a team, by first creating an adjusted plus-minus statistic, then normalizing to minutes played and the value a theoretical minimum-salary player would bring in replacement of that player. An important note is that VORP was created to scale linearly with salary, so that a player with twice the VORP should be worth twice the salary.

### Metric 4: Fantasy Points

The fourth and final non-feature engineered statistic we used was Fantasy Points (FP). Fantasy Points is a simple metric calculated by weighting each of the basic box score statistics (Points, Rebounds, Assists, Blocks, and Steals) as well as turnovers and producing a sum. We chose to use FP because it provides a different angle to player evaluation- one which fans will appreciate. FP is not normalized to minutes played nor pace.

### Metric 5: Basic Percentile

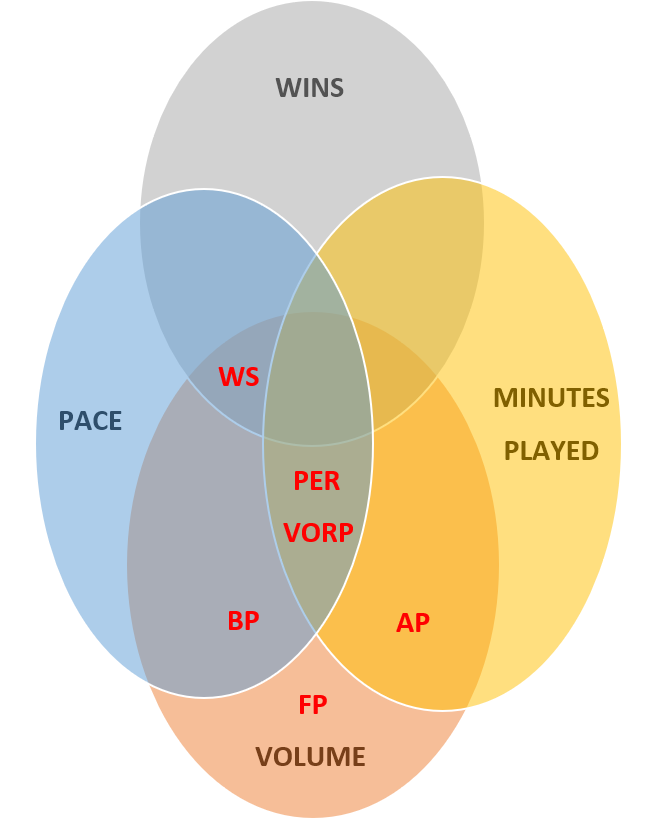
After researching existing metrics and developing an understanding of what already exists, we set out to create our own metric that measures a different variant of player performance. Basic Percentile (BP) ranks a player on their basic box score stats relative to the rest of the league. This means a player who leads the league in Points, Rebounds, Assists, Steals and Blocks would receive a score of 5 for a season. A player first in all categories except third in steals would receive a 7 (1+1+1+3+1). This naturally normalizes to league pace because it considers the raw volume from every other team that year.

Figure 1: Classification of Metrics by Normalization

### Metric 6: Advanced Percentile

Advanced Percentile (AP) follows a similar train of thought, but it instead looks at True Shooting Percentage (TS%), Rebound Percentage, Assist Percentage, Steal Percentage, and Block Percentage. The percentage metrics evaluate the proportion of total box score stats available a player amasses while they are on the court. Therefore, we can say that AP naturally normalizes to minutes played.

### Cumulative vs Per Season

One major decision crucial to the analysis was choosing to look at career totals for each metric, or instead take an average based on the number of seasons data was collected for. Both approaches have benefits in specific situations. By totaling a player’s career metrics, we can more accurately evaluate players with other players. Longevity is a major factor in careers, and many superstars’ numbers are hampered by playing diminished minutes (and therefore lower statistical output) in the twilight of their careers. A cumulative number provides better insight as to the total value of that player’s career.

For other analyses, however, it makes more sense to look at an average per season. When looking at data over a finite time, and ranking players whose careers do not align, it would be unfair to look at ten years of one player’s career and only five of another. Another shortcoming of only looking at cumulative statistics is the unlikelihood that a player will spend an entire career with one team. By averaging out an average season value, teams can approximate player value for trades and signings. Because each approach has benefits, we will switch between the two depending on the scenario.

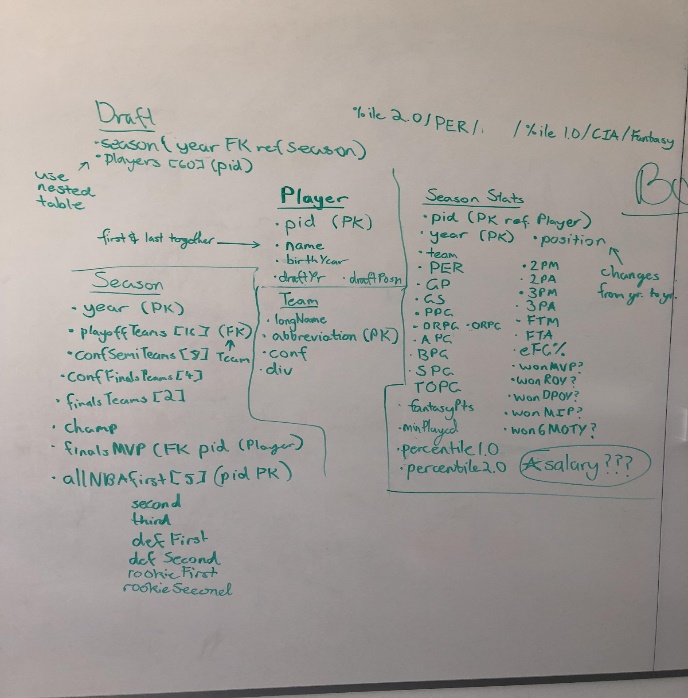
Our first drafted Entity Relationship Diagram is shown to the right. In the end, we modified this slightly to have a Master Season sheet for every year with all player season stats, a Draft sheet with the players drafted that season, and a Master Players sheet with all players and their IDs. With a clear goal with how we were going to structure our data we then set out to collect the data. Initially, we were going to reach out to a 3rd party and obtain a licensing fee for their data. As students, we were not able to access their data, and so we looked to online resources. Using Python 3.7 we incorporated the BeautifulSoup4 library to web crawl the Basketball-Reference.com website. From this website we were able to obtain all of the single season metrics for every player, all single season awards, the draft selections for every player who played in 1990 to 2018 and utilized the player unique id from basketball reference as our own primary key.

Figure 2: Initial Entity Relationship Diagram (Convert to Lucid Chart)

With the help of the Python package Pandas we created dataframes for all the information that was pulled from Basketball-Reference.com. We would then export the dataframes to local spreadsheets so we could access them again in the future. Below is a picture of our spreadsheet for the 2014 NBA season

A picture containing rain, nature

Description generated with high confidence

# 4. Design

# 5. Results

## Evaluating the Evaluators: Analyzing Player Performance Metrics

As discussed in the methods section, a crucial decision in evaluating player value is how ‘performance’ is quantified. The below table lists the top 20 players ranked using our six selected metrics, averaged out over the course of each player’s career.

Table 2: Top 20 Table by Metric

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | WS | PER | VORP | BP | AP | FP | AVG |
| Karl Malone | 2 | 4 | 2 | 4 | 3 | 2 | 2.83 |
| LeBron James | 1 | 2 | 1 | 7 | 8 | 1 | 3.33 |
| David Robinson | 4 | 1 | 4 | 6 | 1 | 5 | 3.5 |
| Tim Duncan | 8 | 6 | 9 | 11 | 7 | 4 | 7.5 |
| Charles Barkley | 10 | 5 | 6 | 9 | 5 | 21 | 9.33 |
| Hakeem Olajuwon | 21 | 10 | 17 | 3 | 2 | 10 | 10.5 |
| Kevin Garnett | 17 | 17 | 12 | 8 | 4 | 8 | 11 |
| Chris Paul | 5 | 7 | 5 | 21 | 21 | 11 | 11.7 |
| Kevin Durant | 7 | 8 | 15 | 13 | 21 | 7 | 11.8 |
| Shaquille O’Neal | 14 | 3 | 18 | 21 | 13 | 3 | 12 |
| Michael Jordan | 3 | 11 | 3 | 21 | 21 | 21 | 13.3 |
| Clyde Drexler | 16 | 19 | 7 | 5 | 14 | 21 | 13.7 |
| Larry Bird | 21 | 21 | 14 | 2 | 9 | 21 | 14.7 |
| Russell Westbrook | 21 | 16 | 8 | 21 | 21 | 6 | 15.5 |
| John Stockton | 6 | 13 | 19 | 21 | 21 | 16 | 16 |
| Scottie Pippen | 21 | 21 | 13 | 10 | 10 | 21 | 16 |
| Chris Webber | 21 | 21 | 21 | 1 | 11 | 21 | 16 |
| James Harden | 12 | 21 | 11 | 21 | 21 | 17 | 17.2 |
| Kobe Bryant | 20 | 15 | 21 | 14 | 21 | 13 | 17.3 |
| Dwyane Wade | 21 | 12 | 20 | 12 | 19 | 21 | 17.5 |
| Stephen Curry | 15 | 21 | 10 | 21 | 21 | 19 | 17.8 |
| Vlade Divac | 21 | 21 | 21 | 20 | 6 | 21 | 18.3 |
| Jason Kidd | 21 | 21 | 16 | 17 | 21 | 15 | 18.5 |
| Dirk Nowitzki | 13 | 18 | 21 | 21 | 21 | 18 | 18.7 |
| Magic Johnson | 9 | 21 | 21 | 21 | 21 | 21 | 19 |
| Dwight Howard | 18 | 21 | 21 | 21 | 21 | 12 | 19 |
| Yao Ming | 21 | 9 | 21 | 21 | 21 | 21 | 19 |
| Allen Iverson | 21 | 21 | 21 | 21 | 21 | 9 | 19 |
| Reggie Miller | 11 | 21 | 21 | 21 | 21 | 21 | 19.3 |
| Shawn Kemp | 21 | 21 | 21 | 21 | 12 | 21 | 19.5 |
| Anthony Davis | 21 | 14 | 21 | 21 | 21 | 21 | 19.8 |
| Gary Payton | 21 | 21 | 21 | 21 | 21 | 14 | 19.8 |
| Chris Bosh | 21 | 21 | 21 | 15 | 21 | 21 | 20 |
| Larry Nance | 21 | 21 | 21 | 18 | 18 | 21 | 20 |
| Christian Laettner | 21 | 21 | 21 | 21 | 15 | 21 | 20 |
| Shawn Marion | 21 | 21 | 21 | 16 | 21 | 21 | 20.2 |
| Paul Millsap | 21 | 21 | 21 | 21 | 16 | 21 | 20.2 |
| DeMarcus Cousins | 21 | 21 | 21 | 21 | 17 | 21 | 20.3 |
| Jeff Hornacek | 19 | 21 | 21 | 21 | 21 | 21 | 20.7 |
| Lamar Odom | 21 | 21 | 21 | 19 | 21 | 21 | 20.7 |
| Amare Stoudemire | 21 | 20 | 21 | 21 | 21 | 21 | 20.8 |
| Marcus Camby | 21 | 21 | 21 | 21 | 20 | 21 | 20.8 |
| Patrick Ewing | 21 | 21 | 21 | 21 | 21 | 20 | 20.8 |

Starting at the top, we can see a moderate amount of consistency. The top three players never drop below 8th for any metric, which is a good indicator that the metrics are in agreement about the players. Beyond this initial agreement, the model completely breaks down. Michael Jordan, for instance, is ranked third in two metrics but fails to make the top 20 in three others. Chris Webber is the best player according to the Basic Percentile metric, but doesn’t crack the top 10 in anything else. Overall, the chart is an excellent way to show just how different the rankings are based on how you evaluate talent. Another thing to consider is that by generating these metrics as per season values, they give less credit to long, consistent careers than short but bright careers.

*So, which metric is the best predictor?*

In a perfect world, this analysis would guide us to a single metric that is the most accurate portrayal of individual player talent. However, after observing how each metric is normalized, and the quirks that come with each approach, we conclude that specific groupings of metrics are more appropriate for use in different scenarios.

For instance, if fans are interested in comparing the careers of two retired players, a metric such as Basic Percentile or Win Shares would be most appropriate. Basic Percentile ranks players relative to the other players in the league during that year, and naturally adjusts for league pace of play. BP provides more insight into how dominant a player was relative to the players he was around at the time. Similarly, Win Shares considers how successful a team was during the regular season, which is often used subjectively in assessing careers. Finally, using cumulative values for these metrics would make the most sense, as both players have completed their careers and therefore a summative value is appropriate.

On the other hand, if an NBA front office is attempting to evaluate the talent loss or gain in a proposed trade, Player Efficiency Rating or Advanced Percentile would be better indicators. The number of minutes a player receives is based highly on the talent level of the other players sharing the position on the team, as well as coaching style. As a result, a player is likely to have a large change in minutes, and therefore statistical volume, when changing teams. Fantasy Points, Win Shares, Value Over Replacement Player, and Basic Percentile aren’t good options, as they don’t consider efficiency. Even though PER and AP are clearly better options for this kind of talent evaluation, care must be taken to ensure that outliers aren’t being generated due to a lack of minutes played. Additionally, it is more appropriate to look at per season values for these metrics (and potentially multiply by the seasons left on the player’s contract) to capture the value a trade adds or subtracts.

Overall, this analysis confirms that there is a reason for having so many individual player evaluation statistics. Therefore, when attempting to quantify draft value, we look at all six metrics.

## Average NBA Draft Pick Value

With a more nuanced understanding of the benefits and shortcomings of basketball player evaluation statistics, we can start to incorporate the draft into the analysis. The below graph uses cumulative career values for each of the six metrics collected. These values were then normalized with the first pick receiving a 100 for their value. By using cumulative values here, we help to display the phenomena that skews a lot of draft data: sunk costs. The first picks in the draft will almost always play large minutes, whereas later picks are rarely good enough to see the court. Because of this, the players who have NBA data in the 50-60 range are often seen as ‘steals’ and don’t accurately reflect the talent level generally attainable at that draft selection. Cumulatively totaling every first draft pick’s season will inevitably give a higher value, because there are more first picks found in the data than sixtieth picks.

On first look, the graph appears to be a jumbled mess of jagged lines. What’s more important than the directions of each line, however, is the separation between each y-value at a given x. At almost all values, Advanced Percentile is the highest value. This indicates that AP values later picks as having more comparable talent levels to higher picks than the other stats. Compare this to VORP, which immediately drops after the first few picks and flattens out after the 30th pick. VORP suggests that talent is heavily front-stacked, and very little value can be found in later selections.

To smooth out this data and better draw these conclusions, we clustered the draft positions into intervals. The results of this are shown below.

This graph gives a much clearer approximation of the talent drop off over draft pick. This graph also indicates the number of players found in each grouping, to further contextualize the metrics calculated. As described earlier, the seven picks from 8th to 14th have 301 players, whereas the cluster containing the fifteen final picks has only 388 players. This is one of the motivating factors behind using the cumulative metrics to indicate the true value of each draft position. Finally, we add the fixed rookie salary scale to provide a guideline as to how the NBA approximates the talent available at each draft position. As only first round picks have guaranteed salaries, the values used for second round picks are simply the league minimum.

## Predicting NBA Performance for NCAA DI players

# 3. Methods

Once we had a grasp on the value of a player and the expected value from a given draft pick we set out to predict NBA performance for NCAA Division I players. To do this, we first needed to gather statistics about al NCAA Division I players. Using the same methodology to pull data from Basketball-Refernce.com we were able to pull the college data from Sports-Reference.com. We were able to pull data from all NCAA division I teams from 2000 – 2018. But due to the need to lack of identifiers for an NCAA player (the ids used in sports reference are not the same as the ones used in basketball reference) we needed to manually enter where a player was drafted and so we focused on college players from 2010 to 2018. When we were evaluating NBA player performance only in game performance was accounted for, but since predicting NBA readiness and expected performance it is also necessary to consider physical attributes. Thus, we also made sure to collect height and weight measurements for all NCAA players. To further investigate how physical attributes play a role into the probability a college player will reach the NBA we also collected data from the NBA combines from 2010-2018.

After collecting all the data that we needed to we integrated Python with sklearn, a machine learning package, to predict whether or not a player would make the NBA. We defined making the NBA as playing in an official game during the NBA season. This excludes players who were drafted and never played a game, as well as those who signed contracts and were on NBA rosters but failed to play in a game. These distinctions echo the distinctions that are enforced on the sports reference page in order for a college player to be considered having gone on to play in the NBA. We created and ran a logistic regression, decision tree classifier, random forest classifier, MLP classifier, and Zero R model to see which model would be best at predicting whether a player would make the NBA. The Zero R model, predicting every player as never making the NBA, was going to be our baseline. Since the vast majority of NCAA DI players never make the NBA, a model that predicts no one will make the NBA is still correct over 99% of the time. But in order to tell a story worth listening to we needed to predict the players who did end up making the NBA.

A close up of a map

Description generated with high confidenceIn an attempt to improve the prediction ability of our model while also using realistic sub sections of NCAA DI players we broke up our dataset into the following categories to test our model with.

Freshmen only: We decided that it would be appropriate to only look at players who were in their freshmen year because the trend of freshmen being drafted, especially in lottery selections, has been increasing. From our previous work on NBA performance and the expected value of a pick it was appropriate to put an extra consideration on lottery picks. In the 2018 draft 11 of the 15 lottery picks were freshmen, the other four being international player at 3, junior at 10, sophomore at 12, an junior at 13. In the 2017 draft 11 of the 15 lottery picks were also freshmen. The other four being international at 8, sophomore at 12, sophomore at 13, and junior15

Last Year of College: We decided that including the last year a player played would be a good sub section of players to consider as well. This is because this subsection inherently captures a players best season or their

Guard:

Forward:

Center:

We also ran the model on all of the players in our dataset.

# 3. Results

Note: due to an overwhelming amount of empty data points in the combine anthropology and agility datasets we didn’t these metrics. This is due to three main issues, the first is that players who attend the combine rarely perform all the tests, the second is often times the most notable college players rarely perform any of the tests if they attend at all (the vast majority of NCAA players also do not attend) and lastly the combine usually occurs only a month before the NBA draft and by then most scouts/ fans have already decided who they feel are most draft worthy. For these reasons we decided that adding the combine metrics to our machine learning models would negatively affect the model’s ability to predict NBA readiness.

A close up of a map

Description generated with high confidence

# 6. Discussion and Future Work

# References