

Introduction

The National Basketball Association, or NBA for short, is a professional basketball league situated in North America. Widely acknowledged as the best basketball league in the world, the intersection of competition and financial profit (8.3 billion in the 2019-2020 season) has led to NBA teams increasingly turning to analytics to gain an edge. The “analytics revolution” in the NBA has resulted in changes to how teams play basketball, changes to the perceived value of certain skills, and the rise of data and metrics that the traditional box score (a summary of results on a per game basis) does not encompass.

Overall the incorporation of data analysis has produced positive results. But there is still dispute on how exactly star players in the league can have their impact quantified and ranked. A variety of metrics aim to encapsulate several different data marks into an all in one evaluation of a player’s performance. Ranking players by these stats can be said to give a rough estimate of how good a player is, especially when compared to their peers. Different metrics do result in a different consensus of a player’s impact, making it hard to ever come to a definitive conclusion as one metric may place more emphasis on certain factors than others.

Furthermore, the efficacy of all statistics is a debated topic. A survey of independent and team affiliated NBA analysts that ranked offensive and defensive stats on a scale of 1-10 concluded that publicly available defensive stats grade out at 3.6, offensive stats grade out at 7.6, and private defensive stats (available to teams through their own data analytics branches) grade out at 5.6. All in one metrics rely on a composite of these offensive and defensive stats to reach their conclusion which compounds the fact that their results may not be entirely accurate.

One contributing factor is that many stats (or their components) may not always represent the reality of what occurred in an actual game situation. For example, a statistic that notes player A scored a basket against player B records this information without underlying context. Did player B have to cover for a defensive breakdown from another teammate? Did player A simply get lucky and score despite good defense from player B? These questions and more require careful study of the game, either in real time or through recordings. The level of effort required to track these details and the subjectivity involved (what defines “good defense” for instance?) over an 82 game regular season and 4 rounds of playoffs with over 500 players is far too great to sustain.

Problem Statement, Goals, and Methodology

The problem this research project is attempting to answer is how can we measure the impact of a basketball player to determine which players, roles, and skills are the most important for winning games. To accomplish this, footage of NBA games and play by play data must be analyzed to extract meaningful statistics that the traditional box score does not encapsulate. While we cannot go over every single game and player, we can better focus our efforts by analyzing star players, as they provide the most impact in winning. We can reduce our sample size of games to the playoffs as the outcome of these games directly leads to winning the NBA championship or not.

Our first step is to isolate star players to analyze and track. Going through NBA All Star lists (participants of the League's All Star exhibition games designed to showcase star players), we can select the following players:

- Kevin Durant
- Joel Embiid
- James Harden
- Kawhi Leonard
- Donovan Mitchell
- Jayson Tatum
- Giannis Antetokounpo
- Stephen Curry
- Luka Doncic
- LeBron James
- Nikola Jokic
- Damian Lillard

The above 12 players have versatile skill sets and numerous playoff runs, which allows for a sufficiently large sample size of data to be collected. Using a traditional role definition as per NBA voting, this list has 5 guards, 5 forwards, and 2 centers. But players like Joel Embiid and Nikola Jokic, both centers, have completely distinct playstyles. Similarly, Luka Doncic and Stephen Curry are both classified as guards, but are utilized differently by their teams. To properly assess value, we need to redefine the traditional role listing into one that more accurately represents a player's skill set and what they are asked to do. For this, we will need to use publicly available data rather than our own collected data (as a proper redefinition of roles would require analyzing more than just the 12 chosen star players). We can run a k-means clustering algorithm to group together players that are more alike to get a better understanding of their roles. From there we can assign role definitions to each cluster based on our own understanding of what players in a cluster are tasked to do in a game.

There are two types of plays in basketball. Offensive plays, which occur when a team has possession of the basketball and is trying to score, and defensive plays, which occur when a team does not have possession of the ball and is trying to stop their opponents from scoring. For our defensive statistics, we can measure the following:

- Help defense/Proper rotation
- Isolation defense
- Shot contests
- Rim protection
- Perimeter defense
- Interior defense
- Offball defense
- Pick and roll defense

While not covering every possible aspect of defensive basketball, these 8 categories provide a good insight into whether or not a player played good defense. Each stat should be measured on a per possession basis, with a grade of -1 (bad defensive contribution), 0 (neutral/good defensive contribution), and 1 (excellent defensive contribution).

Similarly, our offensive statistics can measure the following:

- Contested shot making
- Onball scoring
- Offball scoring
- Onball creation
- Offball creation
- Isolation scoring
- Midrange scoring
- Perimeter scoring
- Rim finishing

These statistics will be gathered by watching games and tallying what the player we are gathering data for did.

In addition to these metrics, we can aid our model with certain publicly available stats, such as measures of scoring efficiency, and rudimentary player impact stats that can help measure teammate value (A logical conclusion to make is that if a player has better teammates, there is less of a burden on them both defensively and offensively, causing an increase in positive statistical measures).

Next, we can also train a classifier to make predictions on the outcome of a game given a set of data modeled with the same features as our collected statistics. We can accomplish this through the use of a logistic regression model, random forest classifier, and a support vector machine. After training all models, we can compare them to see which one is more accurate in predicting the outcome of a game.

Finally, we can create an interactive visual representation of our findings using the Altair visualization library. For instance, we can display the data we have charted in games, as well as how players fit in to the new clusters we have discovered.

Results

In the classifiers notebook, a variety of models and methodologies were tested. A logistic regression model, random forest model, and a support vector machine were trained on recorded player data. Each model was evaluated with a simple train test split of 75% training data and 25% testing data. In addition each model was also evaluated using stratified k-folds in order to create a more robust model.

Ultimately, the support vector machine (using a linear kernel) model had the highest accuracy, at just over 80% accurate. It is possible that the nature of an SVM being suitable for smaller quantities of data could have made it a more accurate model due to the size of our dataset. Recording more player data could prove or disprove this theory. The 80% threshold was achieved with cross validation, which gives us reasonable confidence that it is indicative of the model's true performance given the dataset.

Role redefinition was a slightly trickier subject. Principal component analysis suggested the optimum role classification would be three categories. NBA theory has arrived at this conclusion, with notable figures such as former Boston Celtics coach (and now General Manager for the Celtics) Brad Stevens stating that player positions can be broken down into ball handlers, wings, and bigs. While accurate, we chose to go with the next best clustering number, 6, to try and gain more insight into NBA positions and roles.

To this end, the clusters identified were traditional big men, 3+D wings, star scorers, role players, defensive specialists, and bench players. We can see here that clusters such as role players and bench players are not too informative or help us understand what a player does more. Other clusters however, especially defensive specialists give us a deeper understanding of the game. We can see through the Golden State Warriors and Brooklyn Nets usage of players such as Gary Payton II, Deandre' Bembry, and Bruce Brown, that such a niche in the NBA can be used in unexpected ways to help not only a team's defense, but a team's offense as well.

When looking at defensive and offensive ratings (DRTG and ORTG) in correlation with our recorded stats, we can see that the impact of individual performances are hard to isolate effectively from overall team performance, especially on a per stat basis. Combining all of this information in conjunction with a model does allow us to gain a better understanding of whether or not a player's performance was likely to result in a win though.

The total.html file combines these visualizations as well as a handful of player performances and predicted vs actual results into a webpage where viewers can see the results of this project for themselves.