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

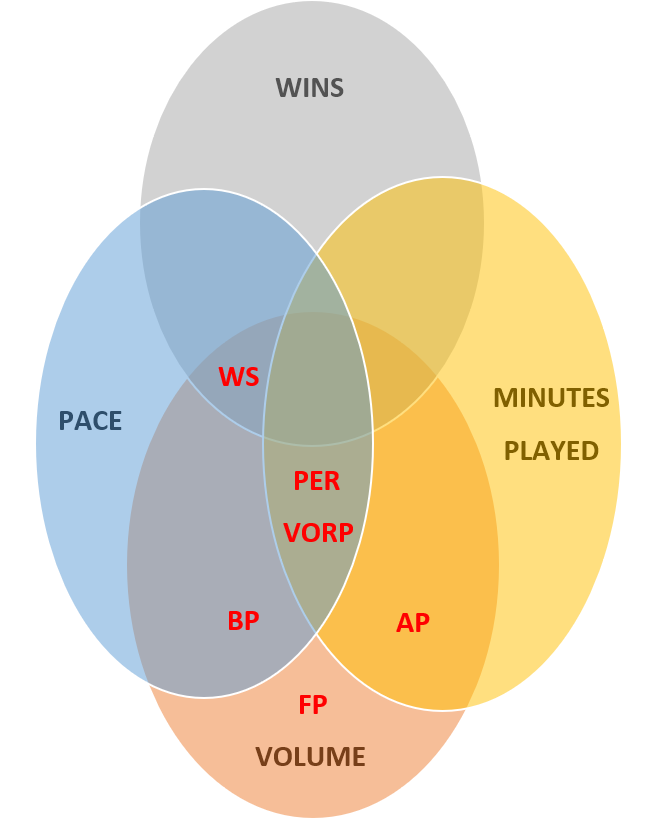
After researching existing metrics and developing an understanding of what already exists, we set out to create our own metric that measures a different variant of player performance. Basic Percentile (BP) ranks a player on their basic box score stats relative to the rest of the league. This means a 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 takes into account 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.

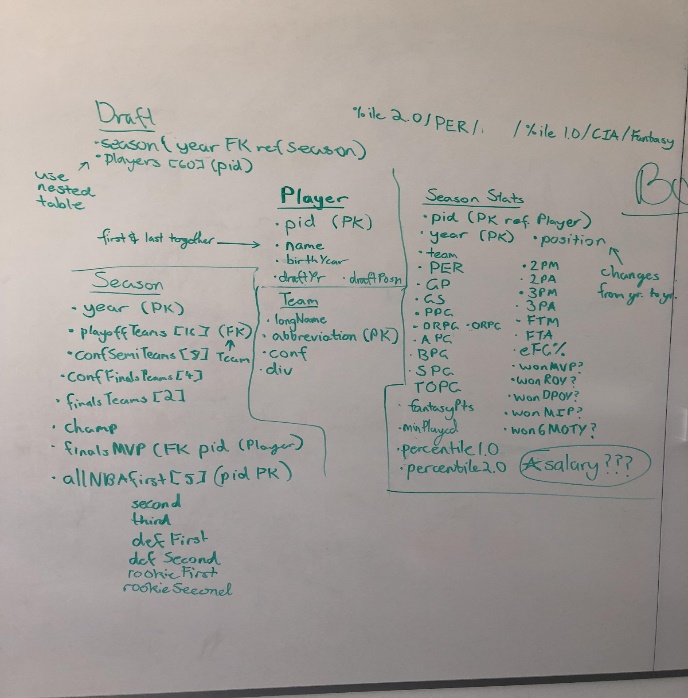
Our first drafted Entity Relationship Diagram is shown to the right. In the end, we modified this slightly to have a Master Season sheet for every year with all player season stats, a Draft sheet with the players drafted that season, and a Master Players sheet with all players and their IDs. With a clear 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)

In order to obtain all of this data we searched through the <table> tags on the webpage, used the <th> tags as column headers, used the <tr> tags as the rows and used the <td> tags as the individual cells of data. At times the tables that contained the information that we needed to pull was within comments on the HTML. In this case we had to simulate a user using the webpage and read through the comments of the webpage. Once this was performed we were then able to pull all of the data that we needed. With the help of Pandas we saved all of the pulled data in excel files as Dataframes. This allowed for easy manipulation and merging of each table.

# 4. Design

# 5. Results

## 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

# 6. Discussion and Future Work

# References