NBA Players Analysis

Executive Summary

Adam Magyar, LaShawn Murray, Xiaotong He, Tanyi Tan and Yue You

Group liaison: Xiaotong He

Historically players have been grouped into one of five categories, or positions. These categories were generally ascribed based on the physical attributes of the player. For example, taller players were Centers, smaller players were either a Shooting Guard or Point Guard and still other players would have either been Strong Forward or Power Forward. Teams have been built based on these five categories. Team executives would look to fill the categories with players that fit the physical attributes needed for each and would look to maximize skill per position. This would then have to be done within the constraint of total salary spent on players. This was then all done in the hopes of maximizing wins.

Our goal is to explore other ways of looking at player attributes and abilities. Can players be assigned to categories other than traditional position types? Can teams be built to maximize wins based on other types of attributes? Are there desirable attributes that command higher salaries or lead to more wins? We believe that through Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Common Factor Analysis (CFA), Cluster Analysis and Ridge Regression, we can help answer some of these questions.

We looked at on court basketball statistics for 446 players in the 2016/2017 NBA season. There were some descriptive dimensions such as position, team name. minutes played per game and age. The remaining dimensions explained a player's productivity from either an offensive or defensive standpoint. We want to be able to take this dataset and apply different methods of multivariate analysis to see if we can uncover some interesting insights around player rankings, salary, games won, or any number of other potential outcomes.

PCA is used to group linearly correlated dimensions into separate principal components. It allows us to reduce the dimensionality, but still have independent attributes that can be used to compare and assess player performance. This is similar to what we'll do with CFA where we look to see what attributes a player has that may lead to an offensive or defensive player. In both cases, we are looking to move away from traditional categories of player position, and rather describe players by the type of their play. A third technique used to determine attributes players have is Cluster Analysis. Similar to PCA and CFA, Cluster Analysis is done to group similar player attributes in an unsupervised manner.

Canonical correlation analysis is used to identify the strength of association between two sets of variates. Each variate is a linear combination. It allows us to understand the how two sets of variates relate to each other. We use CCA to explore the association between players' skillset and performance.

Many of the variables in the data are highly correlated. However, we want to explore what attributes a player has that may lead to higher scoring. To deal with the highly correlated nature of the data, we apply ridge regression. Ridge regression sacrifices some bias to decrease beta variance.

CCA uncovered that a player’s scoring ability was the most important variable when measuring their performance. While there are many aspects to the game, scoring was the most important. Further, their ability to score 2-Pointers weighed more heavily than 3-Pointers. On the defensive side, defensive rebounds weighed more heavily than offensive rebounds indicating that the ability to prevent the other team from scoring was more important than keeping the ball in play for your team.

Ridge Regression Analysis showed that Field Goal Percentage was the most important variable in determining a player’s ability. This was followed by 3-Pointers made, and Effective Field Goal Percentage. There are many ways to evaluate a player’s scoring ability and impact on the game. Ridge Regression was interesting in that it pointed to high percentage made and effective field goal percentage, in addition to three pointers made. Clearly scoring more points per possession is important. This can be done by making three pointers, or by hitting a high percentage of the shots you do take.

PCA removed the traditional categorization of players and instead categorized players by the way they played the game. PCA grouped variables into one of three components. So instead of labeling players as Center, Power Forward, Strong Forward, Shooting Guard or Point Guard, we label them as “Aggressive,” “Shooter,” and “Reliability” This is more about style of play rather than title. This can be significant because teams can look to maximize wins based on these combinations rather than traditional. It could let them exploit inefficiencies in the market.

Factor analysis removed traditional labels and assigned players to types of play rather than title. It identified the Attack factor which mainly reflected player attack ability from aspects of score, assist, steal, offensive ranking compared to teammates and Defense factor which reflected player defense ability from rebound, block shot, personal foul aspects and defensive ranking compared to teammates. It roughly confirmed to past research results. Together with the path model and clustering, factor analysis concluded that attack ability of a player contributes more to the winning of the team. High salary is paid for the excellence of a player’s attack ability. For a team, it should adjust its resources to focus on strengthening the attack abilities of the whole team. Finding a balance of scoring, attacking, rebounding and attacking is necessary in more winning in the game.

Cluster analysis allowed us to delve deeper into the clustering of NBA players bringing through further tones of offensive and defensive performance. The best construction resulted in three clusters which were named High Scorers, Excellent Defenders and Great Attendance.

Much of the work done here was preliminary work to see if we can remove some of the traditional aspects of game and player analysis and instead let the data tell us how to evaluate players. The analysis was limited to this scope. The models may be helpful to acquire talent to fill necessary role but it would be better if including some features to identify the non-measurable qualities of a players to fully evaluate the potential of a player in the future. Now that we feel that we can clearly take a differentiated view of player analysis, future work should concentrate on maximizing the benefits of this work. Teams can take what is learned here to create new types of lineups to maximize wins or decrease payroll.