David Keyer

After having spent time working on my Capstone over the last couple of weeks, a lot has changed in the way I originally wanted to approach the project and how I plan to now finish it. My original problem was a regression problem, and my problem statement focused on effectively being able to predict NBA player salary based on game statistics. One of my most important tasks was how I grouped/cleaned the data. I eventually decided that I would use player data from the past five seasons to predict his salary for the current season. I made this decision for a couple of reasons. First, I wanted to get rows (players) out of the model that could negatively affect it. Generally speaking, higher stats such as Minutes Played, Points, Rebounds, etc. leads to higher salary, and first/second year players typically have lower salaries. By grouping only players who had been in the league for the past five years, I was able to get rid of those rows that could take away from the predictive power of the model(first/second year players with high stats). My second reason for grouping the players this way was for applicability. I designed this project so that we can be able to look at a player's consistent statistics over a significant length of time in his career, and predict salary based on those statistics. So, being able to look at veterans' (players who have been in the league for a long period of time) stats and salaries can help predict current/future salaries. I tried grouping by players who were in the last four seasons and three seasons leading up to the salary data, and there was only a minimal increase in the amount of rows that were in the dataframe, and the metrics were actually worse in both models than when I grouped by five years, so I stuck with five.

Thus far, I have primarily worked with 2019-20 salary data and stats from the five seasons leading up to 2019-20. I plan to work with at least the salary data from the previous seven seasons as well, with the five seasons leading up to all of that salary data. I'll build a function to do this.

The main change that has happened is that I'll be switching from a regression problem to a classification problem. Unless there is a drastic turn and I engineer some feature that gives me tons more predictive power, my regression metrics that I have been using to define success (Primarily r-squared and RMSE) have not been good enough. As a result, I am going to switch my main problem to one of binary classification. To guide my results, I will be using K-means clustering to help choose my classes. I'll also choose classes based on quartiles and choose them myself so as to see the differences in the models. I plan to have five classes overall. As mentioned, I'll be working with salary data from the previous seven years, and trying to classify if players are above/below the mean for that year, based on several years of statistics.