Evaluating if Machine Learning Can Be Used to Classify Non-Pitching Baseball Players by Their Hall of Fame Status and Identifying the Significant Variables that Separate Classes

Christopher Spann

3/12/2023

Introduction

The National Baseball Hall of Fame (HOF) serves as a method for honoring players, managers, and others who excelled above most others in serving and representing the sport of baseball; however, election into the HOF is often debated among the media, fans, and individuals of the baseball community due to the subjective voting process that takes place for inductions. The primary method for induction into the HOF is through the voting process of the Baseball Writers' Association of America (BBWAA). The BBWAA is comprised of journalists specialized in professional sports that were awarded the right to make decisions regarding the worthiness of a player being inducted into the HOF following the player's retirement. BBWAA voting rules limit eligibility for HOF voting to former Major League Baseball (MLB) players that played at least ten seasons at the professional level and have been retired at least five years. The BBWAA creates new voting ballets each year that consider players who are newly eligible as well as players who failed to be inducted from previous ballots.

For the BBWAA voting process, "each voting cycle, qualified members of the BBWAA name no more than 10 eligible players whom they consider worthy of Hall of Fame honors [and to be inducted,] a player must be named on at least 75% of the voters' ballots (*Hall of fame voting procedures*. Baseball Reference, n.d.). Players are removed from the ballot if they are named on fewer than 5% of ballots or if they have failed to be inducted after appearing on the ballot 10 times (*Hall of fame voting procedures*. Baseball Reference, n.d.). If a player is not inducted into the HOF via the BBWAA voting process within twenty years of retirement, that player is only eligible to be inducted into the HOF by the Era Committee, formerly known as the Veterans Committee. The Era Committee consists of sixteen members, made up by members of the National Baseball Hall of Fame, baseball executives, and veteran media members that have all been appointed by the Hall of Fame's board of directors (*Era committees*. National Baseball Hall of Fame, n.d.). Similar to the voting process of the BBWAA, a candidate that receives votes on at least 75% of the casted ballots will earn election into the HOF – each candidate is evaluated based on individual records, ability, sportsmanship, character, and contribution to the game of baseball (*Era committees*. National Baseball Hall of Fame, n.d.).

As of March 2023, the National Baseball Hall of Fame has 342 elected members, with 270 of those members being former major league players (*Hall of famers by election method*. National Baseball Hall of Fame. n.d.). Controversy surrounding the induction process for the HOF focuses on the subjective, human input that is required to make decisions. For example, one voting member might place more significance on individual-oriented statistics such as number of homeruns, batting average, etc. while other voting members might focus on team-based players that contribute sacrifice hits, runs-batted-in (RBI), or assists while in the field.

The hypothesis of this research analysis is that HOF inductions can be predicted using various machine learning models with attributes related to career batting and fielding statistics as well as individual awards earned. The significance of this analysis is to determine if machine learning has the capability to model a process dominated through human decisions; moreover, various classification methods will be evaluated to determine which method performs best based on two different criteria: 1) the model with the highest accuracy, and 2) the model with the smallest percentage of false positives, minimizing the number of players classified as being inducted into the hall of fame that should not be inducted.

In addition to evaluating machine learning methods based on performance, the analysis will seek to determine the most influential factors for classification and compare the results to other literature on the subject. As previously noted, the subjective nature of HOF voting means that individuals must make decisions on which factors are most important for determining the worthiness of HOF induction for a player. The objective of this analysis is to determine the significant attributes that result in a classification status of inducted in order to gain a better understanding of the collective importance of evaluation criteria used by voting members.

Literature Review

There have been several studies conducted previously that focus on similar research goals, most notably William Young II, William Holland, and Gary Weckman's article published in the Journal of Quantitative Analysis in Sports titled "Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network." The authors argued that there has been a lack of research for predicting HOF inductions for baseball using career statistics, and that previous research in literature did not provide the best solutions for classification since they employed linear methods for prediction on non-linear datasets (Holland, Weckman, & Young II, 2008). Holland, Weckman, and Young's research was limited to non-pitchers and the results of the artificial neural network were compared to a baseline of logistic regression. The research found that an artificial neural network resulted in a model that predicted HOF status with an accuracy of 98%, 6% more accurate than a logistic regression model (Holland, Weckman, & Young II, 2008). The research also concluded that the most sensitive attributes for classification included position, number of hits, performance awards, number of errors, number of sacrifice flies, and character awards; furthermore, the authors explained how "the HOF committee does not have a tendency to enshrine players who are solely recognized for their defensive playing abilities […and…] misclassified players have a tendency to receive many performance-based awards like Golden Gloves" (Holland, Weckman, & Young II, 2008).

Another piece of research on HOF voting was done by Brian Mills and Steven Salaga in 2011, published in the Journal of Quantitative Analysis in Sports. This analysis, titled "Using Tree Ensembles to Analyze National Baseball Hall of Fame Voting Patterns: An Application to Discrimination in BBWAA Voting," employed a random forest tree ensemble method that resulted in an error rate as low as 0.91% in the most accurate forecast. Mills and Salaga concluded that a random forest classification algorithm was reasonably successful in predicting the pattern of HOF voting by the BBWAA from 1950 through 2009; furthermore, the research indicates that the number of all-star games and the number of homeruns are the most significant attributes for decisions related to HOF induction (Mills & Salaga, 2011). One of the primary benefits of using a random forest for classification is that it can measure attribute importance, and this analysis supports one of the research goals of this study of determining if machine learning has the capability to model a process dominated through human decisions.

Another piece of literature that supports the goal of this analysis comes from the popular data science site, Towards Data Science. In an article published in February of 2022 titled "Using Random Forests to Help Explain Why MLB Players Make the Hall of Fame," author Chris Kuchar uses random forest machine learning to identify overall variable importance for classification of HOF status in baseball players. This analysis used batting and fielding statistics for non-pitching baseball players, as well as end of season awards, to discover the class-wise basis for variable importance via a random forest algorithm (Kuchar, 2022). The results of Kuchar's analysis indicate that the top five features for being classified as being inducted into the HOF for batters/fielders are the number of times a player was named an all-star, the total number of games played, the total number of sacrifice hits a player hit, the total number of runs batted in a player has, and the number of total home runs a player hit (Kuchar, 2022). When comparing these attributes between classes, those classified as being inducted into the HOF had a mean all-star game count of 8.28, a mean count of homeruns of 220.50, and a mean total for games played of 1713.00. In comparison, non-HOF classified players had a mean all-star game count of 0.09, a mean count of homeruns of 14.29, and a mean total for games played of 245.60 (Kuchar, 2022). In summary, the research explains how random forests show that more years of being named an all-star, more homeruns hit, and more games played are significant predictors for classifying if a player will be inducted into the HOF; moreover, the analysis supports the idea that machine learning can be used to model a process reliant on human decision-making and offers some insight as to the most important predictors for determining HOF inductions.

All three of the literature pieces discussed not only show the capability for machine learning to be used for classification of HOF inductions in baseball, but they also highlight the important features for classification. The significance of this research comes from the fact that it can offer insight into the player attributes that members of the voting committees find most important for deeming a player worthy of the baseball HOF. For example, both Kuchar's analysis and that of Mills and Salaga identified all-star appearances and number of homeruns as significant identifiers for HOF induction. In addition, Kuchar identified the number of sacrifice hits as a significant feature while Holland, Weckman, and Young II identified the number of sacrifice flies as one of the most important predictors. When combining the results of all these analyses, one can argue that machine learning is more than capable at successfully modeling the HOF decision-making of the committee voters; furthermore, the similar results for important predictor attributes offer some guidance for this research project as well as future work to come.

Analysis on the capability for machine learning to be applied to classifying HOF status for baseball players is important because it highlights the ability for machine learning models to accurately represent human decision-making. As noted previously, HOF voting is widely considered a subjective matter than varies on an individual basis. Machine learning of the classification problem offers some detail to baseball enthusiasts on the features collectively agreed upon that make a HOF player in addition to showing current and future players what standards they generally need to aspire to reach in order to also reach the HOF one day. This analysis seeks to confirm results seen in previous literature that machine learning can accurately be used to classify players by HOF status. A number of methods, including a random forest similar to that of previous studies, will be used for classification and key variables will be identified and compared to the studies mentioned in this section.

Data, Methodology, and Analysis

Exploratory Data Analysis and Partitioning

The data used for this analysis comes from the Sean Lahman Baseball Database. The work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License and the database contains pitching, hitting, and fielding statistics for Major League Baseball players from 1871 through 2021. For this specific project, the HallofFame table of the database is used to identify all players who have received HOF votes in baseball. Supplemental data is brought in from the Batting, Fielding, AllStarFull, BattingPost, FieldingPost, and AwardsPlayers tables.

Table 1 – Sean Lahman Database Table Descriptions

AllStarFull: All-Star appearances by player and year
Appearances: details on the positions a player appeared at
AwardsPlayers: awards won by players and year
Batting: batting statistics by player and year
BattingPost: post-season batting statistics by player and year
Fielding: fielding statistics by player and year
FieldingPost: post-season fielding statistics by player and year
HOF: hall of fame voting data by person and year

The base HOF table has a total of 4,191 records detailing HOF voting data between 1936 and 2018; however, this data includes voting for players, managers, umpires, and baseball pioneers. This analysis limits the scope for classification to only focus on players that were not primarily pitchers in their playing career. The important attributes for pitchers are more than likely different than that of outfield players, and grouping these categories together could potentially impact the analysis. In addition, joins were made with other relevant tables to remove any players in the HOF voting dataset that do not have statistics related to batting or fielding. After removing players that are missing data in other tables, the resulting dataset consists of 786 players who have received votes for the HOF and have at least fifty recorded games batting in the appearances tables of the Lahman database.

For the 786 players in the HOF dataset that had been included on at least one HOF ballot since 1936, the supplemental tables were used to create the main dataframe used for analysis. The Batting, BattingPost, Fielding, FieldingPost, AllStarFull, and AwardsPlayers tables were all individually explored for relevant features that could be predictive of HOF status, and then all of the relevant fields were joined together using the common *playerID* field. There were instances of players who did not appear in certain tables, such as the AwardsPlayers table and the AllStarFull table if the player had not appeared in an All-Star game throughout the entirety of their career or failed to receive one of the chosen awards for analysis. For these instances of null values, the nulls were replaced with 0 to accurately reflect the count of the attribute. The final dataframe consisted of 786 records and 83 columns.

To begin the process of exploratory data analysis, summary statistics for each variable was documented to better understand the distribution of data for each attribute; furthermore, evaluating the minimum values, mean, maximum values, and standard deviations of variables helped identify any variables that could be impacted by an outlier. It was clear from the summary statistics that the variables had a wide range of spreads which means scaling might be needed in order to apply certain classification methods. In addition to the descriptive statistics of variables, a correlation heat plot was created using all of the independent variables to see which variables were highly correlated with each other. It was at this point that feature selection began as variables that were highly correlated could be condensed if they ultimately gave the same information to the analysis. For example, the batting statistics included fields for hits, singles, doubles, triples, homeruns, and total bases. The fields for total singles, doubles, and triples don't add much to the analysis when hits and total bases can capture the same information. Another example comes from the popular analytical measures of batting performance. On-base plus slugging percentage is a statistic that measures the sum of a player's on-base percentage and slugging percentage; as a result, the on-base percentage and slugging percentage attributes were removed since the information gain they attribute to the analysis can be found in other variables. A total of 33 variables were removed as a result of exploratory data analysis to limit the features that can be used in classification models. It is important for feature selection and the removal of highly correlated variables to occur because multicollinearity and variables that don't explain much of the variation in the data can reduce the performance of a machine learning model.

Following exploratory data analysis, data partitioning was done to split data into training and testing sets for the purposes of training machine learning models and evaluating performance among different methods. A total of 156 players in the final dataset were inducted into the HOF, which is roughly 20% of the total number of players in the data. Due to this class imbalance, a stratified data partition was done to ensure roughly equal splits between the classes in both the training and testing sets. For the partition, 80% of the data was split to be used for training and the remaining 20% is to be used for testing performance and comparing models. A breakdown of the training and testing datasets can be found below:

Table 2: Data Partition Summary

Training Shape	(628, 49)
Testing Shape	(158, 49)
% HOF in Training	125/628 ≈ 20%
% HOF in Testing	31/158 ≈ 20%

Methodology

Four machine learning methods were used for this analysis to determine which method results in the best performance on testing data. In the end, two different evaluation criteria will be explored for the purposes of identifying the best machine learning method. First, the model with the highest classification accuracy will be noted. The model with the highest classification accuracy will be the model that classifies the most records correctly between the different classes, inducted and not inducted. Second, the analysis will note the model that results in the fewest records of false positives. The reason for this analysis decision is due to the exclusive nature of the baseball HOF. The group is supposed to be reserved for individuals of the highest merit and quality, and it is important to minimize the number of players inducted that should not have been. For that reason, minimizing the number of players classified as being inducted into the hall of fame that should not be inducted will be another performance metric to evaluate the different machine learning methods.

The four methods chosen for this analysis include logistic regression, random forest decision tree, k-nearest neighbors (KNN), and stochastic gradient descent (SGD) classification. All methods aside from the random forest regression can be impacted by the scaling of data. The variable with the largest maximum value is InnOuts, or the time played in the field expressed as outs, with a value of 90,298. Postseason fielding triple plays has a maximum value of 1 and a minimum value of 0, with many other variables having minimums of 0. For that reason, standard scaling was applied to the training and testing independent variables to ensure scaling would not impact the results of models such as KNN and SGD.

In addition to standard scaling applied to the predictor variables of the analysis, principal component analysis (PCA) was run on the data to reduce the dimensionality. PCA is by far the most popular dimensionality reduction algorithm and the process involves first identifying the hyperplane that lies closest to the data, and then projecting data onto it (Géron, 2017). The benefit of using PCA for HOF classification is that it simplifies the complexity of the high-dimensional HOF data while retaining trends and pattern by transforming the data into fewer dimensions.

A total of 9 machine learning models were trained using the HOF training data; furthermore, the machine learning models were then used to classify instances in the testing set to evaluate classification accuracy and performance. For the purposes of this analysis, each model had the following statistics tracked:

- Training Accuracy
- Testing Accuracy
- # of False Positives on Testing Data

Logistic Regression is used in this analysis as a baseline for comparing other classification models. Similar to Holland, Weckman, & Young II, logistic regression is pointed out for comparison purposes, but the assumptions needed to fully support the use of logistic regression aren't necessarily met by the data. Basic assumptions for logistic regression include the absence of multicollinearity in the data, independence of errors, linearity in the logit for continuous variable, and the lack of strongly influential outliers.

KNN is an algorithm based on feature similarity and is being used since it is known for being very simple yet powerful for machine learning applications. The basic idea behind KNN is that the model will locate all of the closest neighbors around an unknown data point and then uses distances to classify which group it is closest to. One of the main benefits of KNN is that it is a non-parametric algorithm that does not make assumptions about the underlying data. In addition, the only inputs needed by the researcher include the number of neighbors to look at, the distance calculation being used, and the method for aggregating the classes of the points. One of the biggest weaknesses of KNN is the susceptibility to class imbalance. As mentioned earlier, only 20% of the data are classified as HOF inducted, so the KNN model might not work as well for this type of analysis due to the class imbalance.

Gradient descent is being used in this analysis as another linear classification method. The general idea is that parameters are tweaked iteratively to minimize the cost function for an analysis, and SGD specifically picks a random instance in the training data at every step and computers the gradients based on that one instance (Géron, 2017). The benefit of SGD is that it can handle large amounts of data and process it very quickly. The type of linear model, either logistic or support vector machine, is not being specified for the parameters of the SGD and instead the goal is to pick the best model for the optimization function.

Lastly, random forest classification was chosen for this analysis due to the benefit of generalization on unknown data as well as the interpretability of feature importance. A random forest is an ensemble of a large amount of decision trees in which every tree is slightly different from the others. The random forest algorithm will make a prediction for every tree in the forest and individual predictions are combined for classification. The random forest model is especially important for this analysis because it does not require normalization of features and can offer insight as to which attributes are most important for being inducted into the HOF. In addition, the analysis will employ cross-validation for the random forest model to prevent overfitting on the training data.

Analysis

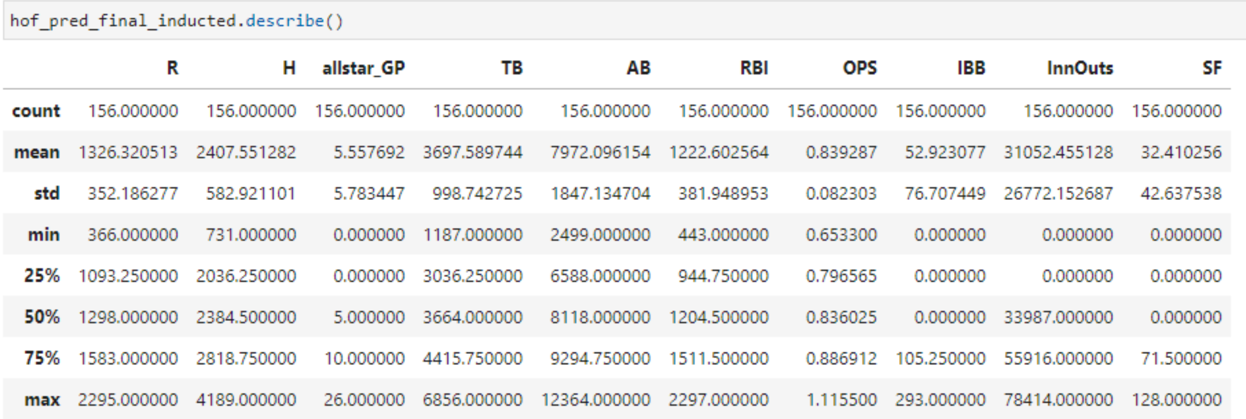
Model	**Training Accuracy**	**Testing Accuracy**	**False Positive Rate**	**False Positive Count**
Logistic Regression Scaled Predictors	0.912	0.905	0.031	4
Logistic Regression with PCA Scaled Predictors	0.896	0.892	0.047	6
SGD Scaled Predictors	0.893	0.886	0.055	7
SGD with PCA Scaled Predictors	0.866	0.873	0.079	10
KNN with k=3 Scaled Predictors	0.920	0.873	0.031	4
KNN with k=5 Scaled Predictors	0.880	0.880	0.024	3

KNN with k= 7 Scaled Predictors	0.869	0.886	0.016	2
Random Forest with 10-fold Cross Validation	0.954	0.918	0.039	5
Random Forest with 10-fold Cross Validation and PCA	0.882	0.804	0.039	5

The model with the highest accuracy on the testing data is the random forest model with the next highest being the logistic regression with the scaled predictors. The model with the worst generalization performance on unseen data was the random forest model that used the reduced data from PCA. For the first analysis goal of finding the machine learning model with the best predictive accuracy, the random forest and the logistic regression ended up performing the best and would be the models of choice for any future analysis that requires the model with the highest accuracy for classifications; however, the other important metric is reducing the number of false positives in the dataset. As mentioned previously, exclusivity is an important part of the HOF and thus minimizing the number of false positives could be seen as the performance metric of interest. In that case, the KNN model with a k value of 7 performed best with only 2 false positives, meaning a false positive rate of 0.016. The next closest model was another KNN model ok k=5, and the worst model for false positive optimization was SGD using PCA and scaled predictors. Overall, the random forest machine learning model that uses 10-fold cross validation seems to be the best model overall in terms of classification accuracy. The model produced an overall accuracy of 91.8%, and there does not seem to be strong evidence of overfitting since the training accuracy is only a few percentage points higher at 95.2%.

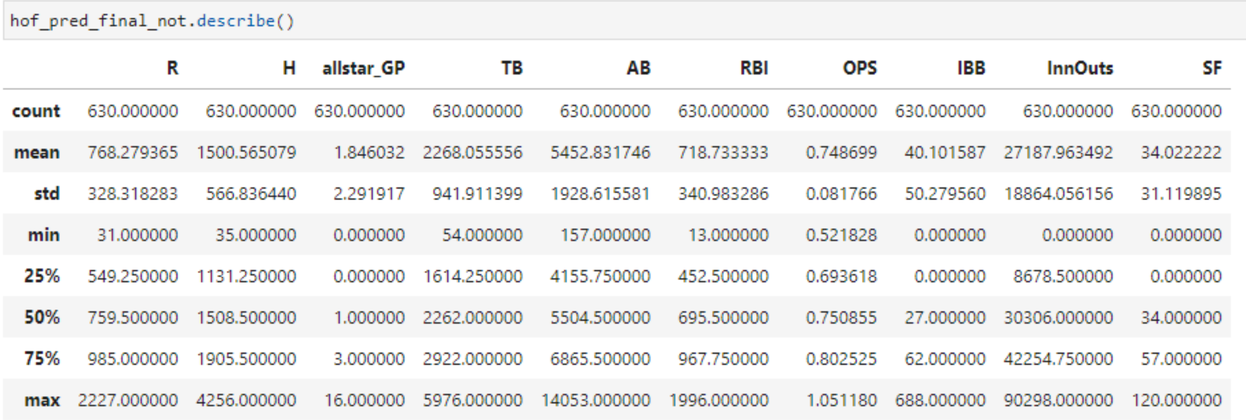
The other intended goal of this analysis was to determine the most important attributes for predicting whether a player will be inducted into the baseball HOF. One of the primary reasons a random forest would be a great choice for a machine learning model to employ consistently for this application is because we can break down which variables are the most important in the decision-making process. Using the random forest that resulted in the highest accuracy, the features that are most important to the classification of a player by HOF induction status are runs, hits, all-star games played, and total bases. The least important features included count of ALCS MVPs, count of NLCS MVPs, postseason triple plays while fielding, and count of rookie of the year awards. The results for feature importance seem to make sense as the players that often make the HOF are the ones that display high performance over a long period of time; therefore, the number of runs, hits, and total bases produced throughout a career would represent how much impact they made. In addition, the number of all-star games played makes sense because it represents the number of times the player was recognized as one of the best players in the league. Interesting to note is that runs batted in (RBI), on-base plus slugging percentage, number of sacrifice flies, number of sacrifice hits, and homeruns are also fairly high in terms of variable importance. The results of this analysis appear to confirm some of the important features seen in other studies. For example, the research of Kuchar, Mills/Salaga, and Holland/Weckman/YoungII all concluded the performance awards, specifically all-star games played, are significant in classifying HOF status and this holds up with the results of this study. In addition, number of hits, number of sacrifice flies, total runs batted in, and total homeruns are all fairly important features for the final random forest model, and all of these were mentioned as significant in at least one of the research pieces mentioned in this paper.

One can see the drastic difference between HOF players and non-HOF players when evaluating the summary statistics between the two classes for the ten most influential predictor variables in the random forest. On average, HOF players have significantly more runs, more hits, more all-star games played, more total bases, more at-bats, more runs batted in, a higher on-base plus slugging percentage, more intentional walks, and more inning outs played. The only variable not drastically different was number of sacrifice flies in which non-HOF players actually have a higher mean value. Overall, the machine learning process and data analysis indicate the players who exhibit strong statistics over a long period of time are the ones most-likely to make the HOF. Output for the summary statistics of the important features can be seen below:



```
hof_pred_final_inducted.describe()
```

	R	H	allstar_GP	TB	AB	RBI	OPS	IBB	InnOuts	SF
count	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000
mean	1326.320513	2407.551282	5.557692	3697.589744	7972.096154	1222.602564	0.839287	52.923077	31052.455128	32.410256
std	352.186277	582.921101	5.783447	998.742725	1847.134704	381.948953	0.082303	76.707449	26772.152687	42.637538
min	366.000000	731.000000	0.000000	1187.000000	2499.000000	443.000000	0.653300	0.000000	0.000000	0.000000
25%	1093.250000	2036.250000	0.000000	3036.250000	6588.000000	944.750000	0.796565	0.000000	0.000000	0.000000
50%	1298.000000	2384.500000	5.000000	3664.000000	8118.000000	1204.500000	0.836025	0.000000	33987.000000	0.000000
75%	1583.000000	2818.750000	10.000000	4415.750000	9294.750000	1511.500000	0.886912	105.250000	55916.000000	71.500000
max	2295.000000	4189.000000	26.000000	6856.000000	12364.000000	2297.000000	1.115500	293.000000	78414.000000	128.000000



```
hof_pred_final_not.describe()
```

	R	H	allstar_GP	TB	AB	RBI	OPS	IBB	InnOuts	SF
count	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000	630.000000
mean	768.279365	1500.565079	1.846032	2268.055556	5452.831746	718.733333	0.748699	40.101587	27187.963492	34.022222
std	328.318283	566.836440	2.291917	941.911399	1928.615581	340.983286	0.081766	50.279560	18864.056156	31.119895
min	31.000000	35.000000	0.000000	54.000000	157.000000	13.000000	0.521828	0.000000	0.000000	0.000000
25%	549.250000	1131.250000	0.000000	1614.250000	4155.750000	452.500000	0.693618	0.000000	8678.500000	0.000000
50%	759.500000	1508.500000	1.000000	2262.000000	5504.500000	695.500000	0.750855	27.000000	30306.000000	34.000000
75%	985.000000	1905.500000	3.000000	2922.000000	6865.500000	967.750000	0.802525	62.000000	42254.750000	57.000000
max	2227.000000	4256.000000	16.000000	5976.000000	14053.000000	1996.000000	1.051180	688.000000	90298.000000	120.000000

Conclusion

The main hypothesis for this analysis is that HOF inductions can be predicted using various machine learning models with attributes related to career batting and fielding statistics as well as individual awards earned. Results from the machine learning process indicate that it is possible to model a decision making process via machine learning and achieve strong accuracy scores. All machine learning models had testing accuracies above 80% with eight out of the nine having accuracies above 87%. The results of this analysis support the conclusions of prior literature that machine learning can be used to classify HOF inductions for baseball players. The best-performing model was a random forest model that employed cross-validation with a final testing accuracy of 91.8%. When evaluating models based on the criteria of minimizing false positive, KNN with a k value of 7 performed the best with a false positive rate of 0.016.

In addition to confirming that machine learning can be used to model the HOF induction decision making process, the results of this analysis support that of previous research in terms of important features for classification. All-star games played, total hits, sacrifice files, and total runs all appeared as important features in other research on the subject; furthermore, a clear distinction can be made between those inducted into the HOF and those not inducted when evaluating the summary statistics of the ten most significant features from the random forest model. The results of this specific analysis highlight the importance of total bases and on-base plus slugging percentage for a player, two fields that were not mentioned in the literature mentioned in this paper.

This analysis can offer insight to baseball enthusiasts on the features collectively agreed upon that make a HOF player in addition to showing current and future players what standards they generally need to aspire to reach in order to also reach the HOF one day. Given this information, players seeking to be inducted into the HOF should strive to achieve the greatest number of runs, hits, and all-star appearances that they possibly can. In addition, the longevity of players who made the HOF is much longer than that of players not inducted, so players should also strive to stay healthy and play for a long time consistently. This analysis confirmed results seen in previous literature that machine learning can accurately be used to classify non-pitching players by HOF status; however, future research on the topic should seek to evaluate the same process for pitchers. As noted previously, the important features that a machine learning model would need to classify pitchers by HOF status is most likely different than that of primarily field players. In addition to expanding the scope to pitchers, future research should also seek to better understand how the predictor variables interact with one another in determining HOF classifications. More advanced models, such as artificial neural networks, could be used and the final model can be used on current players to see if any are projected to be inducted into the HOF.

References

Era committees. National Baseball Hall of Fame. (n.d.). Retrieved March 12, 2023, from https://baseballhall.org/hall-of-famers/rules/eras-committees

Géron Aurélien. (2019). *Hands-on machine learning with scikit-learn and tensorflow concepts, tools, and techniques to build Intelligent Systems* (3rd ed.). O'Reilly.

Hall of Fame Election Requirements. Baseball Writers' Association of America (BBWAA). (n.d.). Retrieved March 12, 2023, from https://bbwaa.com/hof-elec-req/

Hall of fame voting procedures. Baseball Reference. (n.d.). Retrieved March 12, 2023, from https://www.baseball-reference.com/about/hof_voting.shtml

Hall of famers by election method. National Baseball Hall of Fame. (n.d.). Retrieved March 12, 2023, from https://baseballhall.org/discover-more/stories/hall-of-famer-facts/hall-of-famers-by-election-method#:~:text=Contents&text=The%20Hall%20of%20Fame%20is,have%20chosen%20180%20deserving%20candidates.

Holland, W. S., Weckman, G. R., & Young, W. A. (2008). Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network. *Journal of Quantitative Analysis in Sports*, *4*(4). https://doi.org/10.2202/1559-0410.1131

Kuchar, C. (2022, February 3). *Using random forests to help explain why MLB players make the hall of...* Towards Data Science. Retrieved March 12, 2023, from https://towardsdatascience.com/using-local-importance-scores-from-random-forests-to-help-explain-why-mlb-players-make-the-hall-of-aa1d42649db2

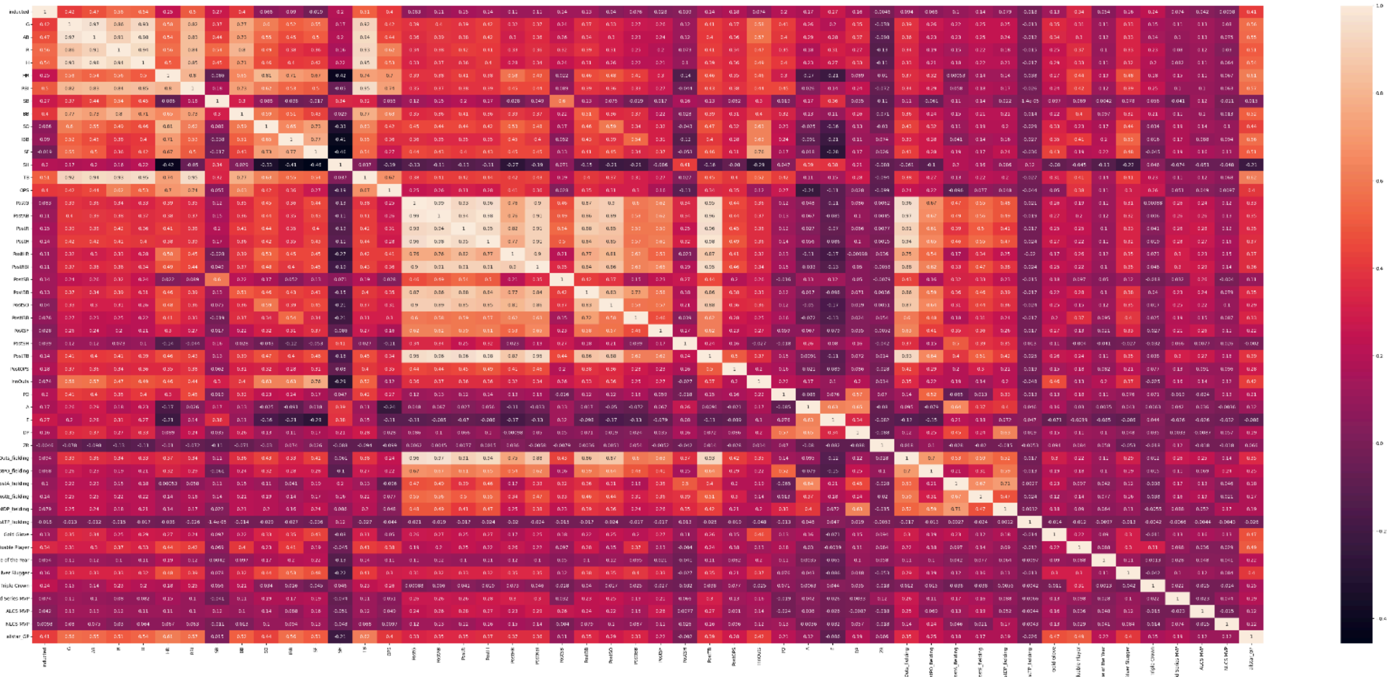
Lahman, S. (2022, March 8). *The Lahman Baseball Database ReadME*. Seanlahman.com. Retrieved March 12, 2023, from https://www.seanlahman.com/files/database/readme2021.txt

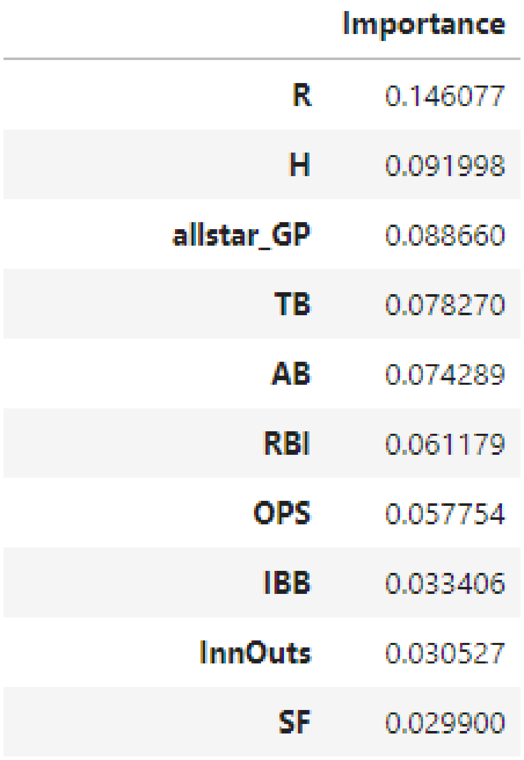
Mills, B. M., & Salaga, S. (2011). Using tree ensembles to analyze National Baseball Hall of Fame Voting Patterns: An application to discrimination in BBWAA voting. *Journal of Quantitative Analysis in Sports*, *7*(4). https://doi.org/10.2202/1559-0410.1367

Appendix:

Item 1: Variable Description and Source
https://www.seanlahman.com/files/database/readme2021.txt%20

Item 2: Correlation Heat Map for Variables

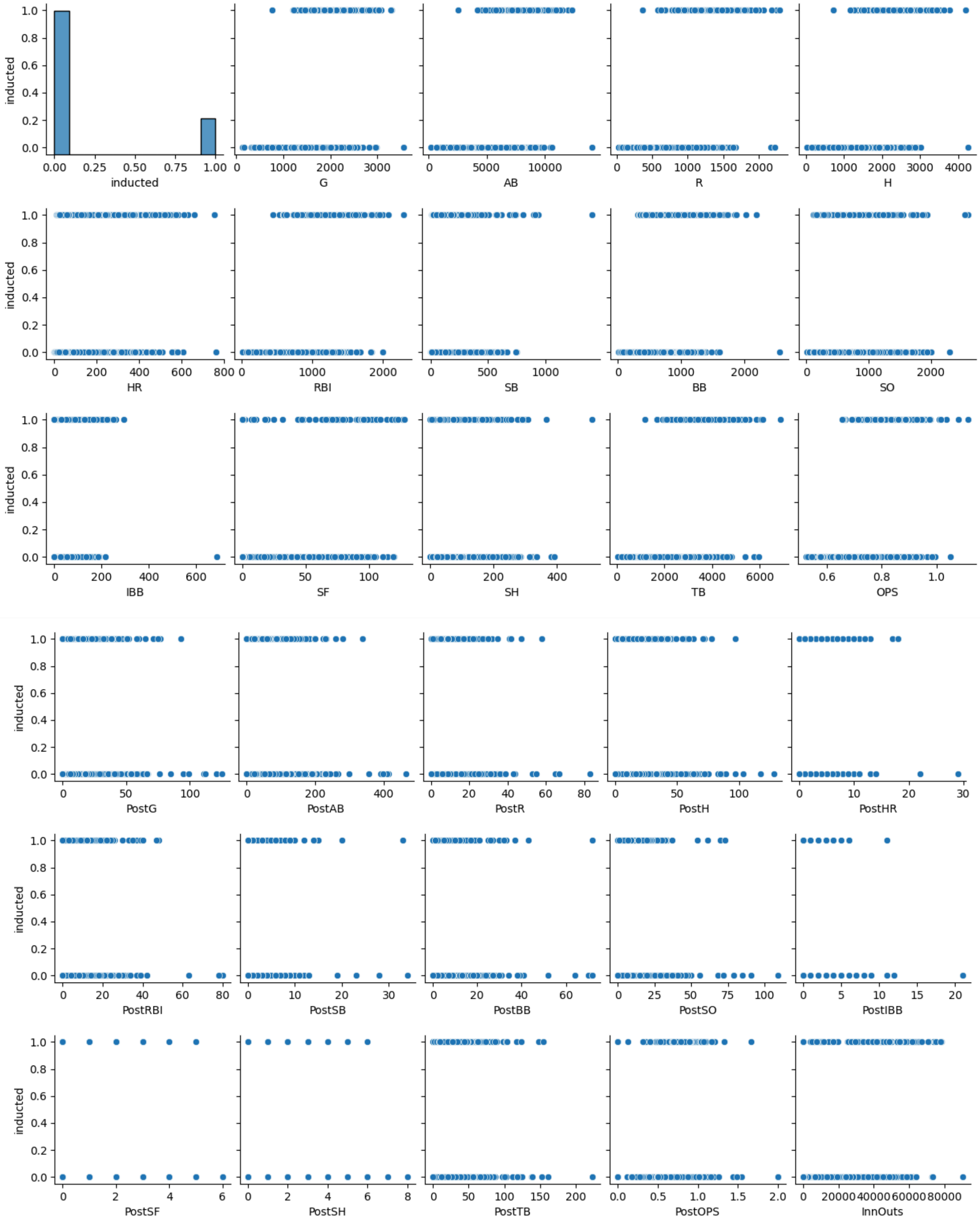


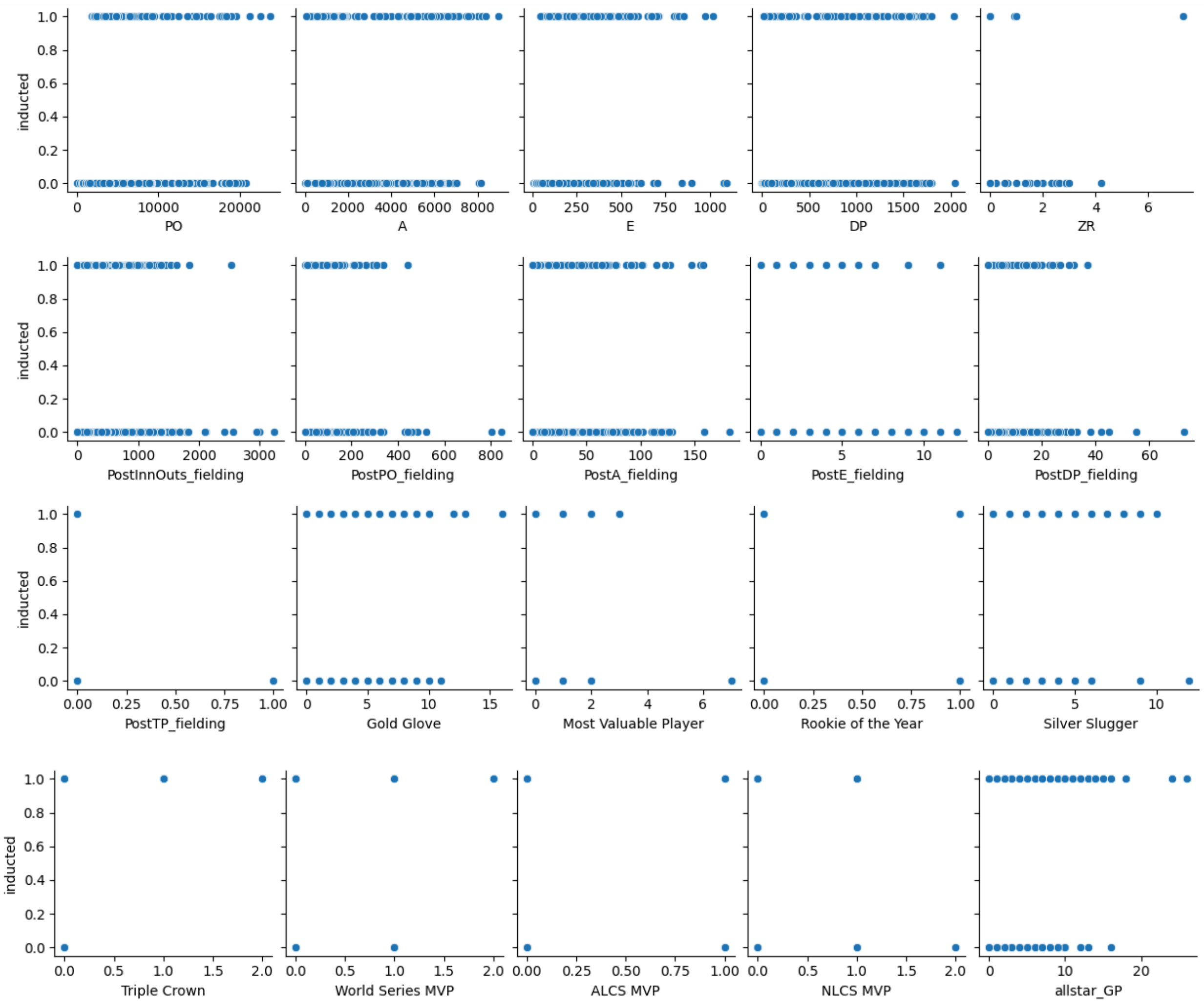


	Importance
R	0.146077
H	0.091998
allstar_GP	0.088660
TB	0.078270
AB	0.074289
RBI	0.061179
OPS	0.057754
IBB	0.033406
InnOuts	0.030527
SF	0.029900

Item 3: Random Forest Variable Importance – Top 10

Item 4: Variable Plots vs Inducted




inducted
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
PO
0
2000
4000
6000
8000
A
0
250
500
750
1000
E
0
500
1000
1500
2000
DP
0
2
4
6
ZR
0
1000
2000
3000
PostInnOuts_fielding
0
200
400
600
800
PostPO_fielding
0
50
100
150
PostA_fielding
0
5
10
PostE_fielding
0
20
40
60
PostDP_fielding
0.00
0.25
0.50
0.75
1.00
PostTP_fielding
0
5
10
15
Gold Glove
0
2
4
6
Most Valuable Player
0.00
0.25
0.50
0.75
1.00
Rookie of the Year
0
5
10
Silver Slugger
0.0
0.5
1.0
1.5
2.0
Triple Crown
0.0
0.5
1.0
1.5
2.0
World Series MVP
0.00
0.25
0.50
0.75
1.00
ALCS MVP
0.0
0.5
1.0
1.5
2.0
NLCS MVP
0
10
20
allstar_GP