Baseball Hall of Fame Report

Andy, Dante, Kai, and Kobe

2023-03-03

Introduction

The data we are working with is career baseball statistics. The data was collected from official scorekeepers who went to every Major League Baseball game and tracked various outcomes players had over the course of each game. For our finished data set, those statistics were added up at the end of each baseball season and also at the end of each player's career. The motivation for the scorekeepers to track the data was that they were paid, either by the home team, the league as a whole, or the local newspaper to track these statistics to keep baseball fans informed.

One motivation for our research question was to better understand which variables are associated with a successful hall of fame induction. Another motivation for our research is an extension of the first, which is for us to predict which players will be inducted into the hall of fame in the future.

As such, our first research question is what factors are associated with a position player being inducted into the hall of fame and whether we can create a model that uses these factors to predict what players will make the hall of fame. Our other research question is what factors are a associated with whether a pitcher is inducted into the hall of fame and whether we can create a model that uses these factors to predict what pitchers will make the hall of fame.

Data Description

Our data consists of mostly quantitative variables mixed in with some categorical variables. Most baseball metrics come in quantitative form so it is often hard to obtain categorical measurements unless they come directly from the quantitative variables itself. However, some categorical variables we may use include factors such as year, league, team, stint, and possibly some other categorical variables that are computed from other quantitative variables. One other potential categorical variable would be a binary variable indicating whether a player was known to use steroids or not.

Furthermore, we are splitting this project into two parts: batters and pitchers. Thus, there is a need to split the data; in other words, we'll have two different data sets, since metrics for batters and pitchers are vastly different. However, we'll still be predicting the same variable: hall of fame. Hall of fame is a binary categorical variable (with the two values being 1 and 0), indicating whether a player was inducted into the hall of fame or not. In addition, we made sure to only include players who were on a hall of fame ballot at some point. In total, we have around 750 observations in the position players training data set and 400 observations in the pitching training data set.

Exploratory Data Analysis

Steroids Clustering

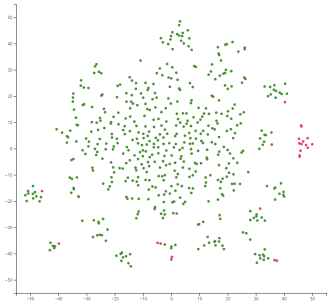


Figure 1: Clustering with Steroids in Red

Methods

For our hypothesis testing, since we use so many coefficients for both models, our family wise error rate would be high. As such, we are going to use the Bonferonni correction. This gives us that the significant p-values are 0.0015 for Position Players and 0.0042 for Pitchers.

Each of our pitcher and batter (position player) data frames consisted of 50 base variables after the conclusion of our various join functions.

- For batters, we began summing across playerID groups in the "Batting" data frame within the Lahman package. This allowed us to see career totals for recognizable rate and counting stats for each observation (player) like batting average, walks, plate appearances, and total hits for each of the four base categories. We then joined a pivoted "award" data frame that added approximately 20 accolade predictors for each observation, this includes total all-star selections, most valuable player awards won, silver sluggers, etc.
- For pitcher, we began summing across playerID groups in the "Pitching" data frame within the Lahman package. This allowed us to see career totals for recognizable rate and counting stats for each observation (pitcher) like innings pitched (a starter would typically pitch ~ 150 in a season), strike outs, earned runs, and we additionally mutated $\frac{9\times ER}{IP}$ and $\frac{W+H}{IP}$ to get the all important earned run average and WHIP rates. We then joined a pivoted "award" data frame that added approximately 20 accolade predictors for each observation, this includes total all-star selections, Cy Young awards won, Gold Gloves, etc.

In addition to the base covariates, a large initial AIC from our step() selection process and some rather large Cook's Distance values indicated the necessity to mutate additional predictors.

- The late 90's, early 2000's Steroids era of baseball altered the way BBWAA voters viewed the careers of some players in spite of otherwise Hall of Fame worthy numbers. And so, using webscraped data from ESPN and Bay Area Laboratory Co-operative (BALCO) court documents, we mutated a binary 0/1 column that indicated if a player was indited on any sort of performance enhancing drug (PED) scandal.
- We then ran our step() model selection process again, and while we obtained better AIC values, we felt we could do better. We understood what the baseball writers considered important to be elected into the Hall of Fame was period and cohort dependent. Therefore, we mutated a column that signified which of two committees elected a player into the Hall of Fame; the Baseball Writers Association of America (BBWAA) or the Veterans Committee. The BBWAA typically votes players who appeal to modern analytics of what makes a player well Hall of Fame worthy, while the Veterans Committee typically votes players based on high batting averages or nepotism from the dead-ball era of baseball (1880-1920).

- For position players specifically, we felt the need to mutate in a variable which identifies the primary position of a player throughout their career. This is important because of something called "Positional Adjustment". Based on the defensive difficulty of each position, a player is typically expected preform at a certain offensive level to be cosidered good. For example, a first-baseman who has what is considered an easy position defensively, is expected to preform far better offensively than a player at the catcher position, which is considered the most difficult defensive position according to FanGraphs Positional Adjustment.
- For pitchers, preliminary parameter selection via the step() function revealed the importance of relief pitcher accolades. Relief pitchers are a subset of pitchers who typically pitch ~ 60 inning a season. We felt the need for our model to be differentiate between a starter and relief pitcher, and so we mutated a predictor which identifies if a pitcher is a starter or reliever based on seasonal innings pitched totals.

For our model building, we used essentially every possible variable that could potentially be useful and did not contain too many missing values. As such, we started our process with even larger models than ones we already had To narrow down the number of covariates, we used step-wise model selection that goes both forwards and backwards, allowing us to test as many models as possible. Since we are more focused on the prediction aspect, we chose to use AIC as our primary criteria for model selection over other criteria such as BIC or K-fold CV. In addition, AIC uses less computational power than LOOCV, which is vital for us given the number of covariates we used in building our models.

Our final models are as follows:

- Position Players:

  ```
  wein ~ all_star + Pos + HR + nice_guy_awards + Steroids + votedBy + RBI + AVG +
  most_valuable_player + R + SB + Pos:HR + all_star:most_valuable_player + Steroids:SLG
  ```

- Pitchers:

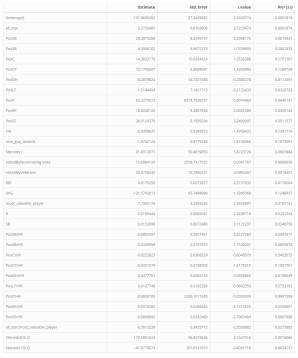
  ```
  wein ~ W + L + SO + SV + Steroids + most_valuable_player + all_star + cy_young_award
  + rolaids_relief_man_award + votedBy + StarterReliever + Steroids:most_valuable_player
  ```

Results

In terms of our model's performance on diagnostics on the training model, we will examine the coefficients, qqplots, and confusion matrices

Coefficients

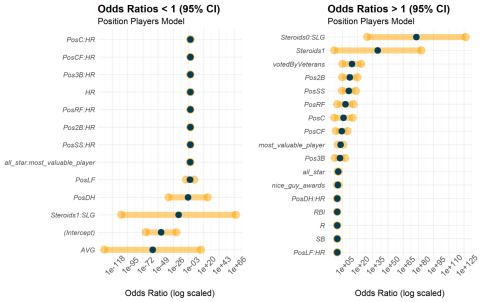
Position Players Coefficients

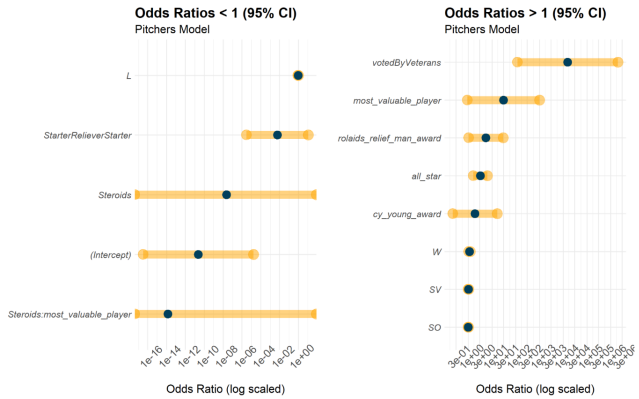


Pitchers Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-24.4649914	6.8720367	-3.5600787	0.0003707
W	0.1262425	0.0382555	3.2999842	0.0009669
L	-0.0544986	0.0281140	-1.9384840	0.0525642
SO	0.0010685	0.0008519	1.2541489	0.2097879
SV	0.0186757	0.0111945	1.6682873	0.0952587
Steroids	-17.5886660	2869.5229192	-0.0061295	0.9951094
most_valuable_player	3.4047341	1.7836545	1.9088529	0.0562811
all_star	1.1730249	0.3428691	3.4212033	0.0006234
cy_young_award	0.6546729	1.0993614	0.5955029	0.5515074
rolaids_relief_man_award	1.7036741	0.8533697	1.9964081	0.0458895
votedByVeterans	9.6297469	2.4845046	3.8759224	0.0001062
StarterRelieverStarter	-5.1514527	3.8447536	-1.3398655	0.1802891
Steroids:most_valuable_player	-31.9055837	7125.9414979	-0.0044774	0.9964276

Position Players Odds	Pitchers Odds

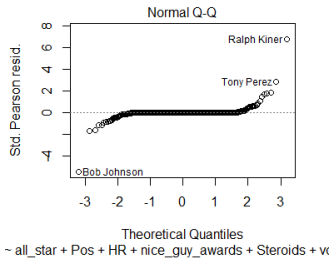


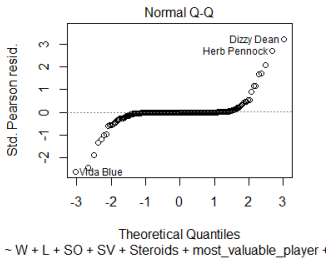


From both of these charts we can see which coefficients, and their corresponding odds ratios, were the largest and most likely to be associated with increasing/decreasing hall of fame likelihood. One thing of note is that the `Steroids` coefficient is positive, but that does not include the interaction of `Steroids` with `SLG`, which means that the overall effect `Steroids` has on the chance a player has to get in the hall of fame is still negative.

QQ-Plots:

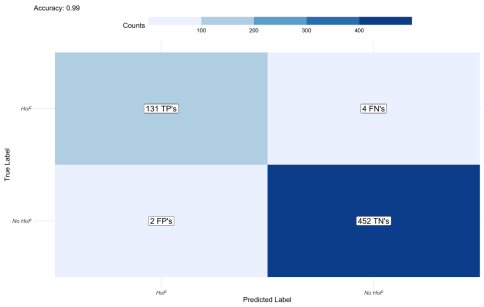
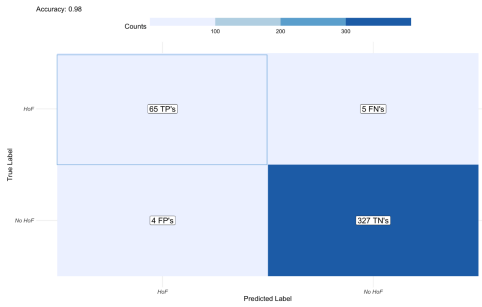
Position Players	Pitchers





As we can tell, for both the hitting and pitching models, the models have a similar fit to the training data. For almost all of the values, they fit right along the line, meaning the residuals are zero or close to zero for most of the data. However, at the extremes of the theoretical quantiles, we do see some variation in the residuals away from the model for both the pitchers and position players (although the pitchers model has more extremes), so the model is not close to perfect.

Confusion Matrix

Position Players	Pitchers
	

We can tell that both of our models have an extremely high accuracy, with very few false positives or false negatives. Thus, both of our models are well-fitted to the training data.

Prediction In addition to how our model performed on the training data, we wanted to see how it would work on future prospective hall of fame inductees. As such, our final prediction results are published on the web application **here**

In terms of our our predictions, we would say that they are, for the most part, accurate. Obviously, our opinion is not a perfect evaluation of whether or not a player will get into the hall of fame, but as people who consume a lot of baseball content, as well as participate in the baseball community, we are likely fairly close to what experts in baseball statistics would look like (not to brag). However, there are a few players in which we disagree with the model's choice in the position player side, namely Adrian Beltre and Yadier Molina. This may be because of overfitting to our training data or it also could be a result of a changing criteria when it comes to what makes player hall of fame worthy.

It is also important to note that these predictions are "now-casting" not forecasting, since we are essentially assuming that all the players that we are predicting are all retired. As such, since our model looks at many counting stats, players who are not retired yet and have low probabilities to get into the hall of fame can, in theory, improve their chances by picking up more stats over the rest of their careers. However, we still believe that our model brings valuable information regarding players who have retired or are at the end of their careers, since they are not likely to improve their chances much, if at all.

References https://library.fangraphs.com/misc/war/positional-adjustment/