MDA 9159 Regression Report:

Using Player Performance to Predict NBA Salaries

By: Pauline Joseph and Khris Wilhelm

**Introduction: Background/Motivation of the data set.**

The National Basketball Association (NBA) has been the largest basketball league in the United States since its inception in 1949, and it has aided in the popularity of basketball worldwide. Being a successful franchise in the NBA can lead to increased ticket and jersey sales, increased fan support, and an increase in elite players who want to play for your team. However, a salary cap exists currently in the NBA. The salary cap sets a threshold on the maximum amount of money a team can spend without being penalized. The salary cap exists to prevent rich owners from signing only elite players and ruining the competitiveness of the league. Thus, to build a great team, it is often necessary to sign great players for cheap contracts. In this project, we aim to use statistics to determine which factors lead to a player earning a high salary. For example, perhaps a player that is a great rebounder will command a high salary. If our model is successful, it will allow NBA owners and managers to decide which players they should target in trades. Fans of NBA teams will also be able to use our model to see if their team has made wise decisions in free agency.

**Description:**

All data used in the report was acquired through web scraping various websites. The advanced statistics of the players in our dataset were acquired from the website basketball-reference.com. This website stores statistics on NBA players and teams, and is an exceptional resource for basketball fans. We acquired all player salary and team payroll data from the website hoopshype.com. Team salary cap data was acquired from the website realgm.com. All financial values were calculated adjusting for inflation to reflect current U.S. Dollars. Each observation in our dataset includes the “advanced statistics” and yearly salary for a player in the NBA, from the 2011-2012 season to the 2019-2020 season. Advanced statistics are used to measure the decisions a player makes when on the floor, and how good a player is at some facet of the game. For example, the advanced stat “3PAr” calculates the percentage of field goals a player shoots that are 3-point field goals. Another example of an advanced stat is “TRB%”, which measures the percentage of available rebounds the player acquired. Typically, tall players who play Power Forward or Center will have high TRB%, while Point Guards and Shooting Guards might have high 3PAr values. A full table of all of the predictor variables and their description is available in the appendix of this report.

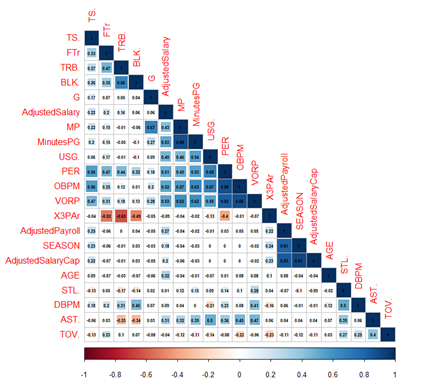
When starting this project, our goal was to only study players who were significantly important to their teams. Thus, we only acquired stats for players who played more than half of the season and who played over 12 minutes per game. The dataset also includes the payroll of the player’s team for the particular NBA season, and the salary cap the league abided by that season. The payroll of a team is simply the total amount of money the team spent on all of their players in the current season. In total our data contained 2,433 observations, and 24 total columns.

**Data Preprocessing**

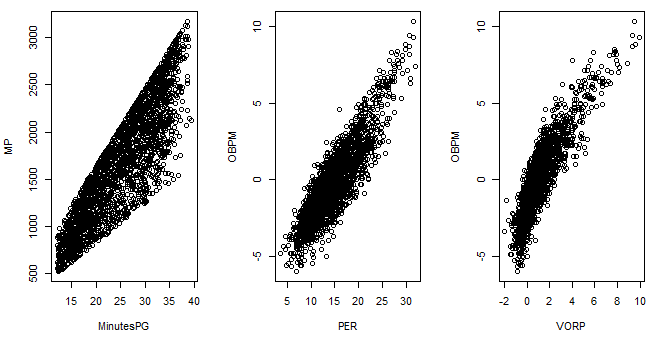
Since the name of the player and the team name are not useful in making the analysis, the two columns, “PLAYER” and “TEAM” are removed. The dataset does not contain any missing or NA values and there are no duplicate rows in the dataset either. There is one categorical column, “POS” and “RookieContract”, these columns are factor predictor variables. The POS variable lists what position the player primarily plays.

**Data Exploration**

Several plots are created to obtain a better understanding of the dataset. To avoid high VIF values that would cause us to draw poor conclusions about our beta coefficients, it is important to understand which predictor variables are highly correlated. The correlation plot pictured below will help us attain this goal.



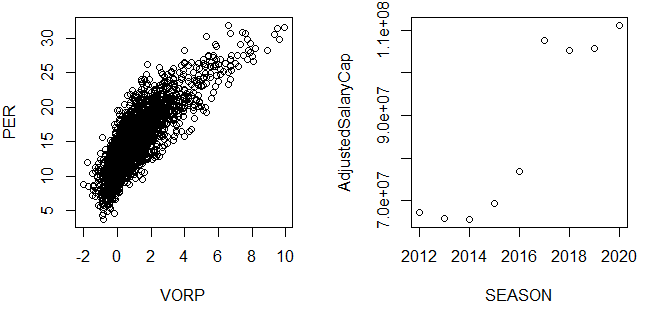
From the plot, it can be seen that the correlation between certain terms are high. The correlation is between “MP” (Minutes Played) and “MinutesPG” (Minutes Per Game) is 0.88. The correlation between “OBPM” and “PER” is also 0.88 and so is the correlation between “OBPM” and “VORP”. Since “OBPM” has a highly positive correlation between both “PER” and “VORP”, it comes as no surprise that “PER” and “VORP” are highly correlated with one another. A highly positive correlation of 0.91 exists between “AdjustedPayCap” and “SEASON”. “OBPM” and “VORP” are also positively correlated. This is pretty intuitive as both features estimate a player's contribution to the team, when the player is on the court. Since “OBPM” and “VORP” are highly correlated and so is “OBPM” and “PER”, it comes as no surprise that “VORP” and “PER” are also positively correlated. A highly positive correlation exists between “AdjustedPayCap” and “SEASON”, this is because the NBA has increased in revenue from 2012 to 2020. Since players are bringing in more revenue for the league, teams must pay players more.

Next, we plotted the correlated features to further understand the extent of their relationship and why the correlation coefficients for the pairs of attributes are so high.

As seen from the plots above, there is a clear linear relationship between MP and MinutesPG. This is due to the fact that MinutesPG = MP / G, where G is the number of games played. Hence, it makes sense that there is a high correlation between the two.

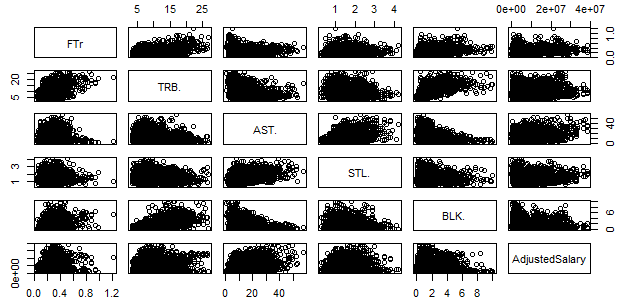
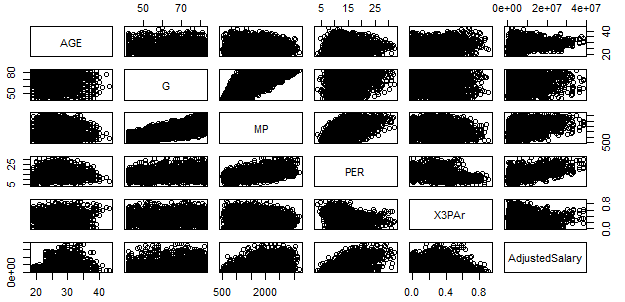
Since both “OBPM” and “PER” determine how good a player is, it makes sense that the highly positive correlation exists between them.

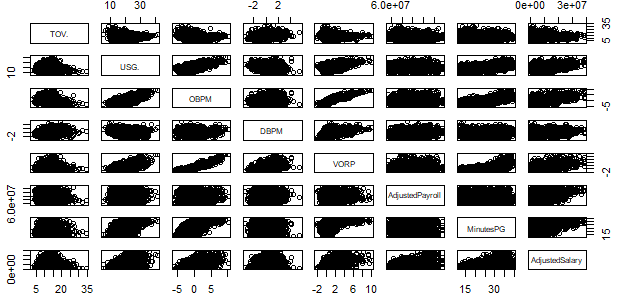
Both “OBPM” and “VORP” estimate a player's contribution to the team, when the player is on the court. So, it is likely that as one increases, the other one will also increase and hence there is a highly positive correlation.



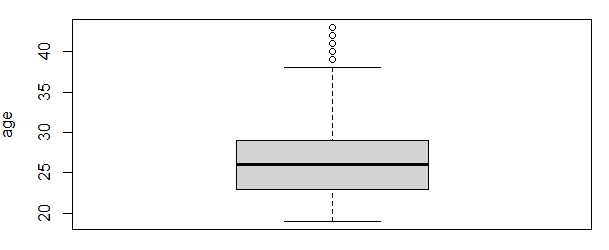
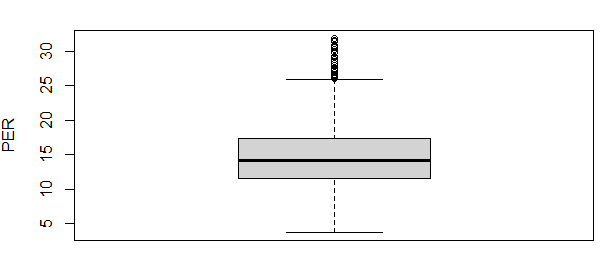
Since “OBPM” and “VORP” are highly correlated and so is “OBPM” and “PER”, it comes as no surprise that “VORP” and “PER” are also positively correlated. A highly positive correlation exists between “AdjustedSalaryCap” and “SEASON”. The NBA sets a cap on the maximum amount a team can pay its players and “AdjustedSalaryCap” represents this amount. The maximum amount changes from season to season. The NBA has increased in revenue from 2011 to 2020. Since players are bringing in more revenue for the league, teams must pay players more. Hence there has been a steady increase in the salary cap from season to season.

Next, to understand the relationship between the response and predictor variables, we plotted scatter matrices.





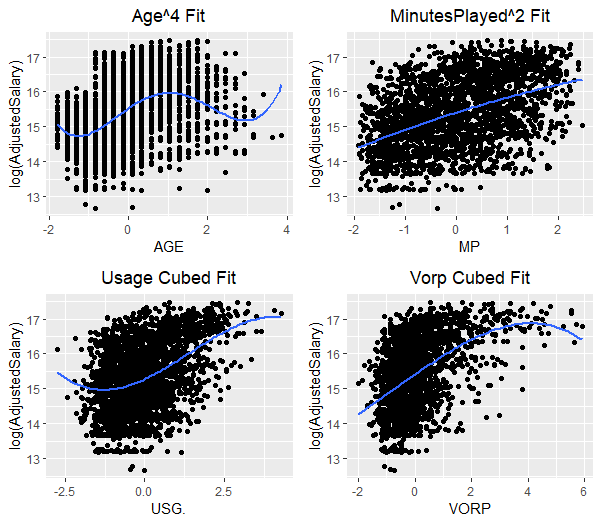
From the plots, there does seem to exist some sort of relationship between the response variable and the predictor variables, however, it is quite hard to determine the exact nature of the relationship between “AdjsutedSalary” and the predictors. After this we plotted boxplots to check the presence of outliers. It was seen that 11 predictor variables contained outliers. Two of these boxplots are shown below.



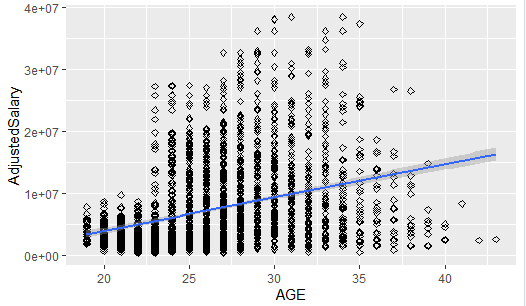
The plot on the left is the boxplot for “PER”. 28 unique “PER” values were identified as outliers by the boxplot. The plot classifies all ratings above 26.1 as outliers.

The plot on the right is the boxplot for age and it identified everyone over the age of 39 as an outlier with the maximum being Vince Carter of age 43. Since Carter retired at age 43, and played in the NBA until then, this is not a simple measurement error and hence should not be removed.

When visualizing the data, we wanted to determine which variables would benefit from having higher-order terms. We made plots of each variable and tested whether adding polynomial terms would improve the model fit. Plots of age, minutes played, usage percentage, and VORP are plotted below.

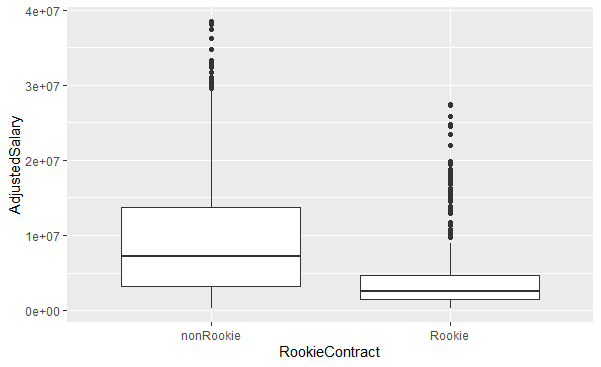


To try and better understand the nature of the relationship between the response variable and the predictor variables, individual plots were made between each feature and the response variable. The plot between “AGE” and “Adjusted Salary” seemed interesting to us as it showed that very young and old athletes are paid less than athletes around 30 years old.



We tried to understand what the reason behind this could be and found that NBA teams give out “Rookie contracts” to players drafted in the NBA draft. These rookie contracts typically pay much less than a non-rookie contract would pay. For instance, Trae Young and Russell Westbrook are both exceptional Point Guards in the NBA. But Trae Young is on a rookie contract which is paying him $6.3 Million in 2020, while Russell Westbrook’s non-rookie contract is paying him $38.5 Million this current season. The average age of players drafted at the NBA draft is 22, and these rookies typically sign with their teams for 3-4 years. So, it is safe to say that generally, athletes under the age of 25 would be signed to rookie contracts. This is an estimate, as we did not have time to go through over 2400 players in our dataset and determine if they were signed to rookie deals or not. From the dataset, it can be seen that the average salary of athletes with rookie contracts is $4.1 million whereas the average salary of players who are not on a rookie contract is $9.4 million. Since there is a significant difference, this implies that whether or not the player is on a rookie contract will determine their salary to a great extent. A new variable called “RookieContract” was created. It is a categorical variable which classifies players as “Rookie” if the player is on a rookie contract and “nonRookie”, if the players are not on a rookie contract.

The boxplot below further illustrates the salary differences between Rookies and veteran players.



One important thing to note from this boxplot is that there are a lot of Rookies whose salaries are outliers. In fact, 74 Rookies have salaries which are deemed as outliers, while only 34 non-rookies have salaries which are outliers. For these Rookies whose salaries are outliers, it is quite possible that in fact they are not rookies.

A model that suits the dataset the best might be one that has interaction terms between different predictors and from the initial correlation plot it is evident that there exists some collinearity among the predictor variables. If there is high multicollinearity between the predictor variables then it can cause problems when we fit the model and interpret the results. To deal with this problem, we have normalised all the numeric predictor variables so as to reduce multicollinearity that is produced by the interaction between multiple predictors.

First, we split the data into a training set and a testing set. 80% of the data is put in the training set and is used for training the model and the other 20% is put in the testing set and is used to check the performance of the model on unseen data.

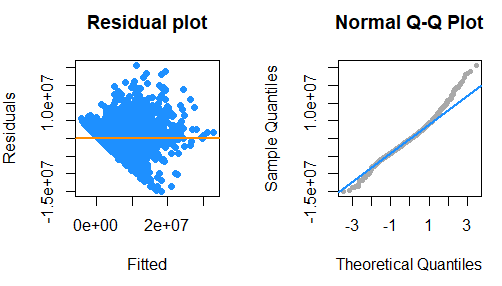
Next, we created a model with just interaction terms. This model had an R-squared of 0.7097 and an Adjusted R-squared of 0.6529.

To see if any of the polynomial terms are significant or not, we ran an anova test between a model with just interaction terms and a model with both interaction and polynomial terms. The p-value returned from the anova test is very small, which means that at least one of the polynomial terms is statistically significant.

To check which of the terms are actually significant, stepwise selection was performed on the model with BIC as the criterion for model selection. Since BIC penalises more complex models, BIC will be less likely to overfit the data then AIC.

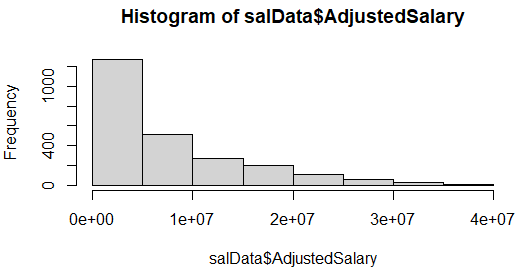
After stepwise selection was performed, a model with 17 predictors (linear, interaction and polynomial) was chosen as the best model. The model has a R-squared of 0.5941 and an Adjusted R-squared of 0.5905.

We then checked to see if the model diagnostics hold on the model that we obtained. The first and second assumptions, ie., linearity and equal variance, residual plots were made with the fitted values on the x-axis and the residuals on the y-axis. To check whether the normality assumption holds, a QQ plot was created.



From the residual plot, it is pretty evident that the linearity assumption does not hold as the mean of the values in each section of the graph does not appear to be 0. Also from the residual plot it can be seen that the Equal Variance assumption does not hold. From the QQ plot, it is pretty evident that the normality assumption does not hold either.

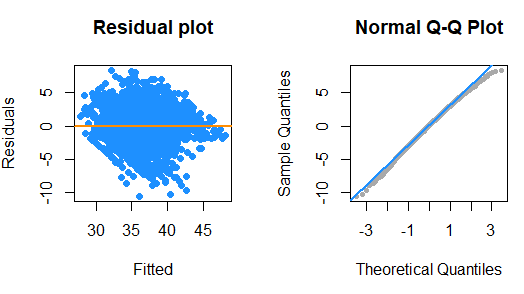
To confirm our findings that none of the assumptions hold true, we performed a BP test to test for equal variance and a Shapiro test to check normality. The p-value from both the tests are very small.



The histogram of the “AdjustedSalary” shows that it is highly left skewed, this means that some sort of transformation may be required.

We transformed the response variable using Box Cox transformation. The Box Cox method found that the best lambda value is 0.1. When building a model with the box cox transformation, it was found that the model had a R-squared of 0.5788 and an Adjusted R-squared of 0.5751.

We then checked the model diagnostics on the new model.



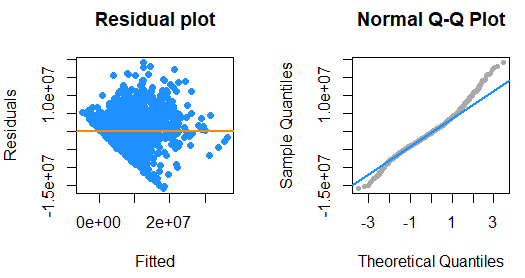
From looking at both plots, it appears none of the assumptions hold.

A BP test was then used to confirm our findings. The p-value from the BP test is very low, confirming that equal variance assumption does not hold.

From the QQ plot, it seems like normality might hold but we couldn’t tell for sure. To get a clearer picture of whether or not normality holds, a Shapiro test was performed. The p-value from the Shapiro test is very small and hence normality does not hold, however both the p-value and the QQ plot after Box Cox is better than the original model.

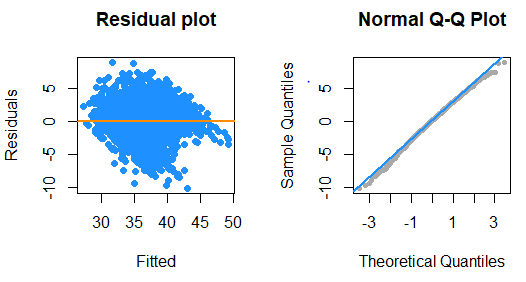
Since BIC places a higher penalty on complex models, it might underfit the data. Thus we decided to use a stepwise selection model with AIC as the criterion. We made both models to compare the results and see which model is better. Stepwise selection resulted in a model with around 40 predictors. The AIC model had an R-squared 0.619 and Adjusted R-squared: 0.6102

We then checked the model diagnostics to see if any of the assumptions were true. We plotted residual and QQ plots:



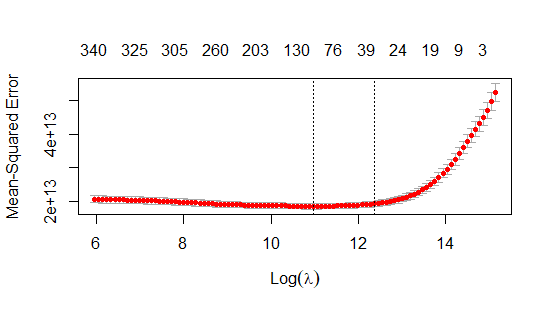
From the plots above, it does not seem like any of the assumptions hold. To reaffirm our belief, we performed a BP test and a Shapiro test. The p-value for both the tests were quite small, confirming our initial hypothesis

To try and get the equal variance assumption and normality assumption to hold, we transformed the AIC model using box cox. With the ideal value of lambda equal to 0.1, we transformed the response variable and checked the model diagnostics again.



Again, from the residual plot it looks like the linearity assumption might hold but not the equal variance assumption. And from the QQ plot, it seems like the normality assumption might hold. However, when we performed a BP test and a Shapiro test, it was seen that neither equal variance nor normality assumptions hold, as the p-values from both tests were very small. The model after the box cox transformation has an R-squared of 0.6042 and an Adjusted R-squared of 0.595.

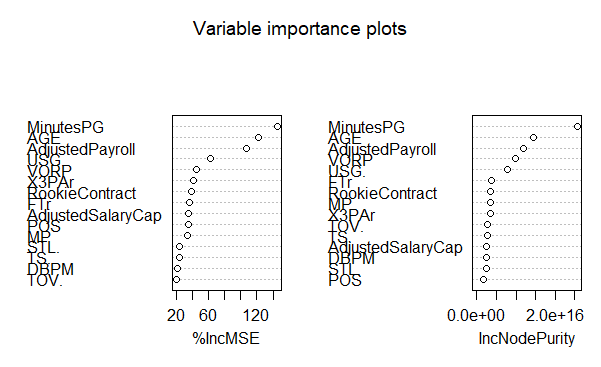
The third model we tried was a model with all interaction terms and polynomial terms. We did not use a stepwise selection process to get rid of any of these terms, instead we chose to perform LASSO regression. One benefit to using LASSO regression is that LASSO regression will perform feature selection for us, which will help prevent overfitting. We performed a five fold cross validation to see which value of lambda was the best. From the plot it was seen that the best log(lambda) is at 11.6. The plot below indicates the best value of lambda.



Using this value of lambda, we fit another LASSO model and evaluated the results. The RMSE on the best fold was $4,379,240. The RMSE on the test set was $4,214,269.

**Random Forest Model:**

When we found that we could not create a model that would pass the model assumptions, we wanted to pivot and try a different model. We chose to use a random forest model because random forests are quite flexible, and do not require any of the linear regression assumptions to be met. To find the predictor variables to use in our random forest, we ran a stepwise AIC and a stepwise BIC feature selection on all of the original predictors in the data set. The number of predictors in the AIC model was 18, while the number of predictors in the BIC model was 16. Then we ran an ANOVA test on both models, and determined that at least one of the predictors that was not included in the AIC model was significant. Thus, we preferred to use the AIC model. When checking the VIF values in the AIC model, we found that no predictors had a VIF greater than 6. This implied that multicollinearity would not significantly affect our random forest. We used 2000 trees for our random forest, but we found that the performance does not improve much after around 500 trees. For the random forest function that we used, a value of mtry needed to be specified. This mtry value specifies the number of predictor variables that each tree randomly samples every time the tree splits. The value of mtry was tuned using the out of bag error, and we found the optimal value for mtry was 10. We then fit the random forest model with mtry = 10 and plot the variable importance plots below. From the plots, it is clear to see that the most important predictor variables are Minutes Per Game and Age, while less important variables are True Shooting Percentage, Steal Percentage, and Turnover Percentage. This intuitively makes sense, as elite players will play a lot of minutes in a game, and thus will demand a larger salary. After evaluating the Random forest on the test set, we found the RMSE was $4,322,731 and the R-squared value was .66. This is a similar result to the linear regression models that were created earlier in this project.



**Results:**

The following results table summarizes the results for all models that were created. We calculated the RMSE when using only the mean of adjusted salary as a predictor, for both the test set and the training set. The RMSE for the test set was $7,448,594, while the train set RMSE was $7,191,733. This would indicate why our model performed better in the test set then in the training set. Given the RMSE using only the mean as a predictor for adjusted salary was $7.4 million, we can say all of our models explained a significant amount of variance of adjusted salary. However, because the variance is not constant for different values of adjusted salary, we don’t think our models can give reliable predictions. Given the results of our models, we can also determine that none of the models show that we have an overfitting issue.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Type | Best Value of Lambda(log) | Adjusted R Squared | RMSE  Training Set | RMSE Test Set |
| BoxCox Transformed AIC Model | N/A | .595 | $4,576,431 | $4,430,128 |
| BoxCox Transformed BIC Model | N/A | .57 | $4,593,891 | $4,482,405 |
| Lasso Regression - Full Model | 11.5 |  | $4,379,240 | $4,214,269 |
| Random Forest | N/A | .63 | $4,353,709 | $4,294,196 |

**Limitations:**

Throughout this project, we realized that more data could have potentially helped improve the quality of our model. In the interest of time and not being overwhelmed with numerous predictor variables, we did not include per game statistics in our model. Per game statistics are used to quantify how well a player performs for the average game. Examples of per game statistics include points per game, rebounds per game, and many other per game values. While we have shown that advanced statistics are certainly useful for predicting a player's salary, it is possible that the addition of per-game stats could have helped to further improve the quality of our model. This will be a further question that we hope to address in the future, once we acquire more data.

One limitation that we cannot rectify with our current data is the supply and demand issue present in the free agent market. For instance, suppose in the free agency window, ten teams decide that signing a small forward would make their team a contender. If there are a limited number of quality small forwards who are free agents in the current season, a bidding war over the available players would ensue, and usually a team would end up “overpaying” for the particular player. In this scenario, the player would be paid much more than their performance would indicate they are worth. This would result in added variance in the players’ salary that we cannot explain with the current data we have.

One further limitation of the dataset was that we did not know when the player signed his last contract. Because the risk of being injured while playing basketball is very high, many players desire security in the form of long term contracts. Thus, the player might desire a contract that is five years in length, instead of a shorter contract. A team might sign a player to a five year deal only for the player to become much worse during the contract, either due to an injury or due to the player aging.

Lastly, the way owners and managers in the NBA choose to structure contracts can add variability that our model cannot explain. Many NBA contracts that players’ sign are “back-loaded” or “front-loaded”. A front-loaded contract occurs when a team pays a large percentage of the players total contract in the early years of his deal, while a back-loaded contract refers to the inverse scenario. This may occur when a team wants to sign an elite player for a large salary, but doing so would cause them to go over the salary cap. Thus, the team would offer the player the same salary, but offer to pay a larger percentage of the salary in the later years of the player's contract. This would allow the team to sign the player they desire while remaining under the salary cap. This phenomenon means that in many situations, the player’s yearly salary would significantly depend on how much salary cap space the team that signed the player has available.

**Appendix:**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Explanation** | **Type of variable** |
| Position | What position a player plays for his team | Categorical. Possible values include “PG”, “SG”, “SF”, “PF”, “C” |
| AGE | How old the player is in the current season | Numeric (non-negative) |
| G | How many Games the player participated in. | Numeric (non-negative) |
| MP | The total minutes the player played for the duration of the season | Numeric (non-negative) |
| PER | A metric that takes in everything a player does well and subtracts everything he does poorly. Average PER for the league is 15, an exceptional PER would be around 27-30. | Numeric (non-negative) |
| X3PAr | The percentage of field goals a player takes that are 3 pointers | Numeric(ranges from 0 to 1) |
| FTr | Total Free throws a player attempts divided by the total Field goals the player attempts. | Numeric (ranges from 0 to 1, can sometimes exceed 1 in rare cases) |
| TRB. | A percentage of the total rebounds a player had while he was on the floor | Numeric (ranges from 0 to 100) |
| AST. | An estimate of the percentage of teammate field goals a player assisted while he was on the floor. | Numeric (ranges from 0 to 100) |
| STL. | The percentage of times the player stole the ball from the opposing team | Numeric (ranges from 0 to 100) |
| BLK. | An estimate of the percentage of two point field goals the player blocked while on the floor | Numeric (ranges from 0 to 100) |
| TOV. | An estimate of the turnovers a player committed per 100 plays | Numeric (ranges from 0 to 100) |
| USG. | The percentage of plays that the team ran for the player. | Numeric (ranges from 0 to 100) |
| OBPM | A box score estimate of the offensive points per 100 possessions a player contributed above a league-average player, translated to an average team. | Numeric (ranges from -6 to 11) |
| DBPM | A box score estimate of the defensive points per 100 possessions a player contributed above a league-average player, translated to an average team. | Numeric (ranges from -3.3 to 5.5) |
| VORP | A box score estimate of the points per 100 team possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season. | Numeric (ranges from -2 to 9.9) |
| SEASON | The current NBA season | Numeric (ranges from 2012-2020) |
| AdjustedSalary | How much money the player made in the current season, adjusted for inflation.  This is our response variable | Numeric (ranges from 3 million to 38 million |
| AdjustedPayroll | The total amount of money the team spent on all of its players | Numeric (ranges from 53 million to 155 million) |
| MinutesPG | The amount of minutes a player played per game. | Numeric (ranges from 12 to 39) |
| TS. | A measure of shooting efficiency that takes into account Free throws, Field Goals, and Three Point Field Goals | Numeric (ranges from .36 to .74) |
| AdjustedSalaryCap | The max amount of salary a team can pay all of its players before going into penalties. | Numeric (ranges from 65 million to 111 million) |

References:

Basketball Reference Data: [https://www.basketball-reference.com/teams/BOS/2020.html](about:blank)

Salary Data: [https://hoopshype.com/salaries/](about:blank)

Salary Cap Data: [https://basketball.realgm.com/nba/info/salary\_cap](about:blank)