Phuc Ton Nguyen

Prediction of salary for MLB player

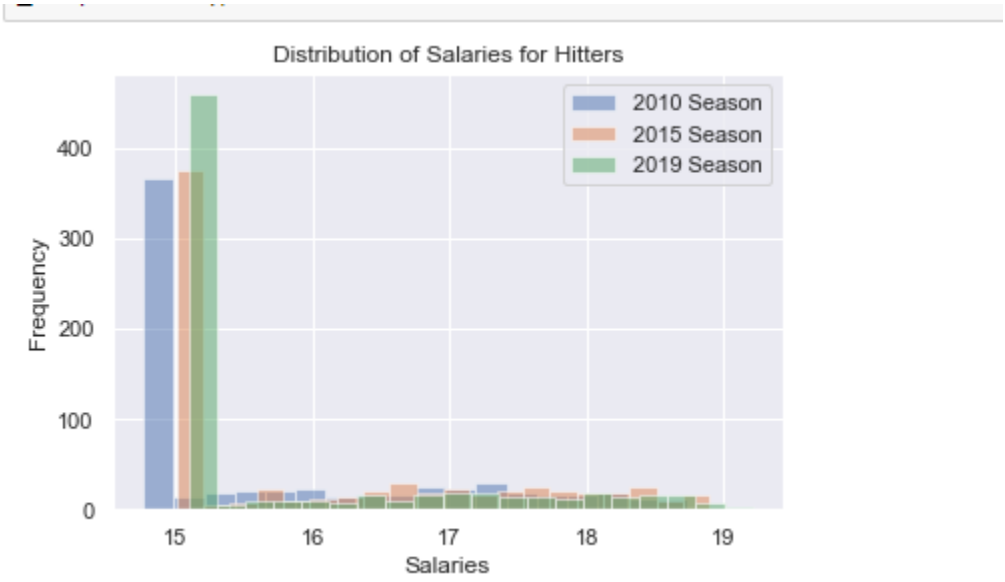
Guided Capstone Project Report

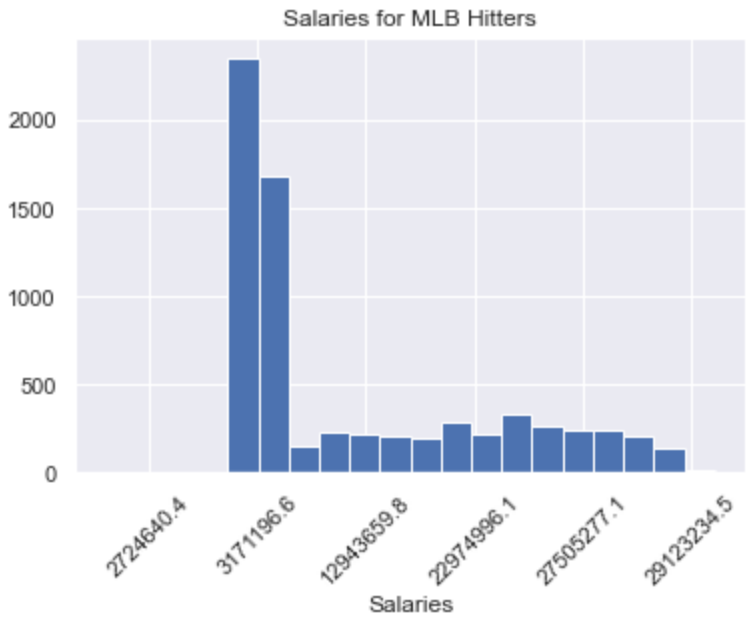
Introduction

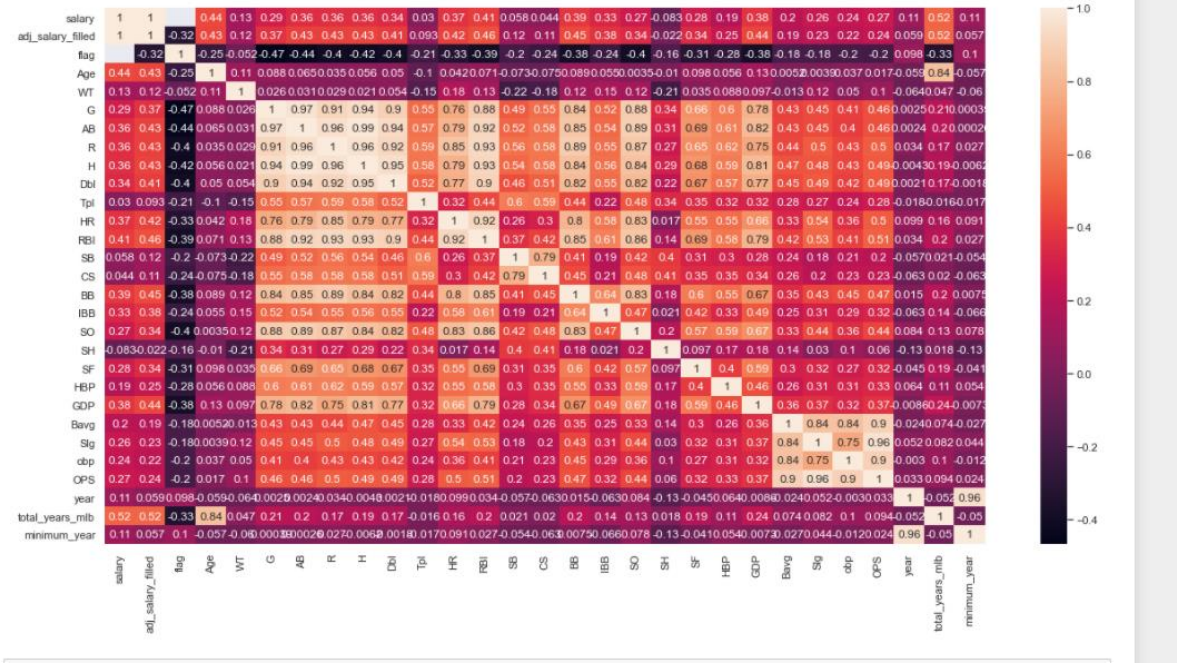
San Francisco Giants wants to buy player and they are asking the evaluation of the players on the market to buy.

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.

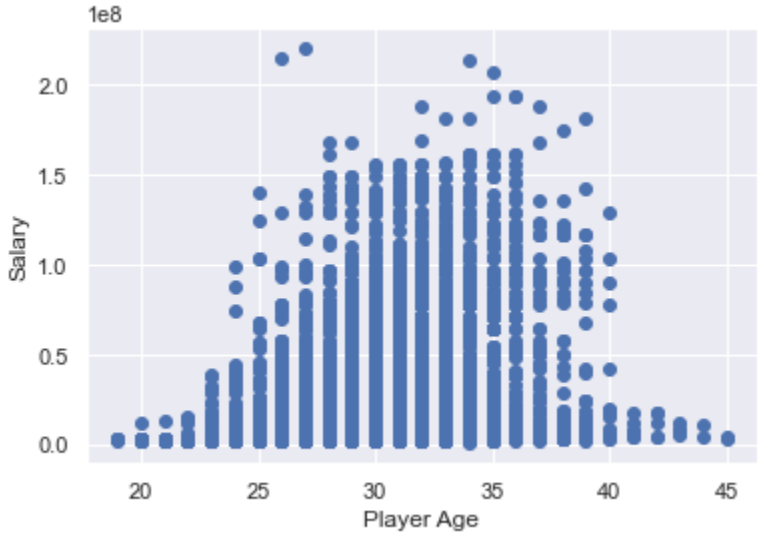


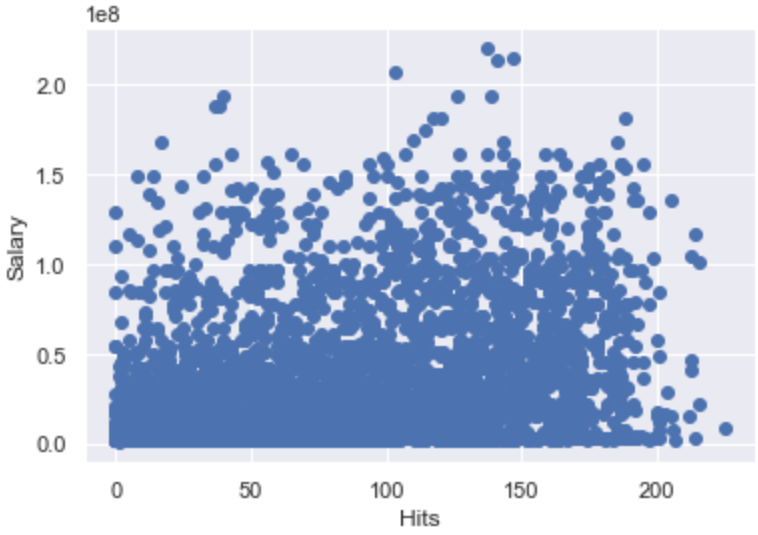

Salaries for MLB Hitters
2000
1500
1000
500
0
2724640.4
3171196.6
12943659.8
22974996.1
27505277.1
29123234.5
Salaries

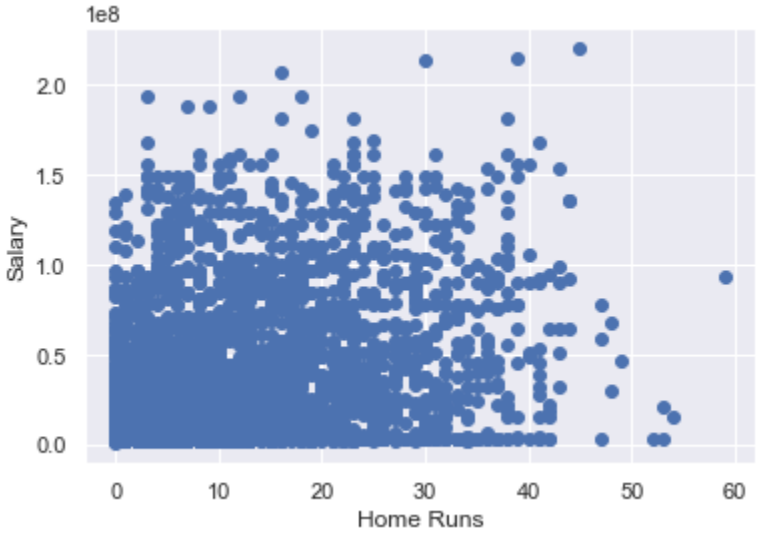

salary
adj_salary_filled
flag
Age
WT
G
AB
R
H
Dbl
Tpl
HR
RBI
SB
CS
BB
IBB
SO
SH
SF
HBP
GDP
Bavg
Slg
obp
OPS
year
total_years_mlb
minimum_year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

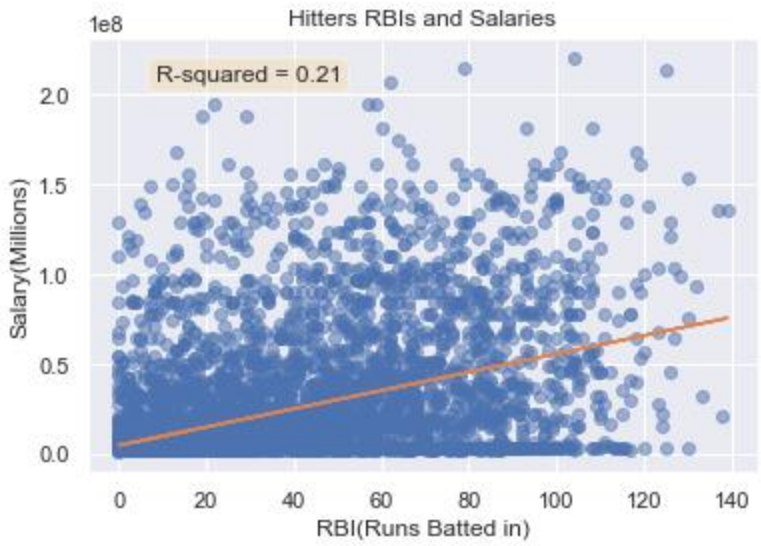
```
salary               1.000000
adj_salary_filled    1.000000
flag                 0.315781
Age                  0.433457
G                    0.374448
AB                   0.426127
R                    0.426275
H                    0.429633
Dbl                  0.407866
HR                   0.423042
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

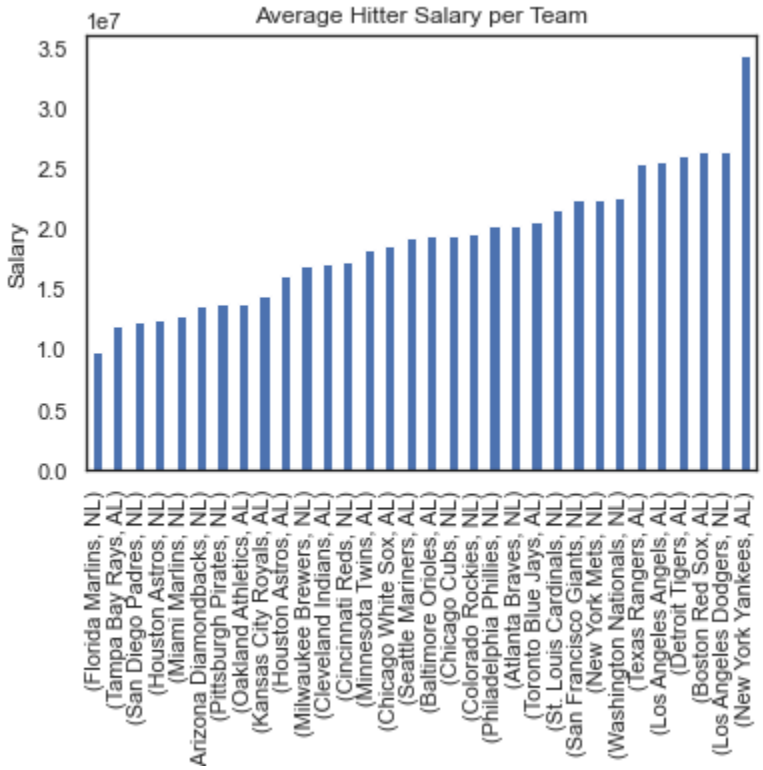



1e8
2.0
1.5
1.0
0.5
0.0
Salary
0
50
100
150
200
Hits


1e8
2.0
1.5
1.0
0.5
0.0
Salary
0
10
20
30
40
50
60
Home Runs



From a hitters perspective what team pays the most?

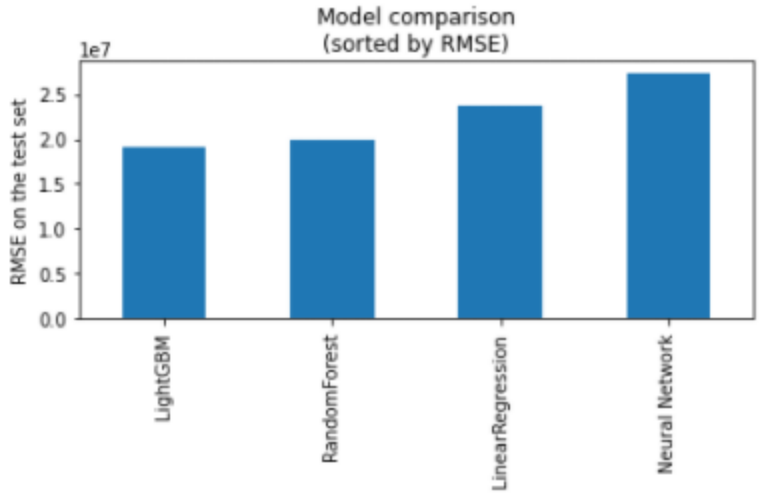


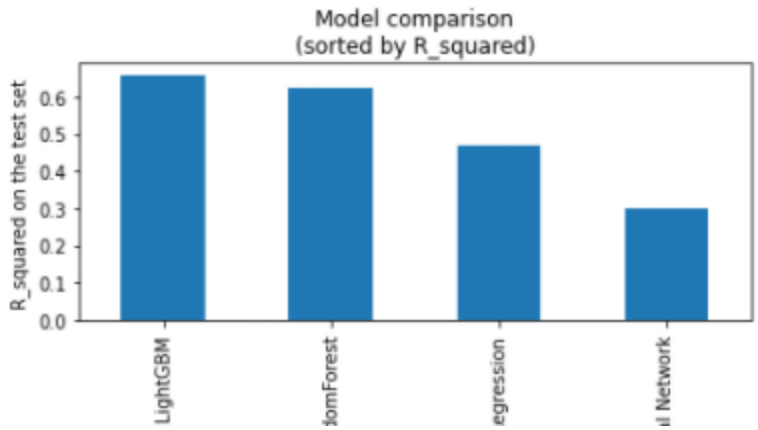
Model

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

Result

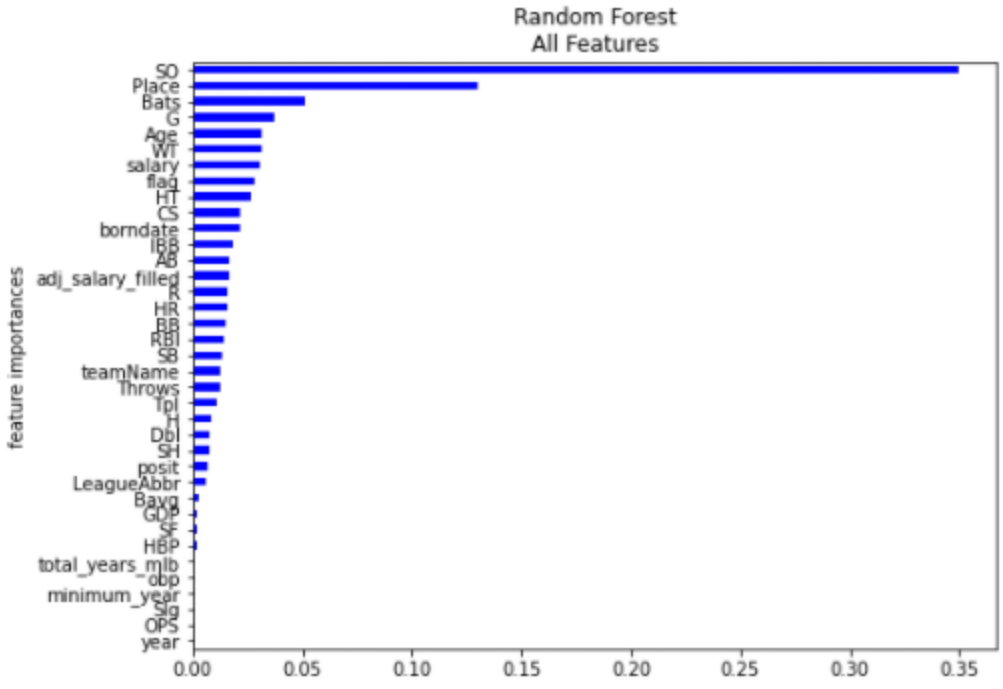
	RMSE_val	RMSE_test	R_squared	Time	Scaling
LightGBM	18248353	19113579	0.6567	10min 31s	True
RandomForest	18853251	19985546	0.6247	3min 6s	False
LinearRegression	22615270	23797151	0.4679	5.98 s	True
Neural Network	26461500	27301891	0.2996	36.1 s	False

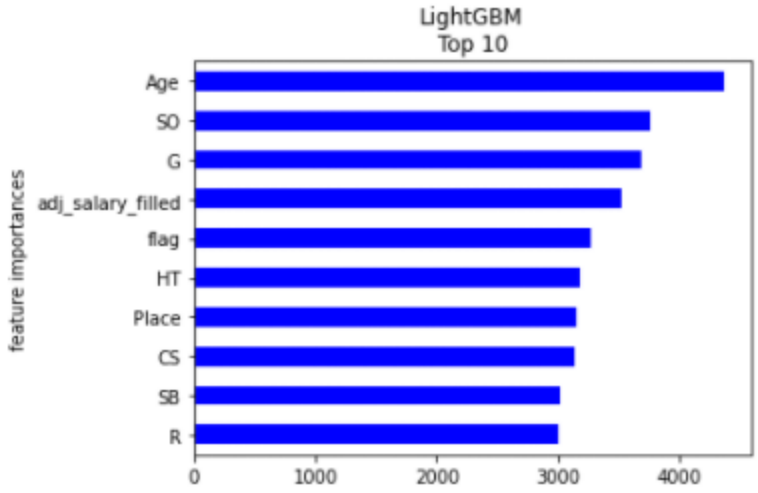




LightGBM is best followed by RandomForest followed by LinearRegression

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.