Projecting NBA Rookie Salaries from College Stats

The NBA draft is one of the most important days for the business. Nailing draft day may decide whether a franchise creates an incredible dynastic future - or prove fatal for the years to come. Predicting the rookie salaries may maximize value to an organization by allowing the organization to optimize their team's make-up with cheaper salaries. Projecting what a player is worth minimizes the chances of being stuck overpaying players.

Data Wrangling

The data I've chosen was the NBA salary data that was pulled from a SQL server from https://data.world/datadavis/nba-salaries (sourced from basketball reference). From here I selected only the necessary columns for my capstone project. These being the player name, position, their statline and the season.

Upon investigation I found NA values in a few rows for draft position, career field goals, 3-point field goals, free throws, player efficiency rating (PER) and effective field goal (EFG). I decided to drop these rows from the dataset because the dataset was large (n = 13626) and wouldn't create a large difference. I filled the missing values in draft picks with "undrafted" to fill out the rest of the dataframe. I was able to get the college statistics from the writer of a blog (dribble-analytics).

I took the classes as a numerical value to sort by values to take the latest college year played. This allowed me to merge these columns back with the updated college dataframe. The next step was to match their latest college year with their respective rookie year in the NBA. The college stats dataframe contained the season abbreviated as "2011-12". In order to match the college stats to the NBA rookie salaries, I took the first four digits of the college stats data (ex: 2011) and added 1 so that it would match their rookie year.

I had to extract the important columns from the NBA salary data. The only relevant columns were the NBA player's name, salary and year played. I merged the college stats dataframe and the NBA salary dataframes.

I figured the college stats needed to be contextualized in order to make them comparable. I took per game averages (points per game, assists per game, etc.) for their stats and transformed their field goals into a percentage to reflect accuracy (FG%). I also added effective field goal percentage (EFG%) - a metric in the NBA to reflect that 3 pointers weigh more than 2 pointers. I then filled the rest of the missing values with zeros - not all players attempted 3 pointers. I also created a completely separate dataframe that would further contextualize these stats into a 36 minute average (Per 36 stats).

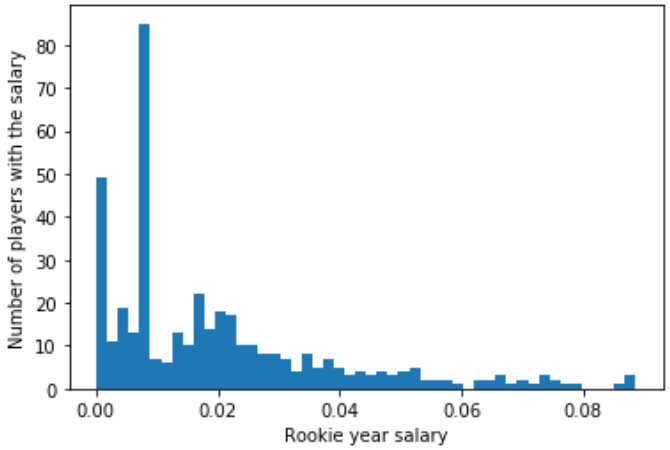
Lastly to account for inflation in salaries throughout the years so I created a new columns that calculated salary as a percentage of the salary cap. The historical salary cap was referenced from basketball-reference. There were only 17, so it wasn't necessary to scrape the web. I added this column by creating a new column that was the season and replaced those values. I then

created a new column which took the NBA rookie salary divided by the salary cap to show the salaries expressed as a percentage of the salary cap.

Exploratory Data Analysis

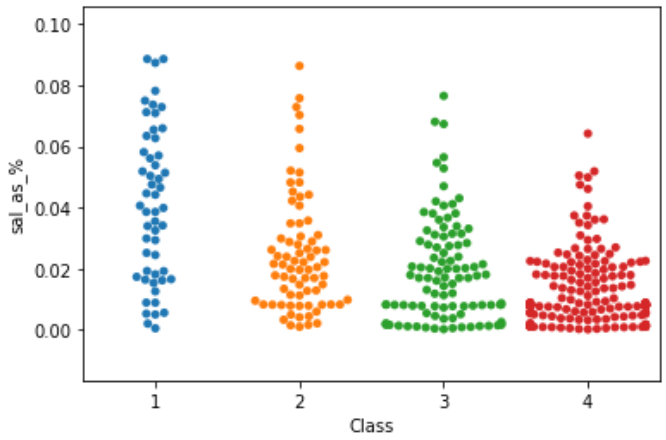
I first created a scatterplot of the NBA rookie salary which first showed a gradual increase of salary throughout the years. In order to control for the inflation of salaries, I calculated salary as a percentage of the salary cap which would normalize the data. I then created a scatterplot of the rookie salaries as a percentage of the salary cap which showed no relation throughout the years.

I plotted the distribution of salary as a percentage of the salary cap. This allows us to visualize rookie earnings:



A majority of the rookie salaries are concentrated under 2% of the salary cap and the distribution is skewed to the right. What exactly separates those rookies with the highest earnings apart from the others? I explored the stats that contribute the most to highest earnings.

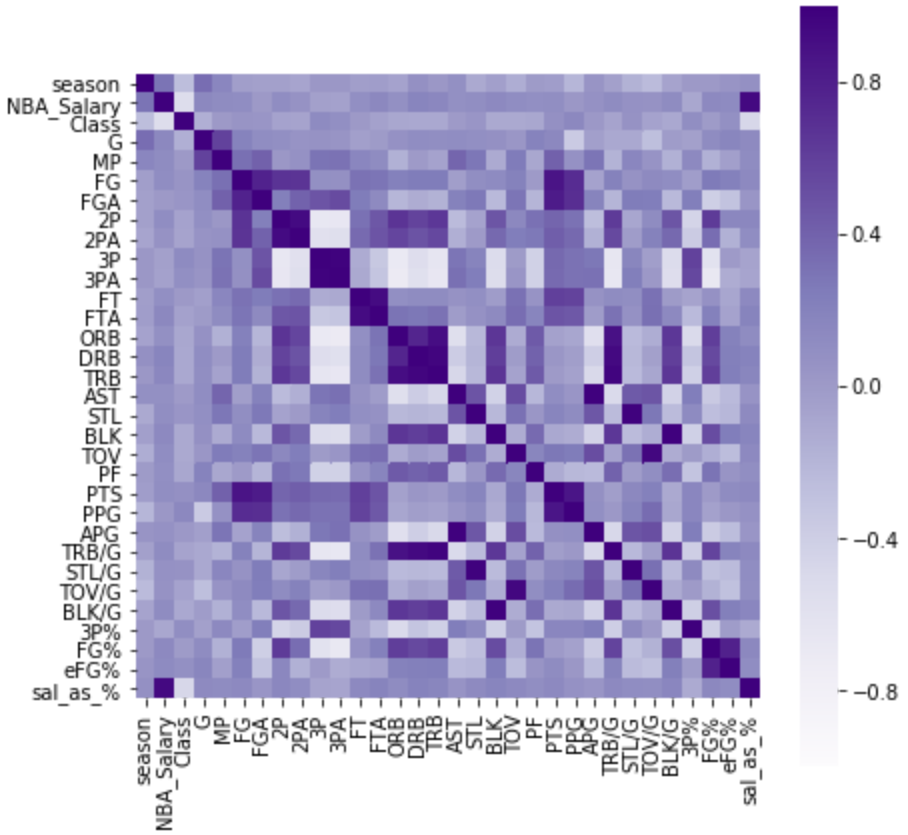
I created a swarmplot to show the distribution of rookies grouped by class. This shows pretty much what I expected - more variation between rookie salaries the less years spent in college. This makes sense since the most anticipated recruits are those who only spend one year in school (one and done) while the ones less sought after are spending more time in college developing their skills. This is backed by the correlation coefficient between rookie salary (% of cap) and class to be (r = -.48).



I decided to query by position and find the spread of the rookie salaries by creating boxplots:

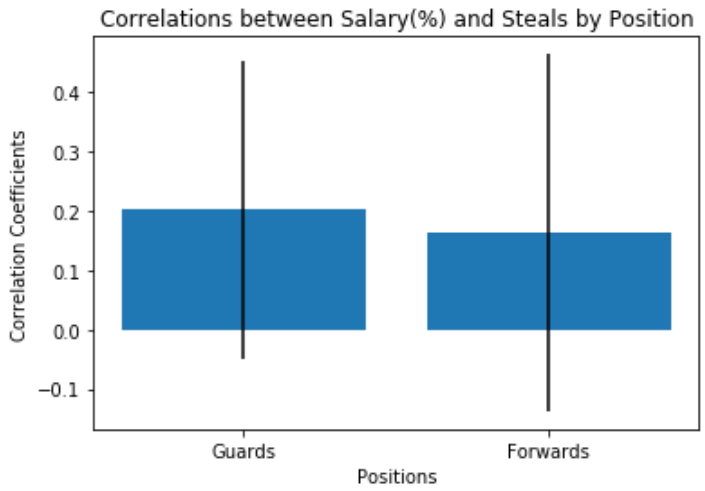
We can see that there are more extremes for forwards and guards. The median is the lowest for guards while having some of the highest rookie salaries and forwards have a middling median. Centers have the highest median, but the highest salary is not nearly as high as the other two positions. Upon further investigation, I found that our dataset included a small number of centers (n = 36). For this reason I've excluded them from the rest of the visualizations as it would create issues with such a small sample size.

To find which stats most contribute to higher salaries, I created a correlation heatmap. Upon my first exploration into the correlation heatmap, I found that there were very little correlations between the stats that set players apart from the rest. The highest correlation from my initial findings (that was not class) was blocks which had a pearson correlation coefficient of r = .18. This shows weak positive correlation between salary as a percentage of the salary cap and blocks during the players' college career.

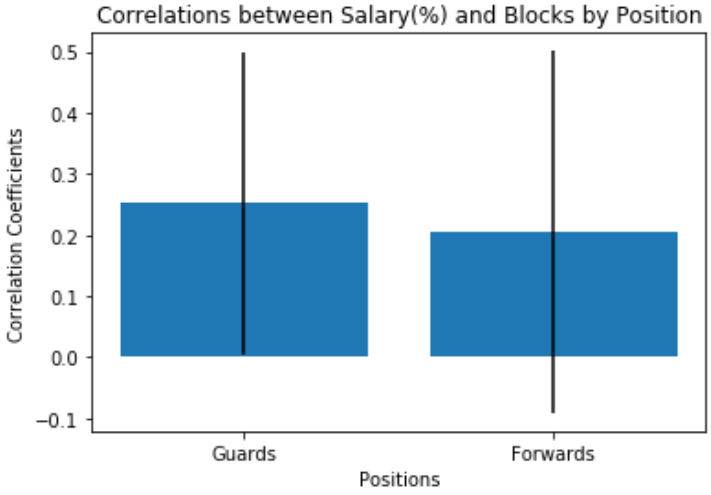


In order to dive deeper I decided to query by position and found the correlation coefficient of each statistic that stands out most:

There is a clear distinction between all positions and their highest correlation coefficients. The highest correlation coefficients for guards were blocks (r = .25) followed by turnovers (r = .24). I don't think that turnovers are what NBA scouts are looking for - rather high turnovers are a result of a high volume of playing time so I will disregard turnovers. The next highest correlation coefficient to salary is steals (r = .202). The correlation coefficients lead me to believe that guards with defensive abilities set themselves apart from the rest. This shows a versatility for guards to defend multiple positions. This hypothesis is backed by the correlation between salary and steals.



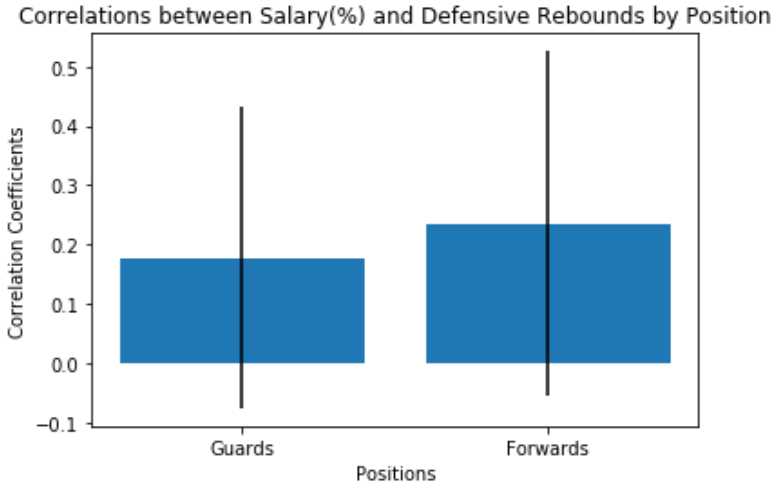
The 95% confidence interval for the correlation between steals on salary(%) by the guard position ranges from .073 to .325. We are 95% sure that the true correlation coefficient of steals by the guard position against salary(%) lies between .073 and .325 with a p-value of p = .002. At an alpha level of .05, we can conclude the correlation between steals against salary by the guard position is statistically different from 0.



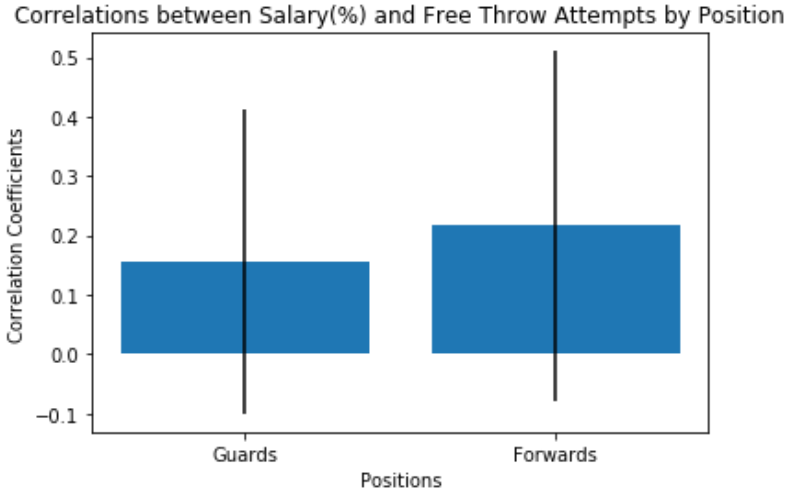
The 95% confidence interval for the correlation between blocks on salary (%) by the guard position ranges from .125 to .371. This means we are 95% sure that the true correlation coefficient for blocks by the guard position on salary (%) ranges from .125 to .371. The p-value for blocks by guards is p = <.001 which we reject at an alpha level of .05. This tells us that the correlation coefficient is significantly different from 0.

The highest correlation coefficient for forwards are defensive rebounds (r = .23) followed by free throw attempts (r = .22). This indicates that forwards who play inside and get to the free throw line set themselves apart from the other players. Their ability to get inside and grab

rebounds or draw contact to take free throws have positive correlation coefficients with higher salaries.



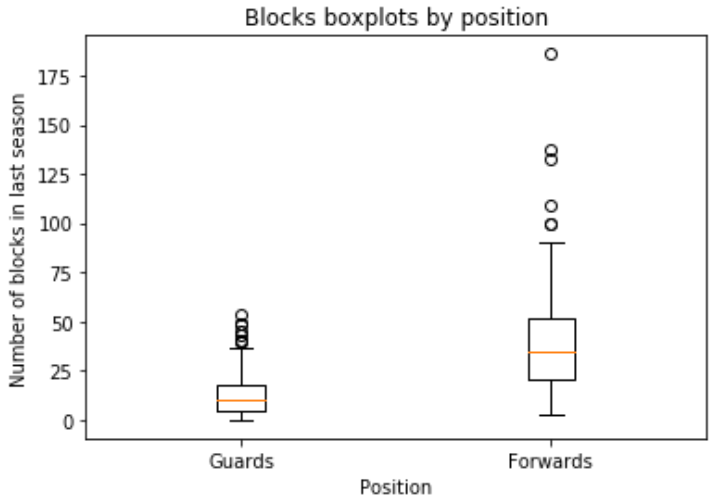
The 95% confidence interval for the correlation of defensive rebounds by the forward position on salary(%) ranges from .084 to .375 with a p-value of p = .002. We are 95% sure that the true correlation coefficient lies between .084 and .375. The p-value of .002 is beyond our alpha level which implies that the correlation coefficient of defensive rebounds by the forward position is statistically different from 0.



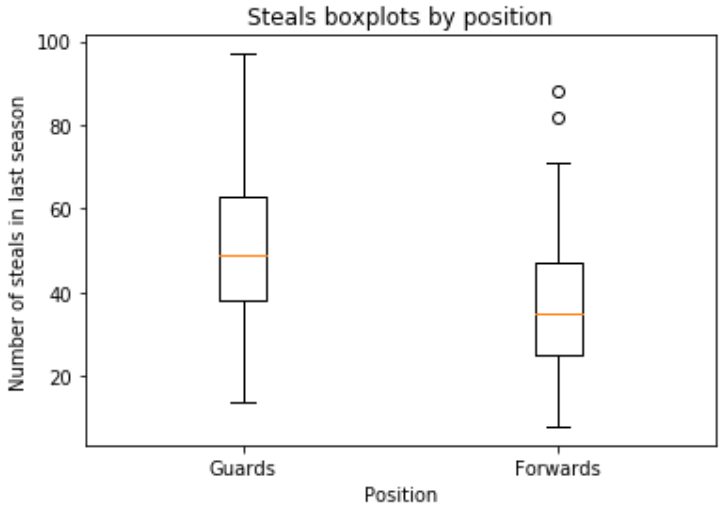
The 95% confidence interval for the correlation between free throw attempts by the forward position on salary(%) ranges from .064 to .36. We are 95% sure that the true correlation coefficient lies between .063 and .36 with a p-value of p = .006. The p-value is statistically significant and means that the correlation coefficient is different from 0.

There are clear distinctions between the stats that affect the salary for each position. From the correlations above, we see that the highest paid basketball players are those who put up stats that aren't typically associated with the position. The highest paid athletes are guards who put up blocks and steals, and forwards who grab rebounds and draw contact.

I created box plot visualizations of the spread of the stats that contributed most to higher salaries per position:

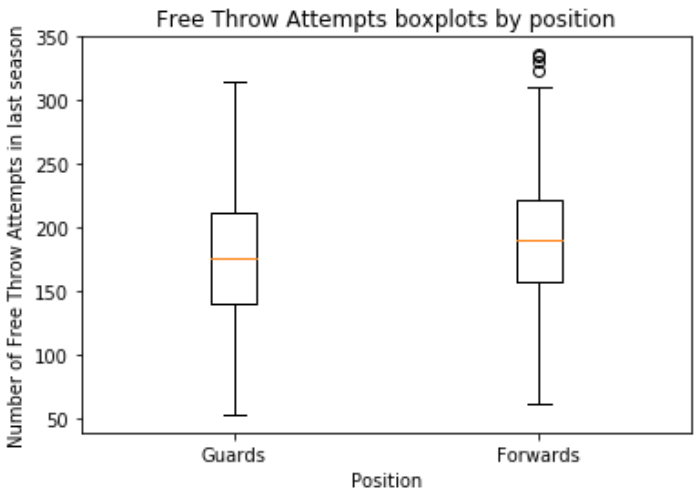


While the median of blocks for guards is relatively low (10), there are guards who clearly outperform the rest of the group. The highest number of blocks by a guard was 54. The spread of blocks for forwards is much wider, with a median of 34.5 and a max of 186 blocks.

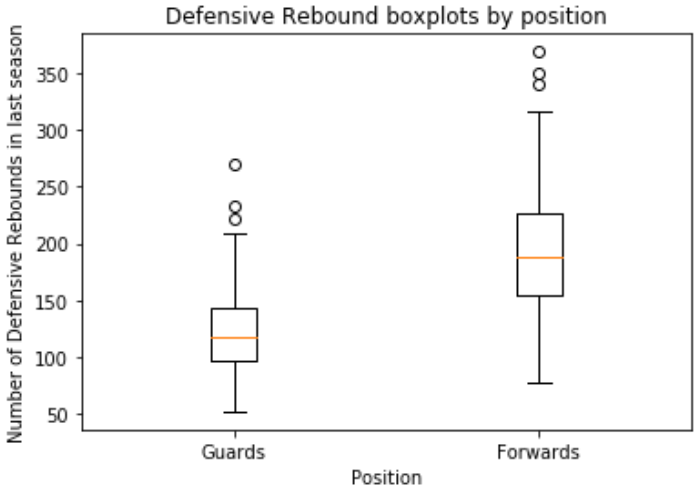


There is a much different spread for steals compared to blocks for these positions. Guards have a higher median than forwards (49 vs 35) and have a much higher maximum than forwards. There aren't as many outliers for steals which show a more balanced spread.

The spread of free throw attempts are similar for each position, however forwards have a lot of outliers on the higher end. This may indicate that forwards with the highest amount of free throws are those who get drafted the highest while those comparable to guards do not.



The spread of free throw attempts are similar for each position, however forwards have a lot of outliers on the higher end. This may indicate that forwards with the highest amount of free throws are those who get drafted the highest while those comparable to guards do not.



The spread of defensive rebounds for forwards is much wider than that of guards. The median defensive rebounds for forwards are 188 compared to guards at 117. The max defensive rebounds for forwards are 369 compared to guards' 188.

I found that there are different correlations to higher salary depending on the position of the player. A guards' defensive abilities such as Blocks (r = .25) and Steals (r = .20) correlated with higher rookie salary. A forwards' defensive rebounds (r = .23) and free throw attempts (r = .22) correlated to higher rookie salaries. To verify the results, I calculated confidence intervals and p-values to validate my results.

I created a confidence interval for all of the correlations between the college statistics and the salary earned during a players' rookie year in order to add a standard error for each correlation.

The confidence intervals calculated above verify that these college statistics have a significant relationship to a players' rookie salary depending on their position.

Applying Machine Learning:

The objective was to predict rookie salaries of NBA players given their college statistics. Because our target variable (salaries) are continuous values this is a regression problem.

Feature Selection

I began by first testing which features out of the 37 should be used for predicting salaries. I eliminated features that I deemed unnecessary for our model - Player name, school and the target variable. I used two different methods for feature selection. The first was recursive feature elimination (RFE). I used RFE to select the most important features from our dataset and queried by only the best features (ranking = 1) which gave a list of 15 features. These features were using in a ridge regression - fitted to a training set and scored on a test set which yielded an r-squared value of .119. The second method I used for feature selection was creating a list of combinations of the features and sorting these by which combination would yield the highest score. I was limited by a loading time so I was only able to get a combination of 8 features. I then fit a ridge regression to the training set using these 8 features and scored it on the test set which yielded a higher score (.1698) than the features selected by RFE (.119). The features selected by the second method were: class, field goals, field goal attempts, 3 pointers, free throw attempts, defensive rebounds, steals and blocks.

Choosing an Initial Model

Now that I had the features selected, I needed to test which model scored the best. The models that I chose were linear regression, ridge regression, lasso regression, elastic net, KNeighborsRegressor and random forest regressor. In order to compare the models, I used the same training and test set and compared which model performed the best. After trying all the models, the ridge regression yielded the highest score at r-squared = 0.1698.

Ensemble Methods

Now that the features have been selected and know which model performed the best, I attempted to improve the r-squared score by pursuing two alternative ensemble approaches:

1. Modeling guards vs. non-guards separately
2. Modeling guards vs non-guards separately for certain features then feeding the output into a new model which predicted salary on the remaining features

Method 1: Modeling guards vs. non-guards separately

My previous work showed that guards and non-guards had features that had different correlations with salaries. Blocks and steals were higher correlated with higher salaries for

guards, whereas free throw attempts and defensive rebounds were higher correlated for forwards. As such, I thought it would be sensible to see if modeling them separately would lead to a better score than a single model for both groups. Then, I tried building a Ridge regression on each, then predicting on the respective parts of the test set, combining the predictions back into a single array, and then getting a score that could be compared to our previous overall approach.

I created a training set and test set for both guards and non-guards for the features, and a training set and test set for the target variable (salary) for each position. I used these to train models for each position. I predicted salaries for each position using these test sets and concatenated both of these predictions into one column. I then concatenated the indexes of both test sets and created a dataframe using the predictions. This dataframe was then merged with the test set before querying to match the predicted salaries with the actual salaries. The last step was to score the predictions against the actual salaries which yielded a score of .167.

Method 2: Modeling guards vs. non-guards separately for certain columns

I tried to improve the score by creating an ensemble model that separated by position only on those columns that had been shown to differ by position in their effects on salary, and wouldn't separate by position for the rest of the columns. For the first set of models, which are separated by position, similar to the previous one, but only using features that are different for these positions - blocks, steals, free throw attempts and defensive rebounds, I predicted on the training set to generate a column of predicted salaries to replace these columns. I concatenated these predictions to the other columns that weren't shown to be different in their effects on salary by position - Class, Field Goals, Field Goal Attempts and 3-Pointers. Another Ridge regression was built on top of this new dataframe which yielded a score which was surprisingly lower than our previous two models with a score of .158.

The first steps of creating a model to predict the salaries of NBA players using their college statistics were to select features. Using a combination method I was able to find a combination of 8 features - Class, field goals, field goal attempts, 3 pointers, blocks, steals, defensive rebounds and free throw attempts. I then verified that the original Ridge regression was the best model by comparing the score against other models which ultimately yielded the highest score compared to the ensemble models created afterwards. As a sanity check, I used all of the available features in a model to test if the r-squared is consistently this low which yielded a score of roughly 0.146 confirming I had improved upon a simple model.

Conclusion

The highest score of the models was not as high as I had hoped. Intuitively, it seemed that college stats could be a good predictor of starting salary, but it turned out that the relationship between stats and salary may not be that strong. This could be due to the highly subjective nature

of drafting. In the end it's made by an organization on a need basis - not simply by who is the best statistically. There are other potential reasons as well as to why the model wasn't as predictive as expected. The first is that the dataset that I was working with is not a complete one. Many of the players that get drafted into the NBA play in leagues overseas. The NBA is an international phenomenon so players are not solely coming from the NCAA - even some extraordinary cases that do not attend college at all and go from high school straight into the league. We were not able to get a large enough sample of centers either (n=36) which affected the predictions of the model. The final dataset that I was working with had roughly 400 unique players which impacts the model's accuracy.