NBA Draft Analysis

A Statistical Breakdown of NBA Player Data from 1990-2020

**Abstract**

In this project I studied and performed statistical analyses on players drafted into the NBA from the years 1990-2020. Data was pulled from Kaggle and a mixture of Python and R were used to perform the statistical analysis. The goal was to search for indicators of a successful career and how players fared in comparison to each other.

1. **Introduction**

With the NBA season at a close, teams start to look toward the draft and free agency in order to prepare for the upcoming season, all with hopes of hoisting the Larry O’Brien trophy. The NBA Draft is always a topic of hot commodity as teams usher in the newest pool of young talent into the league. As the landscape of the game changes so do the type of players selected and where they are selected.

1. **Data**

Data was taken from Kaggle. This data set included all the draft picks from 1990-2020. The stats included within the dataset were from the NBA careers of the players. Various stats including PPG, APG, RPG, career totals of those stats, pick number, team selected to, draft year, college, and other advanced metrics are also included. I added a new column, PickType, depicting if a player was a lottery pick. Furthermore, players from the 2021 draft class were eliminated from the data set as they had an incomplete season at the time of the data set’s release. This was mostly due to skews found within statistical performance across advanced statistical metrics. Furthermore, I wanted to study players that actually played a decent amount of games. A lot of second round picks do not end up playing and would have produced too many null values. I determined that at least half a season played would suffice as a decent amount of games.

Data cleaning was done using Python. A duplicate draft year column was removed, those players playing less than 41 games and the 2021 draft class was removed, and an additional column indicating lottery picks was added. Furthermore, the original data set left foreign or high school drafted players with a blank in the college column, so that was corrected to be listed as ‘Foreign/High School’. The linear regression model used was created in Python to predict the success of a career based on various statistical inputs. **Python libraries used were pandas and sklearn as well as numpy**. All other statistical analysis was done through R using **tidyverse and gt.**

1. **Results and Findings**

**Statistical Analysis**

There is no sure fire way to guarantee the success of a player. Like any sports there are busts and boomers, steals and reaches. Within the NBA, players are graded as first or second round picks, and within the first round, lottery picks. These are the teams that failed to qualify for the playoffs. The players selected within these picks are typically touted as higher end, or even franchise cornerstones. At times, these players can be held in higher regard as a result of where they attend school. Amongst the college basketball world, there are 5 traditional “blue blood” schools- schools with a reputation of continual winning. These 5 blue bloods are Duke, Kentucky, North Carolina (UNC), Kansas, and UCLA. From the last 20 years, I wanted to see what colleges produced the most high end talent, and if it would be dominated by the blue bloods.

| Figure 1: Top 10 Colleges for Producing Lottery Picks. |
| --- |
|  |

According to Figure 1, 4 of the 5 “blue blood” colleges have produced the most lottery picks in the last 20 years. While UCLA still cracks the top ten, their reign of dominance came in the 1960s-1970s. However, despite achieving little program success, they still produced a high number of NBA products.

Lottery picks are often held to a higher standard and deemed ‘higher quality players’. Players taken with these top 14 picks are expected to perform better on the court compared to those drafted later in the draft.

| Figure 2: Lottery Picks | Figure 3: Non-Lottery Picks |
| --- | --- |
|  |  |

As seen in through the stats in Figure 2 and 3, over the last 20 years, lottery picks have had stronger statistical outputs across the board and **typically contribute more positively towards winning with higher plus-minuses (-2.1 averaged for non-lottery picks while lottery picks never averaged a BPM of lower than -1.6)**. As for the individual lottery picks, it was no surprise to see the top 5 picks have the greatest statistical outputs. Variance across the remainder of the lottery can be potentially attributed to the higher frequency of possible busts and boomers. However, good statistical output does not necessarily translate to productive basketball. Using advanced metrics, the value of players drafted at each pick in the lottery were tested. ‘VORP’ or ‘Value Over Replacement Player’ was the selected metric over other advanced metrics such as Box-Plus-Minus accounts for position and team success and individual impact was the center focus. Likewise, ‘win shares’ was not chosen as players taken in the lottery are likely going to tanking teams and may not see a lot of team success. Not everyone is Lebron James or Tim Duncan. Furthermore VORP also accounts for longevity, a key to an individual, successful career. Though one player may not carry his team to a multitude of wins, VORP helps measure the impact a player has while on the court. A player may average a lot of points, rebounds, or assists, but that may not necessarily translate to a positive impact.

| Figure 4: Lottery Picks’ VORP |
| --- |
|  |
| Figure 5: Lottery Pick vs Non-Lottery Pick VORP |
|  |

Through Figure 4, it can be seen that the top 5 picks have had some of the highest impact for their teams. A surprisingly valuable lottery selection appears to be pick 13, a late lottery pick. Their VORP matches that of the 7th overall pick and exceeds picks 6 and 9. This can be attributed to the number of under the radar prospects selected here. These include, Donavan Mitchell, Devin Booker and Zach Lavine. 2 of these players have been repeating All-NBA talents, while the other has been a proven All-Star. Though there is variance amongst the lottery picks, Figure 5 shows that lottery picks are drastically more impactful than non-lottery picks. This further highlights the importance of lottery picks throughout the NBA.

Though finding some gems in the later picks of the draft can be very rewarding, these figures display the importance of capitalizing on early draft picks.

**Regression Analysis**

Using Python and its data analysis libraries a linear regression model was created to predict the success of a player’s career based on certain statistical categories. A successful career was deemed as a career spanning longer than 4 years. This time period was selected as it typically indicates the time that a rookie would be eligible for a new contract. Players playing past this threshold would be considered a successful career. This regression model tested against 3 separate variables: pick selected, VORP, and BPM.When running the model using ‘Pick Selected’ as the independent variable, the model had a 66.5% accuracy rating. The model did a good job of predicting and identifying successful careers with a precision value of 0.69 and a recall value of 0.86, but struggled in predicting unsuccessful careers (precision: 0.56, recall: 0.32). A low accuracy percentage and the model’s inability to predict unsuccessful careers shows that the location of a player’s selection is not the best indicator of a player’s career success. Changing the independent variable to ‘VORP’ improved the accuracy of the model to around 70%. The model improved in its prediction and identification for unsuccessful careers with a precision of 0.57 and a recall of 0.66. For successful careers, precision improved while recall was less than ‘Pick’ (precision: 0.79, recall: 0.72). The final variable tested, BPM, was the strongest indicator of a player’s success according to the model. The model tested its highest accuracy score of 71.8%. Using BPM as a variable created the most balanced model in terms of predicting successful and unsuccessful careers. Precision and recall came in at 0.64 and 0.50 for unsuccessful and 0.75 and 0.84 for successful. Additionally, the pseudo R-squared value was the highest compared to the other variables at 0.25. The high accuracy and association made BPM the strongest variable tested for this model although VORP was also a good indicator. All variables’ x1 values followed logical trends: higher draft picks were predicted for more success, greater BPM and VORP also indicated better success.