PREDICTING MLB HALL OF FAME NOMINATIONS

Kenneth Martin

PROBLEM STATEMENT

Of the 19,000 players that have been a part of Major League Baseball's 150-year history, only a tiny fraction ends up nominated to be put on the ballot to be potentially inducted into the Hall of Fame. Of these nominees, an even smaller percentage (roughly 1% of all players) are actually inducted. But what determines whether a player will even be considered for the Hall of Fame? The rules for being on the ballot are rather arbitrary. The player must have played for at least 10 years and have been retired for at least five, but otherwise it is up to a small committee of six people on the BBWAA (the Baseball Writers' Association of America) to decide who they deem worthy of consideration.

Creating a model to predict Hall of Fame nominations could be useful for the BBWAA and the Hall of Fame to get a better idea of the types of players that are nominated and the biases that affect these nominations. Perhaps a machine learning model could actually direct the BBWAA committee towards a small list of players to consider for nomination based purely on their statistical baseball ability, thus reducing some of the arbitrariness of the selection process. It could also be helpful for teams that are interested in what factors are most indicative of a successful, potentially Hall of Fame career for younger prospects.

We want to consider which factors are most predictive of a player being considered for the Hall of Fame. We will use data from the History of Baseball dataset on Kaggle, which contains 24 CSV files on various MLB data dating all the way back to 1871, including a specific dataset on Hall of Fame nominations.

While there is no specific distinction between pitchers and position players when it comes to the Hall of Fame nomination process, the important statistics are very different for these two types of players. Therefore, I am splitting the players up into "Pitchers" and "Batters" and using different data to train the models for each of these categories. While I would predict that the two models will be similarly successful, it will be intriguing to see which players make the cut and which do not for each of them.

Using a variety of techniques, I was able to create predictive classification models for both pitchers and batters. The most successful model for batters had an ROC-AUC test score of 0.815, and the most successful model for pitchers had an ROC-AUC test score of 0.785. In addition, I was

able to analyze which features were most important in predicting whether or not a player is nominated for the Hall of Fame.

Data Wrangling

The raw data that I started with was five CSV files. These included a file of pitching statistics for the pitchers, a file of batting statistics for the position players, a Hall of Fame file that contains information about which players were nominated to the Hall of Fame each year, and two more files, "player" and "appearances", that contain general information about each player, such as height, weight, career debut date, final game, etc.

I began by creating a binary column in the Hall of Fame dataset that indicated whether or not a player had ever been nominated to the Hall of Fame. I then merged the batting dataset with select columns from the appearances dataset, filtering out seasons prior to 1936. I consolidated the rows from each year of each player's career into a single row with that player's average statistics for their entire career, and then merged this with more columns from the player dataset.

I followed a similar procedure with the pitching dataset, and then merged the batters and pitchers into one dataframe. By creating a career_length column, I was able to filter out players who were not in the league long enough to qualify for Hall of Fame consideration. I then split the dataframe into separate pitchers and batters datasets, dropping the irrelevant columns for each.

From here, I could pull out the players from each dataset who were either still playing or had not been retired long enough to be considered for the Hall of Fame. I now had a final dataset for pitchers and batters, as well as smaller datasets of the currently ineligible pitchers and currently ineligible batters.

Exploratory Data Analysis

After eliminating all the players who were ineligible because their career length was not sufficient or they had not been retired for long enough, I was curious as to what percentage of the remaining, eligible players received at least one Hall of Fame nomination. As it turns out, when only looking at players with career lengths of at least 10 years who have been retired for at least five, roughly one third of them receive a Hall of Fame nomination at some point.

Figure 1

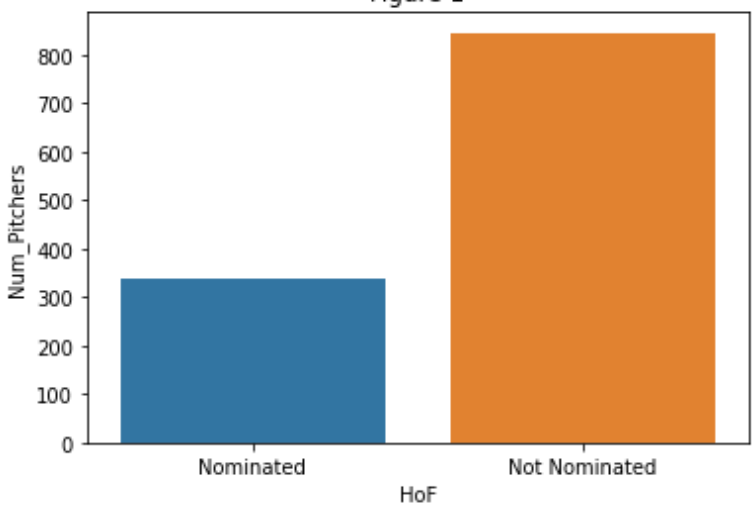
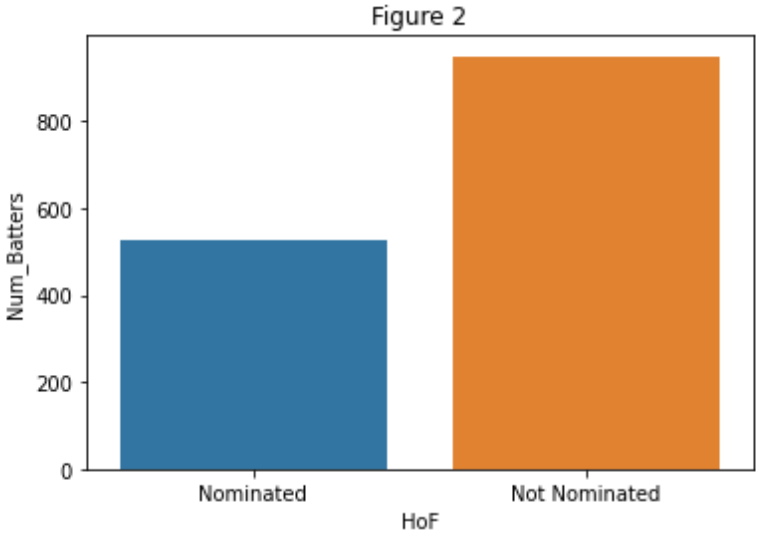
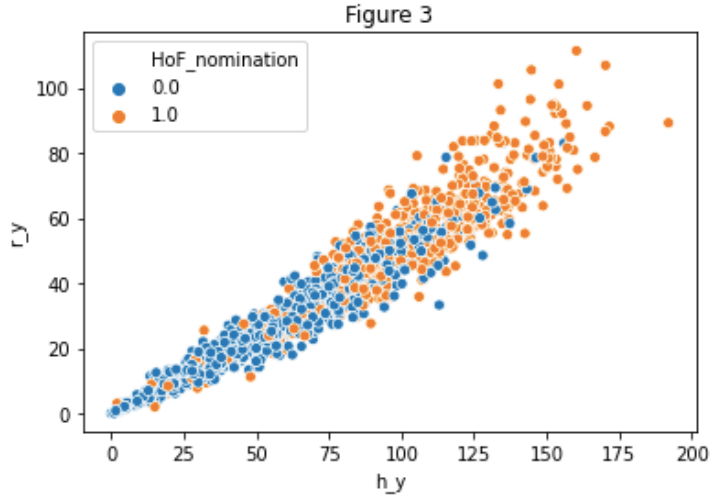


Figure 2



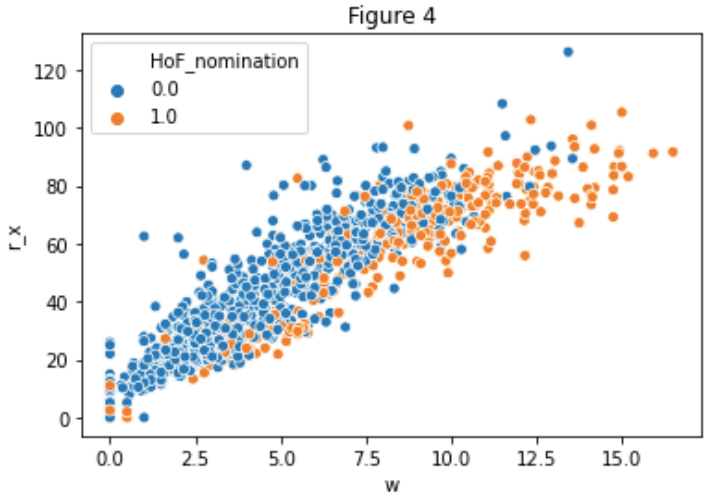
I was also interested in looking at the trends in Hall of Fame nominations based on multiple statistics. In Figure 3, we can see a clear trend in batter nominations based on hits and runs, where the majority of nominations went to players who had high hit and runs scored averages throughout their career.

Figure 3



We see a similar, but slightly different trend in Figure 4 when looking at pitcher nominations based on wins and runs. It is clear that Hall of Fame nominee pitchers tend to average a high number of wins, although average runs allowed seems to be less important. This makes sense, as the best pitchers tend to pitch more and therefore have more opportunities over the course of a season to give up runs.

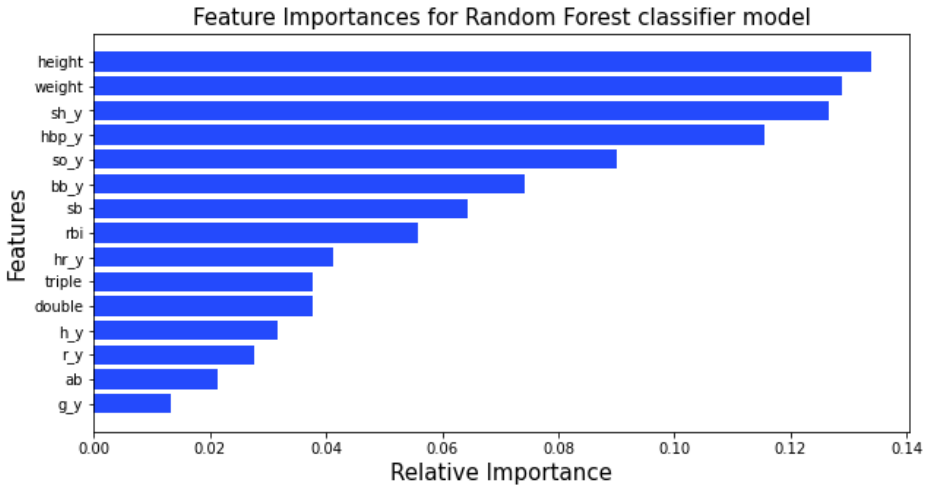
Figure 4

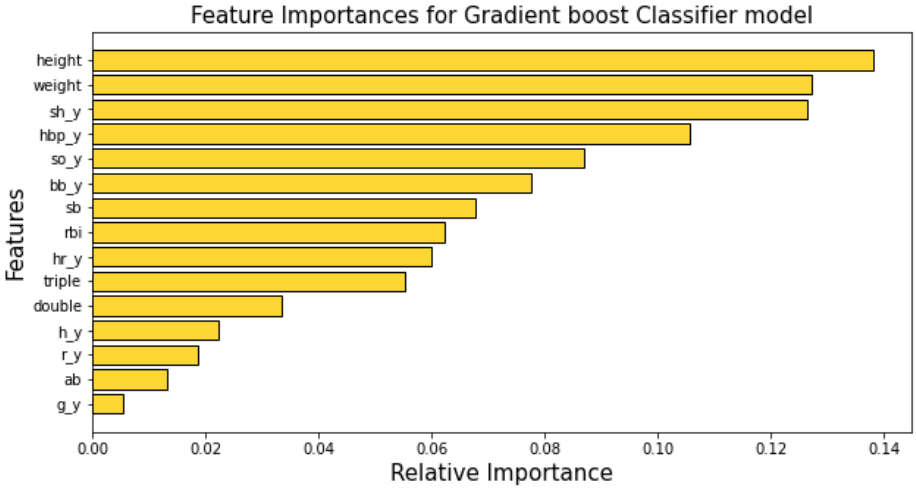


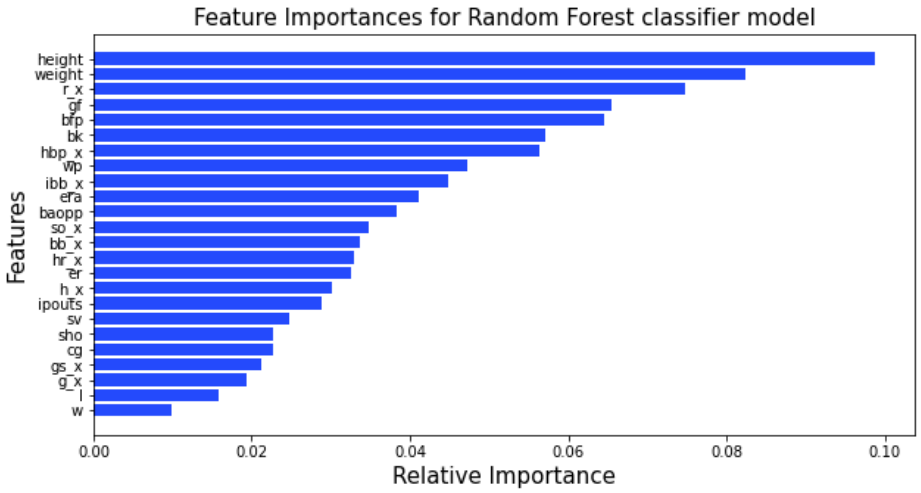
MODELING

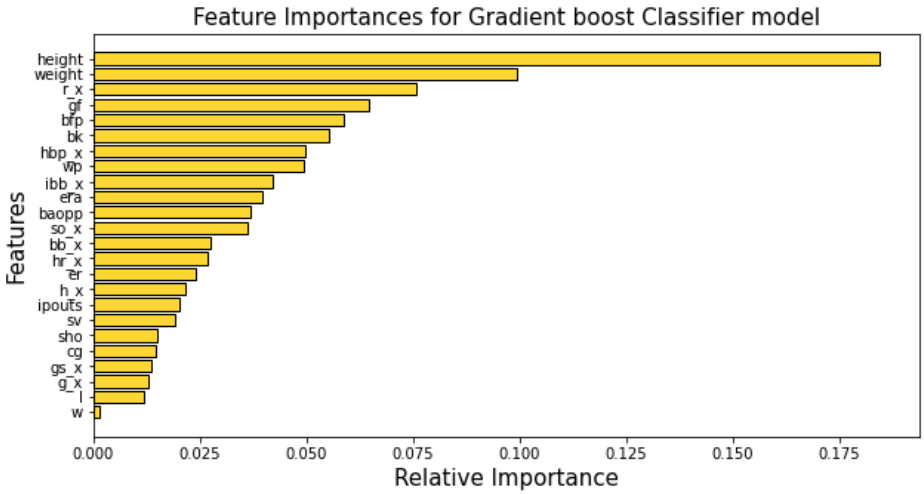
The objective of my modeling process was to create a model that classified each pitcher or batter as a Hall of Fame nominee, or not a Hall of Fame nominee. I tested four different machine learning classification models for both the pitcher dataset and the batter dataset: Logistic Regression, Support Vector Machine, Random Forest Classifier, and Gradient Boosting Classifier.

For the Random Forest Classifier and Gradient Boosting Classifier, hyperparameter tuning was done using grid search cross-validation. The metric I focused on maximizing while building the models was ROC-AUC. After determining the best model for each dependent variable, I plotted the feature importances of that classifier. Below are the feature importances for the batter dataset, followed by those for the pitchers.









After tuning hyperparameters with grid search cross-validation, my most successful model for predicting batter Hall of Fame nominees was the Logistic Regression Classifier with a ROC-AUC score of 0.815. The most successful model for predicting pitcher Hall of Fame nominees was the Random Forest Classifier with a ROC-AUC score of 0.785.

Conclusion/Future Research

Using career average statistics, I was able to create two separate models that were relatively successful at predicting whether or not a given pitcher or batter will be nominated to the Hall of Fame. The model for batters was slightly more successful, with an ROC-AUC test score of 0.815 and accuracy of 0.811, compared to an ROC-AUC test score of 0.785 and accuracy of 0.803 for the pitcher model.

One aspect that I had hoped to incorporate into this project but was unable to due to insufficient data was to include features for the players that are less directly reflective of their abilities on the field. For instance, it would have been interesting to include features such as salary, number of times a player was traded, what teams they played for, etc. While the use of purely baseball statistics (as well as height and weight) was sufficient to create relatively accurate models, there is a degree of arbitrariness to the nomination process that perhaps could be more accurately captured with non-baseball statistics such as these. If I was to continue to try to improve my models, I would focus on finding enough data to be able to add some of these features.

In addition, I would want to do more in-depth analysis of the predictions my models made for the "ineligible" batters and pitchers. After training and testing the models on the eligible players, I used them to predict whether or not each of the hundreds of ineligible players (those who are still playing or have not been retired for at least five years) are on track for a future Hall of Fame nomination. While it seems to have done a good job of classifying some of the more obvious, big name Hall of Fame contenders (Chipper Jones, Derek Jeter, Albert Pujols, Cole Hamels, Justin Verlander, etc.), it would be interesting in the future to see what percentage of the players were actually predicted correctly. Of course, this would not be possible to analyze in its entirety until all of these players have been retired for at least five years (and likely longer, as many players who eventually receive a Hall of Fame nomination are not necessarily nominated the first year that they are eligible).