Projecting Baseball Players' Offensive Outputs During Arbitration Years

By Michael Scognamiglio
nyc-mnhtn-ds-080320

Project Overview and Business Case

- Projecting OPS+ for MLB players during their arbitration years
- During Arbitration, player salaries greatly increase
- For GM's on a budget or trying to win a championship, projecting the offensive output of players is important
- GM's can use model to determine player's value and determine whether other options (Free Agency,Trades) should be explored.
- GM's can use model to lower player's salary during arbitration hearings (if projections are not impressive)
- OPS+ was used as primary target variable (to measure offensive output)

Data and Process

- Dataset was comprised of all players on each MLB Rosters from years 2000-2020
- Initial Data Frame had 140,000 rows
- Each row consisted of a player's offensive stats for a season
- Was cut down to about 9,000 rows
- Dataset was cut down to only include position players who had six years of data for model training

Data Collection

SPORTS REFERENCE

Data Cleaning/EDA

pandas

matplotlib

NumPy

Seaborn

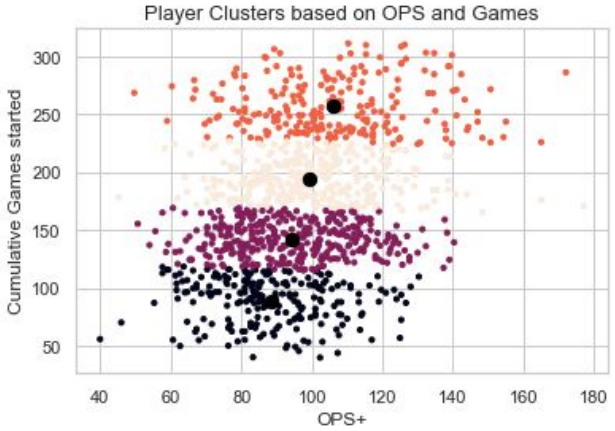
Modeling Process

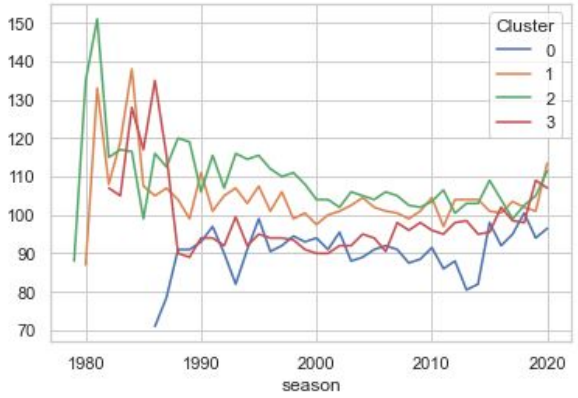
scikit learn

TensorFlow

Clustering

- Clustering was used as a feature engineering technique to give the model's more information on each player.
- K Means Clustering was used for all the players in our dataset.
- Clustering input features were cumulative games started and OPS+ for each player aggregated over their first three seasons

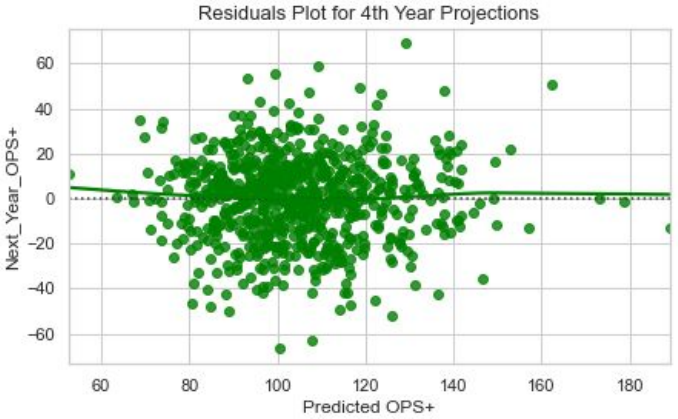


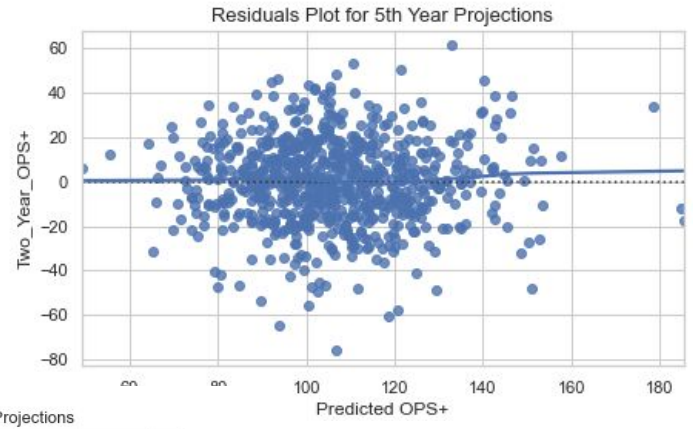


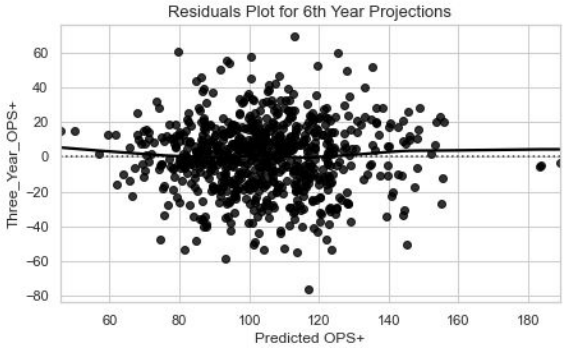
Modeling Selection

Type of Model	RSME	MAE
Simple Linear Regression	**19.30**	**15.1**
Lasso Model	19.33	15.2
Ridge Model	19.31	15.1
Simple Neural Network	22.5	23

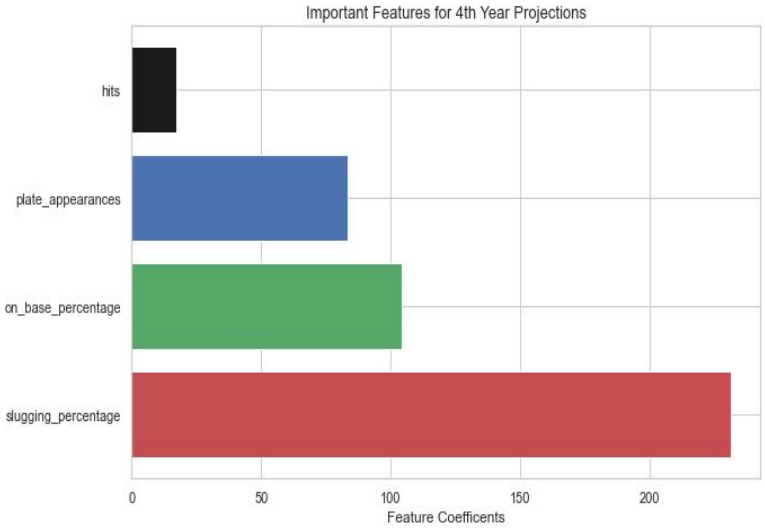
Final Model Evaluation

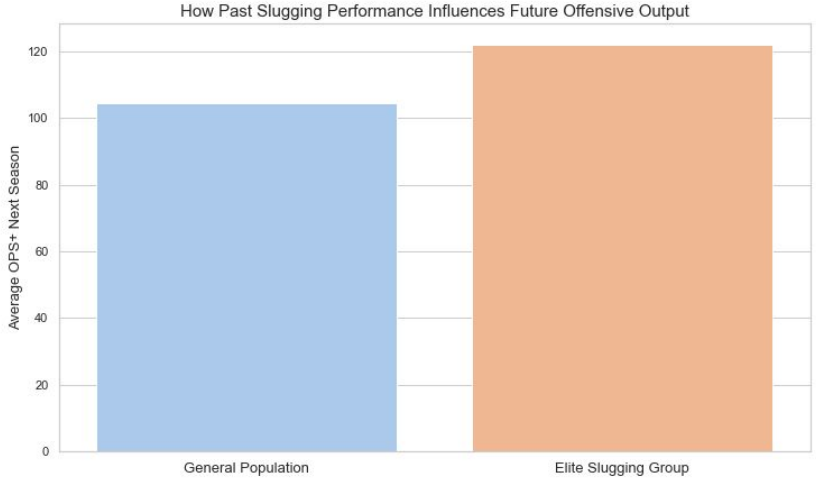






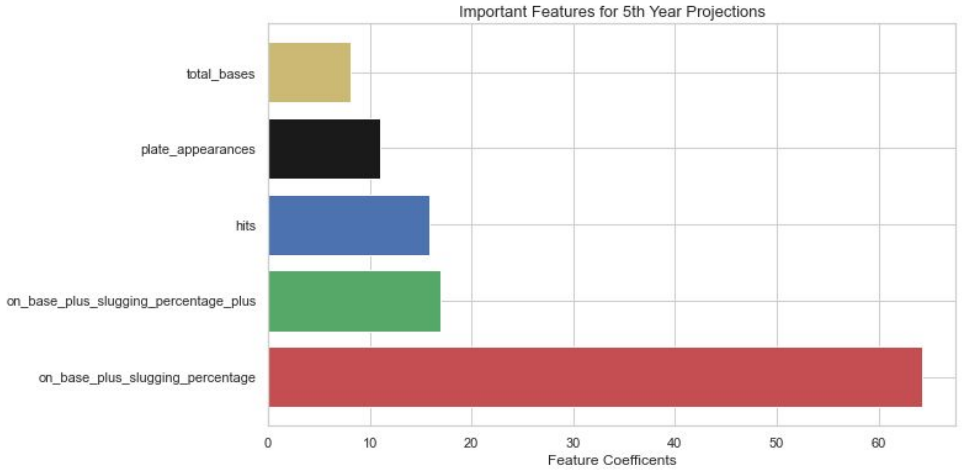
Final Model Analysis

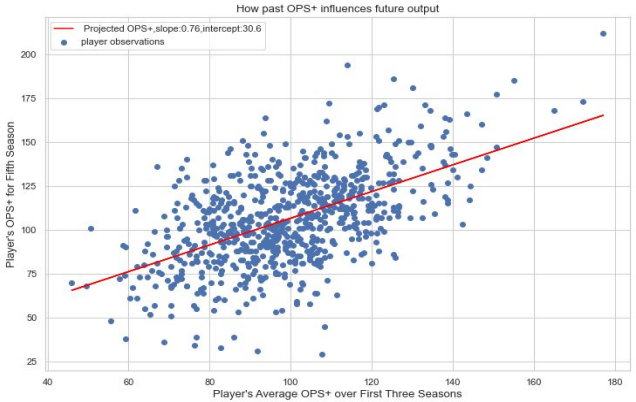




Test Statistic: 9.5 (result was significant)

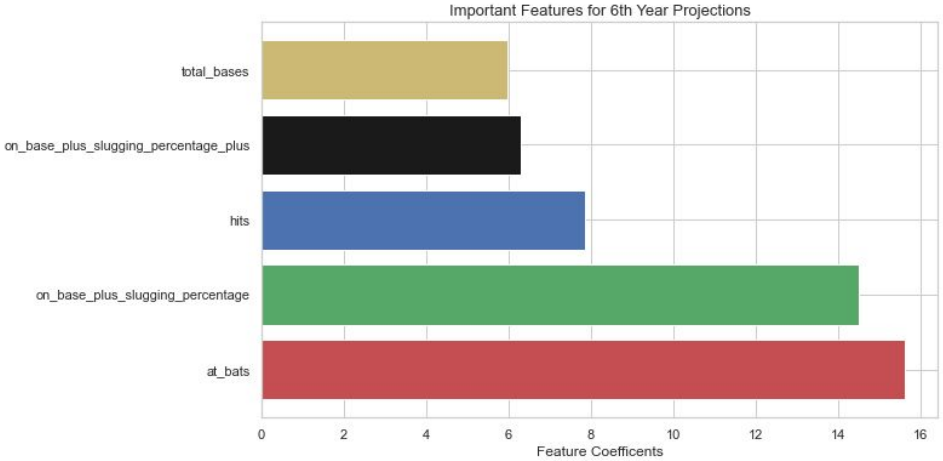
Final Model Analysis Part II

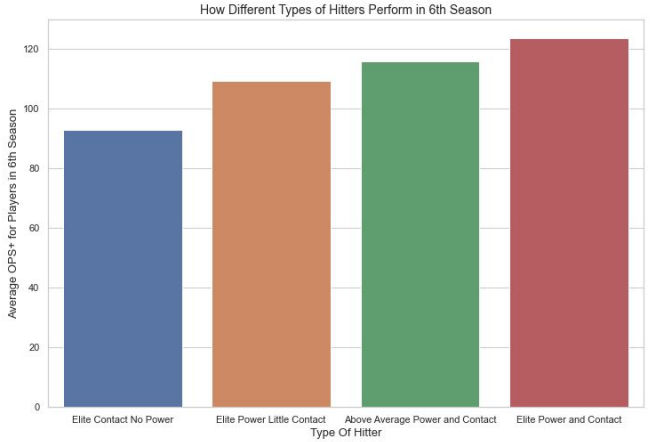




R^2: 0.55,
P-value:: 1.19*10^-60

Final Model Analysis Part III





Test Stat: 8.6
P-value:1.6*10^-5

Conclusions/Next Steps

1. Look for players who have elite slugging percentage, performed on average 22% better than an average player.
2. Look closely at a players average OPS+ over their first three seasons, a R^.2 of .55 between average OPS+ and a player's OPS+ in their fifth season
3. Prioritize Power Hitters above all else, we found a small difference between hitters who can hit for both power and contact and hitters who only hit for power. However, we found a big difference in OPS+ between hitters who can only hit for contact versus hitters who only hit for power.

As a next step, I would like to test this model on free agents to determine whether this model can perform well on that market. I would be interested in building a new model on new data for this market if the current models do not perform well.

Questions??