NBA Draft Day Comparisons

By: Aydin Baradaran-Seyed

Every year, 60 NCAA and European league players are drafted into the NBA. Of those 60, 14 of them are selected in the lottery, in which the teams who were unable to qualify for the best of seven series amongst other teams to compete for the championship are put into said lottery (worse the record, the better odds of a high pick). NBA general members are responsible for drafting the right players to complement their teams or in this case build around. They specifically have scouts who go to different games these players play in and analyze their game. My report is to supplement the scouting reports by providing a more in-depth analysis on what kind of current or past NBA player the up-and-coming prospect is projected to be. As this information is not available to the public, what I have done instead is use other media platforms such as Bleacher Report and Draft Net who specifically write articles on what each year's draft prospects real time NBA comparison is with the help of analysts behind the scenes.

The NBA is a league of 30 teams of which 15 people comprise a team. Team statistics exist, but individuals carry more to the team. There are a vast variety of NBA statistics that range from the old ones in points per game to new ones in player efficiency rating. With these statistics, a player's career is uniquely defined, and as such can be compared amongst others. The data set is comprised of quantitative fields, so models that go hand in hand with say NLP cannot be applied here. What can be, is logistics regression depending on how you phrase the target variable, linear to model increases in one statistic compared to others, to machine learning concepts like clustering in which we group and from there have an easier way of comparing players.

The data sets for all of them but one was found on Kaggle, the other one was done by somebody who managed to scrap what he was able to on basketball reference using R for NCAA statistics. The NBA data sets, 1 is used just for grabbing draft info, and the other for every season every player has played to until 2017. The NCAA table is comprised of a limited scope of statistics that are like the NBA ones. There are 52 (5 as strings, 47 floats) or so columns for the main NBA data set followed by 15,000+ rows, and the NCAA table contains 12 float type statistics columns and 3 or so string columns to identify. Also, to add, a self-made csv file was made to look at each years draft the two media comparisons for easy viewing.

The NCAA and NBA data sets have thousand(s) of NaN values, and as such EDA was needed. All duplicate rows were dropped if any existed, any player name in which it did not have one to identify was dropped. Any column that had less three percent of missing data was removed, while those that did were filled in by their respective positions median.

Linear Regression did not achieve a good accuracy in 5-7% for both adjusted and R^2 (I was not sold on this model and didn't look at QQ plots etc.). Logistics through machine learning libraries achieved 80-85% for train and test scores and in terms of statistics it achieved an 82.5% accuracy

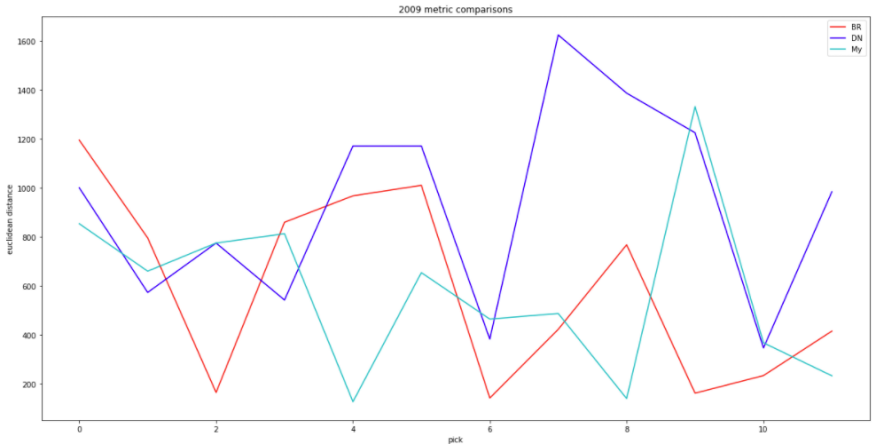
in terms of which statistics determine if a player was in the lottery or not. The machine learning package of logistics regression had a good confusion matrix results for those not in the lottery but suffered for those in the lottery. K nearest neighbors were briefly touched upon and had good results in the low 80%'s. Clustering was ultimately chosen due to the groups being comprised showed noticeable differences between picks in statistics of importance.

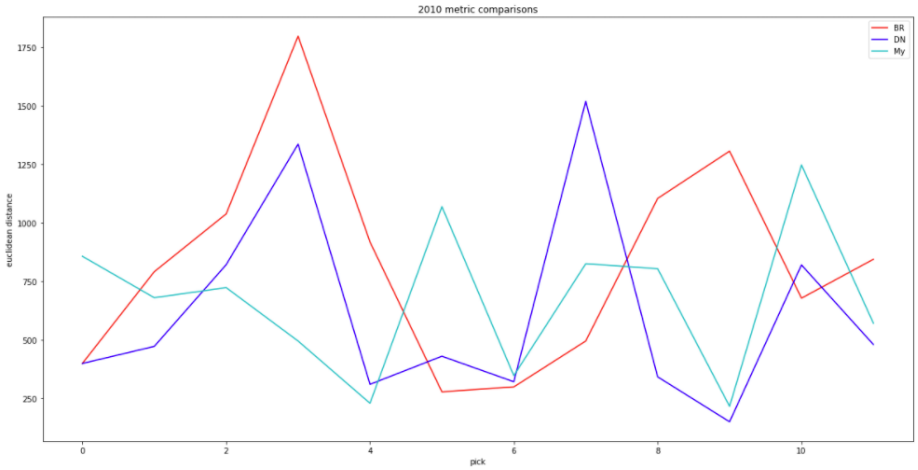
Through clustering, two clusters were created, one for the NBA and one for the NCAA. The NCAA was used to validate previous year's draft classes and give a more accurate comparison to sell the idea that this method and model provided have better results than Bleacher Report and Draft Net. From here, we can see how my model's prediction held up against the two. Getting 5/12 in first, 4/12 in second and 3/12 in being last (closer to the top 10) for the 2009 draft class. These results were done by Euclidean distance measures between points in a cluster, and a comparison was formed after finding the right person close enough for comparison. My average distance came in first between the other two media comparisons for 2009-2011's draft class.

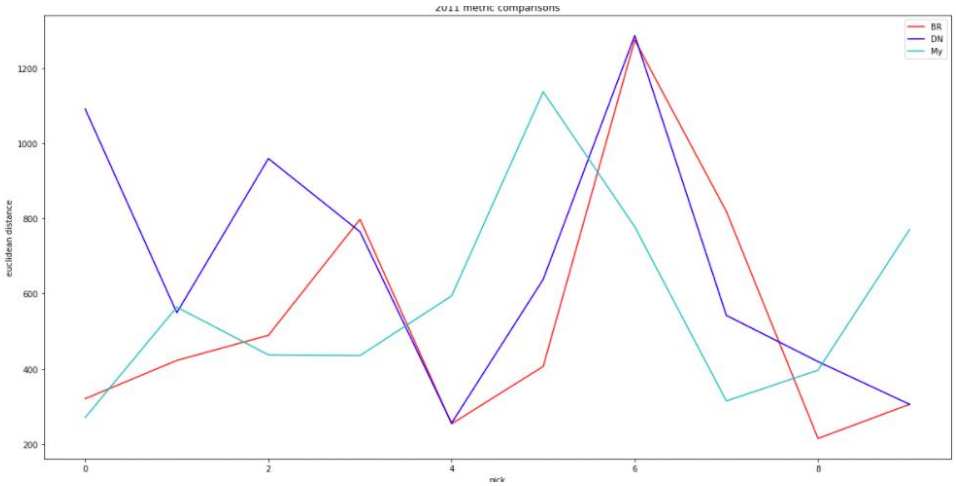
Light blue is my model

Dark blue is Draft Net

Red is Bleacher Report







Of the three Models, 2009, 2011 and 2010 are the in order from best to worst in terms of comparison. I found when my model was right, it was significantly right and when it was wrong it was still close enough to say it was a reasonable comparison to make.

My results were expected since although these sources may have analysts doing these behind the scenes, they ultimately would like to cause controversy by making some bold comparisons, so with using theorems and methods theoretically proven, I would expect my model to have better results in which it did. I was more surprised how much closer Bleacher Reports comparisons were compared to Draft Nets as the latter is a dedicated website towards the NBA draft.

Going forward, this model can be used to sell to NBA franchise if further perfected in the sense that I am given a team's scouting reports over the years and figure out where they went wrong and improve on those, so their reports are more thorough. As well, for the fans, this could be lucrative model as if they can buy the right rookie cards, based on rarity and quantity of a future player, they can make thousands if not multiple millions. Going forward I will improve this model by web scrapping basketballreference.com's website with the help of VPN with Beautiful Soup and Seline. As well, I would also scrap more data, so I am not limited to 2017 for the NBA data set. I will also go further into looking into more drafts to see if my model still ranks first. Lastly, I will explore other machine learning methods to model and predict to see if I can make my average distance even smaller in terms of other metrics (or the same).

Links to the data:

https://www.kaggle.com/justinas/nba-players-data

https://www.kaggle.com/drgilermo/nba-players-stats

https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv