Capstone Project One

Classifying NBA players positions and performance

Milestone Report

by Yiding Weng

January 16, 2019

Problem

Every off-season the management of each NBA team have to find a way to improve their roster, usually through recruiting new players on the free market, or signing new contracts with their current players. But how to decide which new player to sign or which player on the old roster to keep? Will that player be able to provide the desirable contribution for the team? As each team try to optimize the diversity and the integrity of the skill sets of the entire roster, there are problems they have to consider.

NBA teams are increasingly trotting out lineups with five players who can play and guard nearly any position. traditional positions don't accurately explain what a players skillset truly is, they incorrectly oversimplify the skill sets of NBA players. Simply plugging players into one of five positions does not accurately define a player's specific skill set. Moreover, the misclassification of a player's position may lead teams to waste resources on developing draft picks that do not fit their systems.

In light of these changes, we need an effective way to designate positions in the NBA not based on basic physical traits such as height and weight, but in terms of function, such as shooting and defense. A framework for modern NBA positions is important towards our understanding for how players have evolved, and effective roster construction.

Data

Proposed Data Collection:

aggregate NBA player's Statistics for 67 NBA seasons.

Data Location: https://www.kaggle.com/drgilermo/nba-players-stats/data

Historic NBA team records

Data Location: https://www.kaggle.com/druswick/nba-team-records-historic

Approach

The approach to the problem will be considered to have two parts:

1. Find an optimal way to classify players functions based on historical statistics
2. Test the level of completion of each team roster based on the total integrity of skill sets and functions of each individual players.

The first data collection includes the performance statistics of each individual player in each season. Using clustering method to find out the optimal ways of defining categories that specifies function or characteristics of players in the same category, compare that with the traditional way of defining player roles as PG, SG, SF, PF and C.

For the second part understand whether each team has a balanced number of players in each new defining roles, and test if the completion level is correlated to team regular season performance using the historic team records.

Data wrangling

Data cleaning

In dealing with player positions, PF, SF, SG, PG, C are the 5 designated positions for professional players, but a very small portion of players played a role somewhere in between two different positions, so that they were label with 2 positions such as SF-SG, C-PF, SG-PG… due to lack of information of which role these players were more inclined to play, I removed those player information who played double roles.
For the stats that represent total values (others, as TS%, represent percentages), I transformed them into values per 36 minutes. The reason is to judge every player according to his characteristics, not the time he was on the floor.

Deal with missing values

Delete blank columns and rows.
There are significant missing values in 3P (3-Point Field Goals), 3PA (3-Point Field Goal Attempts), 3P% (3-Point Field Goal Percentage), 3PAr(3-Point Attempt Rate). It's because 3-point line was not introduced to NBA until 1979, but for the sake of classifying modern players, I have to assume all players before 1979 have not attempted 3-pointers, so I will fill the missing values of these three columns with 0.
For the missing values in GS (Games Started), since number of Games started for each player does not effect their performance measuring, I will delete the GS column.
Missing values also appeared in ORB, DRB, STL, BLK, TOV and their associated percentages (ORB%, DRB%, STL%, BLK%, TOV%), these measurements are not

included in record before 1974, and TOV data are not present until 1978, These missing values cannot be filled with assumption, but at the same time these columns are too important to be removed from the construction of player classification, therefore all the data before 1978 have to be discarded.

Handle outliers

When using rate statistics (i.e. points per game) or cumulative statistics (i.e. total points) can be misleading when it comes to analysis because these statistics tend to inflate players with lengthier careers. To deal with outliers, I instituted a minimum threshold of 40 games played. Also I keep only players with more than 400 minutes for each season (with a 82 games regular season, thats around 5 minutes per game. Players with less than that will be only anecdotical, and will distort the analysis).

Exploratory analysis

Data storytelling

Basketball players at different positions have very different performance in each category. For both management professionals and basketball audience, they both have certain traditional understanding or expectation on player functions. For example, player in the role of point guard(PG) are the ones with smaller statues, they tend to be the ball handlers most of the time, and often has higher number of assists in the game. And bigger players in the role of center(C) or power forward(Pf) tend to grab more rebounds and blocks since the rim is more accessible for them.

During data storytelling I examined how these common understandings of basketball player function stand firm in the light of real data, and we may find trend of changes in data attribute in relation to player position and time, which help us to see whether the function of each position has evolved into more specialized and distinctive in its own style: Does PG have more assist? Does center grab even more rebounds? Or whether the functions of different position has become more intermingled and interchangeable? Therefore I will divide the cleansed data according to position and according to time. The row data attributes I will look at are:

PTS - Number of points
3PA - number of 3-Pointer attempt
2PA - number of 2-Pointer attempt
AST - number of assists

TRB - number of total rebound
PF - number of personal fouls
FT - number of free throws
BLK - number of blocks

After visualizing the data with scatter plots of individual player information, line plot which shows average number for each position category for multiple attributes and box plot that presents player performance number distribution during different time period, we can draw a quick summary of the changing trend in each data category:

PTS - C and PG took SF and SG's places in leading score positions;
3PA - Overall increase, especially in PG, SG and SF;
2PA - Overall decrease;
AST - Overall stable, PG becomes less dominant in leading the number;
TRB - Overall stable, C becomes more dominant;
PF - Overal decrease;
FT - Overall decrease in FT, C and PG took SF and SG's places in leading number of free throws;
BLK - Overall stable, C becomes more dominant;

Therefore according to certain attributes such as 3PA, TRB and BLK, we can say some functions for each positions have become more distinctive, where on the other end, attribute such as PTS, AST, and FT, suggest some function for certain positions has become less distinctive and can be interchangeable with other positions.

Statistical analysis

In this section we are going to use statistical analysis to explore if the observations from data storytelling can be validated.

Therefore I used scoring points as a case study one of the previous observation. We have already noticed the change of role in leading the scoring data among the positions. Before 1990, players of SG had the highest number in scoring point, and according to the visualization we think SG players have been scoring less. I choose to use the data before 1990s and after 2010s since we want to compare the scoring data of SG from early and recent periods.

Preform hypothesis test

H0: $p = 0$ H1: $p \neq 0$

Null hypothesis:

There is no difference between the scoring points of SG before 1990 and after 2010 (average scoring points difference is 0, $p = 0$)

Alternate hypothesis:

There is difference between the scoring points before 1990 and after 2010 (average scoring points difference is not 0, $\rho \neq 0$)

We get a p-value of 0.0, which suggests that there is a statistically significant difference. We got a difference of 2.314 points between the means. You should combine this with the statistical significance. Changing by 2.314 points in 40 years is substantial.

This case study with one of the data attributes shows that all the validation is verifiable if we preform hypothesis test to each observation we made in multiple attributes.