Final Paper

Noah Edwards-Thro

1/23/2022

Schedule

1/24 - Methods of Sloan Paper 1/31 - New Methods 2/7 - Replicatability Graphic 2/14 - Results Copied from Sloan Paper 2/21 - New Results 2/28 - Clean Up Visuals 3/14 - Discussion Part 1 3/21 - Open Week 3/28 - Introduction/Final Edits Part 1 4/4 - Final Edits Part 2

Methods

To recreate the Sloan model as close as possible, I endeavored to use as close to the same data and methods that Kalman and Bosch used as I could.

Data

Kalman and Bosch manually scraped 10 years (ranging from the 2009-2018 seasons) of player statistics covering advanced aggregation statistics, per possession statistics, and shot distribution statistics. Their data consisted of 5512 observations with 73 variables where each observation was a single season of a single player (if a player played all ten seasons during this range, he would show up 10 times).

While I desired to stay as close to the process that Kalman and Bosch used as possible, I had knowledge of a package in R called nbastatr that pulled data directly from Sports Reference LLC (the same website that Kalman and Bosch manually scraped their data from). In an effort to get more experience working with this package, I decided to pull the data from this package using the `bref_player_stats` function. Additionally, the shot distribution statistics were unable to download via the nbastatr package so I manually scraped them from Sports Reference LLC. Finally, I used the `player_profiles` function in the nbastatr package to download the heights (in inches) for all of the players.

My data consisted of 4760 observations and 136 variables. Through some exploration, I found that the discrepancy in observations has to do with players who are traded. In my dataset, any player who is traded mid-season will still only show up as one row, with that players' team being "TOT" (signaling that the player played for multiple teams). Looking at the data on the Sports Reference LLC website, it seems that Kalman and Bosch likely had a separate row for each team that a player played on in a single season (so if a player played games for three separate teams, he would have three separte rows for that season in Kalman and Bosch's data while he would only have 1 row in mine).

In Kalman and Bosch's analysis, they filtered the data so that only observations with more than 30 games in a season are counted, ending with 3,608 observations. After using the same filter, I ended with 3,676 observations, likely due to the difference in data format with traded players (a traded player who plays 20 games for two different teams doesn't show up in their analysis but will show up as playing 40 games in mine).

Variable Selection

I used the same variables that Kalman and Bosch used in their model. The variables were a combination of offensive, defensive, and aggregate statistics and all statistics were calculated as rates except for player height. The drawback of these rate statistics is that it does not take into account how much a player is on the court. For instance, a player who plays 10 minutes a night could have the same rate statistics as a player who plays 35 minutes a night, but we would consider these players very different in their abilities (largely due to the 35 minute player being able to play efficiently for 35 minutes).