**Define your problem:**

The goal of this capstone project is to be able to predict an NBA player’s first non-rookie salary. For background, when a player is drafted into the NBA they are designated a rookie salary that is determined by where they are drafted. This salary lasts for 4 or 5 years. There’s a great deal of variance in recently drafted players’ skill sets, some quickly become good players, others develop slowly, others do not ever become serviceable NBA players. As a result, an NBA player’s first non-rookie salary is reflective of how that player performed during those first 5 years in the NBA. Our objective is to be able to predict this salary based on the player’s body of work on his rookie contract.

**Describe your client:**

A couple potential clients for this would include: NBA teams and NBA agents. Assuming this model is reputable and does a reasonably good job at projecting salaries, an NBA team could use this when determining how much to pay their players (that are just coming off Rookie contracts). They would be able to determine the “fair price” and ensure that they do not overpay for a certain player. Additionally, they could determine in advance if they will have enough money to realistically sign a player. If it becomes clear they will not have the money to sign a certain player, they could preemptively trade away that player for a different asset.

This would be a similarly useful model for NBA agents representing their clients, the NBA player coming off his rookie contract. This would be a useful negotiation tool to ensure that they are not being undersold by an NBA team.

While not as directly relatable, this could be used as a proxy for overall player quality/talent. It’s important to note that measuring monetary value is not the same as overall player quality because when we model monetary value we are implicitly modeling any market inefficiencies. That said, this would be a useful proxy for overall player quality in lieu of a different model.

**Describe your data set and how you cleaned and wrangled it:**

The dataset I used was from Basketballreference.com. I first needed to create a list of players whose stats I would then go and scrape. I first iterated over the recent NBA draft webpages to get a list of players. From there I had to reformat their names to I could use them to search the player-specific BasetkballReference URLs.

For each player in my list I would go to their personal player specific web page and scrape down 3 tables of data, their traditional stats, their advanced stats, and their contract information. I then concatenated these into one dataframe for player. I later concatenated all these dataframes into one large dataframe which I then later saved to a local TSV file.

I did have to throw out data along the way. If I player didn’t make past their rookie contract in the NBA, or never entered the NBA despite being drafted, I threw out their data. I also had to handle some formatting of the data. In particular, the salary information came in with dollar signs and commas which had to be taken out to be read and used for calculations.

**List other potential datasets you could use:**

The most obvious choice would be NBA.com/stats. While many of their statistics are the same as BasketballReference, the source I used, they do have some different statistics. This site looked a bit harder to scrape which is in part why I did not choose it. Another way I could have gone would have been to supplement my data for college statistics. It’s likely that a player’s draft position and college stats still play a small factor when determining their first non-rookie contract, particularly if that player was injured for some of his first years in the NBA.

**Explain your initial findings:**

I have not yet developed the actual model for prediction but still have had some interesting and important findings.

I first created a dataframe of the weighted averages of all the stats based on minutes played per season. I then appended the first non-rookie contract value to each of these rows. This allowed me to more easily compare correlation levels for the independent variables (all the player stats) to the dependent variable (the player’s first non-rookie salary). I then made correlation matrices to get a rough idea of what statistics might be useful to use in the model. Unsurprisingly, Points was highly correlated, as well as many other advanced statistics that try to approximate overall player quality.

I determined that using salary as a percentage of the overall league salary cap increased correlation levels across the board. This makes sense as salaries are considerably higher in the more recent years due to inflation and other factors like the NBA’s overall success.

After looking at many stats and their correlation levels to salary, I began to break them out by player position. It made intuitive sense to me that Assists might be a driver of salary for Point Guards but not for Centers. I did find that there can be a decent amount of variation for the correlation levels across the positions.

The next step was to determine if the variation in correlation levels that was being seen across positions was statistically significant. Many different statistics were tested but I did find several example of statistically different correlation levels. This is a good basis for creating an ensemble model. The statistics that I have determined are statistically by position (Point Guards’ assists and steals, Big Men’s Blocks to name a few) will be used in position specific models. Then these position specific models will be wrapped up and used in a larger overall model to predict the target, an NBA player’s first non-rookie salary.