ECON3389 Machine Learning in Economics

Course Project Report

Predicting NBA Player Salary

Group Number: XX

Group Members: Kenny Dao, Will Forman, Ian Leissner, Chris Farese

Contents

Introduction

The goal of our research is to accurately predict individual player salary based on our independent variables. Individual player salary is our main outcome variable of interest. Some other key variables that may be needed are height, weight, standing vertical jump, etc. Height is highly correlated with position so this would be important for the algorithm to take into account as certain positions may have higher salaries. Weight is correlated with height so, expectedly, it can also be an important factor in looking at salary. Standing vertical jump is also an indicator for athleticism, which may influence salary. We picked this goal because it represents the challenge that NBA teams face when trying to accurately evaluate the worth of their players. By using a machine learning algorithm that can accurately predict player salary, NBA teams would be able to save on contract negotiation costs. Obviously, a key inference is that we expect players with higher positive statistics (PPG, Rebounds, Assists, etc.) to be paid a higher salary as they will be contributing more to the team with their performance.

Data Summary

Here we will discuss some key variables we included in our predictive model. These include: age, points per game, minutes played per game,

One key variable that influences salary is the age of the NBA player. Age is indicative of experience. Generally, experience is correlated with salary. Although in a physically demanding occupation like playing in the NBA, we suspected that salary would be more correlated with the peak physical condition/age of the player. Given that this is true, it is important to account for age as it is significant in determining a player's salary from an NBA team's perspective. They would not want to overpay for an older player just for his experience. Similarly, they would not want to overpay for a younger player just because of his rising physical condition.

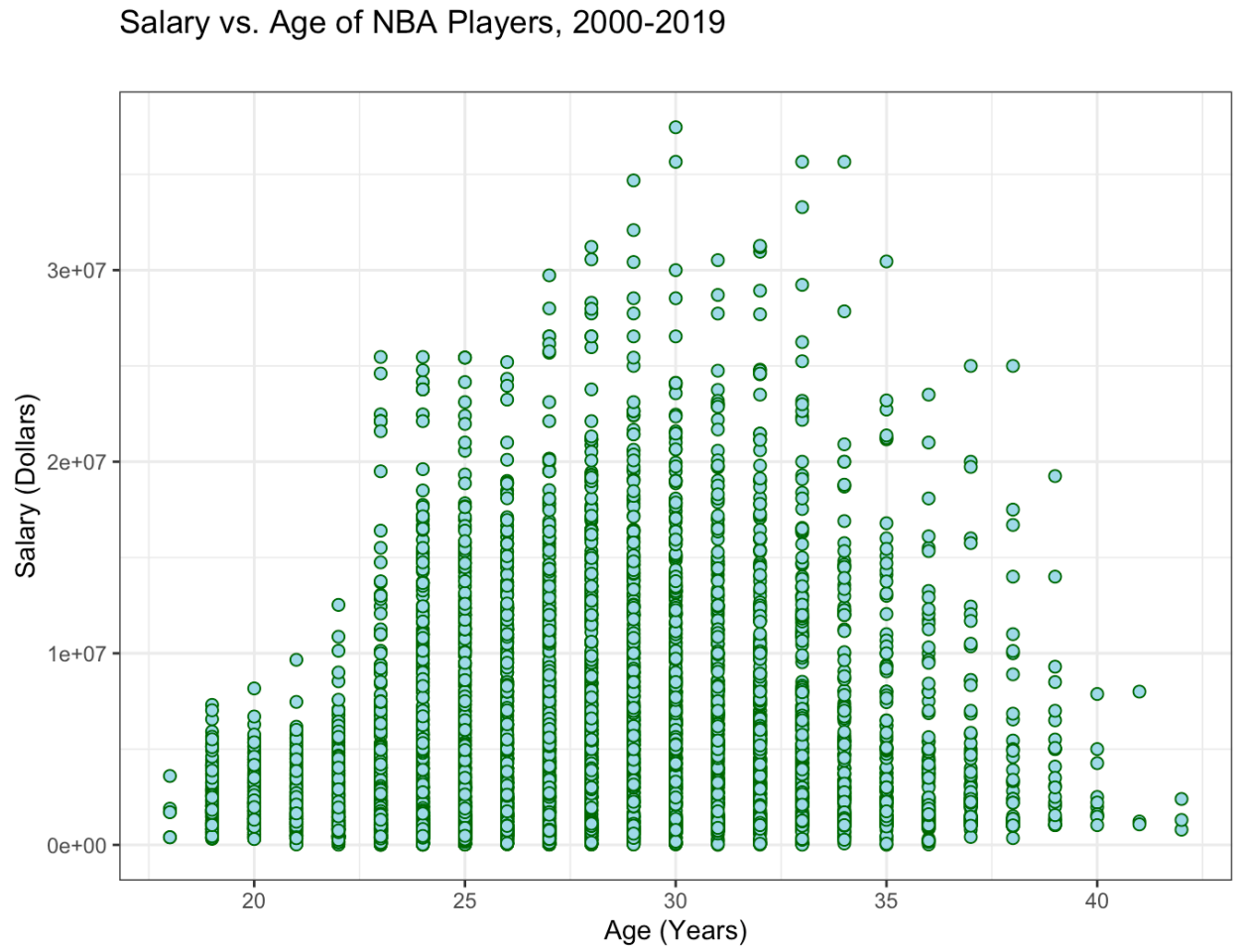
Another key variable that influences salary is points per game. Points per game measures the on-court production of a player. It is also a useful measure to determine how much of an impact a single player has on the team's overall performance. Thus, a team would only want to pay a player more if they are producing more for the team.

Another key variable that influences salary is minutes played per game. Coaches tend to keep the highest producing players on the court; thus, they tend to have the highest minutes per game. If the player is not having a positive effect on the team's overall performance, then the coach will remove them from the game.

Key Variables (Numerical Summary using StarGazer): See Appendix

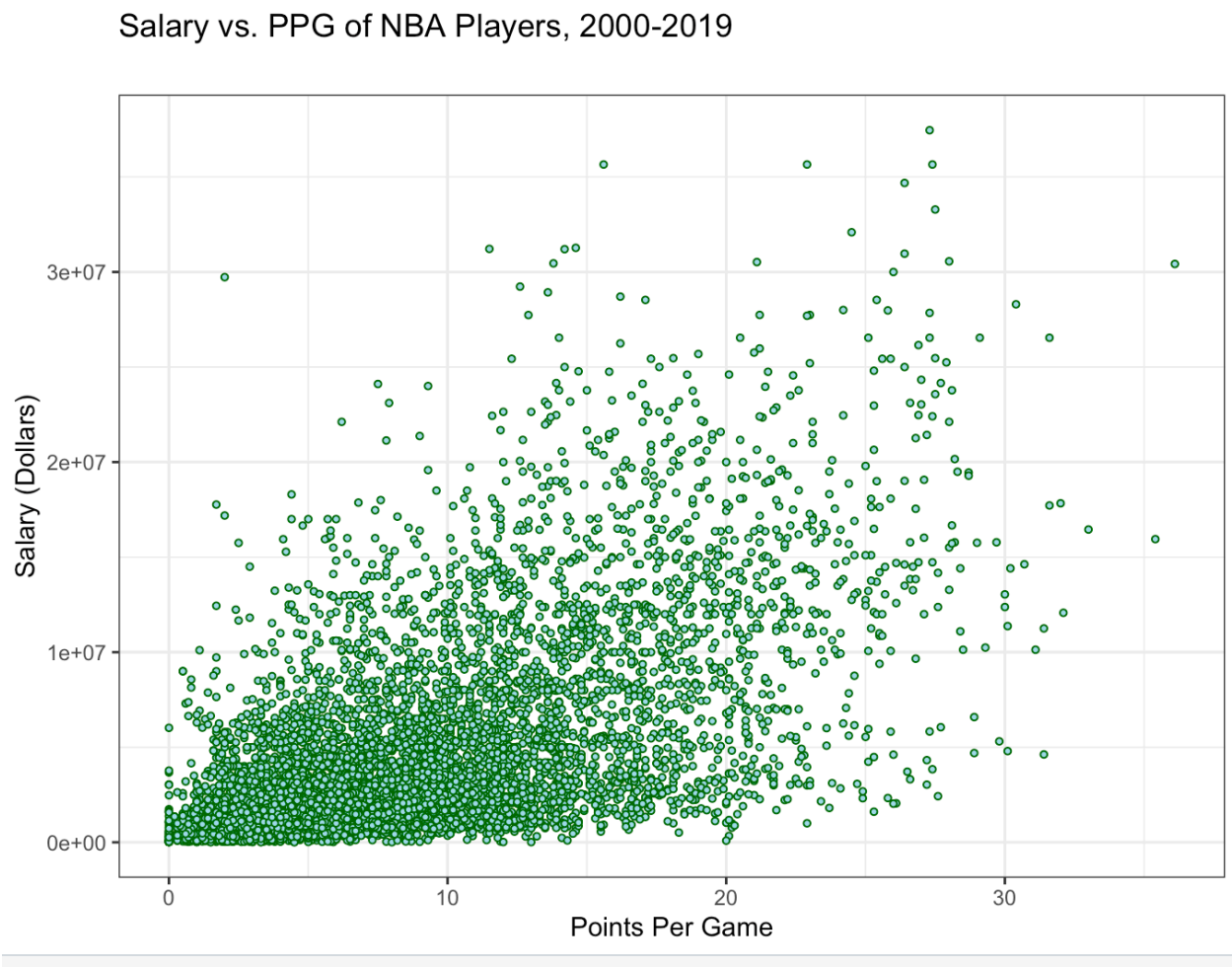
Preliminary Analysis

When conducting a preliminary analysis on age, we found younger players (under 22 years old) and older players (over 37 years old) tend to make the least amount. Younger players tend to be paid a lower salary due to the lack of established skills in addition to their inexperience. On the other hand, older players tend to be at the tail-end of their career; their agility, speed, and mobility decline with age. Thus it makes sense the players with the highest salaries are ages 22-37. The data followed a normal distribution.

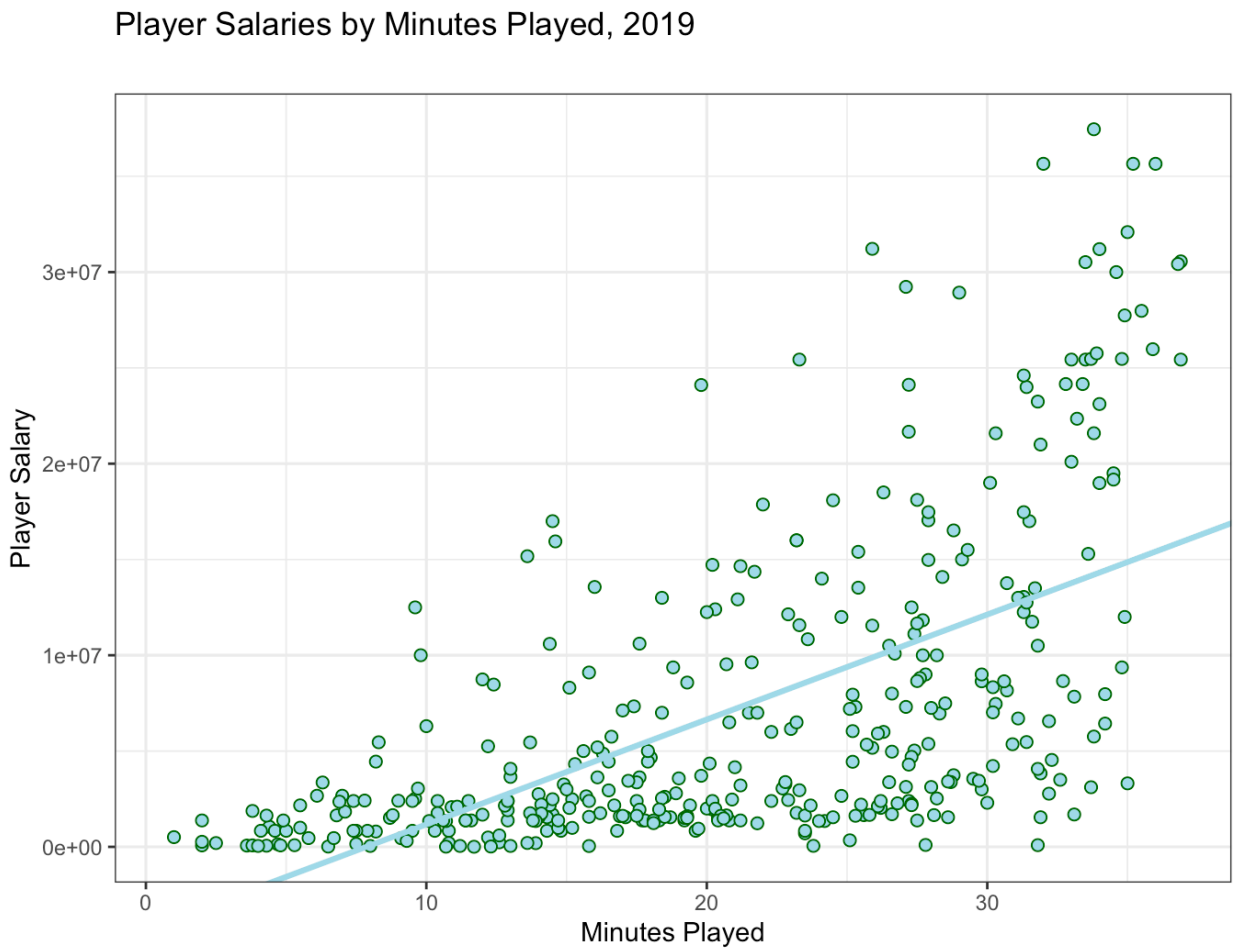


When conducting a preliminary analysis on points per game, as expected, we found players that average more points per game tend to have a higher salary. One intriguing finding was that some players may have averaged less points than other players, but still had the same salary. For example, a player who averaged 10 PPG made a similar salary as another player averaging 20 PPG. When looking further at this, we typically found that the player with a lower average PPG compensated for the lack of production by averaging higher with other meaningful statistics such as assists per game, rebounds per game, blocks per game. Additionally, there were likely

valuable intangible benefits of having the player on the team such as their presence increasing team chemistry, knowledge about other players' playing styles, etc. We found PPG to be positively correlated with salary.



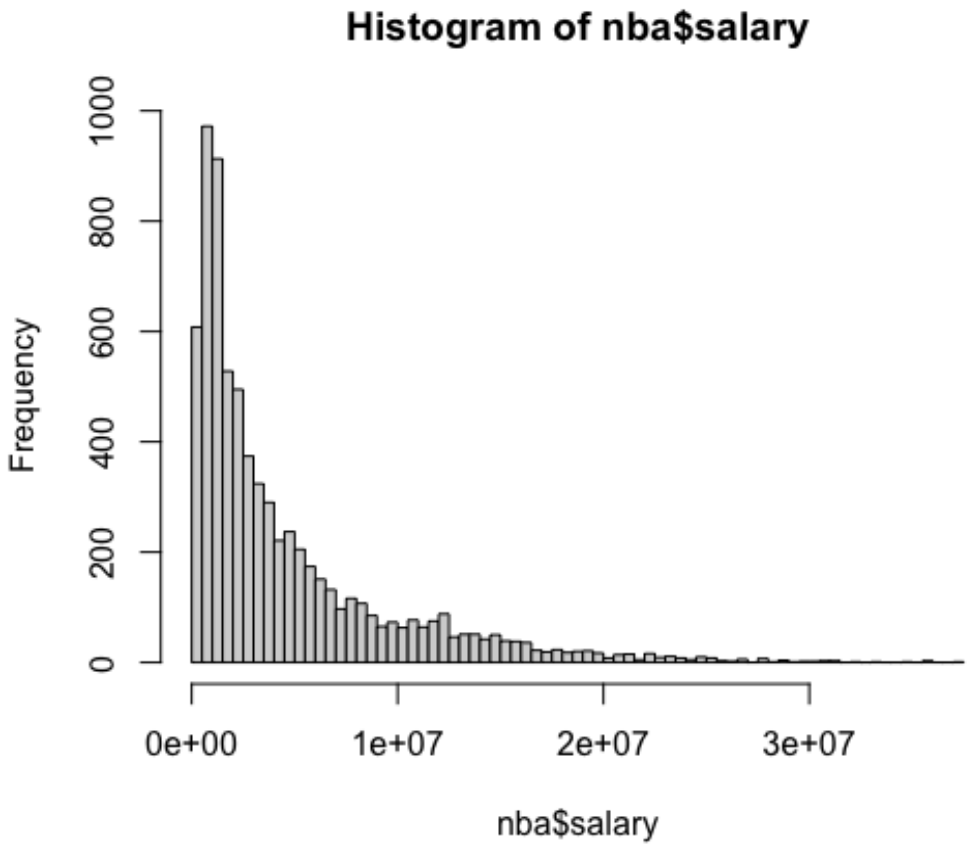
When conducting a preliminary analysis on minutes played per game, we found that higher average minutes played per game is correlated with a higher salary. One important aspect to note is that as the number of minutes increases, the variance between the player salary increases. When looking at this issue further, we determined that certain players serve a different purpose on the court; a player may remain the court for their defensive skills while another player may remain on the court for their scoring abilities. For example, a coach may decide to keep a well-known defensive player like Patrick Beverly on the court for the entire game in order to enhance his team's overall defense. A coach may also keep a scoring player like Kevin Durant on the court in order to enhance his team's offense. Thus, both players remain on the court for the same period of time, yet their salary will be different because the overall team's needs are different; if the team needs scoring rather than defense, then they will be willing to pay a higher premium for a pure score over a defender, and vice versa.



We conducted a summary of the values of salary that appear in the dataset, and found the data to be very right-skewed; the mean salary is $4,653,677 and the median salary only $2,626,473. The difference between these two values implies the existence of large outliers in the dataset for salaries.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4608	1110000	2626473	4653677	6200000	37457154

If one looks at a histogram of salary in the dataset, this inference is confirmed.



Additionally, these outliers result in a very large standard deviation in the data: $5,190,083

Inference model

In order to evaluate how NBA salary figures have been decided upon historically, we considered what variables might be the most important. While there are countless factors that contribute to the "value" of a player, they can all be broken down into two groups: the player's current expected value, and the potential of expected value. We therefore decided that three variables must be included in the inference model: points per game (the most direct impact a player has on the outcome of a game), age (commonly associated with future expected value, as young players are expected to improve, while old players are expected to decline), and number of games started (which should capture a player's value as perceived by his coach). We also evaluated models with these values individually squared, to see how adjusting the weight of PPG, age, or GS would affect our inference. We also tested two more variables, three point field goal percentage and minutes played, and while we are undecided on whether to include them (details below), the model with all five variables has the lowest Adjusted R-Squared value.

```
> summary(mlr1)

Call:
lm(formula = nba$salary.pct ~ pts + age + gs, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21504 -0.03544 -0.00592  0.02402  0.51967

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept) -0.12469472  0.00460307 -27.089  < 2e-16 ***
pts          0.00732164  0.00017525  41.778  < 2e-16 ***
age          0.00499286  0.00016613  30.053  < 2e-16 ***
gs           0.00026532  0.00003537   7.502 7.14e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05774 on 6461 degrees of freedom
  (806 observations deleted due to missingness)
Multiple R-squared:  0.4754,    Adjusted R-squared:  0.4751
F-statistic:  1951 on 3 and 6461 DF,  p-value: < 2.2e-16
```

MLR1: This is our most basic model, a multiple regression including only points per game (pts), age, and games started (gs).

```
> summary(mlr2)

Call:
lm(formula = nba$salary.pct ~ pts + age.sq + gs, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22053 -0.03512 -0.00573  0.02344  0.52001

Coefficients:
               Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.054856753  0.002535740 -21.633  < 2e-16 ***
pts          0.007363266  0.000176272  41.772  < 2e-16 ***
age.sq       0.000086136  0.000002994  28.767  < 2e-16 ***
gs           0.000266162  0.000035555   7.486 8.05e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05804 on 6461 degrees of freedom
  (806 observations deleted due to missingness)
Multiple R-squared:  0.4699,    Adjusted R-squared:  0.4697
F-statistic:  1909 on 3 and 6461 DF,  p-value: < 2.2e-16
```

MLR2: This is the first of three multiple regressions with one factor replaced with its square. age.sq has a slightly lower t value than age (from MLR1), but they are both significant and positive.

```
> summary(mlr3)

Call:
lm(formula = nba$salary.pct ~ pts.sq + age + gs, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22383 -0.03418 -0.00822  0.02302  0.52087

Coefficients:
               Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.097032012  0.004537533  -21.38   <2e-16 ***
pts.sq       0.000259962  0.000006444   40.34   <2e-16 ***
age          0.004974951  0.000167331   29.73   <2e-16 ***
gs           0.000513331  0.000031846   16.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05816 on 6461 degrees of freedom
  (806 observations deleted due to missingness)
Multiple R-squared:  0.4677,    Adjusted R-squared:  0.4675
F-statistic:  1892 on 3 and 6461 DF,  p-value: < 2.2e-16
```

MLR3: This is the second of three multiple regressions with one factor replaced with its square. pts.sq has a slightly lower t value than pts (from MLR1), but they are both significant and positive.

```
> summary(mlr4)

Call:
lm(formula = nba$salary.pct ~ pts + age + gs.sq, data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21384 -0.03502 -0.00623  0.02404  0.51814

Coefficients:
                Estimate    Std. Error t value Pr(>|t|)
(Intercept) -0.1236707896  0.0046214983 -26.760  < 2e-16 ***
pts          0.0074482392  0.0001673052  44.519  < 2e-16 ***
age          0.0050005698  0.0001661683  30.093  < 2e-16 ***
gs.sq        0.0000030798  0.0000004318   7.132  1.1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05776 on 6461 degrees of freedom
  (806 observations deleted due to missingness)
Multiple R-squared:  0.4749,    Adjusted R-squared:  0.4747
F-statistic:  1948 on 3 and 6461 DF,  p-value: < 2.2e-16
```

MLR4: This is the last of three multiple regressions with one factor replaced with its square. gs.sq has a slightly lower t value than gs (from MLR1), but they are both significant and positive.

```
> summary(mlr.fg3.mp)

Call:
lm(formula = nba$salary.pct ~ pts + age + fg3.pct + gs + mp,
    data = nba)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22880 -0.03548 -0.00523  0.02593  0.52128

Coefficients:
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -0.12537661  0.00525409 -23.863   < 2e-16 ***
pts          0.00805868  0.00027990  28.791   < 2e-16 ***
age          0.00530894  0.00018101  29.329   < 2e-16 ***
fg3.pct     -0.05158273  0.00495483 -10.411   < 2e-16 ***
gs           0.00020663  0.00004553   4.538 0.0000058 ***
mp          -0.00004534  0.00021533  -0.211     0.833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05717 on 5522 degrees of freedom
  (1743 observations deleted due to missingness)
Multiple R-squared:  0.498,     Adjusted R-squared:  0.4975
F-statistic:  1096 on 5 and 5522 DF,  p-value: < 2.2e-16
```

Here is the final inference model, mlr.fg3.mp. We chose to experiment with these two variables because they are two commonly valued traits in a player signing a contract. Higher minutes played means a team will get more value out of a player who is on the court during a game longer than another player. 3 point field goal percentage is a more recently (pioneered by Stephen Curry in the past decade) valued statistic, with many NBA superstars earning massive contracts largely because of their accuracy on deep shots. While this model does have the highest Adjusted R Squared value, the value of these factors is dubious. Minutes played is not statistically significant, and even causes “games started” to drop in significance, presumably because players who start more games typically play more minutes per game, leading to multicollinearity. 3pt field goal percentage is even worse for the model, somehow achieving extreme statistical significance while having a negative coefficient, something that most NBA experts would disagree with.

One final consideration with these models involves player age. The main consideration of age into salary is as follows: young players receive a higher salary than they are “worth” today, as teams buy into their potential to improve, while older players of the same skill do not offer the same “upside”. While there is a value placed on veteran leadership, it is usually better for a player’s salary to be younger. With that being said, there is one major exception: the rookie contract. The youngest players in the league, those playing on rookie contracts, are being paid significantly less on average than other players, because rookie draft contracts are capped at much lower figures than typical veteran contracts. This is an interesting question that could be explored further later in the project.

Prediction Model

The first model that we are using is a OLS based model using the Best Subset Selection method. We found that this method was computationally inefficient as it required a large amount of time to conduct subset selection on our large dataset. Because of this inefficiency, we were required to truncated the number of independent variables used by the model to 5. This allowed us to generate a subset that could be used to train our predictive model. The results of our first regression can be seen in Graph 1.The initial evaluation of the model was based on the adjusted R^2. Our Best Subset Selection method achieved an adjusted R^2 value of .5667. This indicates that our model is able to explain roughly 56.67% of the variance of salary in the training dataset.

One advantage of the first OLS model is of high interpretability. The limited number of independent variables allows us to easily interpret the results of the regression such that each one has a statistically significant effect on salary. This is advantageous for managers who are seeking to understand which variables effect the player salary. Moreover, the F-stat of our model is 1359 which indicates that we can reject the null hypothesis that the independent variables of our model are not jointly significant (ie. that the joint effect of our independent variables on salary is indeed statistically significant).

The second model we build also relied on an OLS based regression model. However, this model utilized the Forward Selection method instead of the Best Subset Selection method. We chose to use this technique because of its simplicity and low computational cost. The one drawback to using this method is that in the process of creating the subset, over a third of the observations were lost due to R's matrix size limit of 2^31 - 1 elements. This could be interpreted as lowering the ability of our model to accurately predict the player salary. However, because we were still left with over 4000 observations, we concluded that our estimated coefficients would likely still be accurate enough to provide a robust predictive model.

The second model also suffers from high dimensionality. To perform Forward Selection, the model must first interact each independent variable. This results in a dimensionality that is over n^2 of the original dataset. The total number of dimensions generated by our model is 1850 (compared to 33 for model 1). The high dimensionality therefore results in an Adjusted R^2 value of .1559 which can be interpreted as our model being able to explain roughly 15.59% of the variance in salary. This low value can be explained by the fact that regressions struggle to estimate coefficients when there is high dimensionality. Moreover, the generated model has a reduced degrees of freedom and a statistically insignificant F-stat of 1.978 because of the large number of independent variables included in the model. Our low F-stat indicates that the independent variables of our second model are not jointly statistically significant. This leads us to conclude that this task would likely be better accomplished using a more complex model such as a Decision Tree or Neural Network.

Conclusion

The results of our inference and prediction model both present interesting interpretations. We made 5 different inference models we thought would be most effective. All had adjusted R-squared values between 0.45 and 0.5, proving to be substantially effective at predicting a player's share of the team's total salaries based on the supplied metrics. The model that we found to be most effective, measuring in with an adjusted R-squared value of .4975 used the metrics for points per game, age, 3 point percentage, games started, and minutes player per game. All included metrics were significant at the 0.001% level.

For our prediction model, we found that the OLS model had an MSE that was far lower than the forward model, clocking in at 9,109,623,267,724 and 32,528,332,247,932 respectively. The Forward model's adjusted R-squared value suffered from the high dimensionality, at .1436. The Best Subset Selection method, on the other hand, achieved an adjusted R-squared value of .5667. This model included age, games played, defensive rebounds per game, points per game, and the team's mean salary, and all proved to be significant at the .001% level.

Based on these results, it appears that the predictive OLS model generated by the computer was able to predict player salaries to a high degree, but the forward model was not able to efficiently supply us with accurate results.

Appendix

Some definitions from glossary: https://www.basketball-reference.com/about/glossary.html

- playerid: a unique ID for each player
- yrend: the year in which the season for each observation ends (since NBA seasons begin in October and end in June)
    - Measured in years
- teamid: unique ID for each team/year combination
- salary: dollar amount of player's salary in given season
    - Measured in dollars
- salary.lead:
    - Measured in dollars
- season: years over which the season took place
    - Measured in years
- teamfulsal: full team name
- player: full player name
- Position: abbreviated position
- age: player age
    - Measured in years

- team: abbreviated team name
- g: games
- gs: games started
- mp: minutes played per game
- fg: field goals made per game
- fga: field goal attempts per game
- fg.pct: field goal percentage
- fg3: 3 point shots made per game
- fg3a: 3 point shots attempted per game
- fg3.pct: 3 point shot percentage
- fg2: 2 point shots made per game
- fg2a: 2 point shots attempted per game
- fg2.pct: 2 point shot percentage
- efg.pct: effective field goal percentage
    - Adjusts fg.pct to account for the fact that 3pt FGs are worth 3pts, while 2pt FGs are worth 2pts
    - efg.pct = total field goals made + (0.5 * total 3pt field goals made)total field goal attempts
- ft: free throws made per game
- fta: free throw attempts per game
- ft.pct: free throw percentage
- orb: offensive rebounds per game
- drb: defensive rebounds per game
- trb: total rebounds per game
- ast: assists per game
- stl: steals per game
- blk: blocks per game
- tov: turnovers per game
- pf: personal fouls per game
- pts: points per game
- salary.team.total: total money spent on player salaries for a given team in a given year
    - Measured in dollars
- salary.team.mean: mean player salary for a given team in a given year
    - Measured in dollars
- salary.team.median: median player salary for a given team in a given year
    - Measured in dollars

Key numeric variable descriptions:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
yrend	7,271	2,010.21	5.55	2,001	2,005	2,015	2,019
salary	7,177	4,653,677.00	5,190,083.00	4,608.00	1,110,000.00	6,200,000.00	37,457,154.00
salary.lead	5,545	5,483,953.00	5,503,753.00	8,819.00	1,471,382.00	7,562,500.00	37,457,154.00
age	7,271	26.47	4.32	18	23	29	42
g	7,271	56.12	23.90	1	40	76	82
gs	7,271	28.62	29.98	0	1	58	82
mp	7,271	21.32	9.97	0.00	13.10	29.70	43.70
fg	7,271	3.24	2.21	0.00	1.50	4.60	12.20
fga	7,271	7.21	4.70	0	3.5	10.1	28
fg.pct	7,250	0.44	0.09	0.00	0.40	0.48	1.00
fg3	7,271	0.60	0.70	0.00	0.00	1.00	5.10
fg3a	7,271	1.71	1.83	0.00	0.10	2.90	13.20
fg3.pct	6,266	0.28	0.16	0.00	0.22	0.37	1.00
fg2	7,271	2.64	1.97	0.00	1.10	3.70	11.20
fg2a	7,271	5.50	3.91	0.00	2.50	7.70	23.40
fg2.pct	7,230	0.47	0.09	0.00	0.43	0.51	1.00
efg.pct	7,250	0.48	0.09	0.00	0.45	0.52	1.50
ft	7,271	1.59	1.46	0.00	0.60	2.10	9.70
fta	7,271	2.12	1.84	0.00	0.80	2.90	13.10
ft.pct	7,076	0.73	0.14	0.00	0.67	0.82	1.00
orb	7,271	1.00	0.84	0.00	0.40	1.40	5.50
drb	7,271	2.77	1.83	0.00	1.40	3.60	11.50
trb	7,271	3.77	2.53	0.00	1.90	5.00	16.30
ast	7,271	1.89	1.84	0.00	0.60	2.50	11.70
stl	7,271	0.67	0.45	0.00	0.30	0.90	2.90
blk	7,271	0.44	0.49	0	0.1	0.6	4
tov	7,271	1.23	0.80	0.00	0.60	1.70	5.70
pf	7,271	1.90	0.79	0.00	1.30	2.50	6.00
pts	7,271	8.68	6.09	0.00	4.00	12.10	36.10
salary.team.total	6,465	56,659,916.00	17,789,221.00	12,806,259.00	44,100,475.00	66,672,368.00	121,749,964.00
salary.team.mean	6,465	4,399,197.00	1,297,295.00	1,601,798.00	3,512,108.00	5,036,317.00	10,145,830.00
salary.team.median	6,465	2,819,229.00	1,191,769.00	789,170.00	2,025,220.00	3,403,820.00	8,776,306.00