Baseball Hall of Fame Report

Andy, Dante, Kai, and Kobe

2023-03-14

Introduction

Background and Motivation

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.

Research Question

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.

We'll be using R version `2022.02.3+492` throughout this project, so please update to at least R version `2022.02.3+492` to be able to replicate our process.

Data Description

Our data consist 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 themselves. However, some categorical variables we may use include factors such as year, league, team, 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 publicly suspected to use steroids or not according to ESPN, MLB PED test results, and miscellaneous PED related court documents.

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 will have two different data sets, since metrics for batters and pitchers are vastly different. However, we will still be predicting the same variable: hall of fame. Hall of fame is a binary categorical variable, 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 data set and 400 observations in the pitching data set.

Exploratory Data Analysis

Correlation Plots

Figure 1: Correlation Plots

We can see that a few of our variables have multicollinearity, especially the hitting rate stats which are amalgamations of counting hit stats. However not all predictors are as highly correlated with each other. So, we will carry on with our analysis using most of the variables, since we deem those variables as important factors.

Awards EDA & Steroids SVD plots

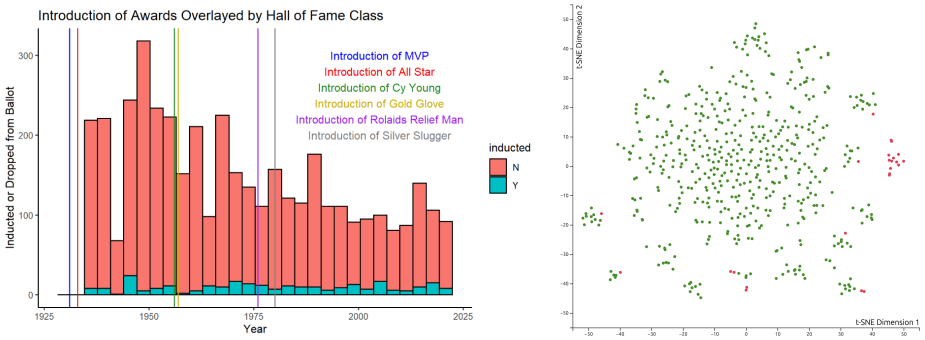

Figure 2: Awards and Election Proportion, tSNE Plot

A time plot showing the total counts of players inducted and players no longer eligible for the normal ballot. Introduction of awards overlaid to explain a lack of significance which we might see for accolades introduced in the mid 20th century after a large proportion of total players who have ever been on ballot had already been inducted or dropped off of the ballot. A player gets dropped off the ballot when they have been on the ballot for more than 10 years or received 5% or less of the votes by all voters. Additionally, a voter can only vote for up to 10 players on any specific ballot. Below are the cumulative percentages of the proportion of players who had either been inducted into the hall of fame or dropped off the ballot by the time each award was introduced. As we see, almost 2/3 of players who had ever been on the ballot had their Hall of Fame candidacy decided before the introduction of the Silver Slugger award.

- Before MVP & All-Star – 0%
- Before Cy Young and Gold Glove Award ~ 33.9%

- Before Rolaids Reliever of the Year and Silver Slugger Award ~ 61%

Using t-SNE, a dimension reduction approach, we can visualize the many variables we have into two principal components, which is what is seen here. As we can tell from this clustering, many of the position players who took steroids throughout their career colored in *red* have eerily similar career statistics. As such, it made sense to make `Steroids` its own variable. The axes have no reasonable interpretation as is with some dimension reduction techniques.

Methods

Family-wise Error

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

Assumptions

We are assuming a constant probability of being elected into the Hall of Fame throughout the history of baseball, a necessary assumption for logistic regression with a binomial random variable response. This seems reasonable according to our `Awards EDA`. Additionally, we are assuming homoscedasticity and normally distributed errors which we will show through `qqplots` after we have filtered our data properly. Lastly, we are assuming each observation is independent of one another, which we must be careful with. Baseball is a series of one-on-one match ups between a batter and a pitcher. It is zero-sum, meaning the success of a pitcher is correlated to the failure of a batter. However, because the number of times any hitter will face the same pitcher is negligible compared to the amount of pitchers the batter will otherwise face throughout their entire career, we will assume independence.

Data Wrangling

- For batters, we summed across playerID groups in the "Batting" data frame in the Lahman package. This allowed us to see career totals for rate and counting stats for each observation (player) like batting average (AVG), walks (BB), and total hits for each of the four base categories (1B, 2B, 3B, HR). We 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 sumed across playerID groups in the "Pitching" data frame in the Lahman package. This allowed us to see career totals for recognizable rate and counting stats for each observation (pitcher) like innings pitched (IP), strike outs (SO), earned runs (ER), and we additionally mutated $\frac{9\times\text{Earned Runs}}{\text{Innings Pitched}}$ and $\frac{\text{Walks+Hits}}{\text{Innings Pitched}}$ to get the all important ERA (Earned Run Average) and WHIP (Walks plus Hits per Inning Pitched) 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.

The initial AIC's (approximately 220 and 80 for our Position Player and Pitcher models respectively) from our step() selection process and some rather large Cook's Distance values indicated the necessity to mutate additional predictors. Players like Barry Bonds, Roger Clemens, and Manny Ramirez were being flagged by our Cook's Distance Plots as uncharacteristically not elected into the Hall of Fame.

- The late 90's, early 2000's Steroids era of baseball altered the way the Baseball Writers Association of America (BBWAA) voters viewed the careers of some players in spite of otherwise Hall of Fame worthy numbers. And so, using web scraped 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 indicted 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; BBWAA or the Veterans Committee. The BBWAA typically votes players 5-15 years after their retirement, 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 add 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 to perform at a certain offensive level to be considered above replacement level. For example, a first-baseman who has what is considered an easy position defensively, is expected to perform 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 variable selection via the step() function revealed the importance of relief pitcher accolades. Relief pitchers are a subset of pitchers who typically pitch ~ 60 innings a season. We felt the need for our model to be differentiate between a starter and relief pitcher, and so we mutated such a predictor. This variable was accurate at identification with a cutoff of 60 innings per season after cross-referencing with 20 starters and 20 relievers.

Model Selection

For our model building, we used essentially every possible variable that could potentially be useful and did not contain more than 30 missing observations. 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 a large number of models. Since we are more focused on the prediction aspect, as we wanted to be able to predict current eligible players, we chose to use AIC as our primary criteria for model selection over other criteria such as BIC.

Results

These tables are the summary information of the models we obtained from many iterations of trying different variables for both pitchers and position players. Further, the first iteration is considered the "base" model which includes all the basic measurements in our dataset as it doesn't include manipulated data like `steroids` and `votedBy`, both variables which needed further research to obtain. The second iteration includes the variables we didn't include in the first iteration whilst the third and final model includes interactions of any variables as well. Thus, the third iteration can be considered the most complex model out of the 3.

One thing that is important to note is how, for both models, the number of observations gets higher with once we arrive at the final model. This is due to some of the variables having missing values in them, so once we eliminated that variable, it allowed us to add those observations back into the model. As a result, our final models use more observations compared to the previous models.

As you can observe, the AIC for the models go down as we clean up the variables and add interactions, so we ended up settling with the last model which gave us the lowest AIC.

Table 1: Pitchers Model(s)

	Dependent variable:		
	wein		
	Iteration 1	Iteration 2	Iteration 3
	(1)	(2)	(3)
W	0.054*** (0.010)	0.076*** (0.016)	0.085*** (0.023)
ERA	−2.048* (1.107)	−3.089** (1.393)	
all_star	0.479*** (0.134)	0.488*** (0.154)	1.048*** (0.275)
WHIP	−4.454 (5.471)	1.672 (6.571)	
SO	−0.0002 (0.001)	0.00004 (0.001)	0.001 (0.001)
gold_glove	0.163 (0.244)	0.093 (0.338)	
SV	0.012*** (0.004)	0.022*** (0.007)	
cy_young_award	−1.025** (0.500)	0.939 (0.797)	1.938* (1.087)
votedByVeterans			9.495*** (2.313)
StarterRelieverStarter			−7.384* (3.983)
SV:StarterRelieverReliever			0.014 (0.010)
SV:StarterRelieverStarter			0.100** (0.043)
StarterRelieverReliever:rolaids_relief_man_award			1.439 (0.972)
StarterRelieverStarter:rolaids_relief_man_award			−0.234 (2,240.016)
Steroids:most_valuable_player			−37.590 (11,865.490)
Steroids		−38.847 (2,007.568)	−15.584 (5,014.066)
most_valuable_player	0.291 (1.072)	0.876 (1.281)	3.915** (1.914)
pitching_triple_crown	1.294* (0.751)	1.168 (0.887)	
rolaids_relief_man_award	1.060** (0.417)	1.078* (0.559)	
nice_guy_awards	−0.632 (0.476)	−1.427** (0.677)	
Constant	−2.057 (4.858)	−11.977* (6.405)	−22.946*** (5.898)
Observations	342	342	401
Log Likelihood	−46.083	−33.545	−19.255
Akaike Inf. Crit.	118.167	95.090	66.511

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Position Players Model(s)

	Dependent variable:		
	wein		
	Iteration 1	Iteration 2	Iteration 3
	(1)	(2)	(3)
G	0.004 (0.003)	0.004 (0.003)	
AB	0.004 (0.003)	0.003 (0.003)	
R	0.009** (0.004)	0.011*** (0.004)	0.011** (0.005)
H	−0.019** (0.009)	−0.019** (0.010)	
HR	−0.010 (0.007)	−0.008 (0.007)	0.031** (0.015)
RBI	0.002 (0.002)	0.003 (0.003)	0.008 (0.006)
SB	−0.003 (0.002)	−0.003 (0.002)	0.011* (0.007)
BB	0.004 (0.004)	0.002 (0.004)	
PrimaryLgNL	−0.206 (0.466)	−0.269 (0.479)	
Pos2B:HR			−0.064*** (0.024)
Pos3B:HR			−0.022 (0.024)
PosC:HR			−0.025 (0.026)
PosCF:HR			−0.033 (0.082)
PosDH:HR			0.021 (0.051)
PosLF:HR			−0.001 (0.017)
PosRF:HR			−0.048** (0.023)
PosSS:HR			−0.079** (0.031)
all_star:most_valuable_player			−0.295 (0.261)
Steroids:SLG			−105.800 (97.506)
all_star	0.494*** (0.089)	0.530*** (0.097)	1.321*** (0.387)
most_valuable_player	0.373 (0.451)	0.200 (0.486)	4.087* (2.150)
AVG	214.493*** (78.055)	212.914*** (80.415)	140.045** (55.842)
OBP	−56.691 (45.829)	−45.906 (47.227)	
SLG	8.154 (19.633)	−0.116 (20.421)	
Steroids		−3.836** (1.663)	41.428 (45.389)
votedByVeterans			23.404*** (7.676)
nice_guy_awards	0.438 (0.408)	0.399 (0.432)	−0.526 (0.883)
Pos2B	2.294** (0.945)	2.087** (0.971)	20.092*** (7.107)
Pos3B	−0.227 (0.971)	−0.502 (1.012)	3.629 (7.117)
PosC	2.150** (1.017)	2.302** (1.051)	9.439 (7.509)
PosCF	0.068 (0.936)	−0.192 (0.949)	2.977 (14.015)
PosDH	1.829 (1.270)	1.806 (1.393)	−2.482 (17.873)
PosLF	1.238 (0.832)	1.067 (0.851)	1.688 (6.026)
PosRF	0.724 (0.822)	0.785 (0.833)	13.536* (7.831)
PosSS	1.130 (0.936)	1.138 (0.964)	19.336*** (7.185)
gold_glove	−0.108 (0.105)	−0.105 (0.103)	
silver_slugger	−0.121 (0.169)	0.003 (0.192)	
hank_aaron_award	0.324 (2.846)	1.006 (1.835)	
Constant	−56.626*** (20.579)	−56.525*** (20.882)	−86.690*** (26.911)
Observations	469	469	566
Log Likelihood	−82.673	−78.995	−19.735
Akaike Inf. Crit.	219.347	213.991	97.469

Note: *p<0.1; **p<0.05; ***p<0.01

Our final models are as follows: