Clarity Kummer, Max Gehred, Eugene Ohba, Jon Karanezi

Introduction

This analysis utilizes Major League Baseball (MLB) Statcast data made publicly available by Baseball Savant to investigate whether Machine Learning techniques, considering only non-outcome-oriented offensive features, can accurately predict or classify an MLB player's offensive value as measured by On-Base-Percentage-Plus-Slugging-Percentage (OPS). This analysis consists of comparing the results of three different linear models: Linear, Lasso and Ridge, choosing the one that best predicts OPS to use on test data, for analysis and interpretation. Further, a classification model is derived that can be used to classify MLB batters as 'At Least Average' or 'Below Average,' based on OPS.

Dataset

The dataset comprises 538 observations and 23 features, encompassing all MLB batters from 2020 to 2023.

- The dataset considers features related to:
    - **Player Info** (first name, last name…)
    - **At Bat Outcome Metrics** (single, double, triple..)
    - **At Bat Quality Metrics** (exit velocity, sweet spot percentage..)

Data Preparation

- Remove duplicate observations, outliers and rows with missing values
- On-Base-Plus-Slugging-Percentage (OPS) is calculated from two other batting statistics, On-Base-Percentage (OBP) and Slugging-Percentage (SLG). To reduce multicollinearity,

this analysis considers 12, of the original 23, features not included in the OPS calculation. The 12 features included are provided in **Table 1** below.

- Train, Validation, Test Datasets: Our analysis encompasses data spanning from 2020 to 2023. We constructed a training dataset using data from 2020 to 2021, a validation dataset from 2022, and a testing dataset from 2023.
- Standardization: The data employed in all tasks was first standardized using 'StandardScaler'

Table 1: Features Included in Analysis

Feature	Description
Strikeout Percentage	Percentage of plate appearances resulting in a strikeout
Exit Velocity Average	Average speed of the ball off the bat
Sweet Spot Percentage	Percentage of batted balls hit in the "sweet spot" of the bat
Barrel Batted Rate	Percentage of batted balls hit with optimal exit velocity and launch angle
Solid Contact Percentage	Percentage of batted balls hit with solid contact
Hard Hit Percentage	Percentage of batted balls hit with high exit velocity
Average Best	Average sprint speed of the player at their best
Average Hyper Speed	Average sprint speed of the player at hyper speed
Whiff Percentage	Percentage of swings resulting in a miss
Swing Percentage	Percentage of pitches swung at
Ground Ball Percentage	Percentage of batted balls hit on the ground
Flyball Percentage	Percentage of batted balls hit in the air

Regression Model Experiments

Linear Model Assumptions

For our linear model comparison, we initially ensured that included variables demonstrated a linear relationship with the target variable, OPS, as illustrated in **Appendix A1.** Further, the response, OPS, is continuous making R^2 evaluations inadequate, therefore, performance is assessed by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Linear Regression

To construct a linear regression model effective in predicting OPS, we compared the results of two models:

- Base Model: including all 12 features considered in this analysis.
- Subset / Reduced Model: including the six most significant features identified through permutation importance shown in **A2.**

Results:

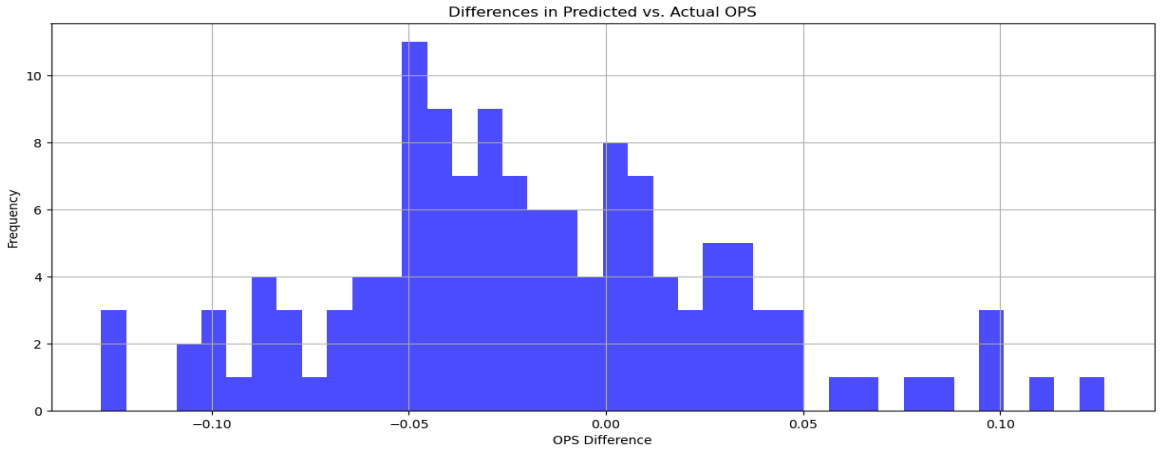
BASE MODEL			
TRAINING DATA		**VALIDATION DATA**	
ERROR METRIC		ERROR METRIC	
RMSE	.058858	RMSE	.063538
MAE	.046586	MAE	.052186
SUBSET REDUCED MODEL			
TRAINING DATA		**VALIDATION DATA**	
ERROR METRIC		ERROR METRIC	
RMSE	.218420	RMSE	.065918
MAE	.047707	MAE	.054388

Table 2: Linear Regression Results; Train and Val

BASE MODEL	
TESTING DATA	
ERROR METRIC	
RMSE	.052448
MAE	.0422228

Table 3: Linear Regression Results; Test

Testing on 2023 Data; Linear Regression



Lasso and Ridge Regression

Utilizing 'GridSearch' with 5-fold cross validation, we investigate two different models:

- Lasso: optimal alpha level = .001
- Ridge: optimal alpha level = 4.94
- Both Lasso and Ridge found the same six most significant features shown in **A3.**

Results:

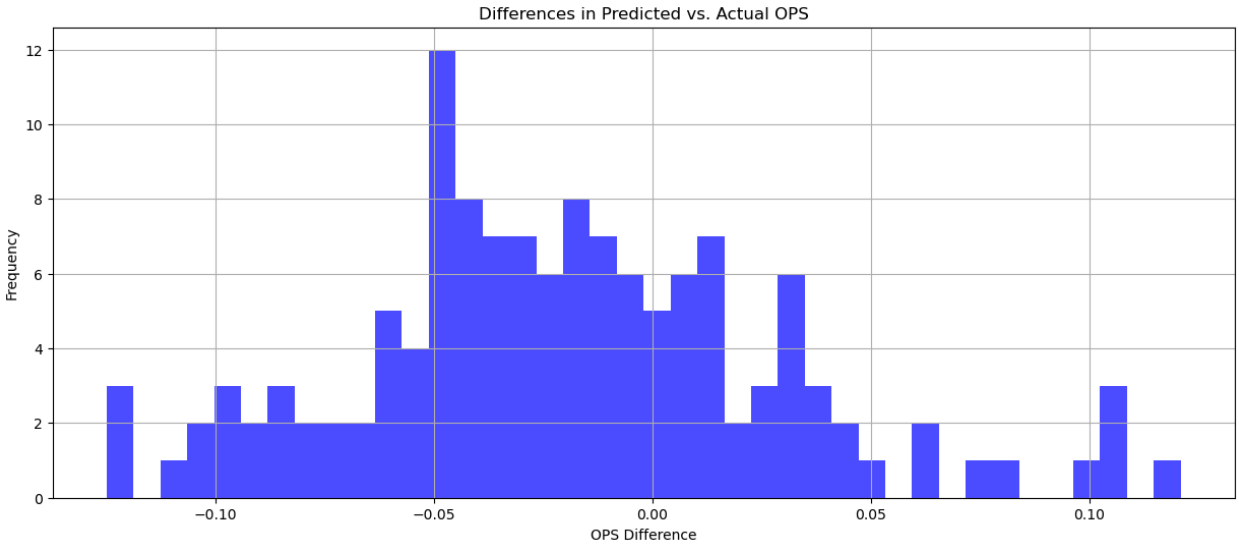
LASSO MODEL			
TRAINING DATA		**VALIDATION DATA**	
ERROR METRIC		ERROR METRIC	
RMSE	.059091	RMSE	.063316
MAE	.046518	MAE	.052077
RIDGE MODEL			
TRAINING DATA		**VALIDATION DATA**	
ERROR METRIC		ERROR METRIC	
RMSE	.059088	RMSE	.063589
MAE	.046727	MAE	.052316

Table 4: Lasso and Ridge Regression Results; Training and Val

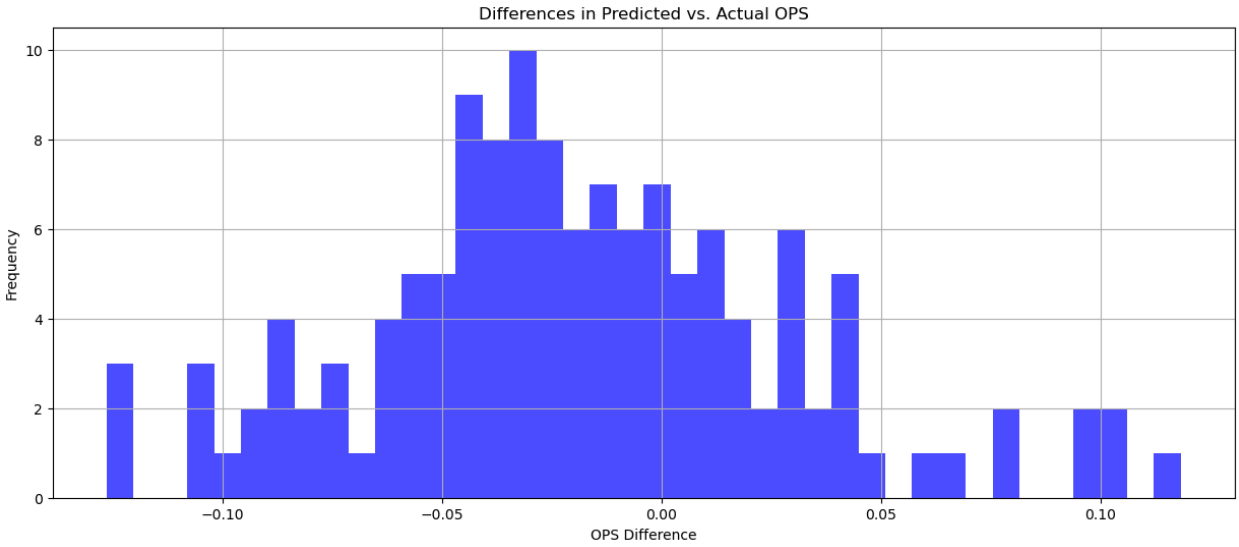
LASSO MODEL	
TESTING DATA	
ERROR METRIC	
RMSE	.051917
MAE	.041456
RIDGE MODEL	
TESTING DATA	
ERROR METRIC	
RMSE	.052425
MAE	.041925

Table 5: Lasso and Ridge Regression Results; Test

Testing on 2023 Data; Ridge



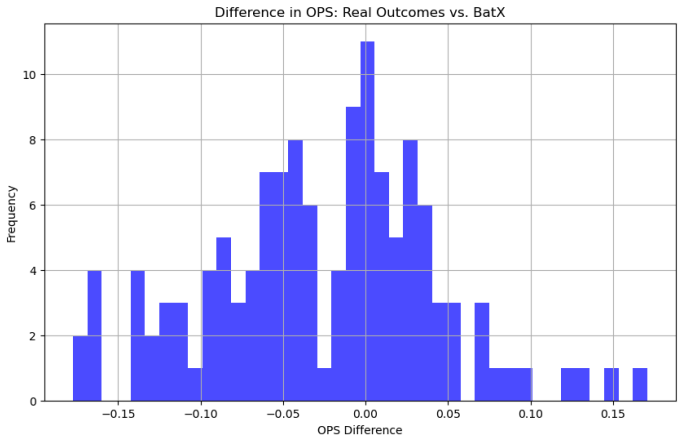
Testing on 2023 Data; Lasso



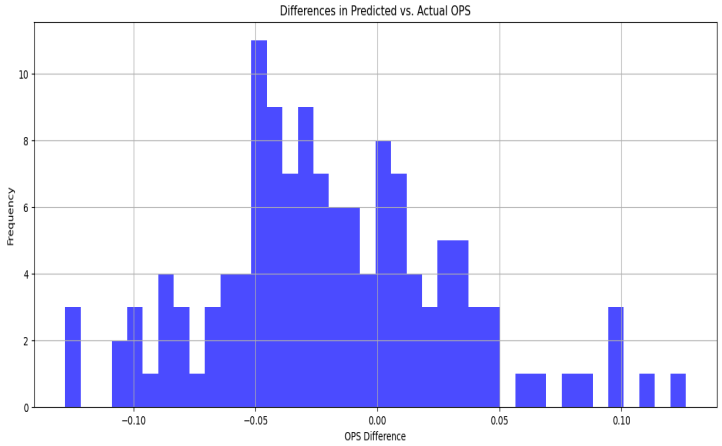
Linear Model Outcomes

Lasso Regression generated the highest predictive performance of all linear models, therefore we compare its predictive ability to professional OPS predictions generated by **BatX.** The Lasso model's predictions were similar to BATX's, on average predicting a **.026 higher** OPS.

BATX compared vs Actual Outcomes



Lasso model vs Actual



Outcomes

On average BATX underpredicted OPS by .027 and Lasso model overpredicted OPS on average by .042

Classification Model Experiments

A logistic regression model for classifying MLB as 'At Least Average' or 'Below Average' categories given their OPS, inspired by Bill James' OPS scale **(A5)**, we established a classification task based on the table below and resolves a class imbalance using 'RandomOverSampler'.

Category	Classification	OPS range
1	At least Average	.7000 and higher
0	Below average	.6999 and lower

- Base Model: including all 12 features considered in this analysis
- Grid Search Model
- Subset Model: including only the six most significant features as identified through permutation importance analysis (**A4**.)

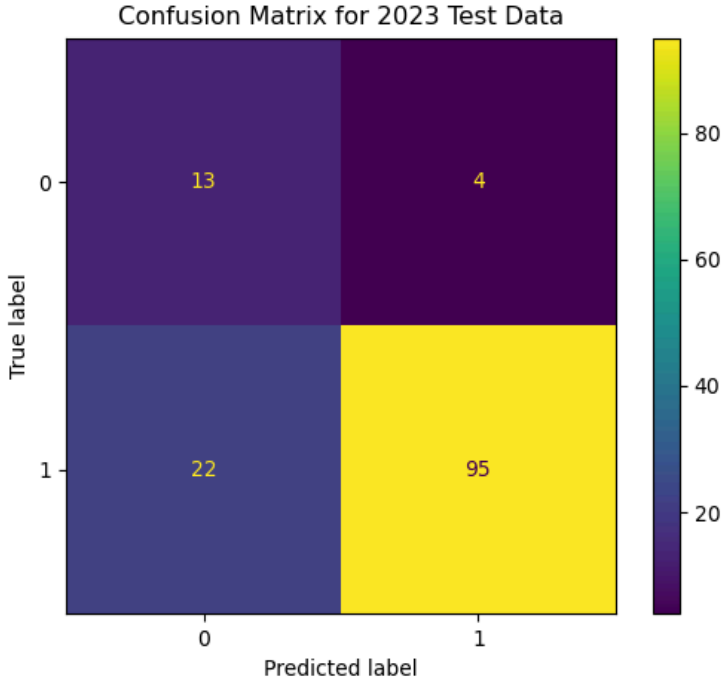
Results: (Performance was the best on the full model, so such model is used in testing.)

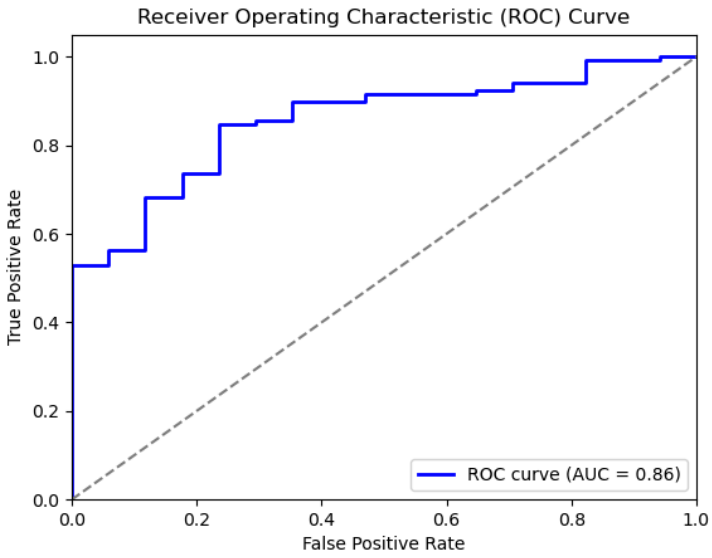
BASE MODEL	
Training Accuracy	.779
Validation Accuracy	.792
Validation Precision	.837
GRID SEARCH: BASE MODEL	
Training Accuracy	.804
Validation Accuracy	.785
Validation Precision	.834
SUBSET MODEL: Feature Importance	
Training Accuracy	.789
Validation Accuracy	.777
Validation Precision	.831

Table 7: Logistic Regression Training and Validation Results; Base, Grid Search, and Subset Model

BASE MODEL	
Testing Accuracy	.806
Testing Precision	.885
AUC	.86

Table 8: Logistic Regression Results; Test





Limitations

Limitations we came across are due to the limited scope of OPS such as not considering pitcher skill, ballpark effect and player health- all of which affect performance and therefore, OPS. Additionally, OPS does not consider league-wide changes, not minor-league minor-league performance data. In this way, disabling new player performance predictions.

Conclusion

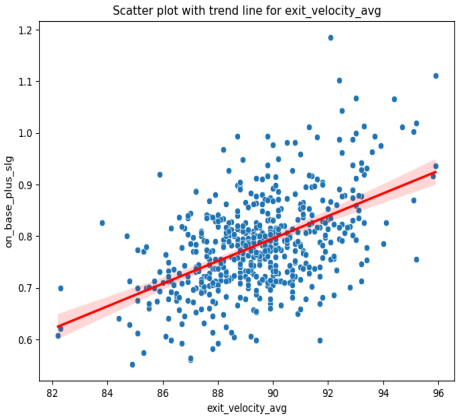
Metrics relating contact quality are crucial predictors of a player's offensive value, measured by OPS. Lasso regression tended to overpredict 2023 OPS measures by an average of 0.042, while logistic regression demonstrated strong performance with an AUC of 0.86. Compared to BATX, our model projected OPS to be 0.026 higher on average, suggesting competitive performance to state-of-the-art models, but indicating potential for improvement.

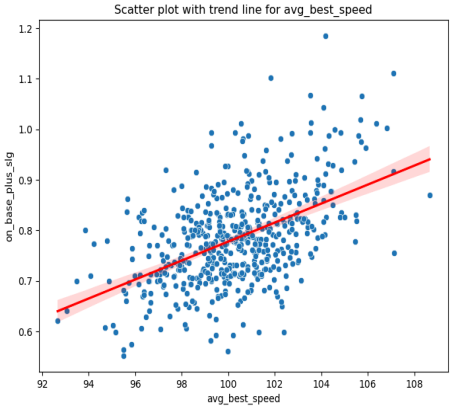
Member	Proposal	Coding	Presentation	Report
Eugene	1	1	1	1
Jon	1	1	1	1
Max	1	1	1	1
Clarity	1	1	1	1

Table 1: Contributions

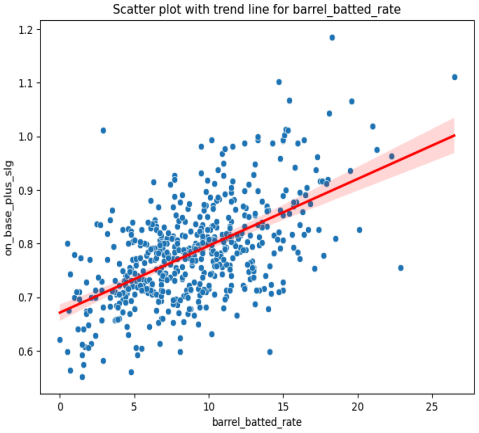
APPENDIX

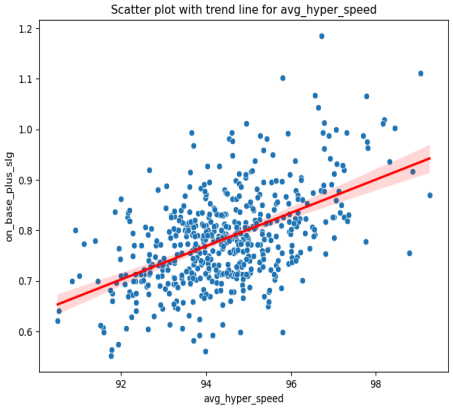
A1)

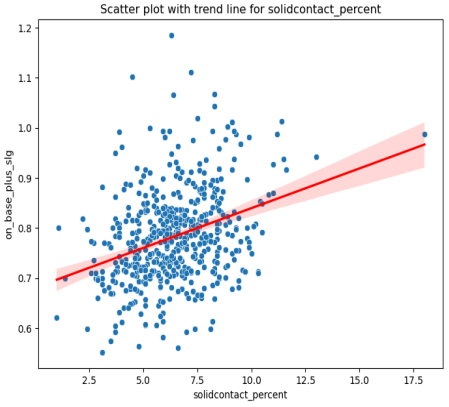

Scatter plot with trend line for exit_velocity_avg
on_base_plus_slg
exit_velocity_avg

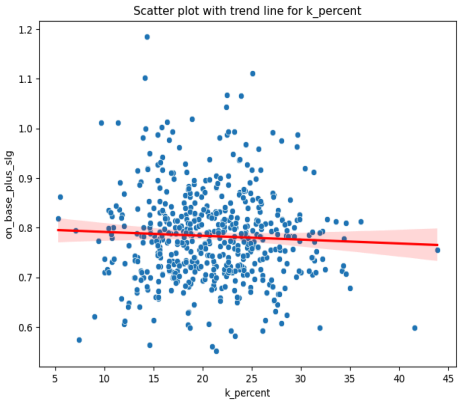

Scatter plot with trend line for avg_best_speed
on_base_plus_slg
avg_best_speed

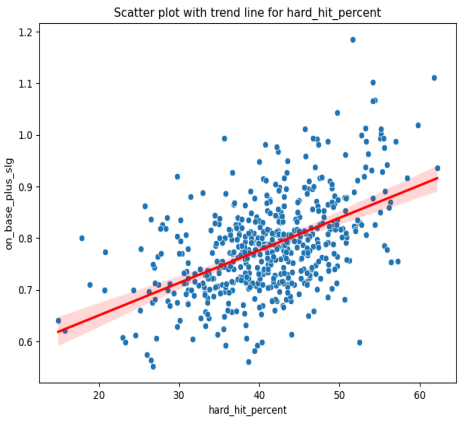

Scatter plot with trend line for sweet_spot_percent
on_base_plus_slg
sweet_spot_percent

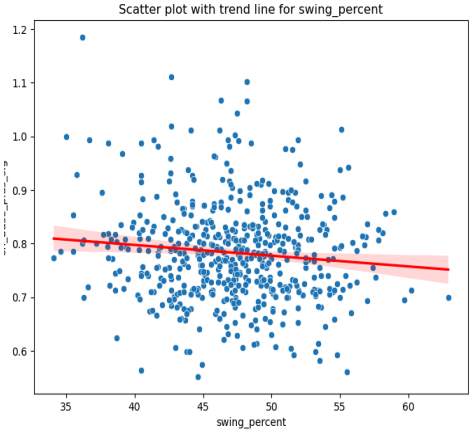

Scatter plot with trend line for barrel_batted_rate
on_base_plus_slg
barrel_batted_rate

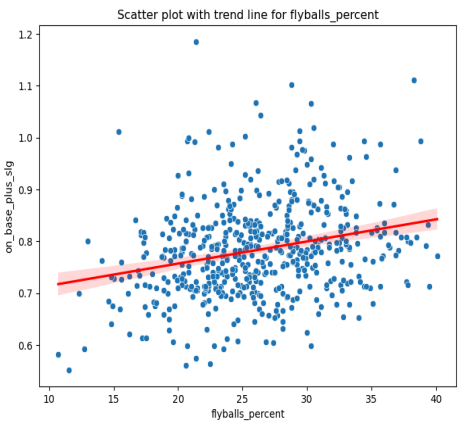

Scatter plot with trend line for avg_hyper_speed
on_base_plus_slg
avg_hyper_speed

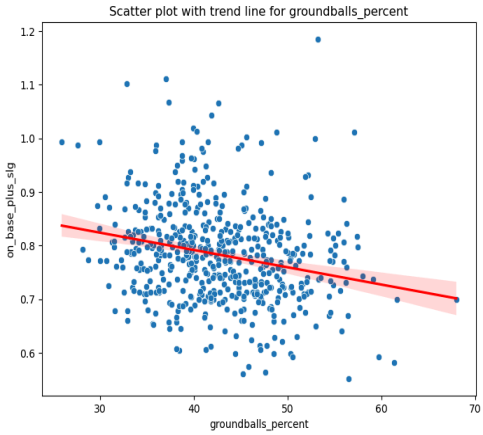

Scatter plot with trend line for solidcontact_percent
on_base_plus_slg
solidcontact_percent

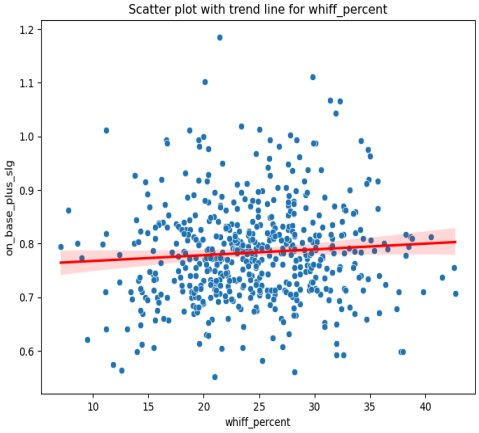

Scatter plot with trend line for k_percent
on_base_plus_slg
k_percent


Scatter plot with trend line for hard_hit_percent
on_base_plus_slg
hard_hit_percent

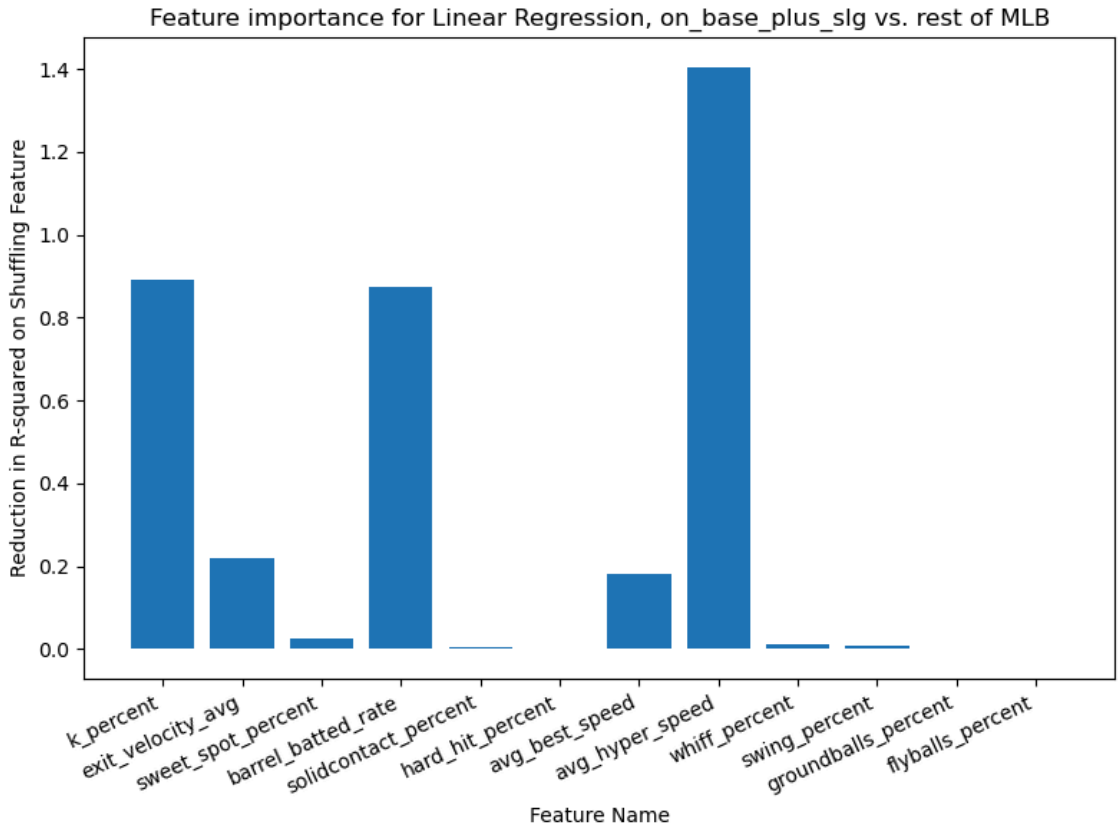

Scatter plot with trend line for swing_percent
swing_percent


Scatter plot with trend line for flyballs_percent
on_base_plus_slg
flyballs_percent

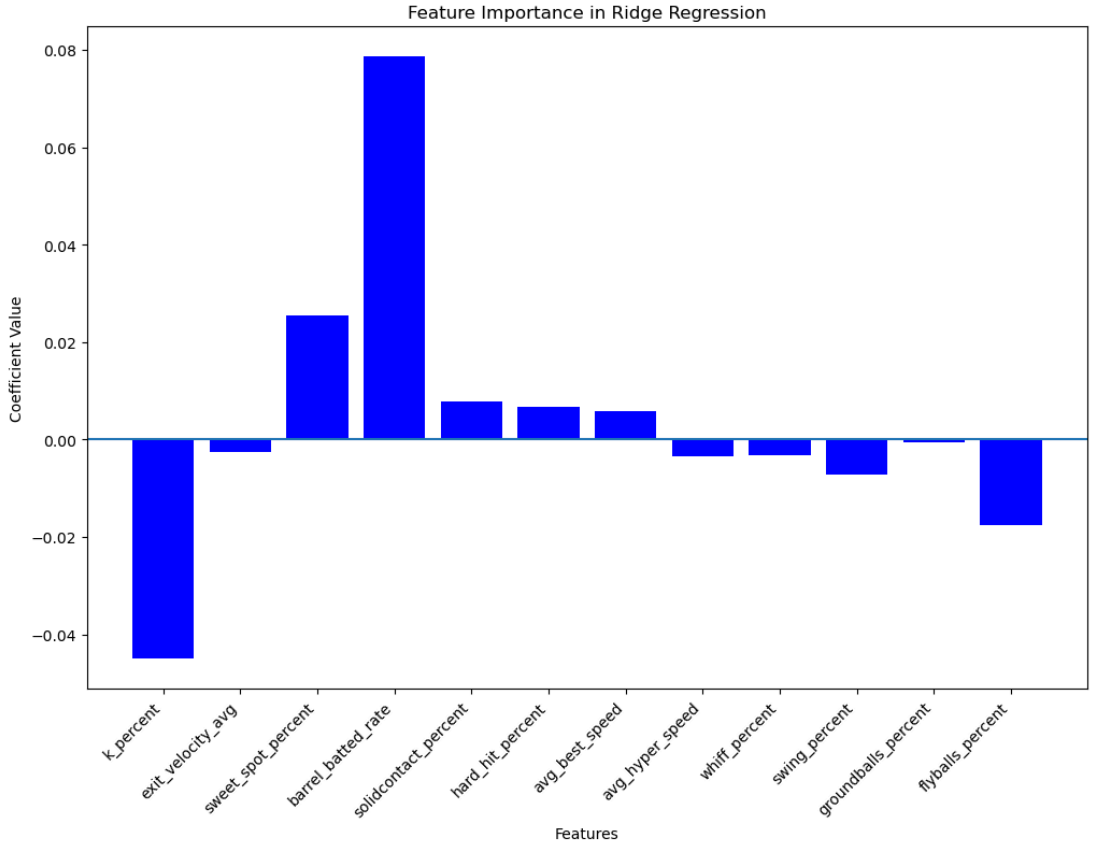

Scatter plot with trend line for groundballs_percent
on_base_plus_slg
groundballs_percent

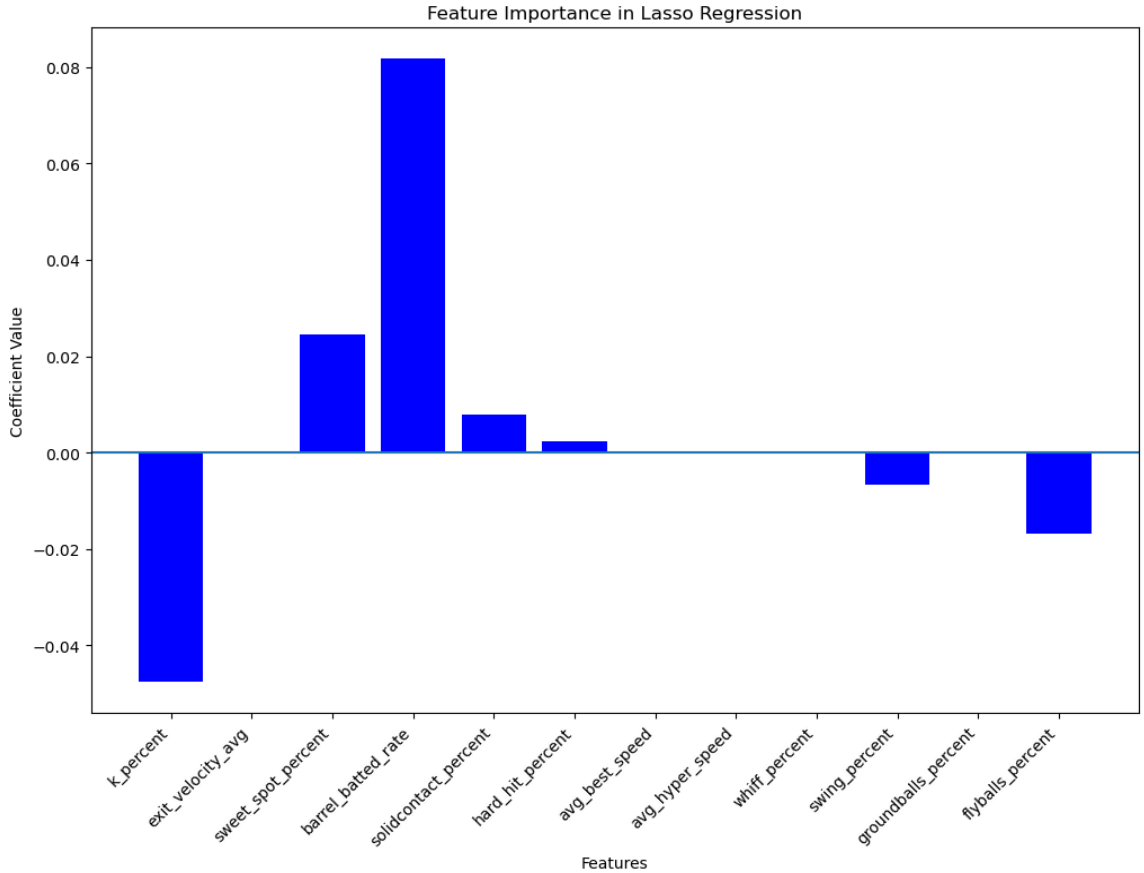

Scatter plot with trend line for whiff_percent
on_base_plus_slg
whiff_percent

A2)

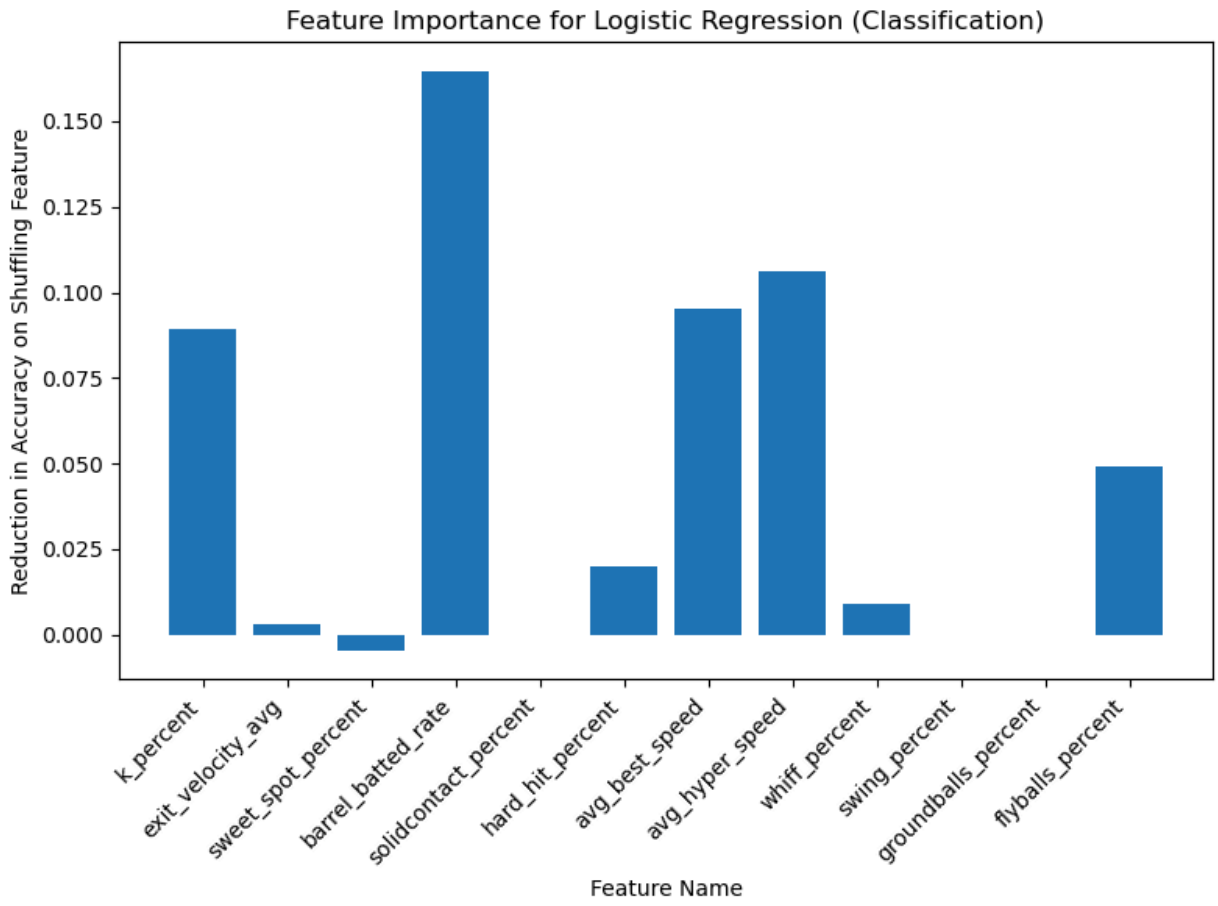

Feature importance for Linear Regression, on_base_plus_slg vs. rest of MLB
Reduction in R-squared on Shuffling Feature
k_percent
exit_velocity_avg
sweet_spot_percent
barrel_batted_rate
solidcontact_percent
hard_hit_percent
avg_best_speed
avg_hyper_speed
whiff_percent
swing_percent
groundballs_percent
flyballs_percent
Feature Name

A3)


Feature Importance in Ridge Regression
Coefficient Value
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
k_percent
exit_velocity_avg
sweet_spot_percent
barrel_batted_rate
solidcontact_percent
hard_hit_percent
avg_best_speed
avg_hyper_speed
whiff_percent
swing_percent
groundballs_percent
flyballs_percent
Features


Feature Importance in Lasso Regression
Coefficient Value
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
k_percent
exit_velocity_avg
sweet_spot_percent
barrel_batted_rate
solidcontact_percent
hard_hit_percent
avg_best_speed
avg_hyper_speed
whiff_percent
swing_percent
groundballs_percent
flyballs_percent
Features

A4)



A5) https://en.wikipedia.org/wiki/On-base_plus_slugging

Category	Classification	OPS range
A	Great	.9000 and higher
B	Very good	.8334 to .8999
C	Above average	.7667 to .8333
D	Average	.7000 to .7666
E	Below average	.6334 to .6999
F	Poor	.5667 to .6333
G	Very poor	.5666 and lower