Paper

Vedanth Pothina

Abstract

This project studies the success of NBA players drafted between the years 1995-2015 by showing their career success relative to their draft position. Using a variety of sources, career points, rebounds, assists, steals, blocks, turnovers, three point percentage, free throw percentage and field goal percentage averages were compiled for each of the players in the dataset. In order to evaluate pure performance without any draft position weight attached to it, a formula was created called PerformanceScore in which it incorporates all of the statistics we created into a formula that encompasses the entirety of a player's career. For each position(G,F,C), the stats are weighted differently using a regularized linear regression model in the PerformanceScore formula. Next, ExpectedScore was made, a score that shows us what a player drafted at a certain pick in the draft expected PerformanceScore should be. Finally, FinalScore was created, a combination of both PerformanceScore and ExpectedScore to get a ranking of every single player drafted from the year 1995-2015's career success relative to their draft position. This analysis provides an insight into under and over achievers throughout this 20 year span while also being able to provide General Managers an understanding for the future on who to draft.

Introduction

NBA teams and fans are in dire need of a singular metric to determine whether a player they drafted underachieved or overachieved relative to their draft position. Understanding this will help them as they go into future drafts as they will see similarities in players from the past they drafted and future players they may want to draft. The FinalScore formula gives these teams just that as it is an unbiased formula to show their success relative to their pick and position.

This study has numerous objectives. It first examines how players drafted between 1995-2015 performed relative to their expectations based on their draft position. In theory, the players that have been drafted in the lottery and the first round should be the best in the entire draft. This is almost never the case. This paper examines the players who are able to overcome the adversity of being drafted in the later-rounds and provide that same superstar effect as the lottery players while simultaneously looking for the lottery players who never reached that height. Using the analysis done during this 20 year period, it will become clear without bias who the true underperformers and overperformers were from draft night.

Questions that surround the topic of expected draft value are questions such as which position provides the best value? Which draft group returns the most value by pick, lottery, mid-first or second round? On prior research, there were not many studies surrounding expected value for pick and rankings, especially not in the time period of 1995-2015. The research question for this project is how can we measure how NBA players drafted perform relative to their expected value by draft pick?

Methodology

To address the question above, the dataset is filtered down by year. For the purposes of this project, the dataset is purely NBA players drafted between the years 1995-2015 which is the true start of the "modern-era" of basketball. To evaluate player performance non-relative to draft pick, PerformanceScore is used, a

formula using all the major statistics of both offense and defense to create a score for each player. Player performance independent of draft pick was quantified using a metric called PerformanceScore, a weighted formula composed of the key offensive and defensive statistics, made using a ridge regression optimization formula. Different weights in the formulas were developed for guards, centers and forwards to account for position-specific roles on the court. The metrics used to evaluate PerformanceScore are points, assists, rebounds, turnovers, steals, blocks, free throw percentage, three point percentage and field goal percentage.

An ExpectedScore was then derived by evaluating the mean PerformanceScore per draft pick, showing the expected PerformanceScore associated with that selection in the draft. Finally, a FinalScore is calculated as the distance from the mean per pick or PerformanceScore-ExpectedScore, quantifying how much a player under or over performed relative to their draft pick.

The data used from this project was collected using basketball-reference.com, a popular basketball statistics site where the players in the dataset's career-wide statistics were collected. Basketball Reference returned the Points, Rebounds, Assists, FG Percentage, Three Point Percentage, Free Throw Percentage, Pick and Name. In addition to basketball-reference, hoopR's API was used to find each player's career steals, blocks and turnovers.

For this project, the sample size consisted of all NBA players drafted from the 1995 NBA Draft till the 2015 NBA Draft, providing a total of 1,216 players without the inclusion/exclusion criteria. Due to roughly 5% of NBA players per draft class never playing a game and about 18% of players not providing statistical importance to the dataset, a filter was applied to remove outliers whose performance was inflated or deflated by limited opportunities in their career. All players in the final leaderboard played at least 2 years in the NBA which is the amount needed to put up statistics that will give a player a viable PerformanceScore.

Some of the key variables in this project were all of the major statistics that went into creating the formulas and the PerformanceScore, ExpectedScore and FinalScore. Players with missing data were either due to the fact that they hadn't played a game and were filtered out or because of the hoopR API. The hoopR API had a different set of names for players with accents in their names than basketball reference so those players' steals, turnovers and blocks were entered in manually.

Players with missing data in the dataset were excluded for 2 primary reasons: either they had been drafted and had never played a game in their career or their statistics were unavailable due to inconsistencies in either hoopR's API or basketball-reference's website. Specifically, there were numerous issues around players with special characters in their name as basketball-reference and hoopR handled name-formatting differently. These happened particularly for players with non-ASCII or accented characters in their names. In these special cases, data was entered manually to ensure data completeness in the dataset.

An extremely critical component in evaluating a player's importance to their team and to their statistics is their availability, which represents how often a player is active on the court. Per-game statistics can be inflated if a player misses games consistently due to injury or personal issues. To account for this, an AvailabilityScore metric was created, a ratio of the games played to the maximum possible amount of regular season games in their career. This measure normalizes PerformanceScore by incorporating reliability, ensuring that the FinalScore also represents how valuable a player is to their team through their sustained presence on the court.

For each position in basketball, the appropriate PerformanceScore formula would be different as players in different positions are drafted for different roles on the team ex. (the guard's role is to run the offense while the center's role is to collect rebounds and score in the paint.) Using that idea, I selected the key offensive and defensive performance metrics, Points, Assists, Rebounds, Field-Goal Percentage, Three-Pointers Made, Turnovers, Steals, and Blocks and applied a ridge regression optimization model to determine the relative weights of each variable. These optimized coefficients were then used to construct a composite formula that produced a single clutch score for every player. They are as follows. The formula for **Guards** is:

$$

$$\text{PerformanceScore}_G = 0.301392942 \times PTS + 0.131585761 \times TRB + 0.245252914 \times AST + 0.071600908 \times STL + 0.088402778 \times BLK - 0.110155566 \times TOV + 0.003795444 \times FG\% + 0.029598577 \times FT\% + 0.018215110 \times 3P\% + 0.015 \times AvailabilityScore$$

$

The formula for **Forwards** is: $

$$\text{PerformanceScore}_F = 0.278032221 \times PTS + 0.035431424 \times TRB + 0.363261918 \times AST + 0.003023791 \times STL + 0.143949521 \times BLK - 0.141206541 \times TOV + 0.001456723 \times FG\% + 0.009143780 \times FT\% + 0.024494081 \times 3P\% + 0.015 \times AvailabilityScore$$

$ The formula for **Centers** is: $

$$\text{PerformanceScore}_C = 0.166423233 \times PTS + 0.099124073 \times TRB + 0.377695524 \times AST + 0.006508985 \times STL + 0.093415676 \times BLK - 0.159690953 \times TOV + 0.054962999 \times FG\% + 0.013530175 \times FT\% + 0.028648381 \times 3P\% + 0.015 \times AvailabilityScore$$

$ To calculate the ExpectedScore, the average PerformanceScore per pick is found.For example, all players at the 17th pick will have the same ExpectedScore. The ExpectedScore also shows that the average player per pick, even in the lottery, is not some sort of superstar but most of the time, a high end role player. Finally, FinalScore is calculated by PerformanceScore-ExpectedScore, essentially finding how far away from the mean per pick the player is.

The coding language used to do this project was R, a data science language in which the data was imported and final leaderboards were returned.