Baseball Statistical Analysis

Eugene Ohba, Jon Karanezi, Clarity Kummer, Max Gehred

Can we accurately predict or classify the offensive value of an MLB player as measured by On Base Percentage Plus Slugging Percentage (OPS)?

Dataset

“MLB Statcast” Dataset made publicly via Baseball Savant.

- **Dataset:**
  - All batters in the MLB over the four years 2020-2023.
    - Reason for choosing four years: Training (2020,2021), Validation (2022), Test (2023)
  - 538 observations on 23 features.

- **What does the dataset look like?**
  - **Player Info:** First and Last Name, Player ID, Plate Appearances
  - **At Bat Outcome Metrics:** Hits, Single, Double, Triple, Home Run, Strike Out Percentage, Walk Percentage, On Base Percentage Plus Slugging Percentage
  - **At Bat Quality Metrics**: Average Exit Velocity, Sweet Spot Percentage, Barrel Batted Rate, Solid Contact Percentage, Hard Hit Percentage , Average Best Speed and Hyper Speed, Whiff Percentage, Swing Percentage, Ground Ball Percentage, Flyball Percentage

Goal

- **Goal**: To create a model that accurately predicts on-base percentage plus slugging percentage, OPS

- **OPS**: provides a comprehensive overview of a player's offensive value. It’s calculated by adding the player’s On Base Percentage and their Slugging Percentage

  - **On Base Percentage (OBP) :** the rate at which a player gets on base
  - **Slugging Percentage (SLG) :** the average number of bases a player records per at bat

OPS and Variable Selection

- OPS = On Base Percentage (OBP) + Slugging Percentage (SLG)

- SLG = (hits+ (doubles *2)+ (triples*3) + (HRs*4)) / at bats
- OBP = (hits + walks + hit by pitch) / (Plate Appearances)
- Ex: 5 plate appearances - Out, Out, Single, Home run, Out = 1.400 OPS

- To reduce multicollinearity,we chose predictor features that are more directly related to OBP and SLG but are not already included in the OPS calculation. The 12 non-outcome oriented features considered are:
    - Strikeout percentage, Exit Velocity Average, Sweet Spot Percentage, Barrel Batted Rate, Solid Contact Percentage, Hard Hit Percentage, Average Best and Average Hyper Speed, Whiff Percentage, Swing Percentage, Ground Ball Percentage, Flyball Percentage

https://towardsdatascience.com/stats-for-baseball-fans-the-single-metric-for-offense-is-ops-fc568af5e87b

Questions of Interest

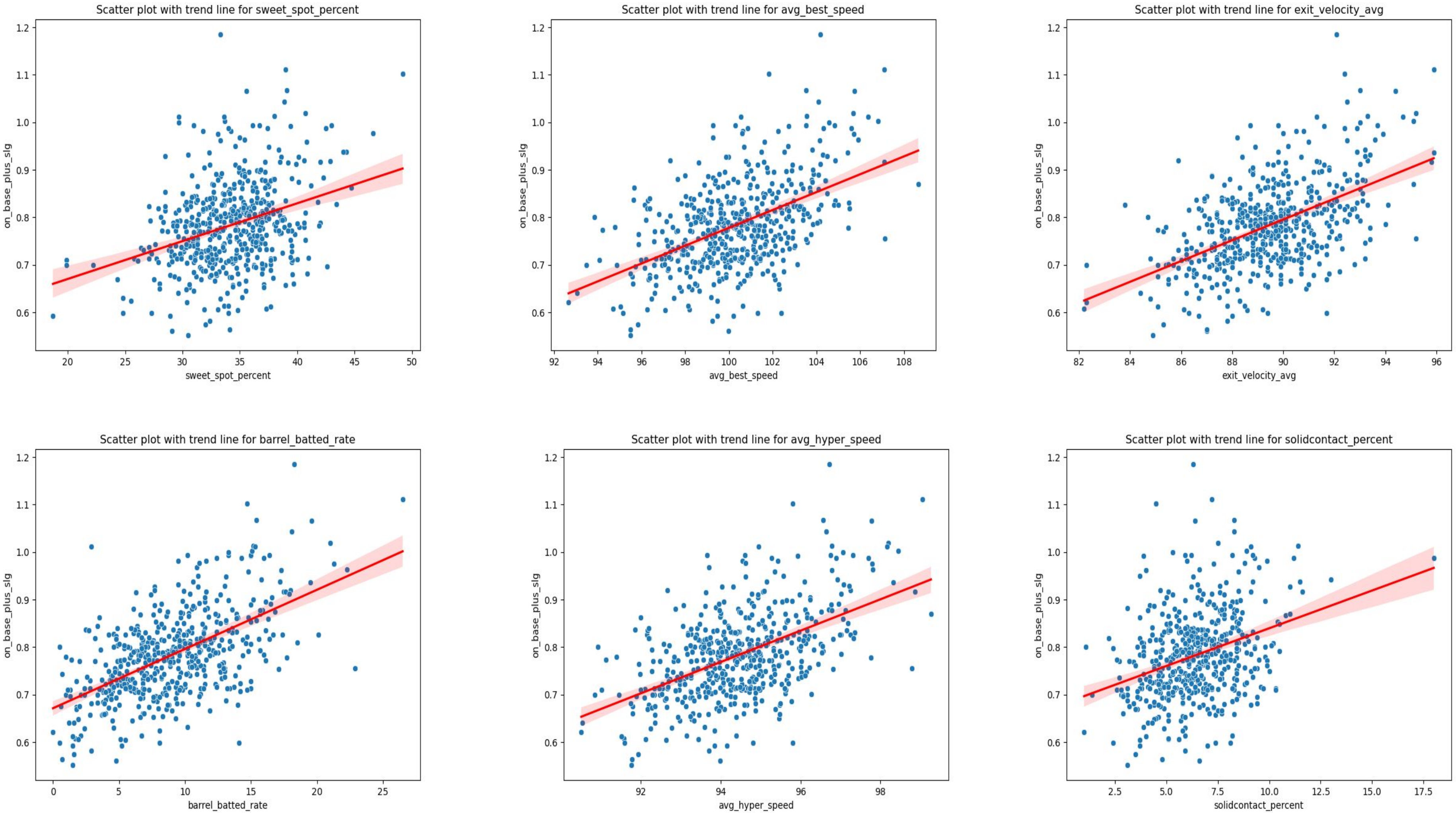
- Which features (of those considered in this analysis) are most predictive and significant on a player's offensive value, as measured by OPS?

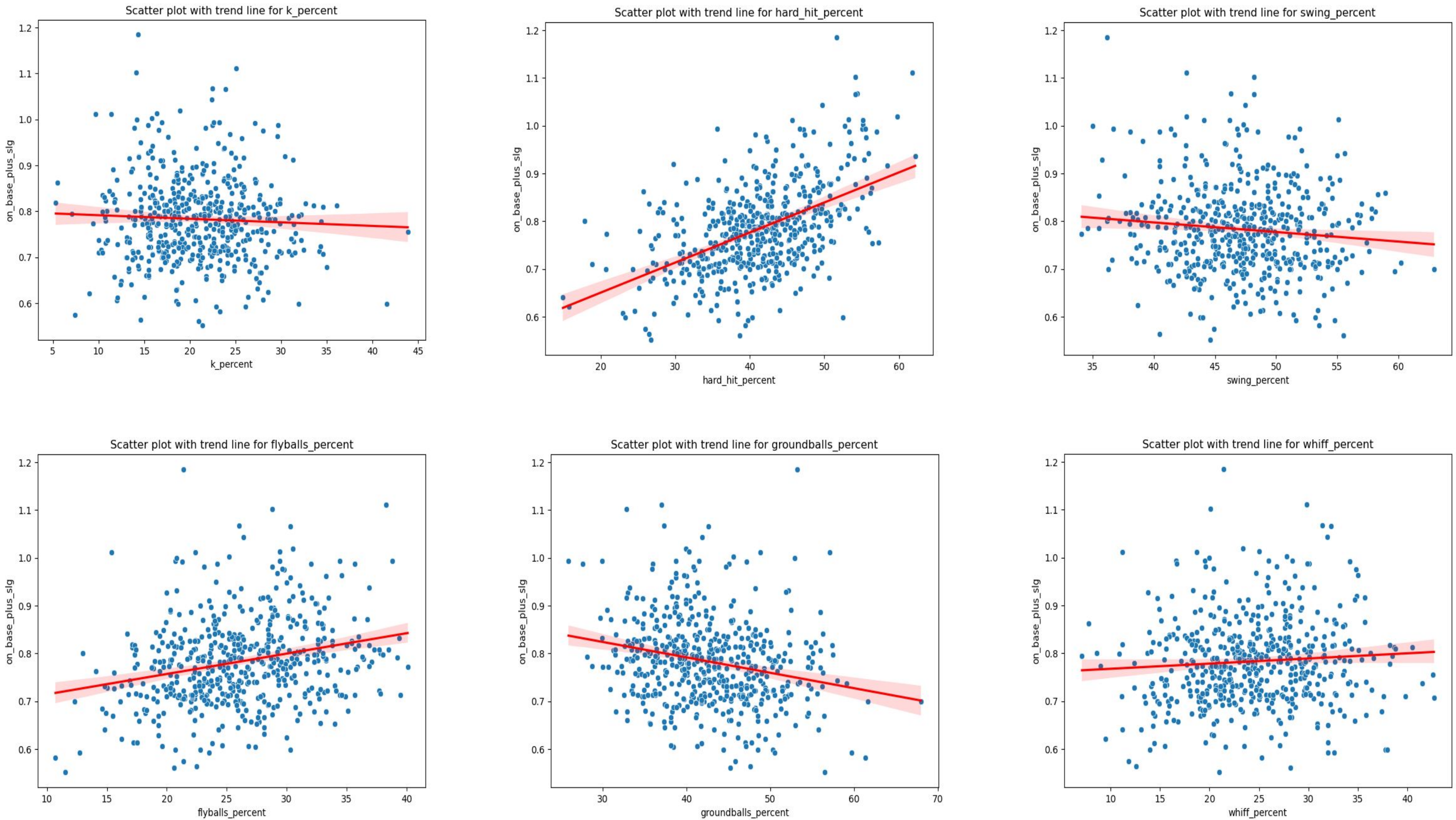
- How closely can non-outcome oriented features predict OPS?

- How does our predictive model compare to other professional projection systems.

Methodologies - Initial Data Discoveries

- **Data Collection :** Compile a baseball hitters dataset, made publicly available by MLB Baseball Statcast over the years 2020-2023, ensure representation across a diverse range of players.
- **Data Preprocessing :** Clean the data for missing values, outliers, duplicate observations, and ensure normalization of features when necessary. Additionally, ensure each feature included is relevant to the task at hand.
- **Exploratory Data Analysis :** Visualize the datasets and subsets of such to identify correlations and underlying relationships present within and between attributes. Additionally, analyze the distribution of each feature and the target variable, OPS
- **Regression Model Comparison:** Compare the results of 3 different linear models: Linear, Ridge and Lasso Regression, choosing the one that best predicts OPS to use on test data, for analysis and interpretation
- **Classification:** Create a classification model that can be used to classify hitters as at least average or below average, based on OPS


Scatter plot with trend line for sweet_spot_percent
on_base_plus_slg
sweet_spot_percent
Scatter plot with trend line for avg_best_speed
on_base_plus_slg
avg_best_speed
Scatter plot with trend line for exit_velocity_avg
on_base_plus_slg
exit_velocity_avg
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

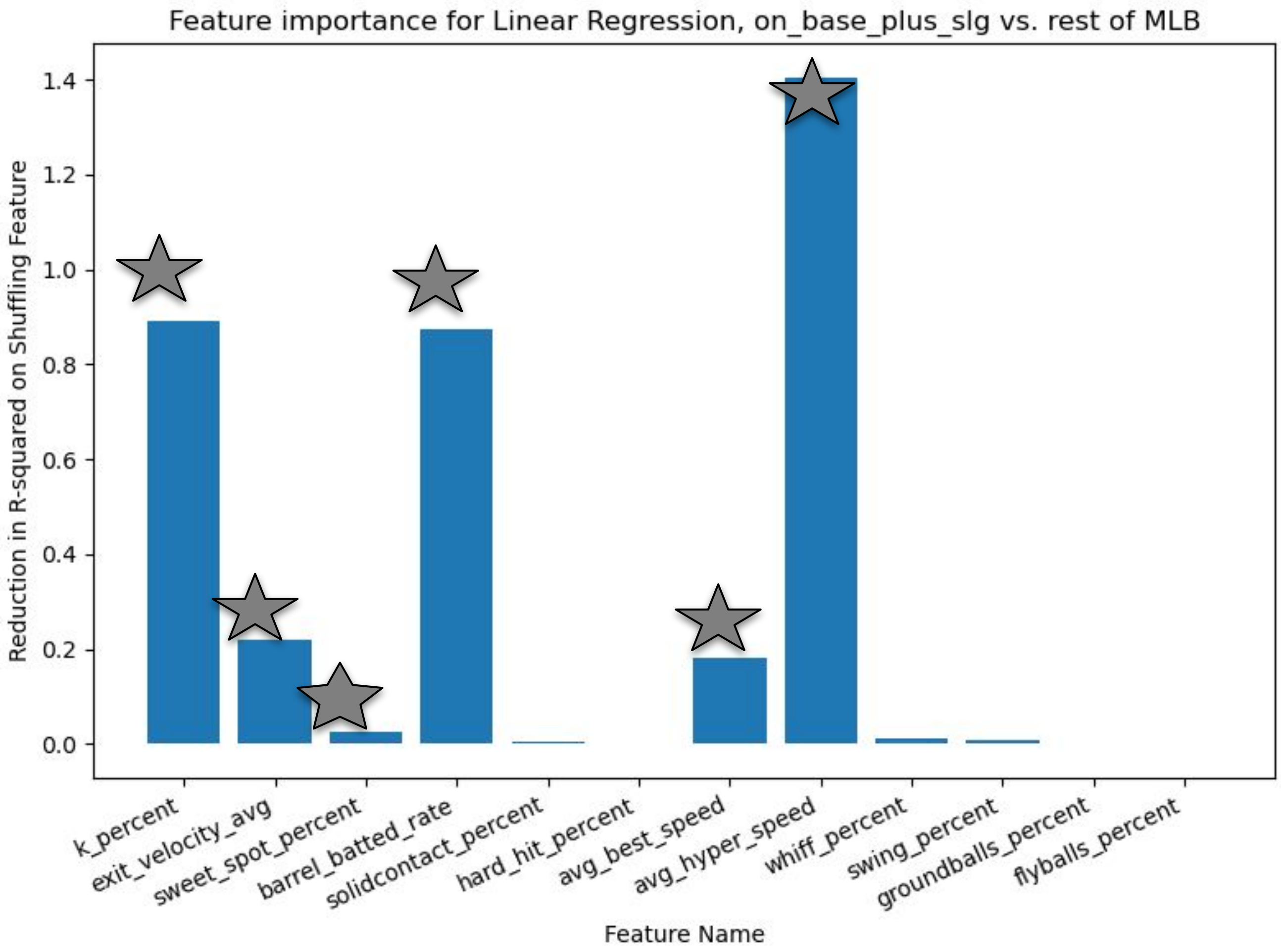


Scatter plot with trend line for k_percent
on_base_plus_slg
k_percent
Scatter plot with trend line for hard_hit_percent
on_base_plus_slg
hard_hit_percent
Scatter plot with trend line for swing_percent
on_base_plus_slg
swing_percent
Scatter plot with trend line for flyballs_percent
on_base_plus_slg
flyballs_percent
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

Linear Regression Methodologies:

- **Create Base Model:**
  - Including variables which demonstrate a linear relationship with OPS
    - (12 features included shown on previous slides)
  - Standardized Data using StandardScaler()
  - Evaluate Performance based error metrics:
    - Mean Absolute Error (MAE)
    - Root Mean Squared Error (RMSE)
    - Accuracy as a measure of R^2 is not as good of a performance measure– response is continuous

- **Feature Permutation for Feature Importance**
  - Using ‘permutation_importance’ to identify a subset of features that contribute most to the model’s performance
  - Create a subset model using only the most important features
  - Compare the performance of the subset model to the base (full) model
  - In this way, we are able to identify the fewest number of features that account for predicting a players OPS



6 most important features identified:

- k_percent
- exit_velocity_avg
- sweet_spot_percent
- barrel_batted_rate
- avg_best_speed
- avg_hyper_speed

Performance Linear Train / Val

BASE (FULL) MODEL

R^2 train = .637

Metric	Value
RMSE TRAIN	0.058858
MAE TRAIN	0.046586

Table 1: Performance Training Data; Base (Full) Model

R^2 val = .417

Metric	Value
RMSE VAL	0.063538
MAE VAL	0.052186

Table 1: Performance Validation Data; Base (Full) Model

SUBSET (REDUCED) MODEL– by feature permutation for feature

Metric	Value
RMSE TRAIN	0.218420
MAE TRAIN	0.047707

Table 1: Performance Training Data; Subset (Reduced) Model

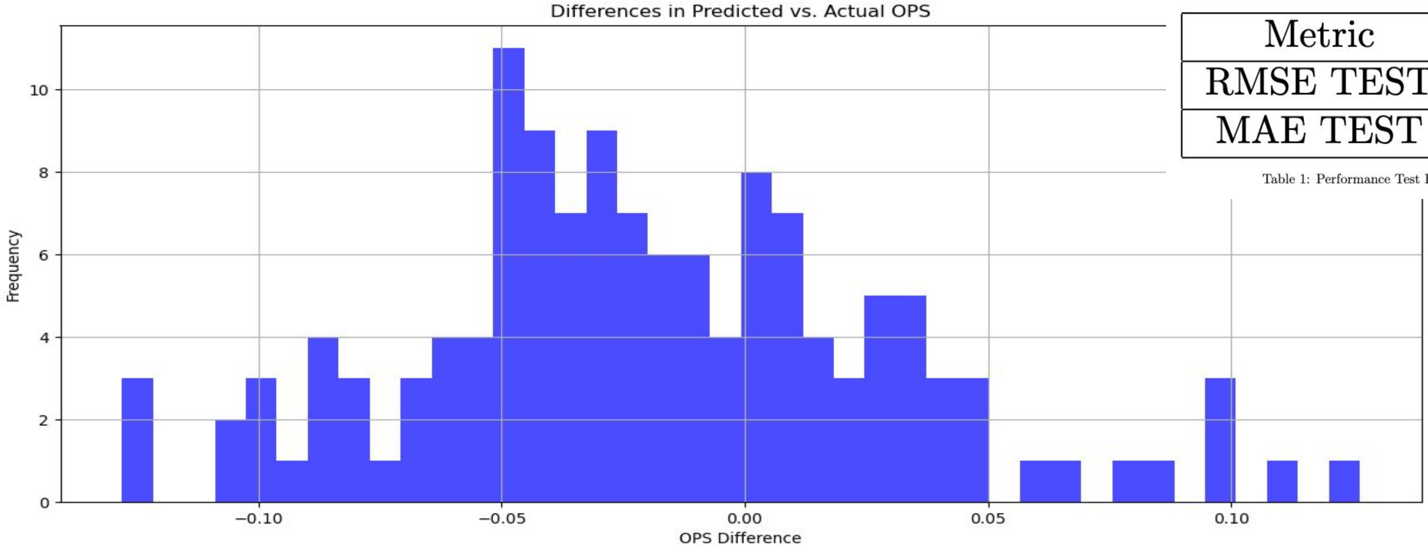
R^2 train = .609

Metric	Value
RMSE VAL	0.065918
MAE VAL	0.054388

Table 1: Performance Validation Data; Subset (Reduced) Model

R^2 val = .373

Outcomes - Linear Regression Model on 2023 Data



Metric	Value
RMSE TEST	.052448
MAE TEST	.0422228

Table 1: Performance Test Data; Full (Base) Model

R^2 test = .585

Outcomes:

| **Mean Absolute Error**: .042 | **Mean Squared Error**: .003 | **Root Mean Squared Error**: .052 |

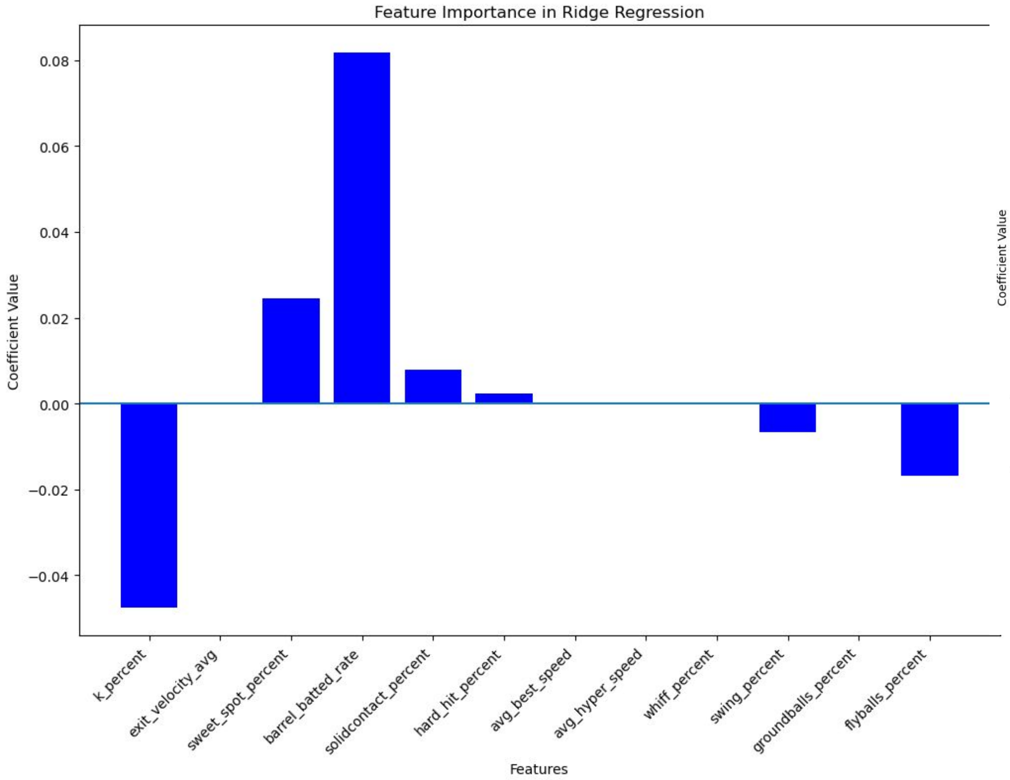
Our model slightly overestimated OPS of 2023 Hitters

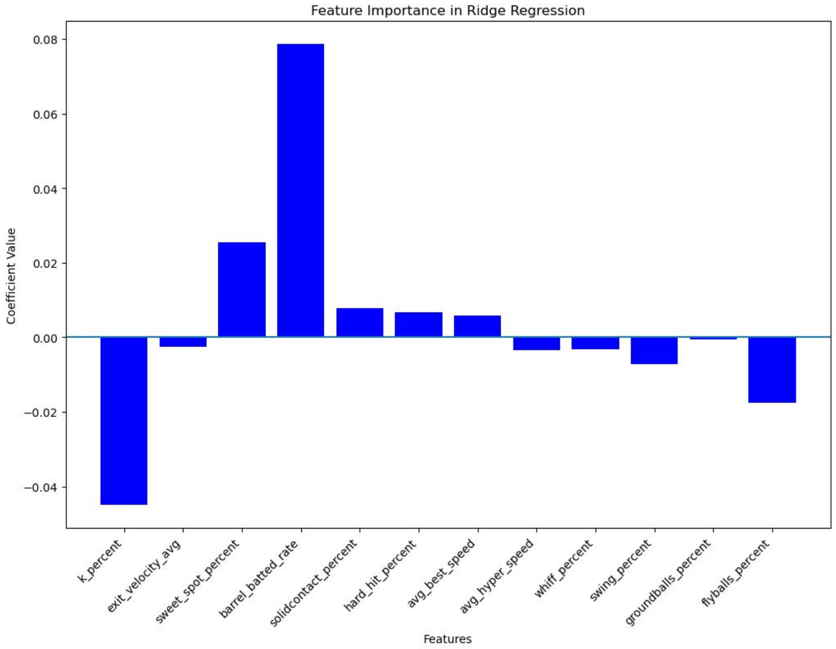
Ridge Regression

- **Create Base Model:**
  - **Features:** Used the same 12 linear features
  - **Standardization:** Utilized standardscaler() to standardize data
  - **Cross-validation:** Utilized 5 folds

- **Ridge for Feature Importance**
  - **Grid Search:** We have used grid search to find the optimal alpha level for our ridge regression.
  - **Optimal alpha level:** 4.94
  - Create a subset model using only the most important features
  - Compare the performance of the subset model to the base (full) model
  - In this way, we are able to identify the fewest number of features that account for predicting a players OPS

Ridge Feature Selection





Performance Ridge Train / Val

Metric	Value
RMSE TRAIN	.059088
MAE TRAIN	.046727

Table 1: Performance Training Data; Ridge Model

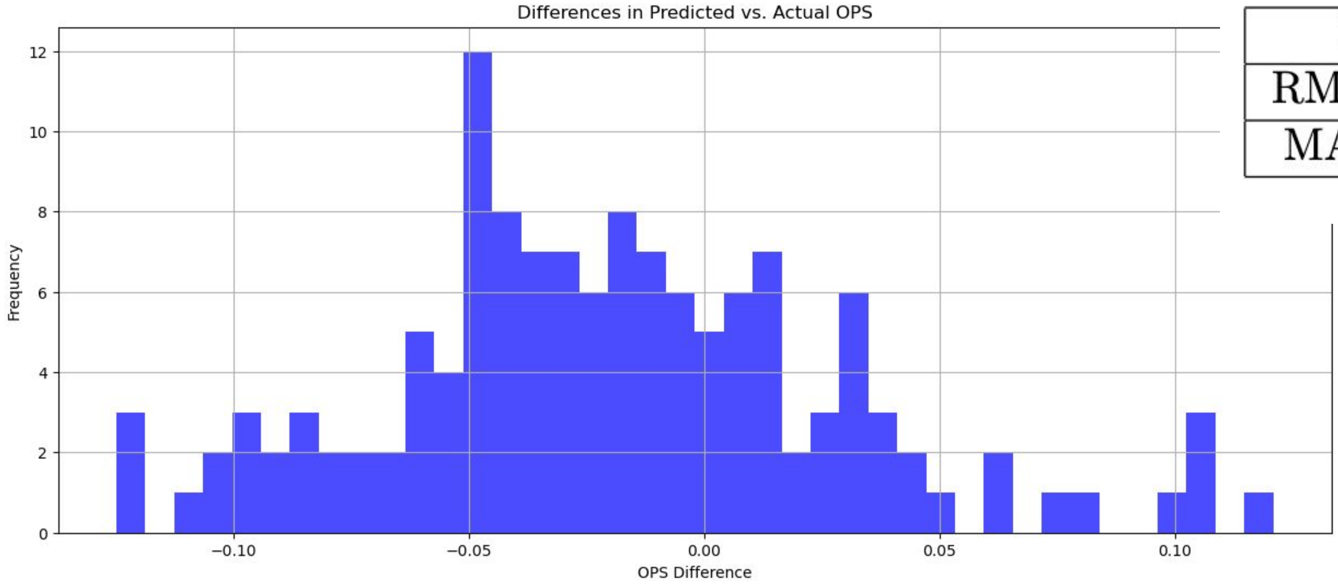
Metric	Value
RMSE VAL	.063589
MAE VAL	.052316

Table 1: Performance Validation Data; Ridge Model

R^2 train = .635

R^2 val = .416

Outcomes - Ridge Regression Model on 2023 Data



Metric	Value
RMSE TEST	.052425
MAE TEST	.041925

Table 1: Performance Test Data; Ridge Model

R^2 test = .586

Outcomes:

| **Mean Absolute Error**: .042 | **Mean Squared Error**: .003 | **Root Mean Squared Error**: .052 |

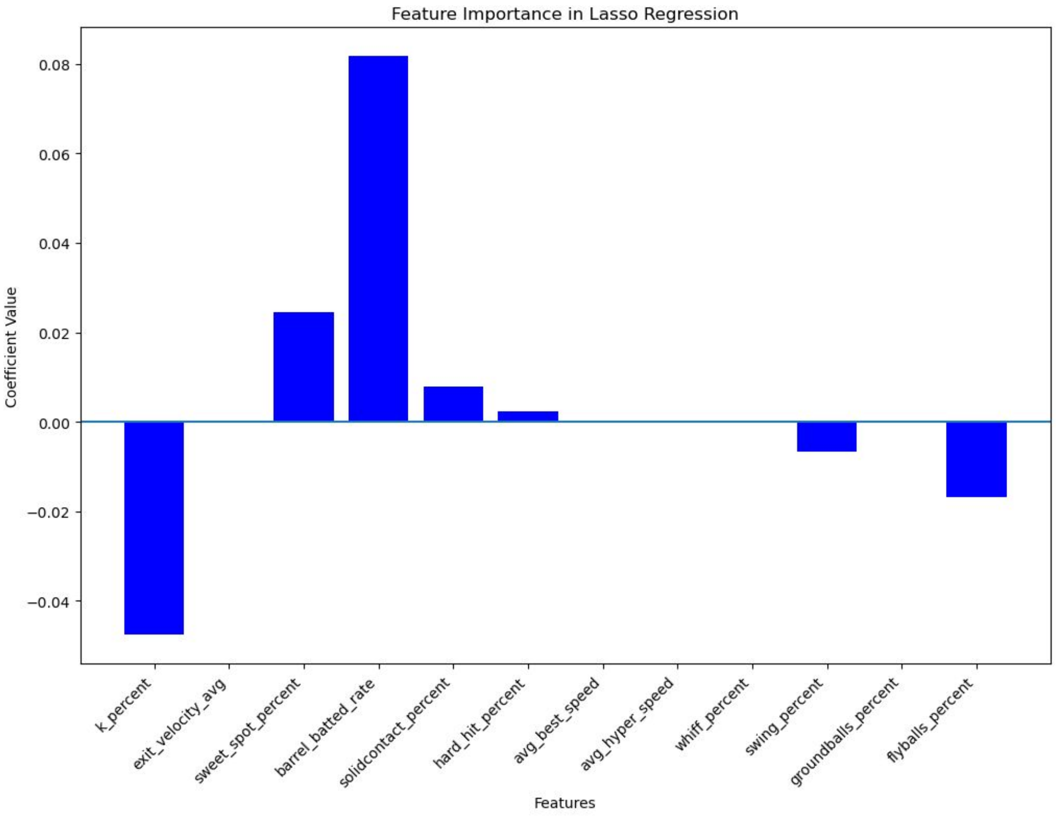
Our model slightly overestimated OPS of 2023 Hitters

Lasso

- **Create Base Model:**
    - **Features:** Used the same 12 linear features
    - **Standardization:** Utilized standardscaler() to standardize data
    - **Cross-validation:** Utilized 5 folds

- **Ridge for Feature Importance**
    - **Grid Search:** We have used grid search to find the optimal alpha level for our ridge regression.
    - **Optimal alpha level:** 0.0010
    - Create a subset model using only the most important features
    - Compare the performance of the subset model to the base (full) model
    - In this way, we are able to identify the fewest number of features that account for predicting a players OPS

Lasso Feature Selection



6 most important features identified:

- k_percent
- sweet_spot_percent
- barrel_batted_rate
- Solid Contact Percent
- Swing Percent
- Fly balls Percent

Performance Lasso Train / Val

Metric	Value
RMSE TRAIN	.059091
MAE TRAIN	.046518

Table 1: Performance Training Data; Lasso Model

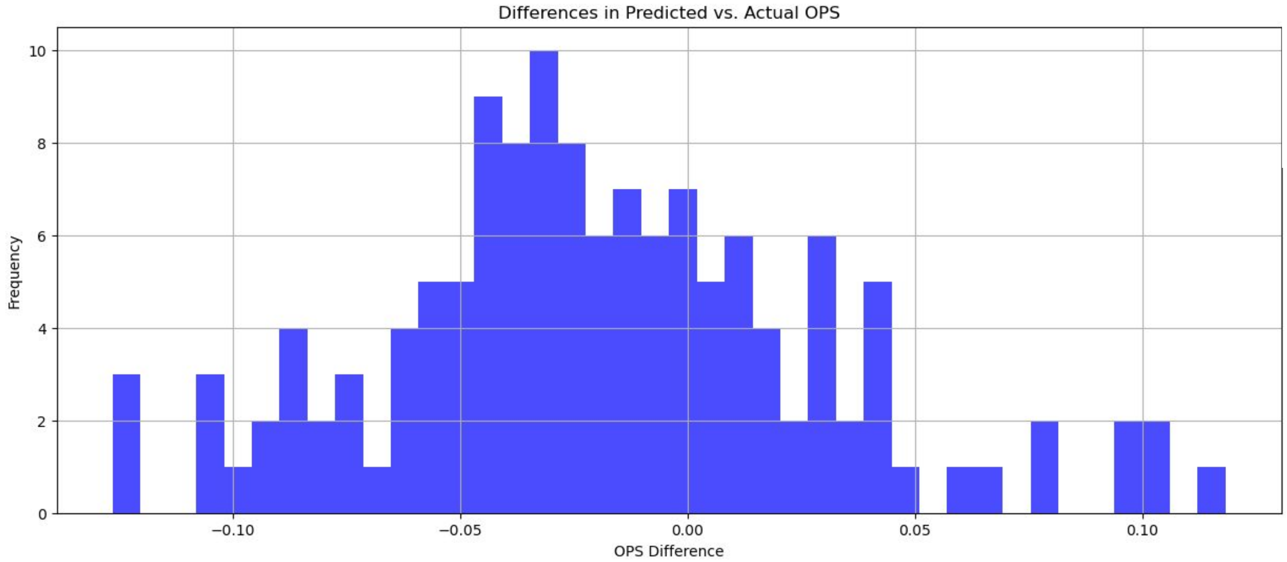
R^2 train = .634

Metric	Value
RMSE VAL	.063316
MAE VAL	.052077

Table 1: Performance Validation Data; Lasso Model

R^2 val = .421

Outcomes - Lasso Model on 2023 Data



Metric	Value
RMSE TEST	.051917
MAE TEST	.041456

Table 1: Performance Testing Data; Lasso Model

R^2 test = .594

Outcomes:

| **Mean Absolute Error**: .041 | **Mean Squared Error**: .003 | **Root Mean Squared Error**: .052 |

Our model slightly overestimated OPS of 2023 Hitters

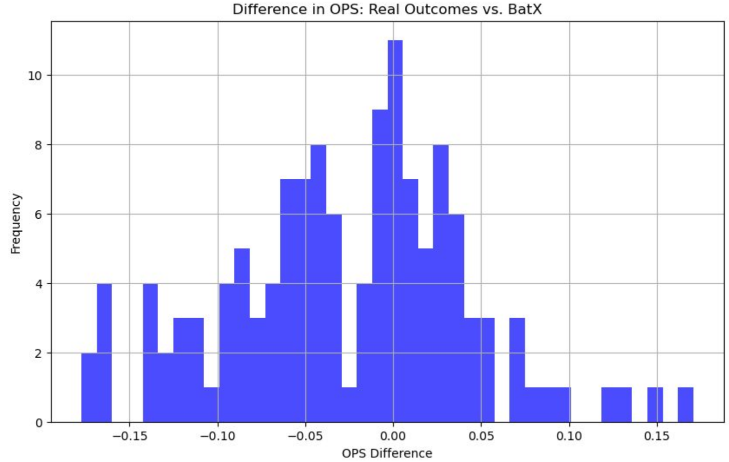
Outcomes - Comparing our Projections to State-of-the Art BATX Projections for 2023 Hitters

Our model's predictions were fairly similar to BATX's, on average we predicted a **.026 higher** OPS than BATX

BATX compared to Actual Outcomes

On average **BATX underpredicted OPS by .027**

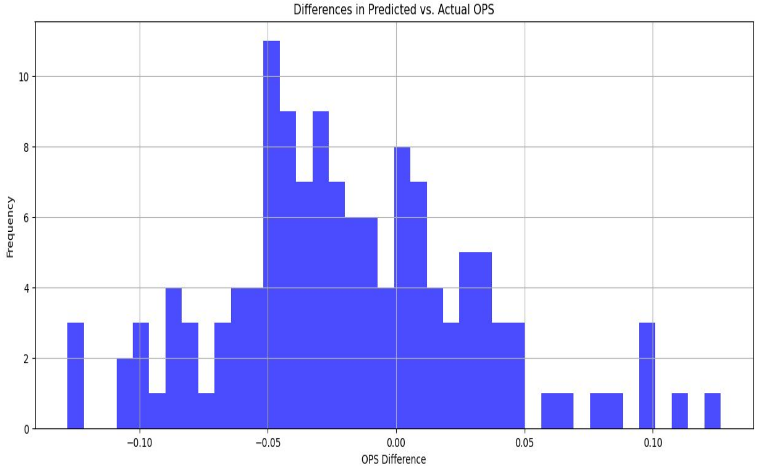
The SD between it's predicted and actual scores is .069



Our model compared to Actual Outcomes

On average **our model overpredicted OPS by .042**

The SD between our predicted and actual scores is .049



Classification Methodologies:

- Bill James, a baseball statistician, developed a comprehensive 7-point ordinal scale to categorize hitters based on OPS
- Drawing inspiration from James' scale, we've crafted a concise 2-point system for classifying hitters by OPS, distilling complex insights into accessible categories.

Bill James' Scale:

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

Scale we manually engineered:

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

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

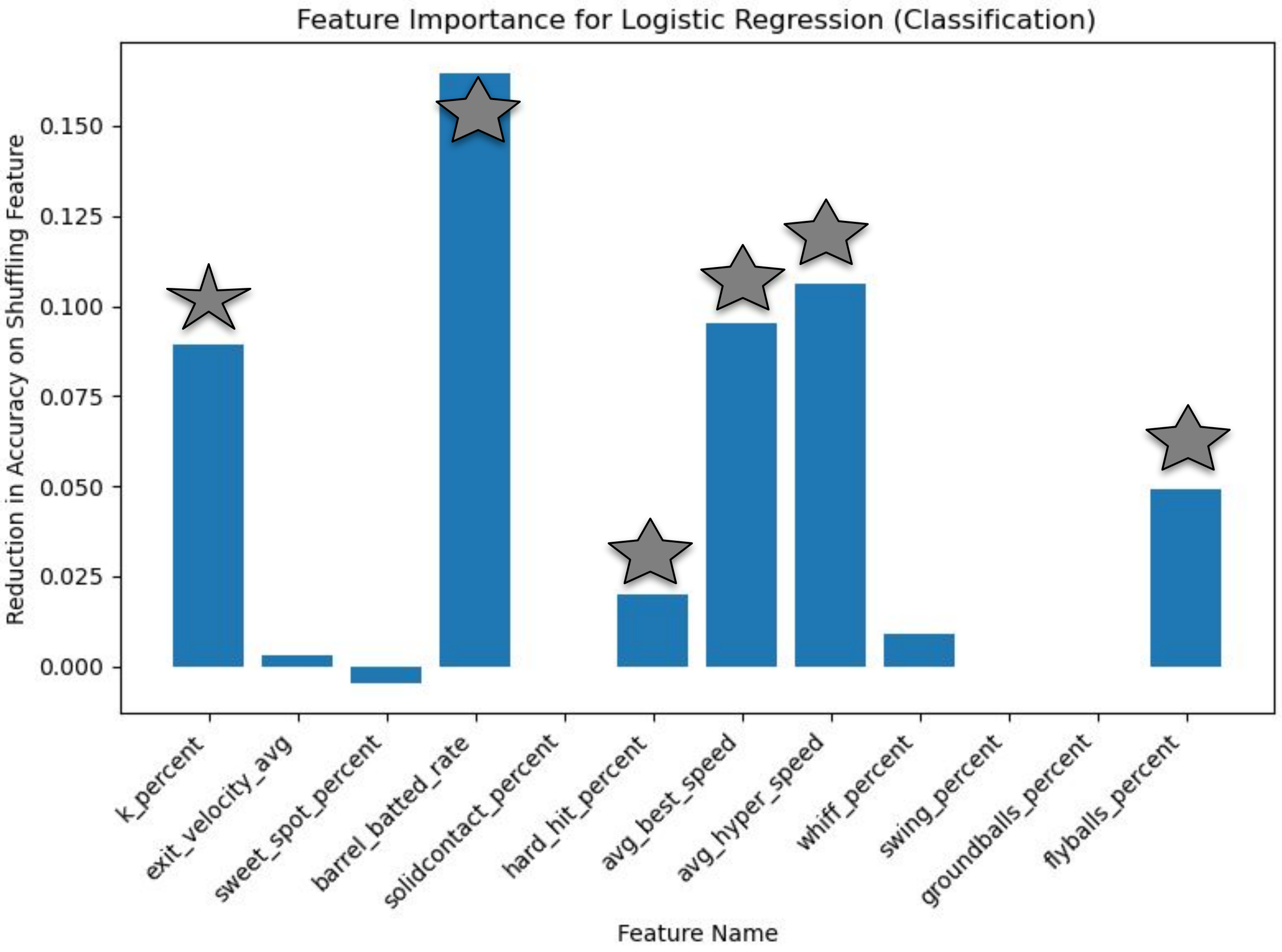
Classification Methodologies Continued

- **Created a column 'Classification'**
    - Player's whose OPS was at least average according to Bill James' scale (.7000) received a 1 in the Classification column, else for OPS values less than average (.7000) they received a 0 in the Classification columns
- **Training:**
    - As consistently throughout the analysis: Training was performed on data from the years 2020 and 2021
    - Normalized the data using StandardScaler()
    - The distribution of class observations in the training set were:
        - Counter({1: 235, 0: 39})
    - To address the class imbalance: fit RandomOverSampler() to the training data, and then resampled such data
    - After resampling (by oversampling) the distribution of class observation in the training set were:
        - Counter({1: 235, 0: 235})

Classification Methodologies

- Created a Base Model Including all 12 features using
  - Model = LogisticRegression()
- Performed Grid Search in aims to increase performance compared to Base Model
  - parameters = [{'max_iter': [1000, 5000, 10000], 'C': [0.01, 1, 10, 1000], 'penalty': ['l2']}]
- Feature Permutation For Feature Importance performed on Grid Search Model
- Compared the performance of Base Model, Grid Search Model, and Feature Permutation in order to determine which model would be best for classification tasks

(Results shown on next slide)



6 most important features identified:

- k_percent
- barrel_batted_rate
- hard_hit_percent
- avg_best_speed
- avg_hyper_speed
- Flyballs_percent

Base Model

Training Accuracy	0.779
Validation Accuracy	.792
Validation Precision	.837

Table 1: Performance Base (Full) Model

Grid Search

Training Accuracy	0.804
Validation Accuracy	.785
Validation Precision	.834

Table 1: Performance Base (Full) Model: Grid Search

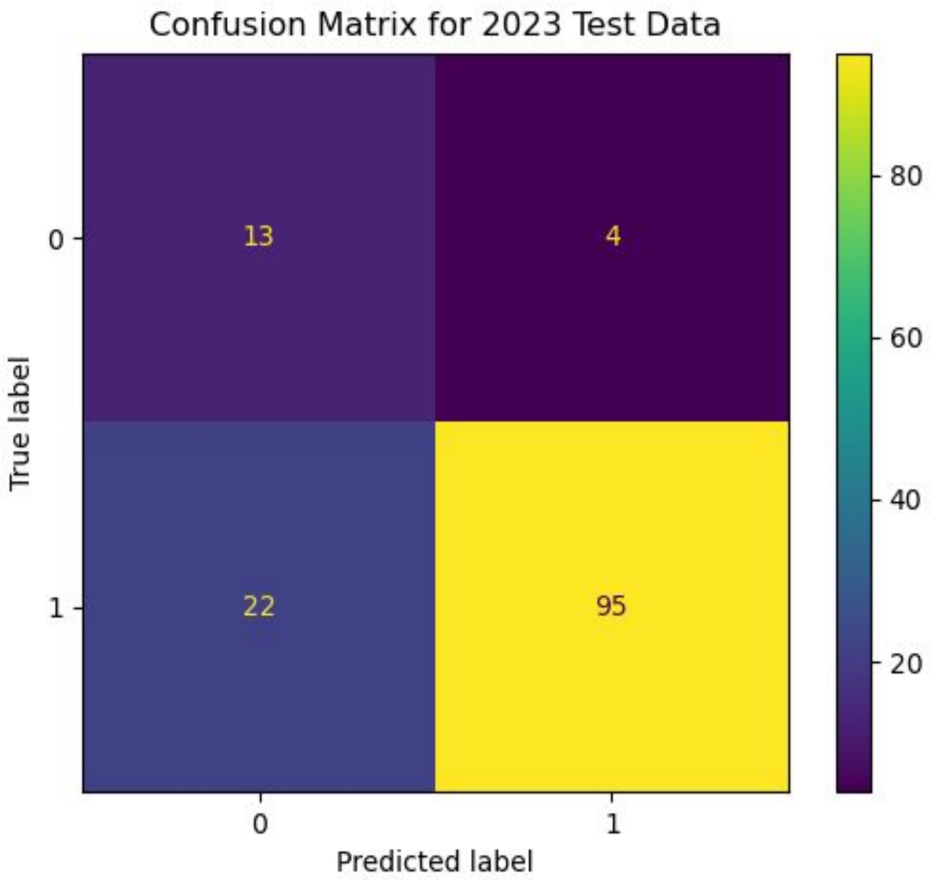
Subset Model

Training Accuracy	.789
Validation Accuracy	.777
Validation Precision	.831

Table 1: Performance of Reduced Model: Feature Importance

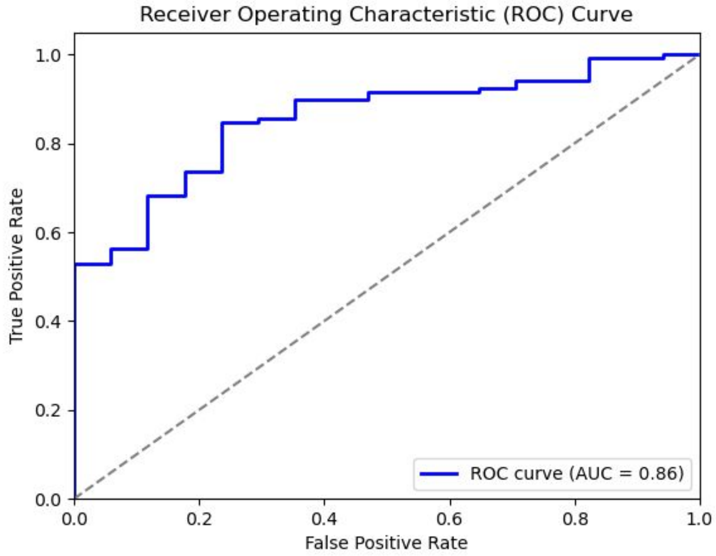
Generalization performance was the best on the base model so we proceed with that model for our classification tasks and analysis.

Subset model → LogisticRegression()



Test Accuracy	.806
Test Precision	.885

Table 1: Performance of Full Model on Test Data



Revisiting Questions of Interest

- Which features (of those considered in this analysis) are most predictive and significant on a player's offensive value, as measured by OPS?
    - k_percent, exit_velocity_avg, sweet_spot_percent, barrel_batted_rate, avg_best_speed, avg_hyper_speed, hard_hit_percent, flyballs_percent

- How closely can non-outcome oriented features predict or classify OPS?
    - Lasso Regression: on average our model **overpredicted OPS by .042**
    - Classification: **AUC = .86**

- How does our predictive model compare to other professional projection systems.
    - Lasso Regression: on average we predicted a **.026 higher OPS than BATX**