Predicting MLB Players Salary

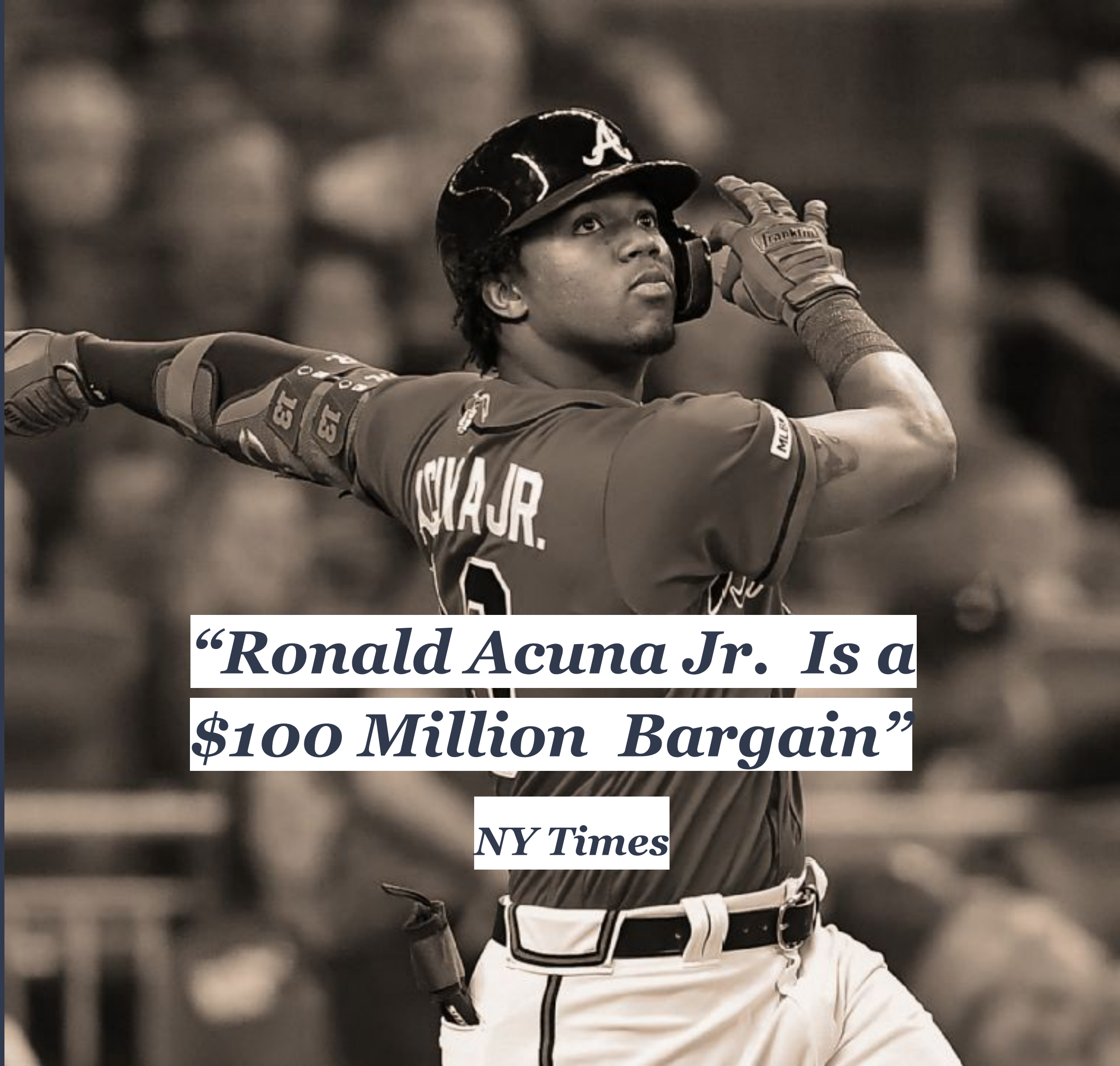
Christopher Feliz
June 2020
Springboard Data Science Fellow

Github: https://github.com/cfeliz3030/Springboard-

"Acuna might have cost himself more than **$15 million** over the lifetime of this deal, and another **$25 million** or more considering the two **$17 million** team options that the team can exercise at the end of the eight years."

-Source

He was **undervalued by his team.**



“Davis signed a deal worth **$161 million dollars** over the span of 7 years and over the last couple of years he has been regarded as one of the **worst players in the league**. He hit just .179 in 2019 and in 2018 he was even worse putting up a .168 average. “

- Source

He was **overvalued by his team.**

Could this have been avoided?

Throughout the years MLB teams have lost enormous amounts of money by overpaying for certain players

There have also been many contract signings where players have been totally undervalued

The Goal

In this project, I plan to build a model to predict player salaries which can be used both by sports franchises, to minimize their risk, and by players, to help them determine their value.

The aim for this project is to accurately predict the salaries for MLB players based on…

- Previous season stats
- Position
- Age
- Length in the MLB

Data Overview

- The data was scraped from 'http://www.thebaseballcube.com/', a baseball data warehouse.
- Two separate data frames were constructed because of the difference in performance metrics between hitters and pitchers.
- Subset the data and used players only after the 2010 season. MLB is an expanding and ever changing sport and a lot has happened over the last 10 years.

Performance Metrics

Pitchers

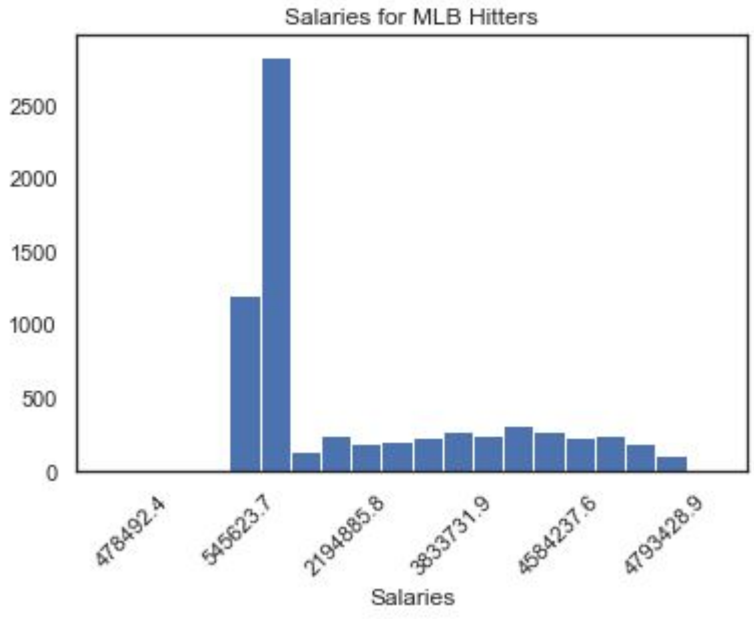
```
Data columns (total 41 columns):
playerName           7696 non-null object
salary               5266 non-null float64
adj_salary_filled    7696 non-null float64
flag                 7696 non-null int64
Age                  7696 non-null int64
HT                   7696 non-null object
WT                   7696 non-null int64
Bats                 7696 non-null object
Throws               7696 non-null object
year                 7696 non-null int64
teamName             7696 non-null object
posit                7696 non-null object
borndate             7696 non-null object
Place                7696 non-null object
LeagueAbbr           7696 non-null object
W                    7696 non-null int64
L                    7696 non-null int64
G                    7696 non-null int64
GS                   7696 non-null int64
CG                   7696 non-null int64
SHO                  7696 non-null int64
GF                   7696 non-null int64
SV                   7696 non-null int64
IP                   7696 non-null float64
H                    7696 non-null int64
HR                   7696 non-null int64
R                    7696 non-null int64
ER                   7696 non-null int64
BB                   7696 non-null int64
IBB                  7696 non-null int64
SO                   7696 non-null int64
WP                   7696 non-null int64
BK                   7696 non-null int64
ERA                  7696 non-null float64
h9                   7696 non-null float64
hr9                  7696 non-null float64
bb9                  7696 non-null float64
so9                  7696 non-null float64
WHIP                 7696 non-null float64
total_years_mlb      7696 non-null int64
minimum_year         7696 non-null int64
dtypes: float64(9), int64(23), object(9)
memory usage: 2.4+ MB
```

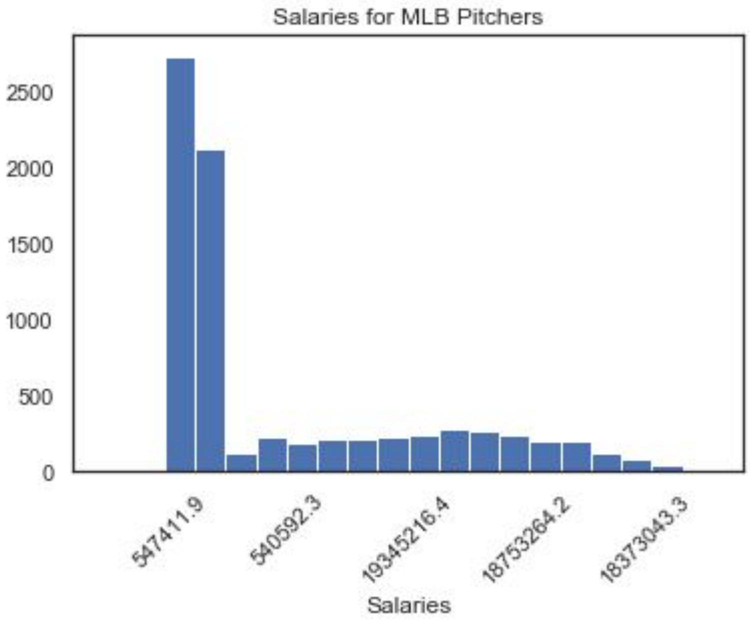
Hitters

```
Data columns (total 38 columns):
playerName           6924 non-null object
salary               5047 non-null float64
adj_salary_filled    6924 non-null float64
flag                 6924 non-null int64
Age                  6924 non-null int64
HT                   6924 non-null object
WT                   6924 non-null int64
Bats                 6924 non-null object
Throws               6924 non-null object
posit                6924 non-null object
borndate             6924 non-null object
Place                6924 non-null object
teamName             6924 non-null object
LeagueAbbr           6924 non-null object
G                    6924 non-null int64
AB                   6924 non-null int64
R                    6924 non-null int64
H                    6924 non-null int64
Dbl                  6924 non-null int64
Tpl                  6924 non-null int64
HR                   6924 non-null int64
RBI                  6924 non-null int64
SB                   6924 non-null int64
CS                   6924 non-null int64
BB                   6924 non-null int64
IBB                  6924 non-null int64
SO                   6924 non-null int64
SH                   6924 non-null int64
SF                   6924 non-null int64
HBP                  6924 non-null int64
GDP                  6924 non-null int64
Bavg                 6924 non-null float64
Slg                  6924 non-null float64
obp                  6924 non-null float64
OPS                  6924 non-null float64
year                 6924 non-null int64
total_years_mlb      6924 non-null int64
minimum_year         6924 non-null int64
dtypes: float64(6), int64(23), object(9)
memory usage: 2.0+ MB
```

Exploratory Data Analysis

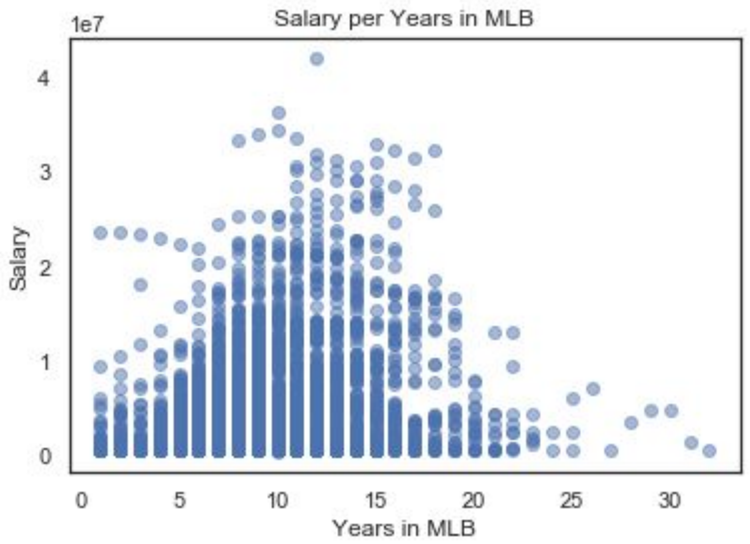
Distribution of Target Variable





Exploratory Data Analysis cont.

- Gradual increase in salary during the first couple of years.
- Players maximum salary tends to peak after 7-10 years of MLB service.
- After 15 years of MLB service a players salary begins to decrease rapidly. This could be due to age and below average performance.



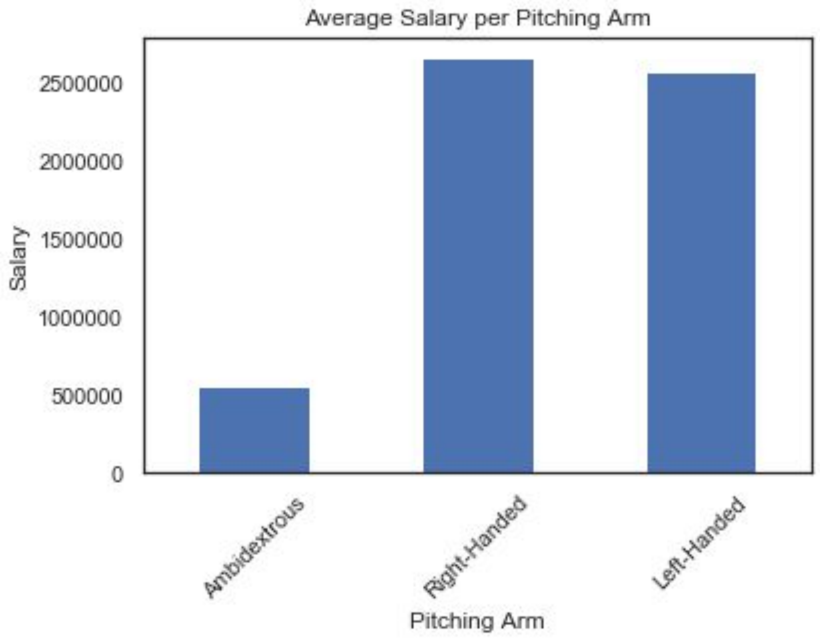
Statistical Inference

Is there a difference between the salaries paid towards right handed and left handed pitchers?

H_0 : The salaries for both left and right handed pitchers are the same.

H_1 : The salaries for left and right handed pitchers are different.

- An independent samples t-test was conducted to compare the salaries paid to left and right handed pitchers
- Results of the t-test indicated that there were not significant differences in salaries paid to pitchers who threw with their left or right arm.



Machine Learning

1. Encode categorical variables using Pandas to dummy function
2. Scale data using Standard Scaler
3. Train-Test-Split Data
4. Tune Hyperparameters using GridSearchCV or RandomizedSearchCV
5. Evaluate models

Model Selection and Performance

Models	RMSE	R2
Baseline Model	$5,397,560	0.00
Linear Regression w PCA	$4,644,008	0.55
RandomForest Regressor	$3,510,004	0.78
XGBoost Regressor	$3,315,287	0.80
HistGradientBoosting Regressor	$3,293,081	0.79
CatBoost Regressor	$3,272,132	0.80
LightGBM Regressor	$3,236,633	0.80

- The best performing model for the **hitter dataset** was the **LightGBM Regressor**. This model's final metrics was an R-squared of 0.80 and RMSE of 3,236,633.

Model Selection and Performance

Models	RMSE	R2
Baseline Model	$4,683,385	0.00
XGBoost Regressor	$2,832,288	0.76
LightGBM Regressor	$2,818,405	0.76
RandomForest Regressor	$2,796,006	0.77
HistGradientBoost Regressor	$2,757,535	0.77
CatBoost Regressor	$2,754,287	0.77

- The best performing model for the **pitcher dataset** was the **CatBoost Regressor**. This model's final metrics was an R-squared of 0.77 and RMSE of 2,754,287.

40%

reduction in the root mean squared error of the baseline models in the hitter and pitcher datasets respectively.

	Predicted	Actual
1375	19076500.5	22215589.6
1376	562280.9	563213.0
1377	675649.4	557998.1
1378	640047.2	555556.3
1379	6832035.1	5399622.4
1380	619921.2	555000.0
1381	532092.0	540592.3
1382	491367.7	534488.5
1383	8945658.7	2969792.3
1384	17156686.0	12959093.7

Final Conclusions

Most of the models had issues with predicting larger salaries given the larger spread in the previous slide. These values could also be deemed as outliers, and correcting this can be done in a number of ways.

For future iterations I will gather data on player injuries and medical history. Collecting information on a player's past injuries will help predict any future injuries as well as affect their total value. Another feature that could be added is a players place of origin. This could be within the states or at the international level.