Using Historical Data to Predict Batting Success

Donna J. Harris (994042890)

Abstract

Major League Baseball is driven by data to make decisions, both for the short-term and the long-term. One powerful way that team management uses data is to predict the efficacy of players based on historical data. In this project, historical batting data was used in an attempt to determine batters' success over the course of their careers. Two approaches were explored, with varying degrees of success: the Hall of Fame Approach and the OPS Approach.

The Hall of Fame Approach extracts a batter's statistics from their first five seasons in the Major Leagues. The player's status as a Hall of Fame member (or not) was intended to help predict high quality batters. In the end, this was a problematic approach which would benefit from considerable adjustments to become viable.

The OPS Approach extracts a batter's statistics from their first ten seasons in the Major Leagues, and for their career. Because OPS is a useful measure of batting production, it was chosen to predict future OPS statistics based on historical values. This approach fared better than the Hall of Fame Approach and can be used to roughly estimate a batter's Career OPS statistic, as well as generally predict their success as a batter.

1. Description of Applied Problem

Baseball is a sport that is driven by data and data analysis. The field of sabermetrics has developed over decades, from the most basic statistics of batting averages and hits to increasingly comprehensive statistics that attempt to bring more meaning to the data and value to both the business and fans of baseball (Wikipedia contributors, 2022b).

Predicting and maximizing results in baseball for teams and individuals is a fascinating topic of study. The goal of this investigation was to identify markers within historical batting data based on regular season plate appearances. Two markers were chosen for exploration: Hall of Fame induction and the On-Base Percentage plus Slugging (OPS) statistic.

1.1 The Hall of Fame Approach

Premier baseball players who were most upstanding citizens within Major League Baseball are often afforded the honour of being inducted into the National Baseball Hall of Fame. While the criteria is not entirely objective (and often a source of debate) induction is considered "the highest mark of

achievement in the game" (National Baseball Hall of Fame, n.d., para. 1).

Hall of Fame members (who were also batters) commonly demonstrated excellent hitting skills throughout their careers in addition to excelling in other aspects of the game. Knowing this, the Hall of Fame Approach took on the goal of predicting Hall of Famers and similar quality hitters using the combination of regular season batting data and Hall of Fame induction data.

The first five seasons of batting data was isolated for each player in the primary dataset and then combined to form an early-career statistical summary. These player statistics were mapped to a Hall of Fame induction classification label. The chosen statistics (described in Section 2.1) were batting focused. Four popular calculated batting statistics were also generated to be included with the summary data.

The pool of players was reduced to focus on players with 2000 or more plate appearances over their first five Major League seasons. This removed the majority of pitchers from the dataset, as well several non-regular players.

The player data was also split into two groups based on the year their Major League careers began. Players with careers beginning before the year 2000 were used for model training and testing. Players whose careers started in 2000 or later were used for independent evaluation throughout.

1.2 The OPS Approach

The second approach taken was entirely different. When asking fundamental questions like: 'Does a player get on base?' and 'Does the player score and also help their teammates to score?', the focus often turns to a calculated statistic called "On-Base Percentage plus Slugging" (OPS). OPS is the sum of two other calculated statistics, which are referenced in its name: On-Base Percentage (OBP) and Slugging Percentage (SLG). The OPS combines information about how often a player gets on base and how effective they are at advancing teammates with extra base hits. The OPS is widely considered a useful statistic for generally comparing batters (Albert, 2010).

For this approach, all of the game data for each player was compiled by season. Players with ten or more seasons in their career to date had their first ten seasons extracted. Then calculated statistics (*i.e.*, SLG, OBP, OPS) were generated for each of the player's ten seasons. In addition to seasonal data, a player's Career OPS was also calculated. The prepared dataset for the OPS Approach compiled the collection of ten season OPS statistics and the Career OPS of each player.

The OPS Approach was partially inspired by the "Predicting OBPs"

section of the book *Curve Ball* where Albert & Bennett explored predicting the following season's on-base percentage with the previous season's (2003, pp. 55–56). Instead of using a single season of OBP data (not to be confused with OPS), this approach selected early season OPS values and explored predicting later season values, as well as Career OPS.

2. Description of Available Data

2.1 Primary Project Data

The primary data source for this project is a Kaggle dataset called "MLB Batting Stats By Game 1901-2021" (HugeQuiz.com, 2021). Its features are are described in Table 1.

Table 1. Description of "MLB Batting Stats By Game 1901-2021" dataset.

Feature	Description
ID	Player ID - Unique to each player
Player	Player Name
Date	Date of Game[1]
Tm	Player's Team
Opp	Opponent
Rslt	Game Result
PA	Plate Appearances
AB	At Bats
R	Runs
H	Hits
2B	Doubles
3B	Triples
HR	Home Runs
RBI	RBIs (Runs Batted In)
BB	Base On Balls (Walks)
IBB	Intentional Base On Balls
SO	Strikeouts
HBP	Hit By Pitch

[1] "Date" may also include an appended value, such as (1) or (2), which indicates the first, second, etc. game of multiple games played that day.

Feature	Description
SH	Sacrifice Hits
SF	Sacrifice Flies
ROE	Reached On Error
GDP	Grounded Into Double Play
SB	Stolen Bases
CS	Caught Stealing
WPA	Win Probability Added
RE24	Base-Out Runs Added
aLI	Average Leverage Index
BOP	Batting Order Position
Pos Summary	Positions Played
DFS(DK)	Daily Fantasy Sports Points - Draftkings
DFS(FD)	Daily Fantasy Sports Points - Fanduel

The dataset includes every regular season game's batting statistics for all players with at least one plate appearance from 1901 to 2021. Its data was collected by HugeQuiz.com from the well respected data website Baseball-Reference.com. The original dataset contains plate appearances for 15985 unique players across 120 years of baseball data.

In the initial data preparation, several features were discarded entirely. Tm, Opp, Rslt, BOP, Pos Summary all represent game-specific details which were not needed. IBB represents intentional walks, which are already included with the standard walk statistic (BB) and not needed on its own. SB, CS are baserunning statistics, which were not needed. ROE is based on chance and baserunning, not on batting. GDP tracks how a player gets called out, but was not needed for any of the calculated statistics as other similarly categorized statistics were. WPA, RE24, aLI are all more advanced statistics than were desired for this project. DFS(DK), DFS(DF) represent points for fantasy baseball and/or sports betting, which did not factor into this project.

The season year was extracted from the Date feature and became its own feature, Season, used to prepare data for both approaches.

A substantial aspect of the project's data preparation was data cleanup and validation. This was necessary to improve the quality of the individual datasets created for each approach.

2.2 Supplementary Project Data

For the Hall of Fame Approach, additional data was obtained directly from the Baseball-Reference.com website, which made it fully compatible with the primary dataset (*National Baseball Hall of Fame Inductees*, 2022). Player IDs are used across the datasets. This made it possible to easily merge the two datasets. The original dataset contains all 340 Hall of Fame inductees including managers, umpires, pioneers, executives, and players. The data is described in Table 2.

Table 2. Description of "National Baseball Hall of Fame Inductees" dataset.

Feature	Description
Year	Year of Player Induction
Name	"{Player}\{ID}" - Identifying Player Information
Unnamed: 2	"{Birth Year}-{Death Year}" - Player Life Span
Voted By	Source of Induction
Inducted As	Basis of Induction
Votes	Number of Votes Obtained for Induction
% Ballot	Percentage of the Ballot Vote Earned

The only data needed from this dataset was a list of player IDs for all players inducted into the Hall of Fame. The player ID was extracted from the Name column and then the list was filtered for inductees classified as a Player under the Inducted As column.

The Hall of Fame inductee data was merged with the player batting statistics for two data groups, separated by career start date. It is interesting to note the group made up of players whose careers began in 2000 or later includes zero Hall of Fame members. This presents a unique opportunity for evaluation, as running this data through a model under development exercises it entirely in prediction mode, suggesting future inductees and candidates for consideration.

3. Analysis Techniques: Hall of Fame Approach

3.1 Initial Exploration

Using all 17 statistical data features, including the four calculated statistics, each player was mapped to a binary label indicating their current Hall of Fame inductee status, where 1 indicated an inductee and 0 a non-inductee. As such, the Hall of Fame Approach is a binary classification problem and the exploration for this approach begins with logistic regression. Additionally, all models were trained using a 70/30 training/testing data split.

The first model, Model 0, used standard scaling to normalize the feature data. The training accuracy was 90.2% and testing accuracy was slightly lower at 87.3%. At a first glance this kind of accuracy seemed encouraging, however when looking at the precision and recall values for the data, a bigger problem began to take shape. For the training and testing data, Model 0 had 80.6% and 40.0% precision, respectively, but 39.2% and 26.7% recall. There is definitely a problem here.

Looking closer at the data's domain, first it is clear that accuracy alone will not be a meaningful measure in this context. For instance, with so few inductees in the entire dataset, having a model incorrectly guess that no players are inducted would still result in a high level of accuracy, but a very poor model.

It is also reasonable to expect with this data that precision would be skewed as well, considering that a number of players should actually be incorrectly classified as inductees (false positives). Some reasons for this include (a) the player has yet to be voted in, but is a likely candidate once eligible; and (b) the player has amazing batting records on paper, but was not a good baseball citizen. Their extra-statistical issues are likely to prevent qualification or acceptance into the Hall of Fame. This doesn't make them a bad hitter, but it does make them an unlikely Hall of Fame inductee.

These observations shifted the examination of evaluation metrics to recall. Because recall reflects a model's success at correctly identifying players already in the Hall of Fame, it can be viewed as highly valuable in this context. If the prospective models continue to do poorly in this regard, it may be a sign of deeper issues with the overall approach.

As another dimension of evaluation and experimentation, Model 0 was run with the independent test dataset including recent and active players. The surprising outcome from this was the model predicted the entire list of players as inductees. This result made no sense and further confirms that Model 0 is not effective.

As an exercise in curiosity, Model 1 also used logistic regression with the same features, but without scaling them. This did not see huge improvements but generally the metrics were higher and also had less variance between training and testing datasets, as can be seen in Table 3.

Table 3. Comparing Evaluation Metrics for Models 0 and 1.

	Accuracy	Precision	Recall
Model 0 Training	0.902	0.806	0.392
Model 1 Training	0.902	0.867	0.351

Model 0 Testing	0.873	0.400	0.276
Model 1 Testing	0.896	0.556	0.333

Running the independent test once again, Model 1 predicted a list of six players, most of whom are known as above average hitters and some of whom are considered future Hall of Famers. While a list of six players wasn't compelling enough on its own, it was more valuable than a list of all possible players!

Seeing this difference between scaling and not scaling the features, the option to move forward with non-scaled data was generally preferred.

Models 2 and 3 were attempts to improve Model 1 using cross-validation techniques. However, these did not impact the evaluation metrics in any significant way.

Before pursuing different algorithms, efforts were taken to reduce the number of features and search for positive impact. Two techniques were attempted: primary component analysis and feature selection, using feature importances..

3.2 Primary Component Analysis

This technique was applied with the intention of reducing the number of dimensions to improve the model. Plotting the primary components (see Figures 1 and 2 in section 5.1) showed that the data significantly overlapped and no clear separation between classes was made evident. As a result, this approach was abandoned, seeing this as another hint towards a flawed approach.

3.3 Feature Selection

Random forest classification was used to identify feature importances. This indicated that OPS, AVG, R, SLG, and H (set 1) were the five most important features from the original 17. Since three of the features were calculated, the importances were also identified without them. This resulted in R, HR, RBI, H, and AB (set 2) being ranked as most important. (See plots in Figures 3 and 4 of section 5.2.)

Models 4 and 5 were attempts to see if Model 1 would fare better with either reduced five-feature set. Model 4 (using feature set 1) had major problems with both precision and recall, predicting no inductees at all. Model 5 (feature set 2) did not have the problems of Model 4 but its evaluation metrics were comparable (or worse) than earlier models. (See Table 4.)

Table 4. Comparing Evaluation Metrics for Models 4 and 5.

	Accuracy	Precision	Recall
Model 4 Training	0.859	0.333	0.014
Model 5 Training	0.882	0.690	0.270

Model 4 Testing	0.888	0.000	0.000
Model 5 Testing	0.896	0.571	0.267

3.4 Further Exploration

Model 6 used a support vector classifier with a polynomial kernel. The original features were used, without the calculated statistics. Removing them should not be a problem, as the data that calculated them remained in the feature set.

The accuracy for Model 6 was expectedly high for both training and testing data (0.919/0.866) but the large variance remained between testing and training data for precision (0.970/0.333). Most problematic of all, the model had low recall scores for both datasets (0.432/0.200). Overfitting and struggles identifying existing inductees continued.

One last model was attempted. Model 7 used a K-Nearest Neighbours algorithm with k=1 and the original features, excluding the calculated statistics. The results were a severely overfitting training model (with all evaluation metrics reaching 100%) and a very poor generalization on the testing data (accuracy: 0.836; precision: 0.182; recall: 0.133).

Out of curiosity, the independent test dataset was run on Model 7, which resulted in a list of 24 (out of 234) predicted Hall of Famers. An interesting thing about this list is that roughly three-quarters of the names were either future Hall of Famers or well respected hitters. However, the KNN model amplified the overall trend of the Hall of Fame Approach. When looking at the nearest neighbours of, presumably, some legitimate Hall of Fame prospects, some of those players would not even be remote considerations for the Hall of Fame.

The data, as modelled here, has not been useful for predicting Hall of Fame induction or general batting success. And yet, there are hints of new directions, which are discussed briefly in section 6.

4. Analysis Techniques: OPS Approach

4.1 Initial Exploration

The OPS Approach is a regression problem, calculating a predicted OPS value based on previous values. To develop this approach, the first goal was to identify what to model. While ten seasons of OPS data was generated, the intention was to use data from within the first five seasons to predict an OPS from a later season or the Career OPS.

In order to choose the direction, the relationships between features were plotted against prospective labels to find the strongest relationships. This process is expanded on in section 5.2. (Also see Figure 6.) This resulted in a focus on the third, fourth, and fifth season OPS values for prediction. Before committing

to the label for the OPS Approach, models were developed for each target to obtain evaluation metrics for each.

Model 1 (Season 6), Model 2 (Season 10), and Model 3 (Career) were each built with linear regressors and default parameters. Additionally, all models were trained using a 70/15/15 training/validation/testing data split. Note that with the OPS Approach, only one dataset was prepared, resulting in more data to work with. (The data was only split by career start season for the Hall of Fame Approach.) The summary of results for these three initial models is presented in Table 3.

Table 3. Summary of Error Scores and R-Squared for Initial Models.

	MAE	RMSE	R^2
Model 1	0.098	0.144	0.589
Model 2	0.144	0.204	0.372
Model 3	0.052	0.089	0.771

Model 3, which predicted the Career OPS, has both the lowest error scores and the strongest R^2 value, making Career OPS the target to pursue for predicting batting success. (See section 5.4 and Figure 7.)

4.2 Improving on Model 3

Next, a stochastic gradient descent regressor was used to predict Career OPS, which was optimized by using grid search. This only slightly improved the evaluation metrics (0.052, 0.087, and 0.779 for MAE, RMSE, and R^2, respectively) over Model 3's initial results. Cross-validation on R^2 only increased it marginally, with an average score of 0.786.

The next attempt at improving Model 3 was using a gradient boosting regressor, which resulted in slightly lower error scores and a notable increase in R^2 (0.047, 0.070, and 0.858 for MAE, RMSE, and R^2, respectively) when compared to the original Model 3. Cross-validation on R^2 did not have any significant impact, including running the model using K-Fold validation.

For one more attempt, a support vector machine regressor with a linear kernel was used. Optimized parameters found through grid search were also used. An improved result over the original Model 3 was found (0.052, 0.079, and 0.818 for MAE, RMSE, and R^2, respectively) but not better than when the gradient boosting regressor was used.

Table 4. Summary of Error Scores and R-Squared for Improving Model 3.

	MAE	RMSE	R^2
Model 3	0.052	0.089	0.771
SGD	0.052	0.087	0.779
Grad Boost	0.047	0.070	0.858
SVR (linear)	0.052	0.079	0.818

Using Table 4 to compare the improvements side-by-side, it is clear that the gradient boosting regressor improves the model the most and in all areas. Error is lessened and the overall fit of the model is improved.

With the improved model built, the data that was held back for testing was run through with similar results. This showed low variance between the datasets and good generalization for unseen data, as can be seen in Table 5.

Table 5. Final Model with Testing Data Evaluation Metrics Compared.

	MAE	RMSE	R^2
Training	0.047	0.070	0.858
Testing	0.046	0.064	0.853

4.3 Generalized Evaluation

An additional measure of evaluation was derived for the final model, providing a a more general answer to the question: "Will this batter be above average over their career?" Using the OPS scale provided by Wikipedia contributors, an above average (or better) hitter is one whose OPS is above 0.7666 (2022a). Using this measure, one can examine the model's overall success at correctly classifying a player as above average or better.

Evaluating this for all players in the test data, it was determined that the final model accurately identifies players as either above or below average 91.1% of the time. While it isn't an overly powerful metric, it does verify the model is generally accurate for the problem.

5. Visualization Techniques[2]

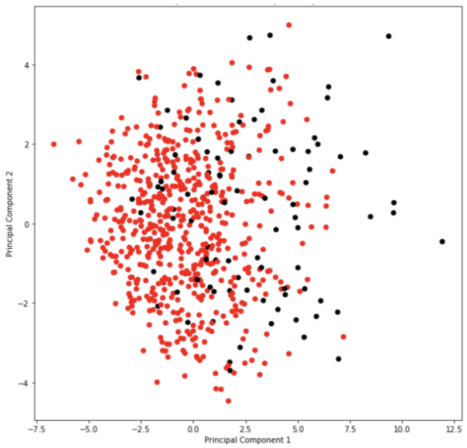
5.1 Primary Component Analysis in the Hall of Fame Approach

In Figure 1, the black data points represent Hall of Fame inductees. The points favour the right side of the two-dimensional graph and generously overlap with the non-inductee data.

It was hoped that examining the same three principal components in three-dimensional space would show a separation of the classes occurring on a different plane. However, upon examination of the 3D graph (see Figure 2) it is evident that the data points are still overlapping.

Figure 1

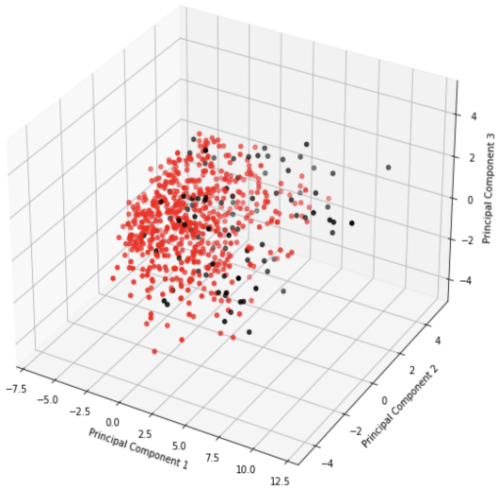
Visualizing Principal Components in Two-Dimensional Space.



[2] Please refer to the project code for full-size visualizations. The Step 3 and Step 5 Notebooks include the Hall of Fame Approach and OPS Approach visualizations, respectively.

Figure 2

Visualizing Principal Components in Three-Dimensional Space.

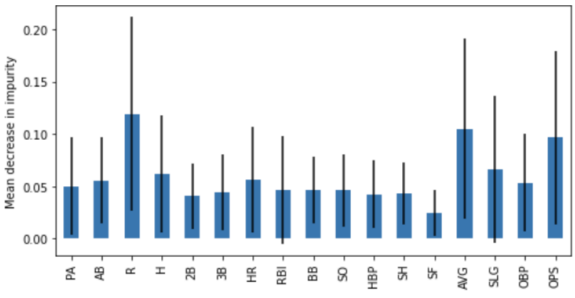


5.2 Feature Selection in the Hall of Fame Approach

Using the feature importance details from a random forest classifier, it was possible to extract the features which were determined to factor most prominently in the classifications. Figure 3 shows from where the selections of the five most important features were derived.

Figure 3

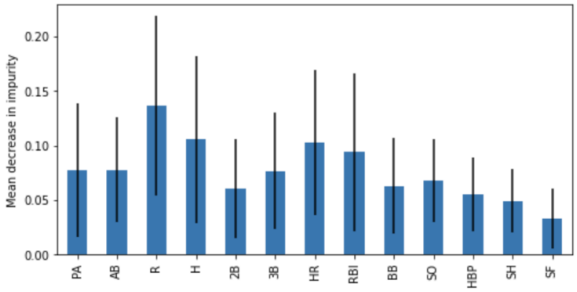
Visualizing Feature Importances.



Based on the height of the plotted bars, R, AVG, OPS, SLG, and H were identified as of highest importance from Figure 3. However, after seeing three of the calculated statistics rank so highly, it was deemed worthwhile to carry out the same process excluding the calculated statistics. (See Figure 4.)

Figure 4

Visualizing Feature Importances, without Calculated Statistics.



When the calculated statistics were removed from the feature set, R, H, HR, RBI, and AB were ranked as most important.

5.3 Model Evaluation for the Hall of Fame Approach

Visualizations featuring recall were chosen, as that metric was most valuable. Figure 5 (at the end) presents the confusion matrices and precision-recall curves for the most important models. The side-by-side comparisons demonstrate the failures common between the models, including overfitting.

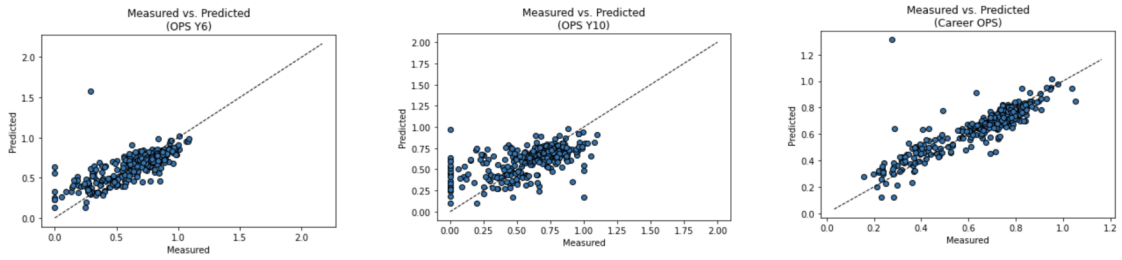
5.4 Feature and Label Selection for the OPS Approach

Scatter plots were important for visualizing correlations between features (the OPS values for seasons one through five) and each of the possible labels (the OPS values for seasons six, ten, and career). Figure 6

(at the end) presents those graphs. Take note of how the strongest relationships were involving the Career OPS, as well as the later seasons of OPS. This shifted the focus to using only the third, fourth, and fifth season OPS values for model predictions.

Figure 7

Comparison of Predicted and Measured OPS Targets.



As seen in Figure 7, plotting the predictions of Models 1, 2, and 3 against the actual values confirmed that predicting the Career OPS would be the strongest target for this approach.

6. Reflections and Improvements

6.1 Hall of Fame Approach

As mentioned throughout, the Hall of Fame Approach was flawed. The exploration highlighted some of the problems and also demonstrated how it fell short, losing focus on the ultimate goal: predicting future batting success. In retrospect, extracting specific Hall of Fame members who were known specifically for their successes in batting and using them to build a profile predictions could be based on would be a better direction to take. The exploration needed to steer away from predicting induction alone.

6.2 OPS Approach

One challenge with the OPS Approach is the exclusion of other dynamics which it cannot account for. For instance, a player may not have a high OPS but still can be a very effective batter that contributes well to their team. This model would not be able to predict those player's successes at the plate because this examination was focused on a single measure. Finding another statistic – perhaps a more advanced sabermetric – which reflects another aspect of a player's hitting might be a helpful improvement to this approach.

Another improvement might come from separating the two types of Career OPS data. In this dataset, Career OPS could represent either a player's in-progress career to date, or their historical, completed career. It might help make predictions more accurate to include only completed Career OPS for training purposes.

Additionally, it might be helpful for the OPS Approach to differentiate between OPS values based on a very low number of plate appearances within a season, where these OPS values may appear to be inflated and not always reflect a players overall ability at the plate.

6.3 Could a Combined Approach Possible?

A final idea which might be worthy of further exploration is the possibility of combining the improvements of these two approaches in order to create a superior approach.

Would taking an approach using the history of specifically-selected Hall of Fame batters and combining it with a Career OPS predictive type approach lead future models closer to the goal? While such an exploration is beyond the scope of this project, it holds some future promise and could make for an interesting hobbyist pursuit.

7. References

Albert, J., & Bennett, J. (2003). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* (Softcover reprint of the original 1st ed. 2001 ed.). Copernicus.

Albert, J. (2010). Sabermetrics: The Past, the Present, and the Future. In J. A. Gallian (Ed.), *Mathematics and Sports: Dolciani Mathematical Expositions #43* (pp. 3–14). Mathematical Association of America.

HugeQuiz.com. (2021, October 28). *MLB Batting Stats By Game 1901–2021* [Dataset]. Baseball-Reference.com. https://www.kaggle.com/datasets/darinhawley/mlb-batting-stats-by-game-19012021

National Baseball Hall of Fame Inductees. (2022). [Dataset]. Baseball-Reference.Com. https://www.baseball-reference.com/awards/hof.shtml

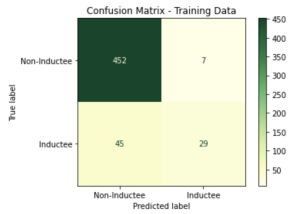
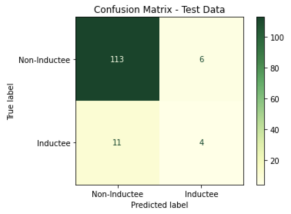
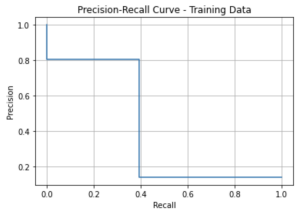
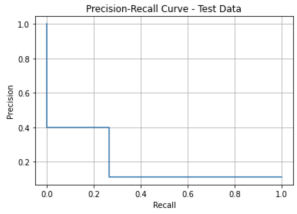
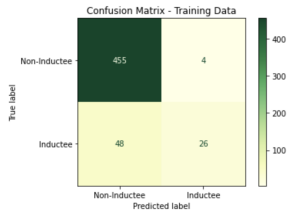
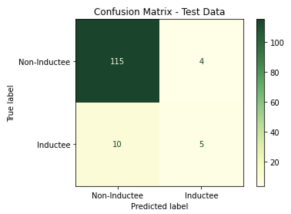
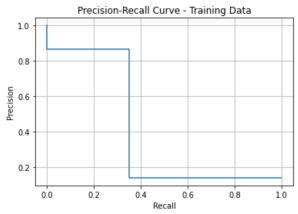
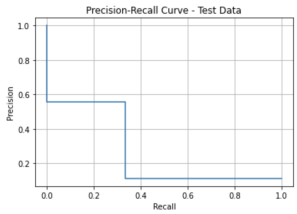
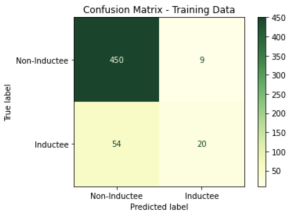
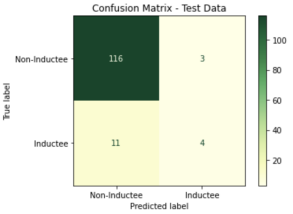
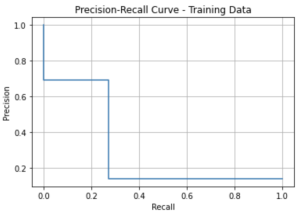
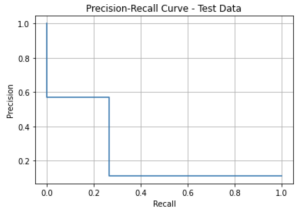
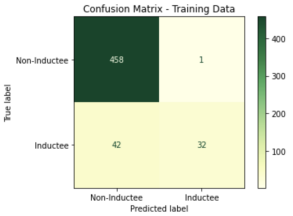
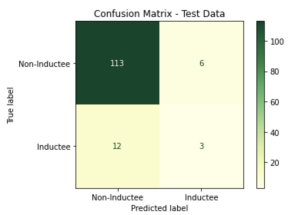
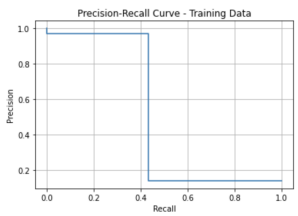
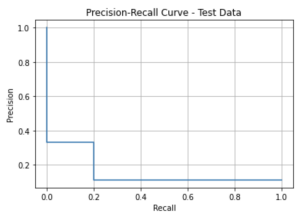
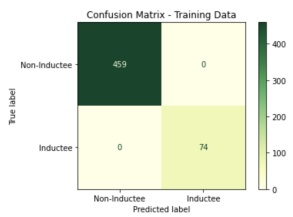
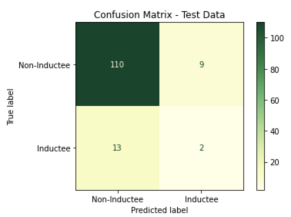
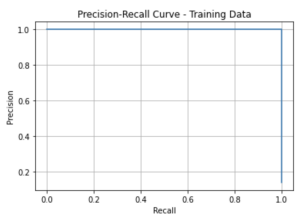
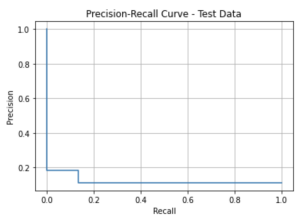
National Baseball Hall of Fame. (n.d.). *Hall of Famers | Baseball Hall of Fame*. Retrieved August 2, 2022, from https://baseballhall.org/hall-of-famers

Wikipedia contributors. (2022a, March 31). *On-base plus slugging.* Wikipedia. Retrieved July 26, 2022, from https://en.wikipedia.org/wiki/On-base_plus_slugging

Wikipedia contributors. (2022b, May 7). *Sabermetrics.* Wikipedia. Retrieved May 26, 2022, from https://en.wikipedia.org/wiki/Sabermetrics

Figure 5

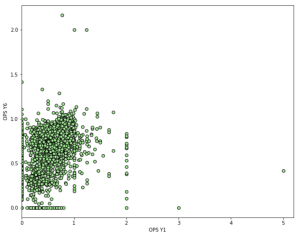
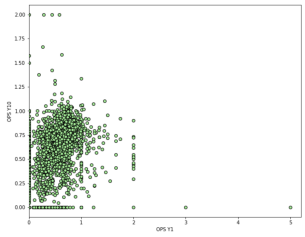
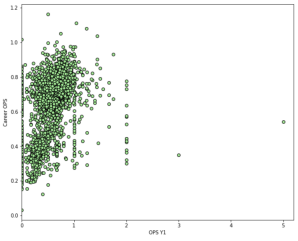
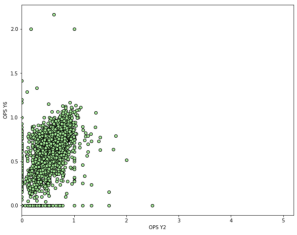
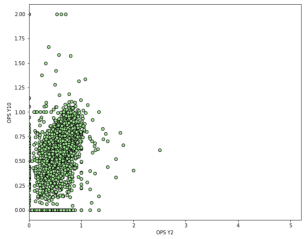
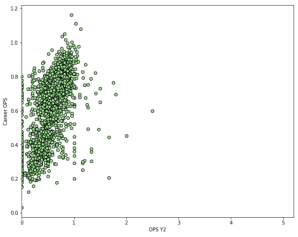
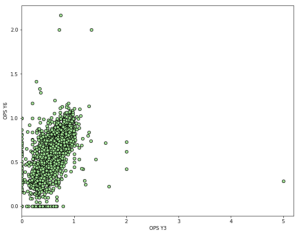
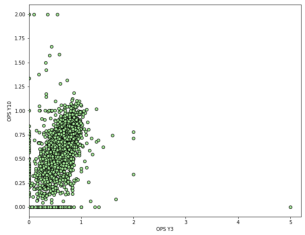
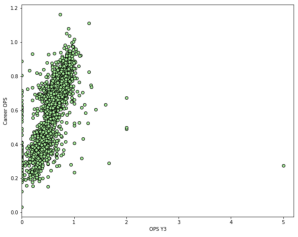
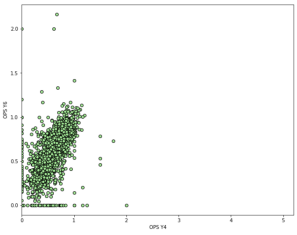
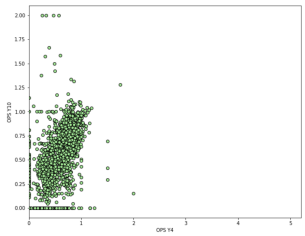
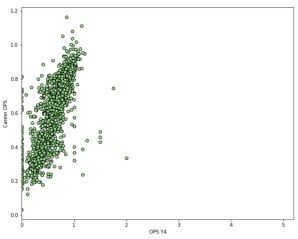
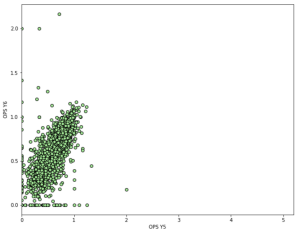
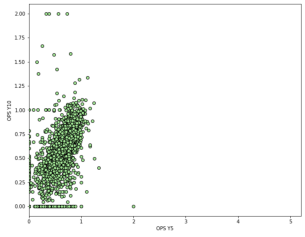
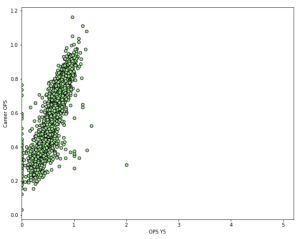
Comparison of Selected Hall of Fame Models using Confusion Matrices and Precision-Recall Curves for Training and Testing Data.

	Confusion Matrices (Training Data)	Confusion Matrices (Testing Data)	Precision-Recall Curves (Training Data)	Precision-Recall Curves (Testing Data)
M0[3]				
M1				
M5				
M6				
M7				

[3] MN is a short name for the model identifier, where N is the number. *E.g.*, M0 is Model 0.

Figure 6

Correlations of Proposed OPS Features and Targets.

	OPS Y6[4]	OPS Y10	Career OPS
OPS Y1			
OPS Y2			
OPS Y3			
OPS Y4			
OPS Y5			

[4] OPS Y<u>N</u> is the short name for the feature or label column in the dataset, where N represents the season number. *E.g.*, OPS Y6 represents the OPS from the player's sixth season.