Project Proposal: Using Recent Year's NBA Advanced Statistics to Classify Player Types and Overall Impact

Prathit Shukla, Michael Gagin

10/25/2019

This project is for Lab Section 007 with Yangyi Lu and we will both present it together.

The dataset we are planning to use is mined from the NBA reference on the website Basketball-reference, which has statistics for All NBA Players for the last 40+ years with few exceptions and bad data. For each player after 1979, there are over 30+ individual statistics availible with 25 of them helpful, numeric quantities including Points, Rebounds, Assists, to more advanced metrics such as effective Field Goal Percentage, Win Shares per 48 minutes. Additionally this large dataset provides us with the player's Name, Age, Minutes, and Seasons Active, as auxiliary data. We may also want to extract new variables and use these instead such as Minutes per Game from Games Played and Total Minutes as we don't want to penalize a player for being injured.

The goal for this project is being able to classify players into specfic types of players similar to how players are classified into position such as PG, SG, SF, PF, C. We want to go beyond this and classify a player as a Point-Forward, 3pt Specialist, Post Player, Rebounding Specialist, Score-first Guard and similar types. This will be the supervised learning approach of our project, and we also want to use unsuperivised learning techniques to see if we can create distinct groups based on just the statistics alone and no pre-defined labels. This is an interesting project because not only are we replacing something players, fans and coachs have done for years and automating this process, but also making a model that can apply to other purposes such as college scouting for potential NBA players and fit. From this clusters or classes of players, we can also see which players are the most unique in the NBA. A great example of this would be Russel Westbrook as he has recorded such outstanding stats such as a 30 points, 10 assists, 10 rebounds average season, but with our classification we can see truly how unique his numbers are in the NBA.

At first, our analysis is going to be limited to one season as this will make our dataset small to the extent of 540 datapoints. However, if our initial attempt is successful for this, we want to expand the range of our problem and see how the NBA's definition of Guards, Forwards, and positions have changed over times in recent years. We can iterate over years of data with a master data set from the same website, Basketball-Reference, with over 10000 data points and select the ones for the years that we are looking for.

Using selection techniques we want to narrow down the statistics first manually, such as remove similar data, or data that can be derived such as Assists Turnover Ratio or Shot percentages that can be derived from Taken and Made. We also want to remove insignificant predictors using either Best Subset Selection or Forward Stagewise selection to narrow down to a more reasonable amount of variables.

For our analysis, we are first going to normalize the data we are receiving by the minutes played per game, and get a new dataset with variables for the variables that can be normalized. We can start off with supervised learning techniques and use perhaps LDA,QDA, or Naive Bayes to see how well we can classify Player Position based on the distribution of statistics.

We are also looking forward to using K-means clustering to see how the algorithm will be able to divide into specific groups of players, maybe even more in depth than just the position but maybe play-style. From this we can just select the number of groups.

From this analysis after we have done a successful anaylsis of one year, because similar years are not supposed to be very different (however, we must confirm this with explanatory plots or other initial conditions). This can be used a test data set, or better yet, we can feed it information about this upcoming year, which will have around 20-30 games done by the time of this execution, to not only gather important data about current players, see if their classification has changed, or also if our training data set is identifying this well.