Baseball Hall of Fame Classification

By: Mihir Singh, Dennis Fang, Justin Pan, Siraj Rayamajhi

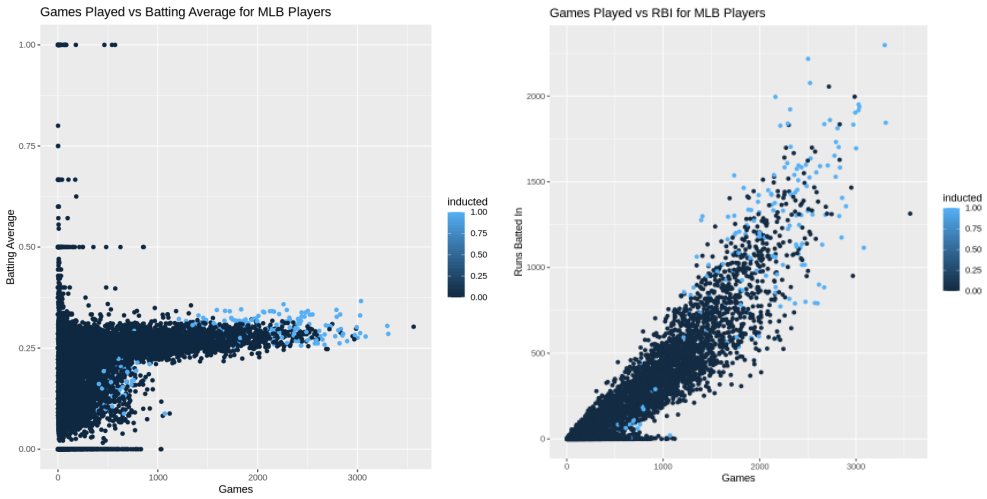
Background

The Baseball Hall of Fame is an institution that honors the greatest players, coaches, and other figures in baseball's history (National Baseball HOF 2024). Admissions to the Hall of Fame (HOF) are determined by the Baseball Writers' Association of America, who vote every year on which recently retired players should be admitted, and the Era Committee considers players not eligible for election by the Baseball Writers' Association of America. The voters take individual statistics, the longevity of a player, a player's peak, and their team's success into account when determining whether to admit a player into the HOF. Being admitted into the HOF is one of the greatest honors a player can receive. Of 20,459 players who played Major League Baseball (MLB), only two hundred seventy-three players have been admitted into the Hall of Fame. We will use MLB players' career statistics to predict which players will be elected into the Hall of Famers and determine which statistical model performs best.

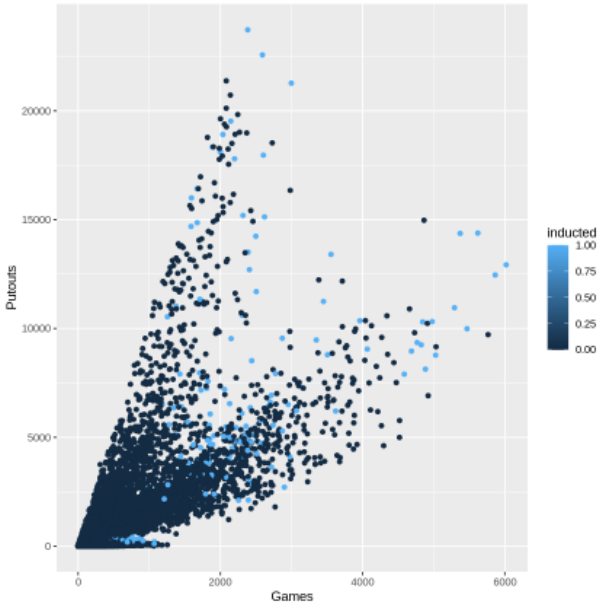
Data and Exploratory Analysis

We collected the data from the Lahman History of Baseball database on Kaggle, and we used the batting, pitching, and fielding statistics for every player in the MLB. The dataset names are: 'batting.csv,' 'pitching.csv,' 'fielding.csv,' 'hall_of_fame.csv,' and 'player.csv.', and the file names for the R document are 'batting_clean (1).csv,' 'pitching_clean (1).csv,' and 'fielding_clean (1).csv.' Since the data comprised the year-by-year performance of every baseball player, we added up all of the rows that shared the same player ID, which returned a data frame of each player's career stats. Then, we removed the 'year,' 'team_id,' 'league_id,' 'player_id,' 'name_first,' 'name_last,' 'stint,' and 'pos' values from each data set and the 'baopp' column from the pitching dataset. Then, we calculated the 'era' column for the pitching data, the 'bavg,' 'hrr,' 'sor,' and 'bbr' columns for the batting data, the 'fpct' column for the fielding data, and the binary 'inducted' variable for all of the datasets. Finally, we removed all rows where the above variables had missing or infinite results. The resulting batting database has 16666 observations of 24 numeric values, the fielding database has 17713 observations of 16 numeric variables, and the pitching database has 9171 observations and 28 numeric variables.

We explored the relationships between variables we hypothesized were significant for our analysis. For batting, we plotted the relationship between RBI (Runs batted in), batting average (hits per at-bat), home runs, strikeouts, and walks vs. at-bats, with the inducted variable as a secondary indicator. We plotted the relationship between ERA and games for pitching based on the inducted variable. For fielding, we plotted the relationship between putouts (the number of times a fielder gets a batter out) and the fielding percentage (putouts + assists) / (putouts + assists + errors) vs. games based on the inducted variable. We posited that the above variables would be significant for predicting the Hall of Fame induction status of batters, pitchers, and fielders, respectively. To support our hypothesis, we used the importance function for bagging and random forest decision tree methods and LASSO regression's variable shrinkage to determine which variables were most significant in predicting Hall of Famers.

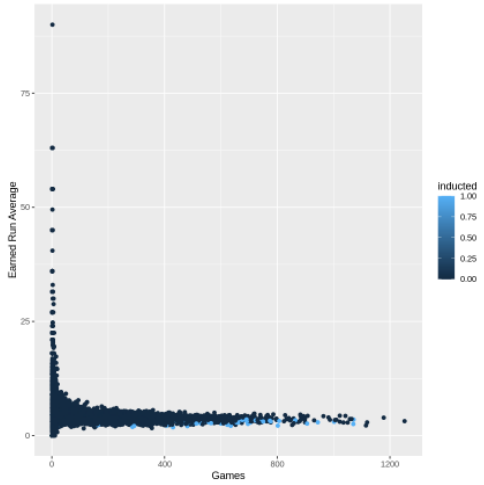


We noticed that batters with a higher batting average, RBI, and more games played in the MLB are more likely to make it into the Hall of Fame. For the batting average graph, as the number of games increases, most players have a batting average of around 0.250, an average to great batting average. However, even though most of the data is blended in, we still see that the Hall of Fame players, represented by the light blue points, stick out even more to the right, which suggests that Hall of Fame players maintain a high batting average for longer. This hypothesis is further revealed in the RBI vs. Games graph, where we noticed a positive correlation between the two variables. Most Hall of Fame players played more games and recorded a larger RBI than regular baseball players. However, some Hall of Fame players near the bottom of these graphs do not follow these trends, and they were likely inducted into the Hall of Fame due to their pitching prowess. Hence, their batting statistics are weaker than those of their batting-focused counterparts.



For fielders, the variable of interest lies in the putouts variable. Plotting the relationship between putouts and the number of games played yields a rotated v-shape, with the left end showing fewer games but more put-outs and the right end showing many games played but fewer

put-outs. The left side could have a larger positive correlation because the players represented were mostly infield players, which makes the most sense as they are the players who end up with the most put-outs. This hypothesis led us to believe that the right-hand side of the V shape represents outfield players' stats, as they rarely see opportunities for put-outs compared to infield players. Similarly to the batting graphs, the Hall of Fame players tend to stick out and appear near the top of the charts, showing a larger rate of putouts as they play more games. Finally, there is a line of players with almost 0 putouts and under 2000 games. This trend most likely represents the pitchers who play fewer games than other players as they need rest time after every game. Pitchers also record fewer putouts than other players, as balls are rarely hit directly back to the pitcher.



Exploring the ERA variable versus Games Played graph for pitchers, we see that it follows the same characteristics as the Batting Average versus Games Played graph. Some players who have played a few games have extremely high ERAs. However, as the number of games increases, most players have an ERA of around 3. Even though almost all the players have about the same ERA, we noticed how many Hall of Fame players tend to have many more games played while maintaining a lower-than-normal ERA. This conclusion suggests that the ERA variable is an indication of great pitchers. Although ERA can explain Hall of Fame induction for pitchers, it also depends on their longevity and how long they can maintain their stellar performance.

Methodology

The four statistical methods used for this experiment on the batting, fielding, and pitching data were logistic regression, LDA/QDA, Ridge/LASSO regression, and decision trees. To determine the model's accuracy, we focused on the model's Mean Squared Error (MSE) and the associated confusion matrices. The MSE is critical to our analysis because it shows the average squared difference between predicted and actual values in the data; lower values indicate a more accurate model. Meanwhile, confusion matrices show the correct predictions from executing a statistical model, showing correctly and incorrectly predicted values based on the model's performance. We will focus on the specificity, the probability that a Hall of Famer is predicted correctly. (False negatives vs. true positives) An important distinction is that players are only inducted into the Hall of Fame if they are no longer in the MLB. Since some Hall of Fame-caliber players are still playing, they haven't been inducted since the dataset's completion.

We conducted logistic regression to predict the inducted variable based on the games played, RBI, batting average, and home runs/strikeouts/walks per at-bat for our batting data, the games played, putouts and fielding percentage for our fielding data, and the games played and ERA for our pitching data. Then, we used LDA and QDA to group the baseball players by the inducted binary variable using the same response variables as the logistic regression. In LDA and QDA, we generate a linear/quadratic decision boundary in which Hall of Famers are on one side and non-Hall of Famers are on the other. Since most of the boundaries between Hall of Famers and non-Hall of Famers in the data we found were non-linear, we hypothesized that the QDA would be more accurate for this project. The next statistical analysis methods that came into play were the ridge and LASSO regression models, which we used to prune unnecessary columns from the dataset and only conduct regression analysis on the most important variables that can predict induction. Ridge regression would reduce unimportant variables to very small values, whereas LASSO regression would eliminate them, promoting a 'sparse' solution of only the most important predictors. Finally, we used decision trees to represent the decisions made within variables regarding classifying Hall of Fame inductees. (Separating dataset observations through inequalities (Batting Average > 0.3?, Put-outs > 14500?, etc.)) We used four different types of decision trees: regular decision trees, pruned trees that calibrate for the best number of branches based on variables, bagged trees (trees that sample observations equivalent to the number of observations in the response set), and random forests (merging the outputs of multiple normal decision trees to get a single response) to determine which method works best for each dataset.

Results

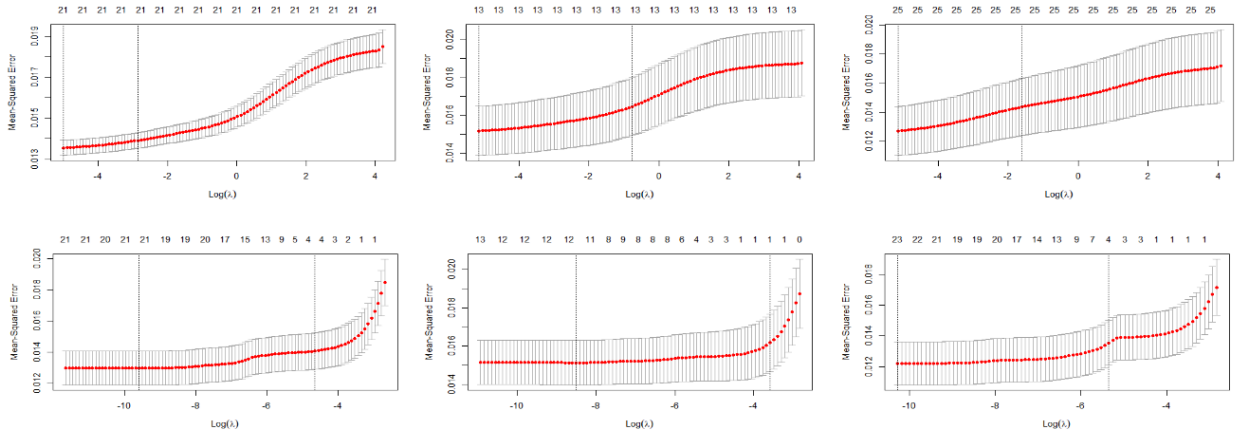
For the logistic regression, we noted that 58% of Hall of Famers and 99% of non-Hall of Famers were properly predicted using our batting data, 44% of Hall of Famers and 98% of non-Hall of Famers using our fielding data, and 29% of Hall of Famers and 99% of non-Hall of Famers using our pitching data. However, the associated MSEs of the logistic regression were 0.015 for the batting data, 0.027 for the pitching data, and 0.011 for the fielding data, all of which are greater than the MSEs for other statistical analysis models. Thus, logistic regression is less accurate in predicting hall-of-fame status than other models.

For the LDA, we found that the specificity, or the probability that a Hall of Famer is predicted correctly, for the batting, fielding, and pitching data to predict was 59.62%, 61.54%, and 95.24%, respectively. The sensitivity or the rate of accurately predicting non-Hall of Famers was 95.62%, 93.14%, and 93.39%, respectively. Meanwhile, the associated MSEs were 1.1129, 1.2024, and 1.1974, respectively. For the QDA, we found that the specificity was 71.15%, 63.46%, and 95.24%, while the sensitivity was 89.9%, 91.23%, and 92.44%. Meanwhile, the associated MSEs are 1.2990, 1.2596, and 1.2255, respectively. Our results show that the specificity was the highest for the QDA, proving our original hypothesis. However, the false positive rate, or the rate at which non-Hall of Fame players are incorrectly assigned, is higher for the QDA than for the LDA, as shown by the low sensitivity percentage and the high MSEs.

The Ridge and LASSO regression models returned very high general accuracies (Above 99% for all datasets using both Ridge and LASSO) and low test MSE values but low predictive accuracy for Hall of Famers. For instance, Ridge regression showed a specificity of 19.23%,

3.85%, and 23.81% and MSEs of 0.0082, 0.0084, and 0.0034 for batting, fielding and pitching, respectively. Meanwhile, LASSO regression showed a specificity of 23.08%, 5.77%, and 52.38%, and MSEs of 0.0087, 0.0073, and 0.0041 for batting, fielding, and pitching, respectively. While Ridge and LASSO regression performed well for batting and pitching data, it failed to classify the fielding data accurately. The low specificity across all fronts means that LASSO and Ridge regression misclassified many players who made it into the Hall of Fame. Still, the high accuracy and low MSE mean it was great at classifying people who did not enter the Hall of Fame. Finally, LASSO regression promotes sparse solutions and removes extraneous variables. The variables that LASSO regression found most important to predicting Hall of Famers are runs, triples, home runs, and runs batted in for batters, games for fielders and losses, completed games, shutouts, and strikeouts for predicting pitchers.

Out of the four types of decision trees, we calculated the MSEs as follows: 0.0106, 0.0096, 0.0093, and 0.0096 for the batting data for normal trees, pruned trees, bagging, and random forests, respectively, then 0.01, 0.011, 0.0067, 0.0068 for fielding, and finally 0.0034, 0.0039, 0.0032, 0.0034 for pitching. Based on this information, bagging yields the lowest average MSE for the three data sets. In addition, we found that the specificity of the confusion matrix is the highest with the random forest method, yielding 71.15% for batting, 61.91% for pitching, and 30.77% for fielding. Finally, using the importance function, we found that the most important variables for determining Hall of Fame induction status are games played and batting average for batting, wins and ERA for pitching, and putouts, games, and assists for fielding.



These are the Ridge and LASSO plots for the Batting, Fielding, and Pitching datasets. The mean-squared error for the training data over different lambda values in the regression is shown. The gray dotted line indicates the best lambda value, as found by cross-validation.

	Reference	
Prediction	0	1
0	7637	42
1	17	10

Example: A confusion matrix for ridge regression was applied to the batting dataset with an accuracy rate of 99.23%. However, its specificity is only 19.23%, making it inadequate for properly predicting Hall of Famers.

Conclusion:

This study shows us that the most accurate statistical method for predicting all baseball players is the Ridge and LASSO regression methods because they were able to accurately predict non-Hall of Fame players while still being able to predict some Hall of Fame players. However, this high accuracy value compromises the specificity of the models. We also found that QDA is the best at predicting Hall of Fame induction status, although it sacrifices accuracy in predicting non-Hall of Famers and the test MSE. Finally, we found that the bagging method in decision trees yields the lowest test MSE for the fielding and pitching data, while Ridge Regression yields the lowest MSE for the batting data. To answer our initial question, however, QDA best predicts Hall of Famers.

To answer the question of which variables are most important to Hall of Fame status, LASSO regression found that runs, triples, home runs, and runs batted in for batters, games for fielders and losses, completed games, shutouts, and strikeouts for pitchers were the most important predictors. At the same time, our decision tree importance analysis shows that games played and batting average are the best predictors for batters, wins, and ERA for pitchers, as well as games played, putouts, and assists for fielders. The two models greatly disagreed on which variables were most important. Overall, our hypothesis that batting average, home runs, runs batted in, and games would be significant for predicting batters, that ERA would be significant for pitchers, and that putouts would be best for predicting fielders was backed up by one of the models. However, we found that strikeouts and walks aren't significant predictors for batters and that fielding percentage isn't a significant predictor for fielders.

Discussion

A primary weakness of our study was that the proportion of normal baseball players versus HOF players is very skewed towards normal players, with 20,459 MLB players in league history and only 273 Hall of Fame players. This imbalance in the dataset likely causes bias in our statistical analysis, which in some cases generates high accuracy by correctly predicting non-hall-of-famers in most instances and ignoring the hall-of-famers entirely. To rectify this imbalance, we could use SMOTE (Synthetic Minority Oversampling Technique) to generate artificial hall-of-fame data based on existing hall-of-fame members' stats and other attributes. Another method that could work for our data is bootstrapping or sampling 273 non-Hall of Famers simultaneously to match them equally with the Hall of Famers. However, this approach might not capture the differences in variables between non-Hall of Famers and Hall of Famers.

Another weakness we noticed in our study is that some people inducted into the Hall of Fame for their pitching prowess had mediocre-to-poor batting stats, which likely influenced the statistical analysis of the batting average. To alleviate this weakness, we would separate the pitchers in the data from the non-pitchers to more accurately analyze and predict the pitchers' and batters' hall-of-fame statuses separately. We could do this by inserting the position column back into the batting, fielding, and pitching data and then conducting plots of the data and statistical analysis by grouping relevant batting, fielding, and pitching variables by the position data or using the position as a categorical variable in our models.

In addition, our statistical analysis doesn't account for narratives, off-the-field cultural impact, team success, playoff stats, or individual regular season stats. Hall of Fame admission is determined through voting, and individual voters value different things. A classic debate is peak versus longevity: is the career of a player who was good for a long time better than one who was great for a short time? Our dataset just contains career statistics and can't capture that nuance. Team success and playoff performances are also extremely important in creating iconic moments where voters weigh more.

Also of note to our analysis is that baseball continues to change, and the statistics of Hall of Famers in the past may increasingly differ from Hall of Famers today. For example, steroids used to be legal for players to use and gain an advantage in batting. However, players who use steroids, such as Barry Bonds, are often left out of the Hall of Fame despite having Hall of Fame statistics. In addition, advancements in movement tracking and aerodynamics have led to an increase in the strength of pitchers. Pitchers can throw faster and more accurately, leading to decreased batting and increased pitching statistics. In addition, to preserve pitcher health, fewer pitchers throw complete games than in the past to wear out their arms less. Shutouts, which are when pitchers throw complete games without giving up a single run, and complete games are both variables that LASSO regression found most important to determining pitching Hall of Famers. The model accuracy may decline as fewer pitchers throw complete games.

In the future, improvements can be made to the dataset to account for more than just individual statistics. A player's best statistical season could be incorporated, a statistic accounting for team success (such as winning percentage or wins above replacement), and a statistic that compiles accolades won through a points system. For example, a player could get 5 points for winning the Most Valuable Player award, 2 for making an all-MLB team, and 1 for an all-star game appearance. The above system would account for more of what voters consider when selecting prospective Hall of Famers. These additions will likely improve our model's Hall of Fame player classification and lead to higher accuracy and specificity. In tandem, we could also improve the accuracy of our LDA and QDA methods using the variable importance methods we found through LASSO variable shrinkage and the importance() variable for the batting and random forest decision tree methods. In addition, other machine learning models could be used in addition to the ones here. Forward and backward selection could be used on logistic regression to select better variables and find the most important ones. Another classification method, such as K nearest neighbors, could cluster the dataset into two distinct groups.

Citations

- National Baseball Hall of Fame Explorer. 2024. https://baseballhall.org/hall-of-fame/hall-of-fame-explorer
- Path to the Hall of Fame. 2024. baseballhall.org/hall-of-fame/election-rules
- The History of Baseball. 2019. https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball/data