COGs 109 Final Project
Golden State Spenders

Mitchell Gault, A14796194
Anthony Martinez, A13378551
Michael Mech, A15325910
Jeong Lee, A92055207

Abstract:

We aim to gather basketball statistics of NBA players in order to create a model that will predict their salaries, and subsequently compare predicted salaries vs. real-life salaries to evaluate which players are paid in/appropriately. Possible implications to follow up on are figuring out which statistical variable contributes most to salary(most variance), which 'variable' is the most overrated, and classifying players into tiers(possibly comparing our classification to already-established classifications).

Introduction:

We are trying to determine if NBA players are paid proportionally to their statistics throughout the playing season. We want to create a model to predict an NBA player's salary and compare it to their actual salary to see if they are over or underpaid. Since we are big basketball fans, the motivation behind this problem is that we seem to notice that many players are given large contracts, but do not perform well during the playing season. Similarly, there are extraordinary players that are paid very little. We want to see if we can create a model that predicts what a player should make based off of their statistics.

We will be using three data sets for this project. The data was collected from Basketball Reference. 2017-18stats.csv contains all relevant playing statistics for each player in the NBA during the 2017-2018 season. Some of the playing statistics include: points per game, rebounds, assists, blocks, steals, and more. Included in this data set is also the salary of what each player made during that season. 2018-19stats.csv contains the same thing as the previous file, but the statistics are for the 2018-2019 NBA playing season. Nba2019-20.csv contains only the salary made for each player in the NBA for the 2019-2020 season. We are not including player statistics in this data file since the season is ongoing and the player statistics are constantly changing.

Hypothesis: A players salary should be solely determined by certain statistics, therefore we predict that 95% of players are paid within these salaries. We can test for this by using our model that we create. Purely based off of statistics, players should be making salaries within certain ranges. If players have certain statistics(i.e. 25ppg, 5 rebounds, 5 assists), then they should be making a certain amount of money a year(i.e. 28 million dollars/year). We can test for how much a player should make based off their statistics by using a model that we create. We hope the model allows us to determine a set salary that depends on a player's statistics.

Another hypothesis: Points per game is the most influential predictor when determining a player's salary. We can test for the most influential predictor on a player's salary by finding which predictor has the lowest MSE. We hypothesize that points per game is the most influential because it is the most popular statistic in basketball. Fans watch the sport to see players score.

Data Characteristics:
We have collected the average statistics of a player per game in a season and matched it to their salary for that season. We have dropped the columns highlighted in red, as to avoid multicollinearity.

Data Columns:

- Player - Player Name
- ID - Abbreviated name
- Pos - Position
- Salary - Salary for that season in USD. Not adjusted for inflation
- Age - Age
- TM - Team
- G - Count of Games
- GS - Games Started
- MP - Minutes played per game
- FG - Field Goals per game
- FGA - Field Goal attempts per game
- FGper - FG/FGA
- ThP - 3 pointers per game
- ThPA - 3 pointer attempts per game
- ThPPer - FG% on 3 pointer FGAs
- TwP - 2 pointers per game
- TwPA - 2 pointer attempts per game
- TwPPer - FG% on 2 pointer FGAs
- eFGPer - Effective FG%
- FT - Free throws per game
- FTA - Free throw attempts
- FTper - 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
- PV - Personal fouls per game
- PTS - Average Points per game

The 2017-2018 data set includes 540 unique individuals, while the 2018-2019 data set includes 530 unique individuals. The data sets includes multiple copies of the same individual if they played on different teams/positions, as their relevant statistics per team is recorded separately. This adds slightly more bias, as each copy of the individual has the same salary. However, we deemed it important to keep them as separate entities as they have different records.

Methods:

Data Cleaning

From the original csvs taken from Basketball Reference, we cleaned up the dataset by removing redundancy (percentages are accounted already by other variables i.e FTA and FT already account FTper). We also changed the arabic numerals in the original dataset (i.e 3PA is ThPA) to their letter representations as its easier to handle the data in python this way.

Linear Regression

One of the model types we will be using is multiple Linear Regression; this is because the high interpretability of Linear Regression can help us isolate the truly relevant predictors. The clarity of which parameters work best together for the “best” model lets us know what combination of statistics work best in predicting a player’s salary.

We are also able to classifier variables such as position (PG/PF/SG/SF/C) to compare players by those variables.

Forward Stepwise Selection

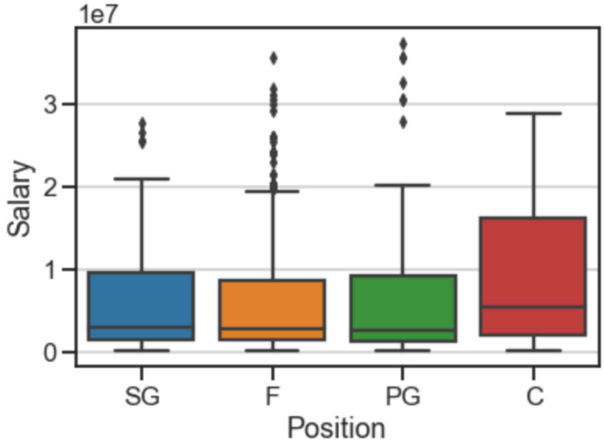
We decided to use Forward Stepwise Selection to compare the multitude of models (2^p where p = 29 as there are 29 parameters). Best Subset Selection seemed like it will be fairly computationally complex, while Forward Stepwise Selection is a lot less computationally (time) demanding.

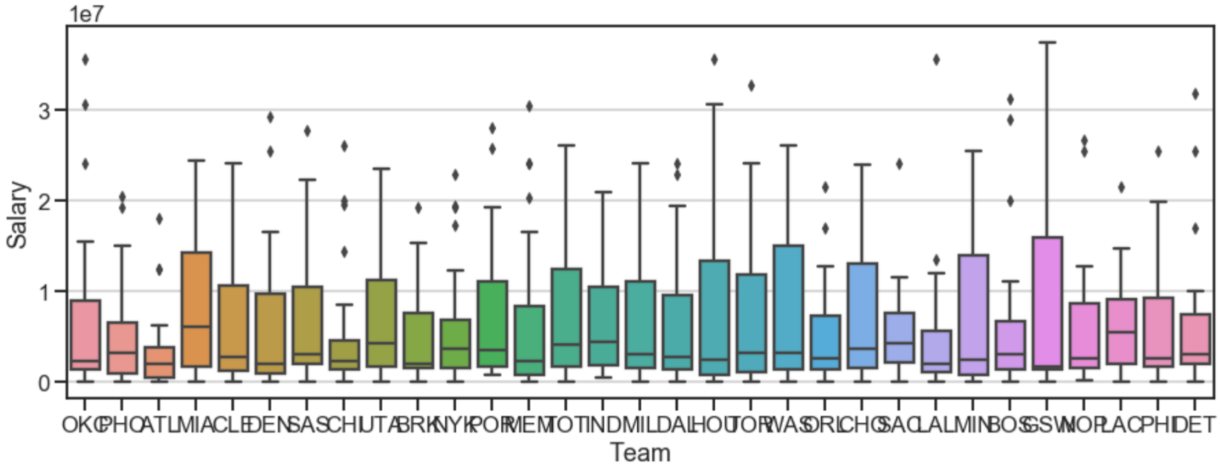
Principal Component Analysis

We used principal component analysis in the form of single-value decomposition to figure out the basketball statistics(variables) which would contribute most to the top principal components, and hence, contribute to the most variance seen within the data.

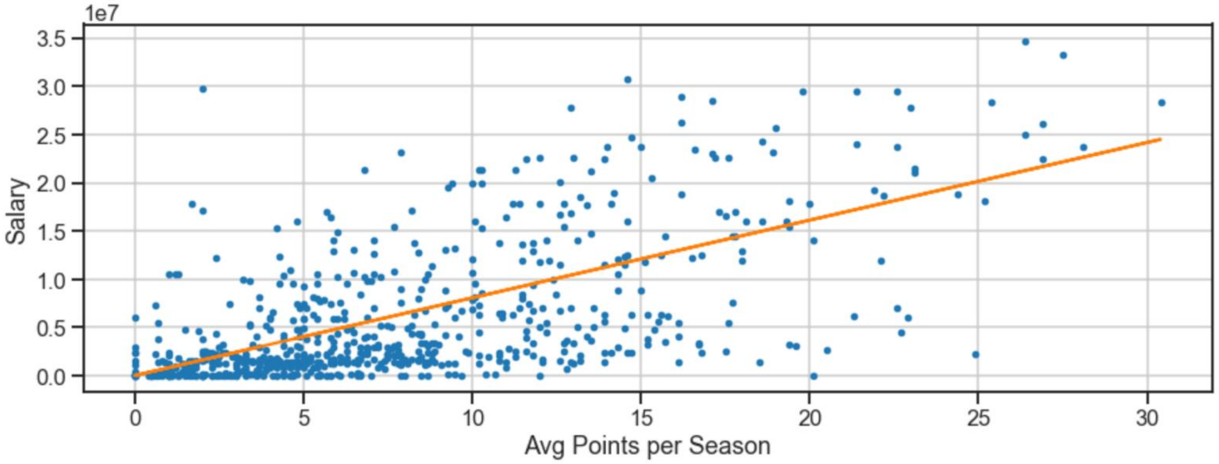
Results

Initial analysis showed that on average there is no difference in salaries across positions or teams:



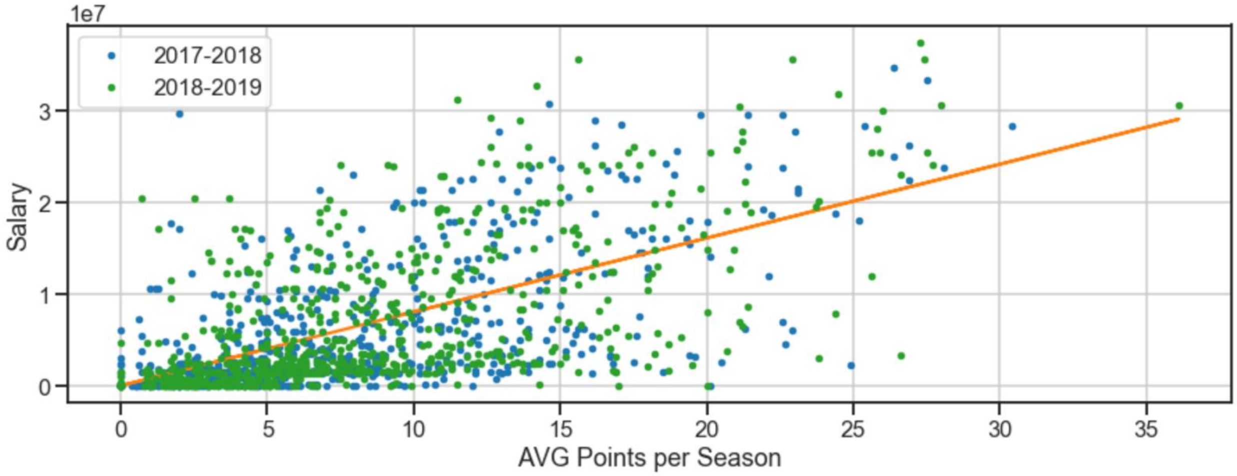


After calculating for the best singular predictor (via minimizing mean-squared-error) we found that Average Points Earned per season is the best predictor for salary. This regression has an R-squared of 0.401, and a coefficient standard deviation of 39388.45313897598.



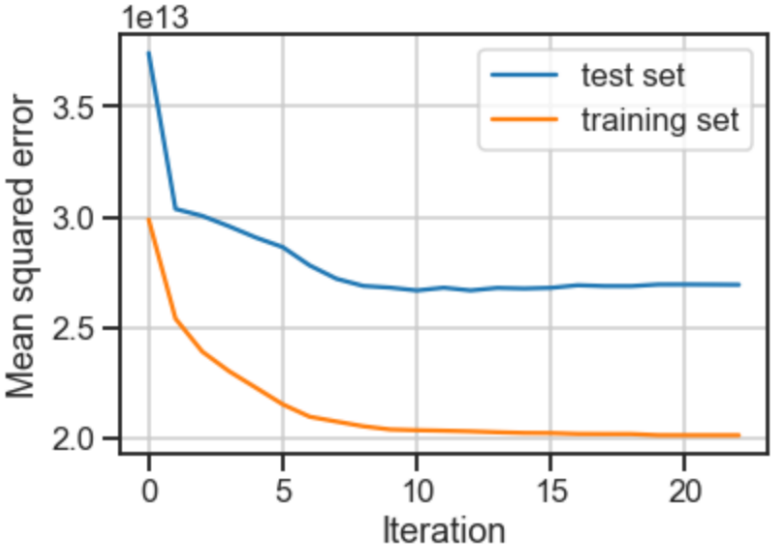
Using Forward Stepwise Selection, we decided to use the 2017-2018 statistics and salaries as the training data, and the 2018-2019 statistics as the testing data. This cross-validation should help us get an accurate model that helps predict the next season's salaries. Using Average Points Earned per season as the only indicator and both data sets, we get:

- Training MSE: 29892249963209.125
- Testing MSE: 37427635366392.055



After using Forward Stepwise Selection, we found that the best predictors for minimizing the testing set is:

Salary ~ 1 + PTS + Age + GS + Tm + Pos + G + AST + TRB + PF + FT + TWPA

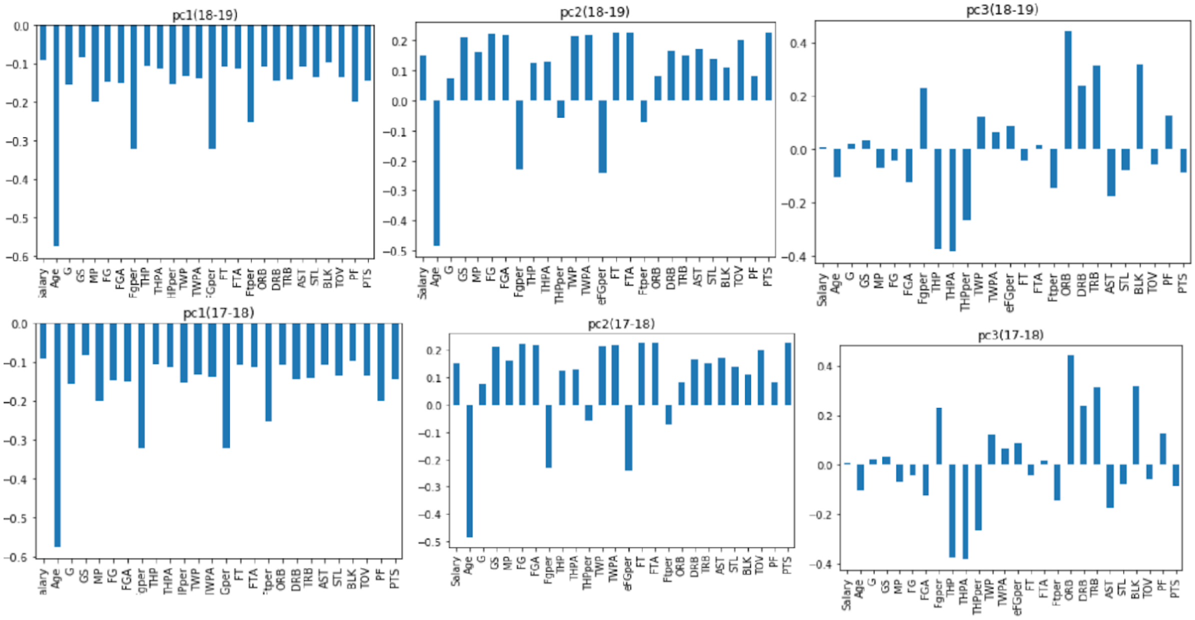


This formula results in:

- Training MSE: 20356960706108.863
- Testing MSE: 26681163196357.33
- R-Squared: 0.575

Looking at the first three principal components of both the 2017-18 and 2018-19 data, we find that the variables which contribute the most to the variance within the datasets are respectively given by age for both 1st principal components, free throws/free throws attempted and age for both 2nd principal components, and three pointers attempted and offensive rebounds for both

3rd principal components.



Conclusion

As shown from the data, only about half of the variance in player salary can be attributed to key player metrics from the previous season. Additionally, when using a 95% confidence interval on the coefficients, less than half of the data falls within that interval. Average Points scored per game is the single most important coefficient, yet still accounts for only ~40% of the variance. In summary, there are some external factors other than player performance that contribute to their salaries next season.

As for the principal component analysis, regardless of trying to predict salary, the statistic which provided most variance overall was age. It would make sense that the most stratified parameter be age, which accounts for great variability in the sense that younger players(rookies) may only contractually have certain salaries, while older players have a veteran minimum. Another statistic of great variance revealed in the principal component analysis were that of free throws, which raises the possible implication that superstar players are more likely to not only make more free throws, but have more opportunities to do so(due to usage rate, minutes played etc.),

Another thing to note is that the data weighted disproportionately leading to more bias. The vast majority of players with a lower salary get less playing time, thus pushing their statistics lower. This shifts the whole model downward, resulting in bias. Additionally, there are more players with lower salaries as teams prefer to have roster backups. The presence of rookie salaries in the league, especially for star players also may skew the salaries. Star players like Joel Embiid are paid disproportionately lower compared to other superstars (<7 mil compared to someone like Steph Curry at 35+ mil) even when their overall statistics are comparable to each other.

Some other variables that may be explored in the future is player presence (i.e. social media presence, screen time, etc.). Another future analysis could include restricting the data to only players with more than *x* playing time, so that the data can more accurately reflect the impact player statistics have on salary.