DSC 423 - Winter 2021

Sports/Video Game Group: El Dorado

Ken Thomas Wei Tong Su (James Su) Zhong Xie
Chunnan Liu Devin Carroll

Assessing and predicting player performance is an important task when building an NBA roster. Usage Rate, field goal percentage, and player efficiency rating (PER) are three key indicators of a player's impact and quality. Our goal for this project is to attempt to develop a regression model which will help us predict how a player will come out in those metrics.

Player movement is at an all time high in the NBA, particularly among star players. The recent blockbuster James Harden deal is just the most recent example of the NBA's biggest talents combining forces. Typically when three star players are on one team, at least one of them has to change their game up a bit to make the pieces fit. When Lebron James and Chris Bosh joined Dwayne Wade in Miami, Bosh was able to change his style of play to allow all three generational talents to share the court. Chris Bosh's volume of shots took a small dip, but his shooting percent went up. Conversely, however, when Joel Embiid, Ben Simmons, and Jimmy Butler tried the same thing in Philadelphia, the team got worse and the output by all three suffered, to varying degrees. Eventually the 76ers traded Butler for Josh Richardson then Seth Curry. Although Seth Curry is arguably the least talented of three, his three-point shooting ability has given Philadelphia the spacing to thrive. Which way will things go for the Brooklyn Nets with three high volume scorers like Kyrie Irving, Kevin Durant, and now James Harden? Our group plans to examine either field goal percentage, usage rate, or PER and form a regression model in an effort to predict those metrics for players when their circumstances change.

At present there are two datasets being considered for use. Both datasets contain a comprehensive list of features and response variables, with approximately 70 years of data between them. The group is examining three different response variables for use: Field Goal Percentage (FG%), Usage Percentage (Usg%), and Player Efficiency Rating (PER). Before a decision can be made some preliminary research about the viability of using calculated response variables will need to be completed - both Usg% and PER are calculated variables. The impacts of using a feature in a model that is also used to calculate the response variable are currently unknown. For this reason, the group is also considering FG%, which is an uncalculated, continuous response variable. For modeling the response variable, there exists a wide array of features that can be chosen. Once the final response variable is chosen, the direction for which features to choose will become more clear, but at present features such as age, minutes played, weight, height, and draft position all intuitively seem to affect the chosen response variables. Blocks, steals, turnovers, and assists are also explanatory variables we anticipate including in our model. The more interesting discoveries will be revealed once the final response variable is chosen and the modeling process has begun.

The first dataset we will be looking at is, "Basketball Players Stats per Season – 49 Leagues" which can be found at: https://www.kaggle.com/jacobbaruch/basketball-players-stats-per-season-49-leagues. This dataset was scraped from the website https://basketball.realgm.com/ using Python using the Pandas packages. The second dataset we plan on analyzing is, "NBA Players stats since 1950" (https://www.kaggle.com/drgilermo/nba-players-stats?select=Seasons_Stats.csv). This dataset

focuses on NBA players dating back to 1950 and was taken from www.basketball-reference.com/, which is a very thorough repository of basketball statistics. It is commonly cited by nationally recognized NBA analysts and pundits. Each of these data scraping efforts were collected into two respective CSV files that we will mesh together to create our complete dataset.

The international basketball league dataset includes 14,582 unique players recorded from the NBA and all of the international leagues spanning from the year 1999 to 2020. This dataset has nearly 53000 observations, where each observation represents the season long stats for a specific player. There are a total of 34 variables, 26 numerical and 8 categorical variables. While this data set has a greater depth of observations it also has fewer variables for each observation. Two important variables this dataset does contain are draft round and draft pick, which we will likely condense into one variable in the cleaning process. These variables were not present in our second dataset.

The NBA focused dataset has 3921 unique players from 1950 to 2017 and a total of more than 24000 observations. Each observation represents a season of stats for a specific player. Total of 52 variables, 49 numerical and 3 categorical. The variables in this data set are more specific, for example this dataset differentiates between three point and two point attempts. It also contains advanced metrics which were not present in the other dataset.

In the dataset “NBA players stats since 1950”, we observed that there are categorical few variables that have missing data such as college, birth city and birth date. These values are likely missing because they weren’t recorded at the time. There are a number of nulls in some numerical variables because those stats were not recorded or considered important until the 1980s. Some of the noticeable numerical variables with missing data include turnover, steals, blocks, rebounds, and advanced stats that require defensive stats for calculation. Additionally, there are a number of players missing values for 3-point shooting because the league did not have a 3 point line until 1979.

Comparing the dataset “NBA players stats since 1950” to “Basketball players stats per season – 49 leagues”, we found that the dataset “NBA players stats since 1950” has a greater variety of features which focus on player information. While the dataset “Basketball players stats per season – 49 leagues”, has a greater number of observations on players performance.

In summary, we have two large datasets with a variety of variables related to on-court player metrics, measurables like height and weight, and draft information. We will use these explanatory variables to predict either Field Goal Percentage (FG%), Usage Percentage (Usg%), or Player Efficiency Rating (PER) as a measure of that player’s on court effectiveness. With this regression model, we hope to gain an understanding of how general managers might identify useful pieces for their team or how to get the best out of a specific player.