PREDICTING MLB BASEBALL SALARIES FOR OFFENSIVE PLAYERS USING LINEAR REGRESSION.

- GURUNADH PARINANDI

- October 9, 2020

MOTIVATION:

- During the year 2019, Major League Baseball made an annual profit of 10.9 Billion Dollars.
- In 2018, the league spent 54.2% of their revenue on player compensation.

OBJECTIVE:

- Using MLB data and linear regression techniques, we:
  - Build a model to predict an offensive player's salary using traditional MLB features.
  - Try to understand what features affect an offensive player's salary the most.

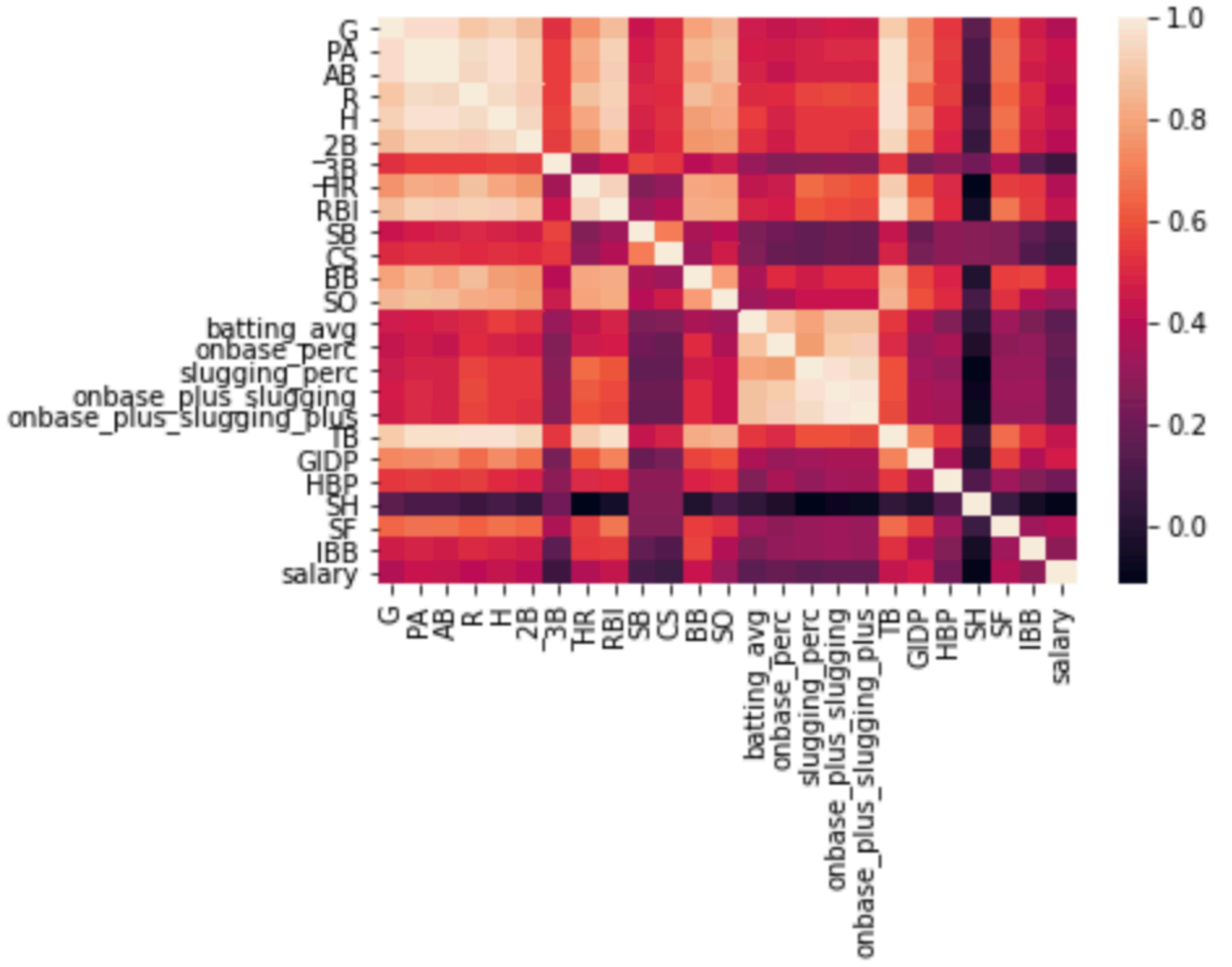
METHODOLOGY:

- Due to the COVID pandemic and the shortened baseball season, we focused our initial analysis on 2019 data.
- Scraped 2019 offensive player data from:
  - https://www.baseball-reference.com
- Scraped 2019 offensive player salary data from:
  - https://www.spotrac.com/mlb/rankings/salary
- Historical MLB player salary data from 1985 to 2015 can be found on:
  - http://www.seanlahman.com/baseball-archive/statistics/

BASEBALL FEATURES USED IN ANALYSIS:

- G – Games Played
- PA – Plate Appearance
- AB – At Bat
- R – Runs
- H - Hit
- 2B - Double
- 3B - Triple
- HR – Home Run
- RBI - Run Batted In
- SB – Stolen Base
- CS – Caught Stealing
- BB - Walk
- SO – Strike Out
- Batting Average
- On Base Percentage
- On Base Plus Slugging
- On Base Plus Slugging Plus
- TB – Total Bases
- GIDP – Ground Into Double Play
- HBP – Hit By Pitch
- SH – Sacrifice Bunt
- SF – Sacrifice Fly
- IBB – Intentional Base on Balls.
- Salary

HEATMAP OF FEATURE CORRELATION:



Predictive Model (FIRST PASS):

- 457 observations were used in this study (out of a total of 636 offensive players in 2019).
- All features from slide 4 were used.
- 70% of data used in training and 30% used in testing.
- For Lasso CV, n_splits = 20.
- Results from modeling were not that good.

MODEL	TRAINING ERROR	TESTING ERROR
Linear Regression	0.37	0.20
LassoCV	0.36	0.21
LassoCV w./ Polynomial Features	0.34	0.25

Issues with Predictive Modeling:

- Interactions between features is causing the initial predictive score not to be optimal.
  - Examples:
    - Batting Average = Hits / At Bats
    - On Base Percentage = (Hits plus Walks plus Hit by Pitcher) divided by (At Bats plus Walks plus Hit by Pitcher plus Sacrifice Flies).
    - Total Bases = Hits plus Doubles plus (2 times Triples) plus (3 times Home runs)

Best Subset Selection:

- Used Best Subset selection method to find infer top four features that collectively affect salary the most.
- **GIDP**, **BB**, **SO**, AND **3B** are the features that affect salary most.
- Prior to 2015, **RBI**, **HR**, and **R** were the most important features affecting salary.
  - Game is being managed much differently now compared to years past.

Best Subset Selection - Results:

	GIDP	BB	SO	3B
Linear Regression Coefficient value	482010.21	102787.74	-25830.90	-436303.59

	TRAINING ERROR	TESTING ERROR
Linear Regression	0.287	0.287

Next Steps:

- Aggregate data from 2016 to 2019 to increase power of the model.
- Analyze historic data (in particular mid-1990's to 2010) to understand how trends in managing the game have affected salary distributions.
- Use this methodology to study what features are important in predicting MLB Pitcher Salaries
  - Pitcher salaries are continuing to rise.
  - Most pitchers are signed to long term contracts after arbitration.

Thank you!

- Special thanks to my group members:
  - Michael Green
  - Jay Park
  - Albert Lee
  - Duncan Sweeny
  - Andrew Zhou
  - Metis Instruction Team

Residual Plot of LassoCV with Polynomial Features.