DS 3000 Basketball Position Predictor: Project Phase 4

Joshua Liu, Christina Long, Michelle Meredith, and Brandon Onyejekwe

Abstract

The National Basketball Association (NBA) is a professional basketball league and one of the major North American sports leagues. As of 2021, there are 30 active teams, each with a maximum of 15 players on their roster. Players typically take on different roles:

Guard: Point guards control the floor and flow of the game, are good ball handlers and passers, take quick shots, and are skilled at assisting other players. Shooting guards are good at positioning around the three-point line and maintaining movement without the ball. They are typically skilled three-point and long mid-range shooters.

Forward: Small forwards are versatile. They are skilled at drawing fouls and shooting foul shots, and are generally good shooters. Power forwards powerful scorers, taking shots close to the basket as well as mid-range shots. They have the strength and athleticism to guard bigger players on defense.

Center: They tend to score close to the basket, are skilled at getting both offensive and defensive rebounds, setting screens for other players, and contesting shots.

Although there can be overlap both among and within each position, the distinct characteristics of each position can be used for a wide range of purposes, including statistical analysis, player development, and fantasy sports. To investigate these characteristics, we trained a model to predict a player's position based on their statistics.

Introduction

Problem Definition

In all sports, stats are used heavily to both categorize and rank players based on their role. Based on basketball's main positions, we wonder if there is that much distinction between different positions or if the roles are just formalities. From an outsider's perspective, it can be hard to distinguish between each of these roles. Yet roles are important for fantasy basketball, analysts, and enthusiasts alike. These days, though, the NBA is largely a positionless league with many players getting assigned multiple roles. According to ESPN, "[t]here is no shortage of PG/SG, SG/SF, SF/PF and PF/C players in ESPN leagues". With the large overlap and players flexing between multiple roles, is the classification of roles outdated?

To answer this question, we have decided to incorporate our knowledge of data science into a machine learning based solution. We plan to build a model that will attempt to categorize player's based on their stats to answer the following: *is there enough of a difference between each of the roles such that a machine can guess what position a player plays?*

Motivation

We chose this topic because we are personally interested in basketball and because there are a lot of interesting problems to explore relating to sports statistics. Basketball is also very popular across the U.S. and worldwide, so this topic has both national and international relevance. Sports have become very prominent in data science due to its general popularity, with data science being a big part of sports betting, fantasy sports, and statistical analysis. We were deciding between making a playoff predictor and a position classifier, but we ultimately decided on making a position classifier because it is a unique problem that not as many people have previously looked at.

Goals and Objectives

Our primary goal is to explore whether a player's statistics can be used to accurately designate their position on a team. We aim to achieve this by creating a machine learning model that takes a player's statistics and classifies them by their correct position. As it is impossible to make this totally accurate due to the issue of positionless players in today's NBA, our goal is to make it as accurate as possible. We want to use our model to see how distinct each position truly is in terms of their role on the team, and we can run our model for past years and compare it to the current year to see how this has changed over time.

Related Work

Classification: Football Player Position Prediction (Part 2)

This article describes a model that was made to predict a player's position for football (American soccer). Rather than using players' game statistics, however, physical attributes were used in the model.

Players, Positions, and Probability in the NBA

In the article above, the author also wondered if basketball players could be classified into roles based on their statistics. The author designed a model that would produce percentages of what position the model thought a player would be. An interesting point is that the author viewed each player as a single observation (i.e. 2017-18 LeBron would be different from 2015-16 LeBron) as players could undergo various offseason changes. Furthermore, the author also discussed using different model types and ended up using the 'Support Vector Machine' model, as we are still learning about Machine Learning, this article will be a great resource for us to explore different options and also provide some guidance. Lastly, the author also analyzed the accuracy of his model to be right approximately 73% of the time, we are curious if we could also look to improve this perhaps by weighting certain attributes.

Methodology

Our methodology centered around training our model to predict what position a NBA player plays based on their stats. Once we decided the general direction for our project we explored various datasets to find one that would include player positions alongside as many unique player stats as possible. We chose a season's worth of data from "nbastuffer.com" as it was the only dataset we could find that included position titles for all players.

To prepare our newfound dataset, we first worked to change the column names from their verbose given labels to acronyms which would be easier to handle. Next we also simplified the positional categories in our dataset. Our dataset originally contained combination positions such as "F-G" that represented a primary and secondary position a player played, but as our planned role evaluation did not account for these combination positions nor did we know what ratio of time the player spent in each position (which could confuse the model), we decided to only account for a player's primary position (the first listed character). Lastly we cleaned up the dataset by removing any rows that contained nulls.

Regarding model selection, at the time of writing our proposal and exploration we only knew one or two machine learning algorithms, K Nearest Neighbors(KNN) and Random Forest. We felt that KNN would be best suited for our application as we hypothesized that if player positions could be guessed via their stats then the neighboring points for any given player should be those of the same position. Additionally we wanted to choose an algorithm we felt comfortable with for debugging and time reasons, thus we chose KNN for our model.

Impacts

Basketball is a very popular sport across the United States and worldwide, so our model is very useful and relevant. Those who are new to basketball or are trying to learn more about it will benefit greatly from being able to tell what the different positions do or what a specific player's role is on their team. Our model could help fans acquire more knowledge about the sport and receive more enjoyment from watching basketball games by understanding the differences between these positions.

Results & Evaluation

Results

We found that our model is best at predicting if a player is a guard, as the precision of our model is 0.69 and the recall is 0.67, which are our highest values for each of these metrics. A precision of 0.69 means that for all players classified as a guard through our model, 69% are actually guards. A recall of 0.67 means that for all players that are actually guards, we correctly classified 67% of them. 69% and 67% are in the precision and recall range that we were looking for, so we feel that our model is very accurate for the guard class. For the forward class our model is not as accurate as for the guard class, but it is still fairly accurate. The precision of our model for the forward class is 0.49 and the recall is 0.59. For all players that our model classified as forwards, 49% are actually forwards. For all players that are actually forwards, we correctly classified 59% of them. For the center class our model is very accurate in precision with a value of 0.67, but it is significantly less accurate in recall with a value of 0.35. This means that for the players that our model classifies as centers most of them are actually centers, but our model leaves out a lot of the other centers. Basically for the center class our classifier is accurate in terms of correctness but not completeness, as it doesn't actually classify a lot of the players that are centers.

Evaluation

It makes sense that it is easiest for our model to predict the guard class, as guards generally have a more defined role in basketball. On the other hand, forwards are generally much more versatile and have a more balanced role in the sport, which is why it makes sense for it to be harder to classify forwards. Similar to guards, centers also have a more defined role, so it logically makes sense for our model to have higher precision when classifying centers.

Classification Report:

	precision	recall	f1-score	support
C	0.67	0.35	0.46	17
F	0.49	0.59	0.54	51
G	0.69	0.67	0.68	63
accuracy			0.60	131
macro avg	0.62	0.54	0.56	131
weighted avg	0.61	0.60	0.59	131

Conclusion

We were able to meet our goals and objectives; our model found success in predicting positions, although there were varying levels of accuracy for each position. To improve our model, we would potentially like to explore different data sets that further break down player positions with more consistency, such as point guards vs. shooting guards. Additionally, it would be interesting to expand the number of seasons with which we analyze to explore different trends amongst the positions over time. It would also be interesting to evaluate these metrics for international leagues to compare position predictability in different countries. Overall, though, we are pleased with our results and consider our work to have successfully laid a foundation for future research.

References

2018-2019 NBA Player Stats
2020-2021 NBA Player Stats
NBA Frequently Asked Questions
The best strategies to dominate your fantasy basketball drafts