Danny Arenson

04/10/2017

Capstone 1: Data Wangling

For my first capstone project, I decided to work on NBA data. Specifically, predicting the “2016-2017 All- NBA Team Awards”, based on an innovative method of evaluating player effectiveness by using calculated variables to compute a “Scorer Rating” and “Playmaker Rating”. In effect, this will be providing a better weighted assessment of that player’s performance. Furthermore, I will then calculate Scorer and Playmaker ratings for the last 16 seasons (2000-2016) to compare/contrast any similarities/differences between the All-NBA teams that were voted in by the associated press and which players maybe should have been awarded solely from a metrics standpoint.

The standard metrics such as: *Points, Rebounds, Assists and Player Efficiency Ratings*, although the predominantly conventional route to determine a player’s ranking, do carry some inherent inaccuracies in weighing the individual player’s overall performance from a quantitative standpoint. These tend to provide a one-dimensional detailed explanation, but to provide more volume and in-depth look at player performance, further metrics need to be taken into account.

This revised format would help add more accuracy to which players are the best at what position. Here are some of the factors (not in any particular order) that structure a player’s Scorer and Playmaker Rating:

|  |  |  |  |
| --- | --- | --- | --- |
| Games Played | Three’s Made/gm | Two’s Made/gm | Assisted Three’s% |
| Team FGA/gm | Team FTA/gm | Team Turnovers/gm | Team Points/gm |
| Assisted Two’s % | Free Throw% | Off. Reb % | Off. Team Rating |
| Team points/possession | Assisted Points/gm | Unassisted Points/gm | FG missed/gm |
| Pts/gm | Ast/gm | TO/gm | FTA/gm |
| FGA/gm | Usage |  |  |

At first glance this seemed like a daunting task, particularly with how/where I would be able to find this type of data, but after giving it some thought, and doing a bit of further research, I realized that I could calculate most of these variables by using basic player statistics, then use the amended variables to formulate my Scorer and Playmaker Ratings. Essentially, my plan was to take the basic player stats, use them to calculate the more complex variables, and take those variables that were derived and apply to the Scorer Rating formula.

I was able to find all of my data online with the various popular NBA stats websites, but they didn’t consist of data put together on one table. My next plan was to scrape the player data from each API by using Pythons package “Beautiful Soup”. I was able to find various helpful YouTube links which provided the proper steps for web-scraping. I seemed to be making some progress in web scraping, but for one reason or another, eventually hit a roadblock with being able to complete this task. I brought this up to my mentor in case he would be able to provide some insight and troubleshoot this issue. He then brought up an interesting idea, perhaps I was overthinking this and if I had the ability to copy and paste the data on to a spreadsheet, then I should just give that a try, especially for time efficacy purposes. So I did this, and fortunately it worked. It took making 5 separate tables from three websites to painstakingly combine all player and team data together.

I took the combined initial player data and imported as a csv on python where I would then have my table and use the various python packages to evaluate the numbers and do any further data cleaning. I noticed that the outliers consisted of mostly unpopular players that I know should not necessarily have a particularly high rating. After looking a little closer, I realized that these players had a ridiculously high rating because they had not played in many games this season and when they did play, their game stats were based on very low averages with higher numbers. Had they theoretically played in as many games as a starter or role player, these outliers would have eventually regressed towards the mean. To solve this problem, I decided that to qualify for the rankings, players must have suited up in at least 55 games by the end of the regular season.