Phuc Ton Nguyen

Prediction of salary for MLB player

Guided Capstone Project Report

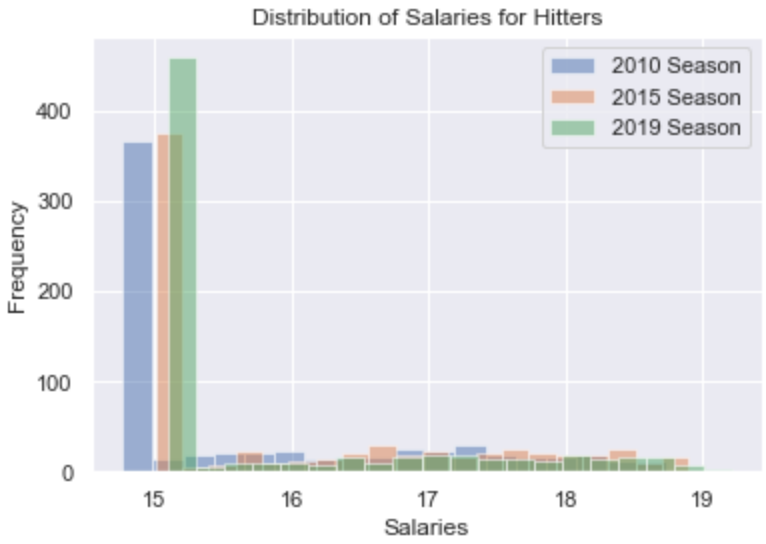
Introduction

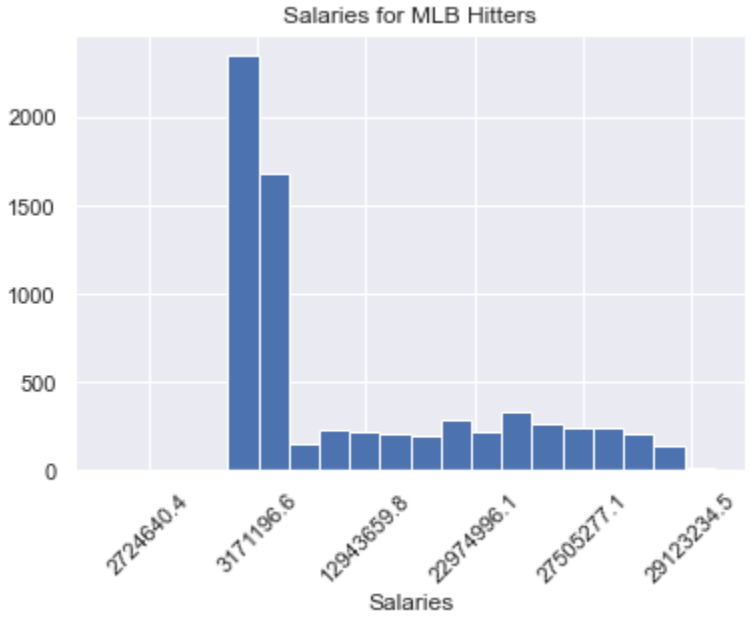
San Francisco Giants wants to buy player and they are asking the evaluation of the players on the market to buy.

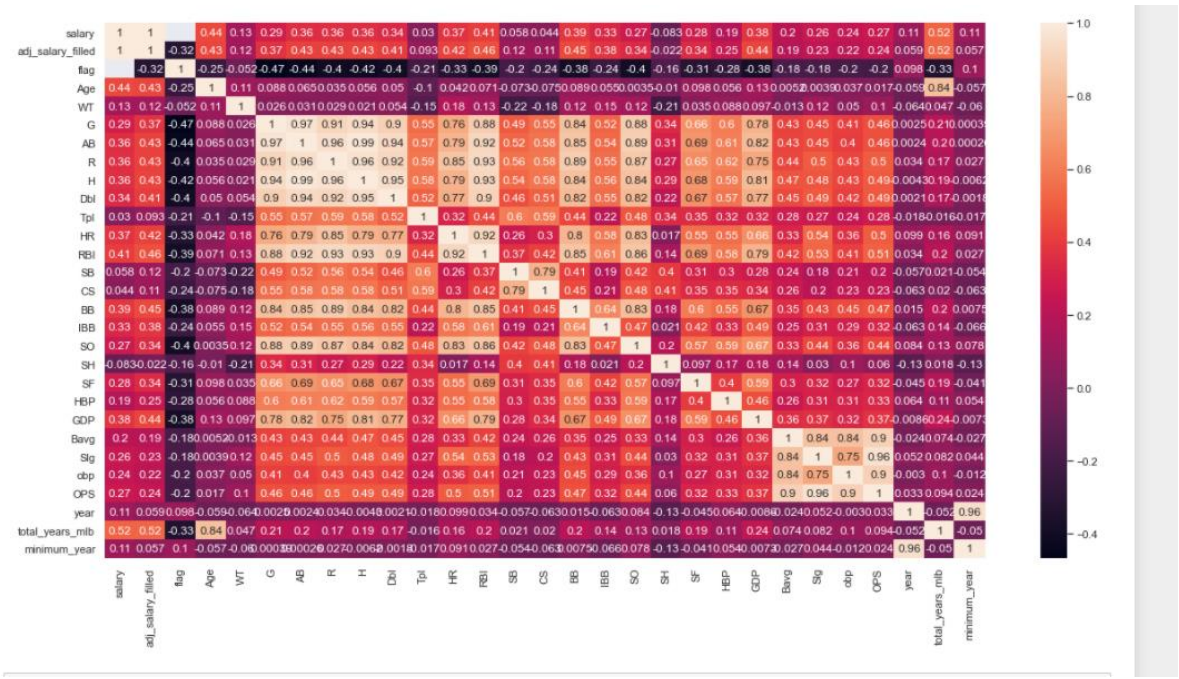
Exploratory Data Analysis

The cleaned and merged dataset contained 6928 observations (players) each with 38 features. I did univariate, bivariate and multivariate analyses along with visualizations. I also performed some hypothesis testing for more statistically rigorous statements. The followings are the summary of what I found through exploratory data analysis (EDA) and inferential statistics.

Hitter

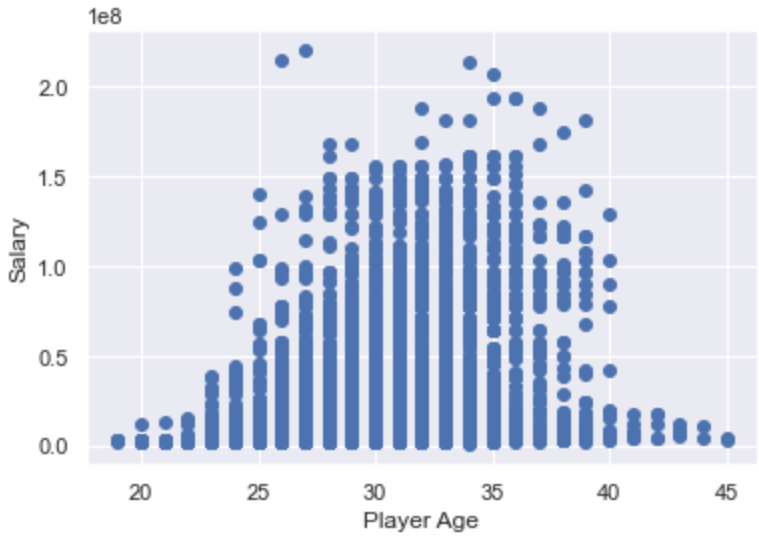

Distribution of Salaries for Hitters
2010 Season
2015 Season
2019 Season
Frequency
400
300
200
100
0
15
16
17
18
19
Salaries

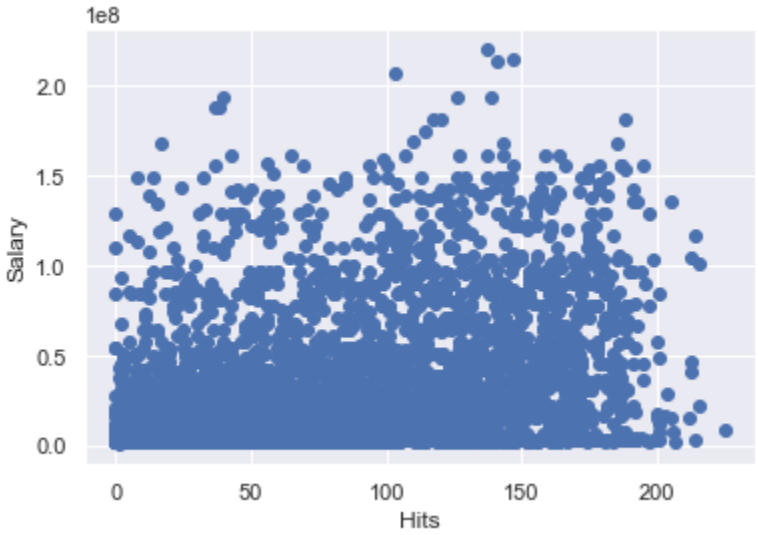

Salaries for MLB Hitters
2000
1500
1000
500
0
2724640.4
3171196.6
12943659.8
22974996.1
27505277.1
29123234.5
Salaries

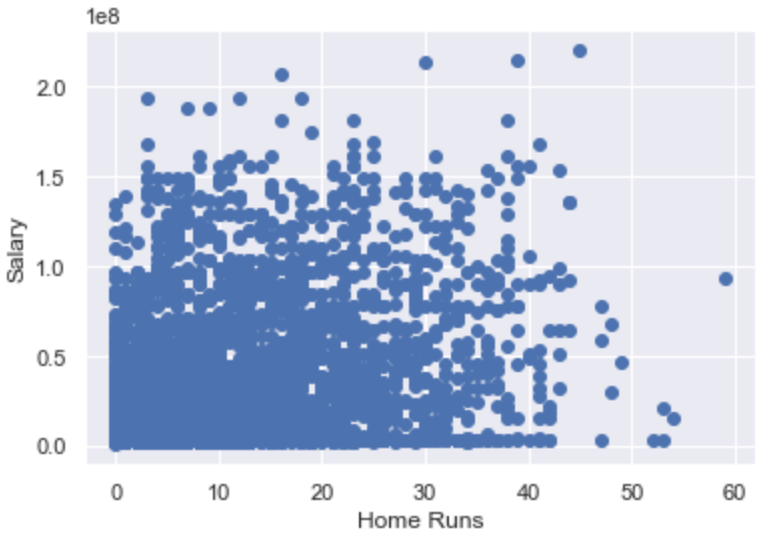


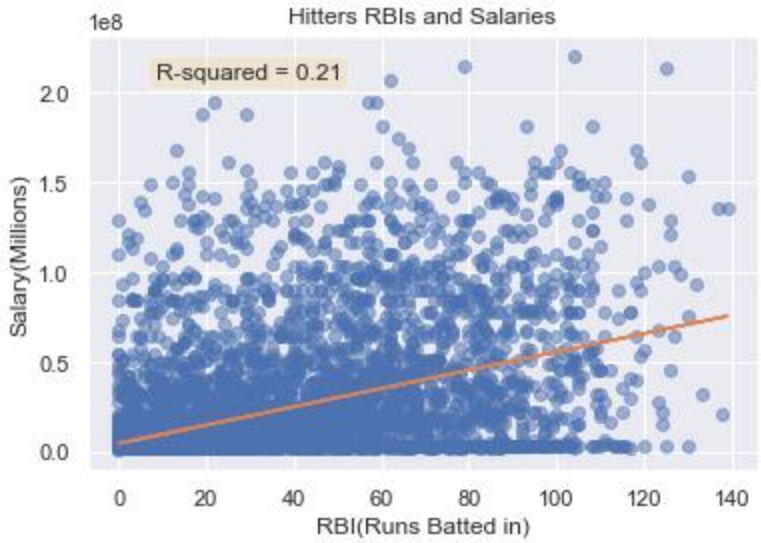

```
salary                1.000000
adj_salary_filled     1.000000
flag                  0.315781
Age                   0.433457
G                     0.374448
AB                    0.426127
R                     0.426275
H                     0.429633
Dbl                   0.407866
HR                    0.423042
RBI                   0.463561
BB                    0.447756
IBB                   0.377933
SO                    0.343339
SF                    0.344241
GDP                   0.438088
total_years_mlb       0.524767
Name: adj_salary_filled, dtype: float64
```

Independent Variables Vs. Dependent Variable Scatter Plots

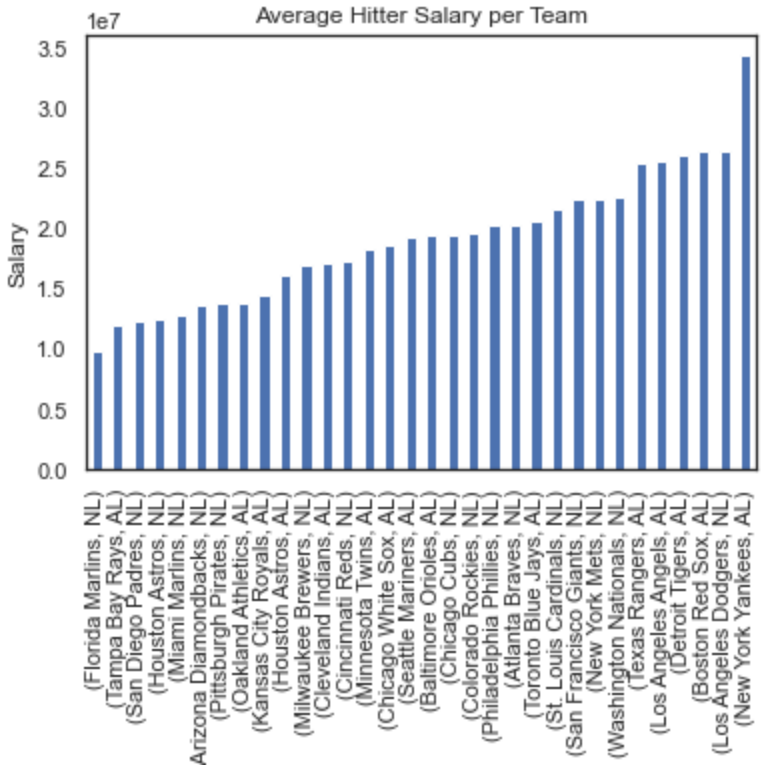

1e8
2.0
1.5
1.0
0.5
0.0
Salary
20
25
30
35
40
45
Player Age


1e8
2.0
1.5
1.0
0.5
0.0
Salary
0
50
100
150
200
Hits

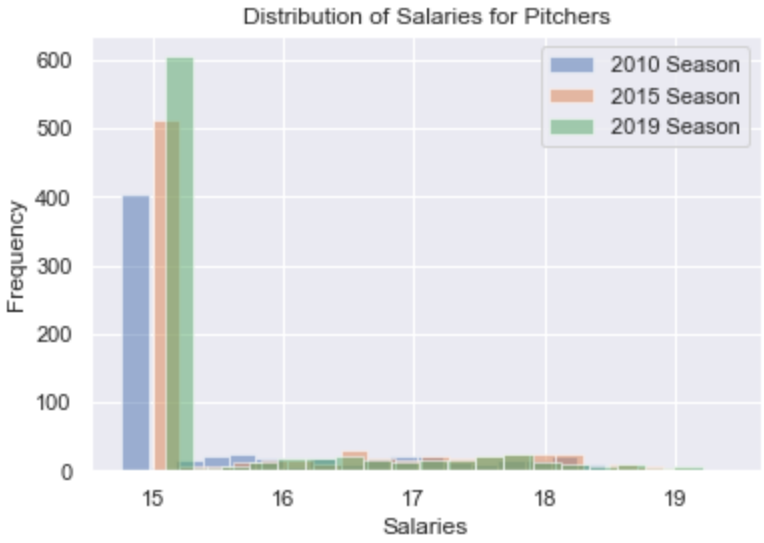


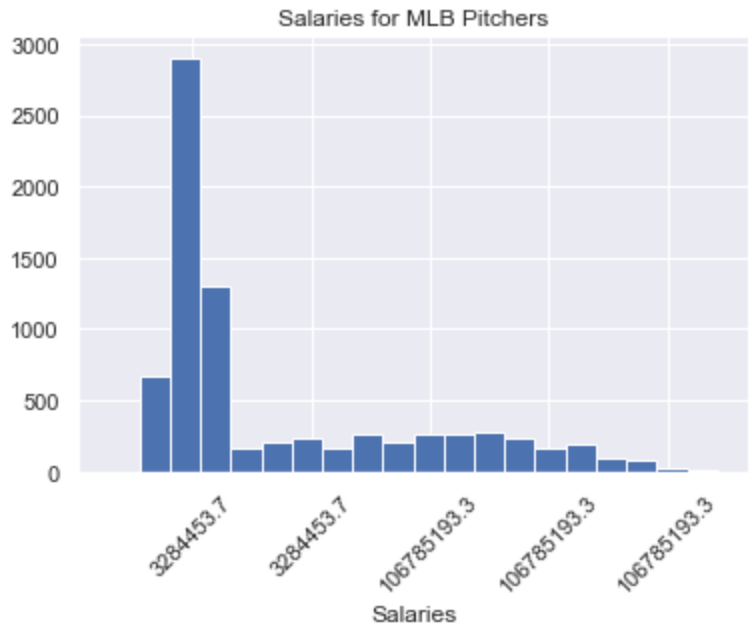


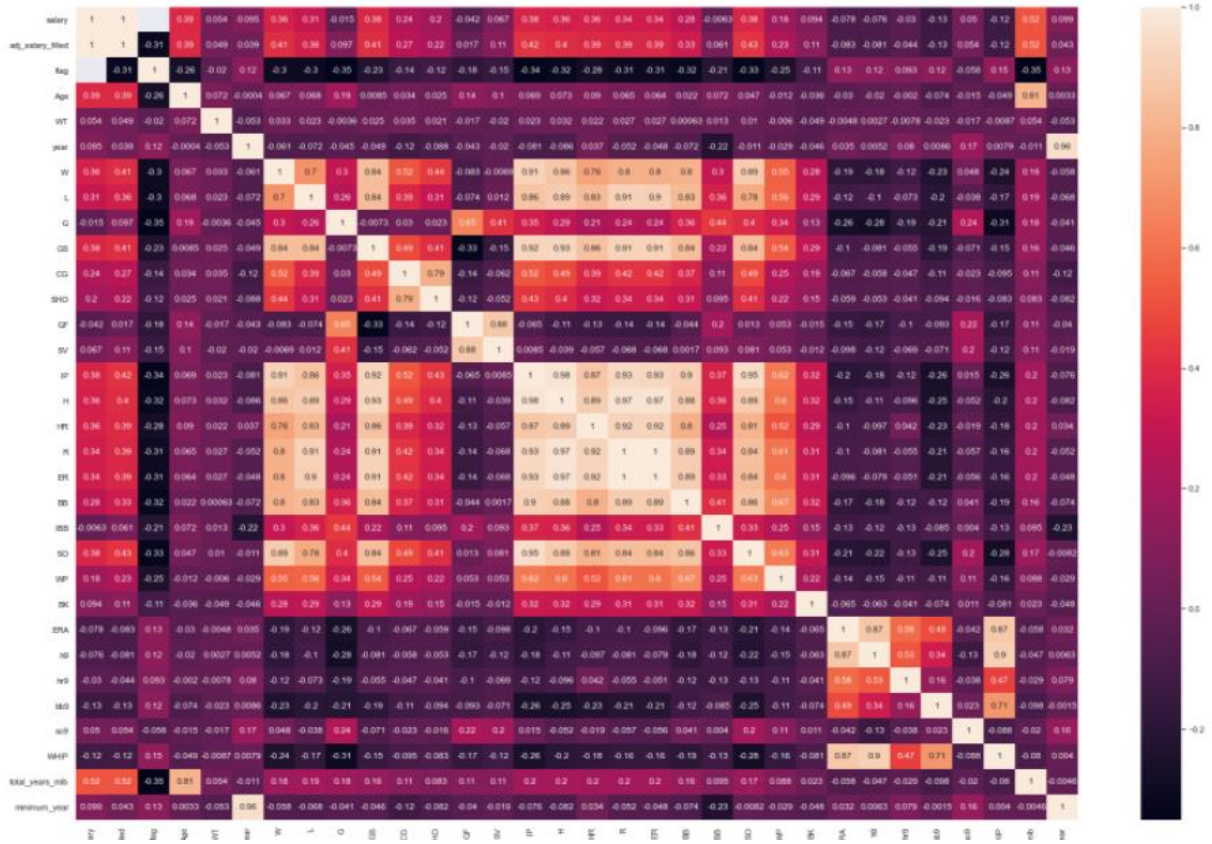
From a hitters perspective what team pays the most?


Average Hitter Salary per Team
1e7
Salary
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
(Florida Marlins, NL)
(Tampa Bay Rays, AL)
(San Diego Padres, NL)
(Houston Astros, NL)
(Miami Marlins, NL)
Arizona Diamondbacks, NL)
(Pittsburgh Pirates, NL)
(Oakland Athletics, AL)
(Kansas City Royals, AL)
(Houston Astros, AL)
(Milwaukee Brewers, NL)
(Cleveland Indians, AL)
(Cincinnati Reds, NL)
(Minnesota Twins, AL)
(Chicago White Sox, AL)
(Seattle Mariners, AL)
(Baltimore Orioles, AL)
(Chicago Cubs, NL)
(Colorado Rockies, NL)
(Philadelphia Phillies, NL)
(Atlanta Braves, NL)
(Toronto Blue Jays, AL)
(St. Louis Cardinals, NL)
(San Francisco Giants, NL)
(New York Mets, NL)
(Washington Nationals, NL)
(Texas Rangers, AL)
(Los Angeles Angels, AL)
(Detroit Tigers, AL)
(Boston Red Sox, AL)
(Los Angeles Dodgers, NL)
(New York Yankees, AL)

Pitcher

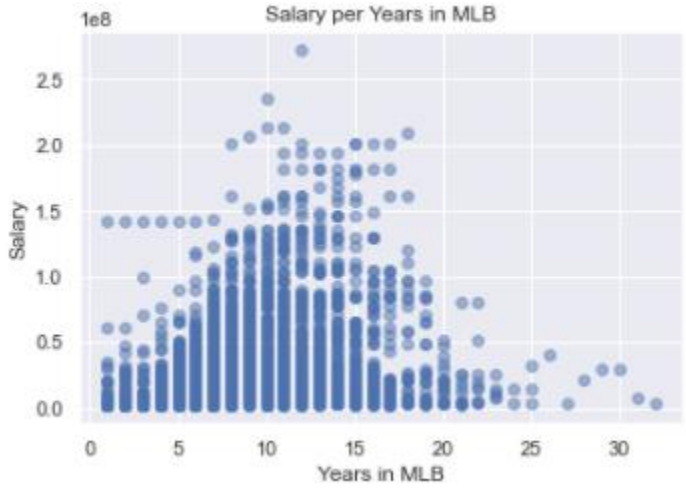

Distribution of Salaries for Pitchers
2010 Season
2015 Season
2019 Season
Frequency
600
500
400
300
200
100
0
15
16
17
18
19
Salaries

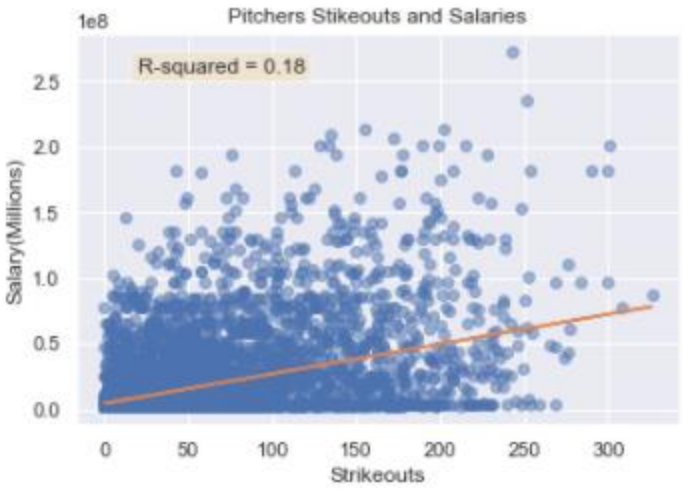

Salaries for MLB Pitchers
3000
2500
2000
1500
1000
500
0
3284453.7
3284453.7
106785193.3
106785193.3
106785193.3
Salaries

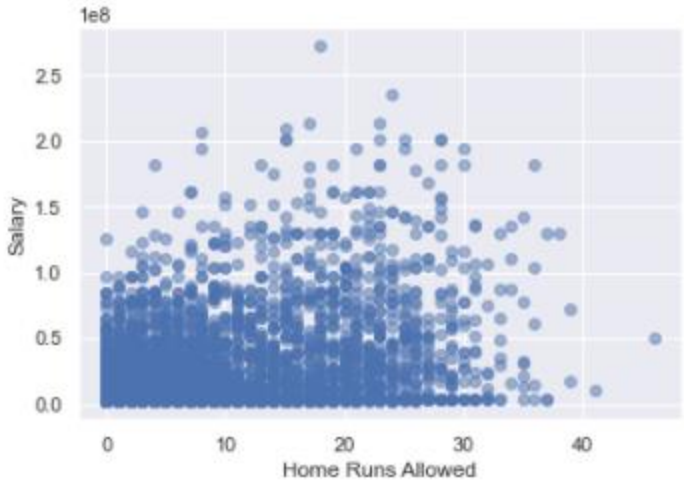


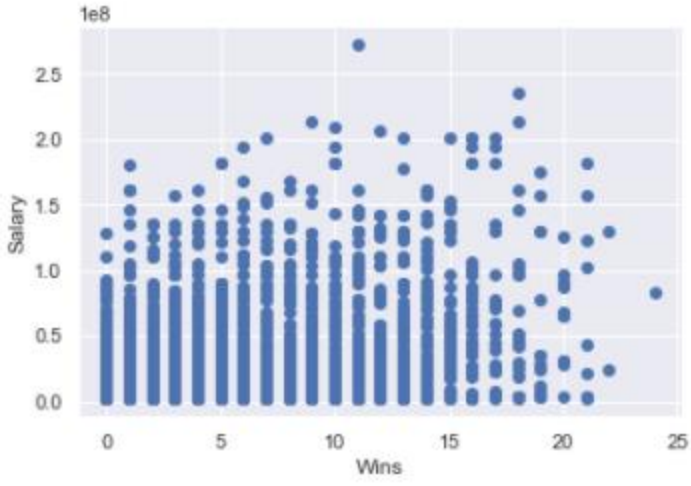
```
salary               1.000000
adj_salary_filled    1.000000
flag                 0.310272
Age                  0.391690
W                    0.406481
L                    0.355426
GS                   0.407395
IP                   0.420481
H                    0.404449
HR                   0.392769
R                    0.385521
ER                   0.385179
BB                   0.333458
SO                   0.426906
total_years_mlb      0.515814
Name: adj_salary_filled, dtype: float64
```

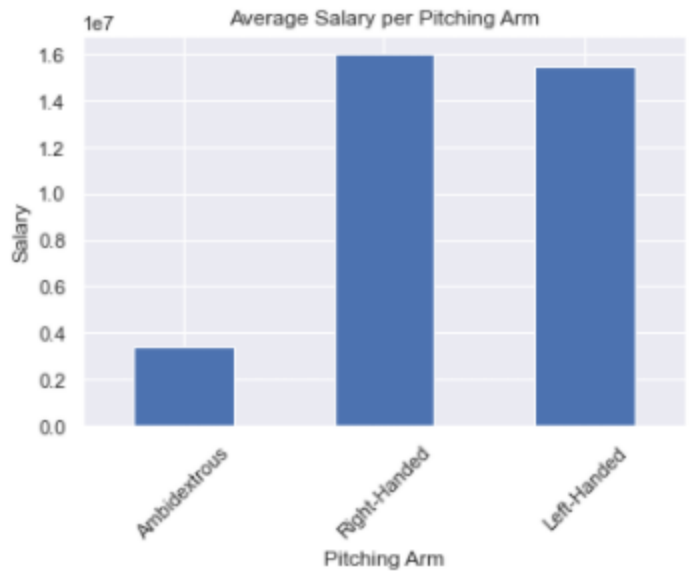
Independent Variables Vs. Dependant Variable Scatter Plots



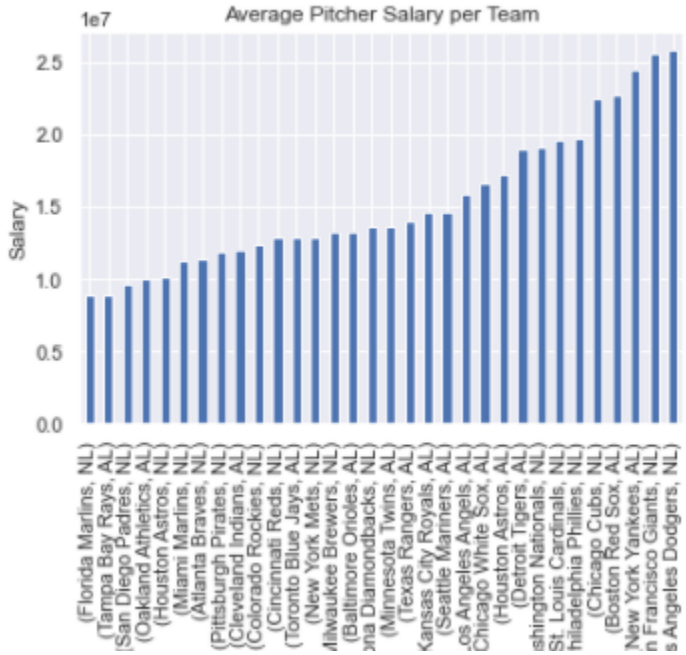

Pitchers Stikeouts and Salaries
1e8
R-squared = 0.18
Salary(Millions)
Strikeouts
0.0
0.5
1.0
1.5
2.0
2.5
0
50
100
150
200
250
300


1e8
Salary
Home Runs Allowed
0.0
0.5
1.0
1.5
2.0
2.5
0
10
20
30
40


1e8
Salary
Wins
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20
25



From a pitchers perspective what team pays the most?

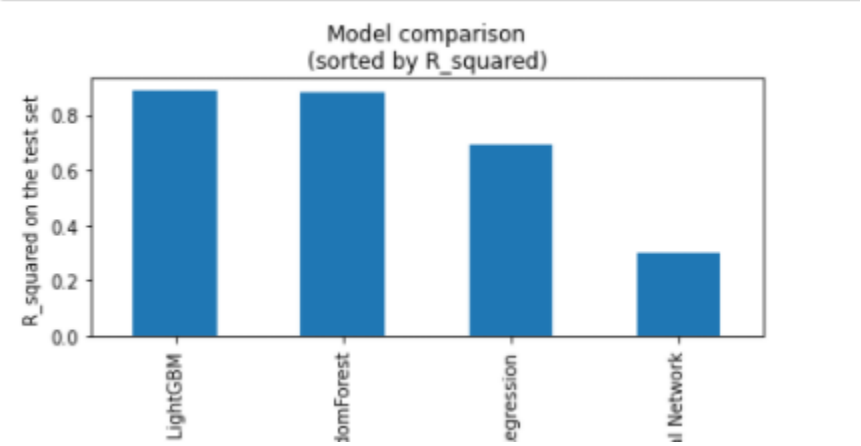


Model

I used 4 models Linear Regression, Random Forest and Light Gradient Boosting, and Neural Network to evaluate the models.

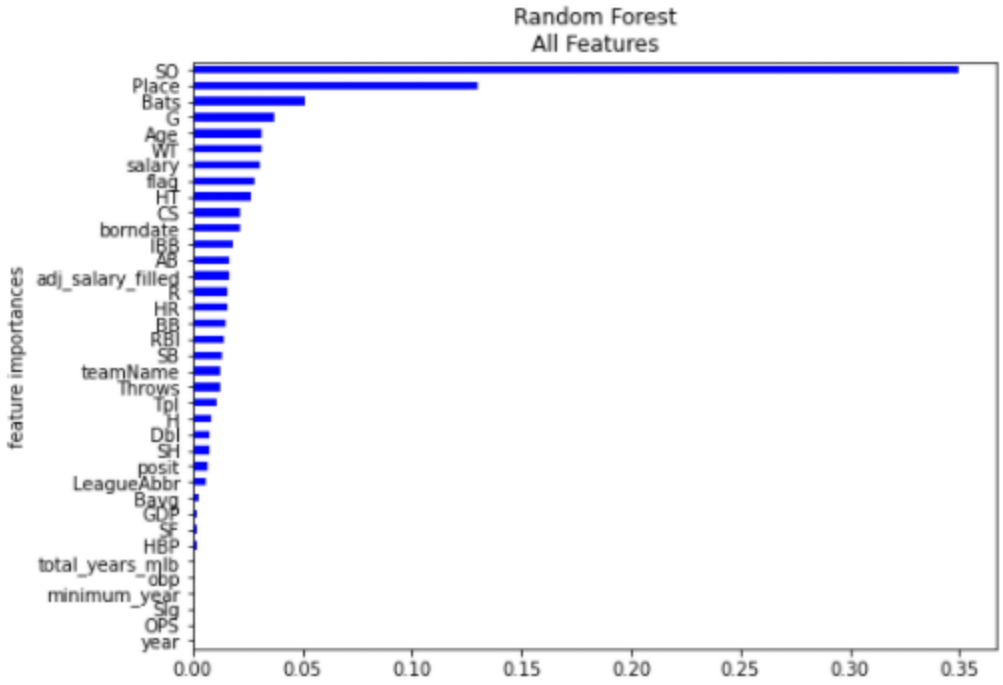
Result

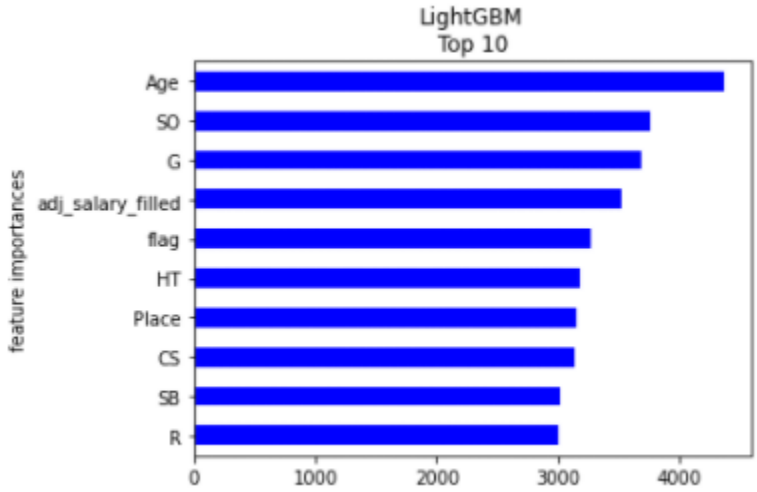
	RMSE_val	RMSE_test	R_squared	Time	Scaling
LightGBM	120472	120537	0.8883	10min 31s	True
RandomForest	131979	124013	0.8818	3min 6s	False
LinearRegression	202941	200704	0.6904	5.98 s	True
Neural Network	26461500	27301891	0.2996	36.1 s	False



LightGBM is best followed by RandomForest followed by LinearRegression

Feature Importances





Conclusion

I have found the best MLB salary prediction models is the LightGBM models which showed high speed and best performance in RMSE. If one model should be selected I would recommend to use the LightGBM model since it is faster and it makes fewer outliers.