Baseball Hall of Fame Report

Andy, Dante, Kai, and Kobe

2023-03-12

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

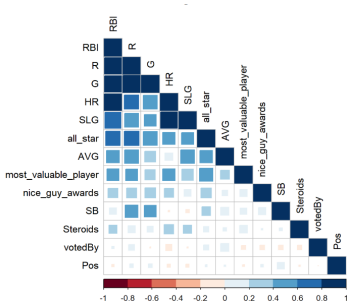
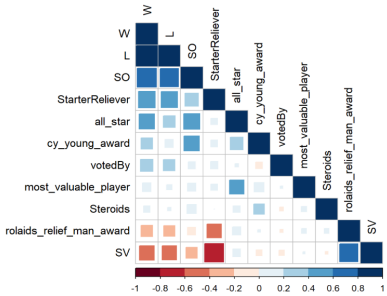
Our data consist of mostly quantitative variables mixed in with some categorical variables. Most baseball metrics come in quantitative form so it is often hard to obtain categorical measurements unless they come directly from the quantitative variables themselves. However, some categorical variables 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.

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, 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.

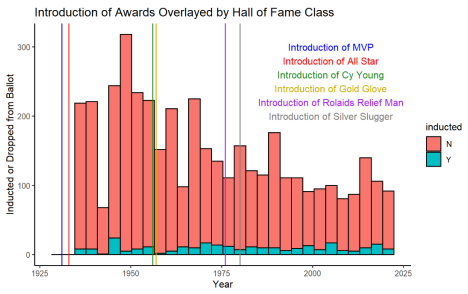
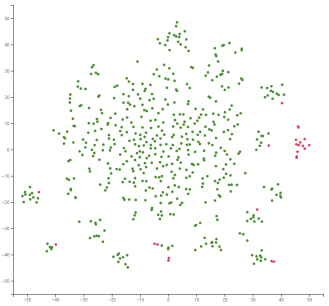
Exploratory Data Analysis

Correlation Plots

Position Players	Pitchers
	

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'll carry on with our analysis using most of the variables, since we deem those variables as important factors.

Awards EDA & Steroids SVD plots

Awards and Elected Counts	Steroids Clusters
	

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. 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 have eerily similar career statistics. As such, it made sense to make `Steroids` its own variable.

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 homoskedaskity 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 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.

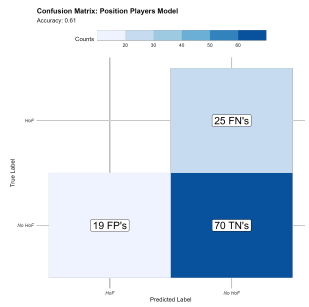
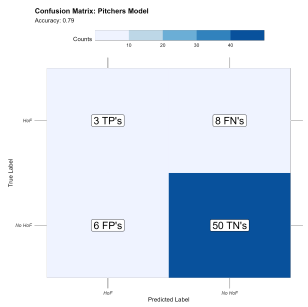
In addition to the base covariates, a large initial AIC from our step() selection process and some rather large Cook's Distance values indicated the necessity to mutate additional predictors.

- 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 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 inducted 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, 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 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 perform at a certain offensive level to be considered good. 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 a predictor which identifies if a pitcher is a starter or reliever based on seasonal innings pitched totals.

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.

In addition, we attempted to utilize AIC over K-fold CV and LOOCV because of the following:

Position Players	Pitchers
	

Our position player model seems to be identifying the players that don't make it to the hof correctly but is not able to correctly identify any hof players. This could possibly stem from the problem that there is an imbalance in the

data set as there are many more non-hof players compared to hof players (~ 18% of players are in the Hall of Fame in both batter and pitcher data frames) Thus, CV may not be practical for our data and make because of its expected behavior.

Results

Table 4: Pitchers Model(s)

	Dependent variable:		
	wein		
	Iteration 1	Iteration 2	Iteration 3
	(1)	(2)	(3)
W	0.117*** (0.043)	0.172*** (0.064)	0.126*** (0.038)
L	0.002 (0.039)	0.016 (0.042)	−0.054* (0.028)
G	−0.004 (0.005)	−0.004 (0.005)	
IP	−0.001 (0.004)	−0.005 (0.005)	
ERA	−4.504** (2.090)	−4.268** (1.975)	
all_star	0.399** (0.182)	0.347* (0.188)	1.173*** (0.343)
WHIP	3.279 (9.587)	2.470 (9.207)	
SO	−0.0004 (0.001)	0.0003 (0.001)	0.001 (0.001)
gold_glove	0.291 (0.780)	0.158 (1.275)	
SV	0.034*** (0.013)	0.037*** (0.014)	0.019* (0.011)
cy_young_award	1.188 (0.981)	0.784 (0.993)	0.655 (1.099)
Steroids		−19.433 (2,477.702)	−17.589 (2,869.523)
most_valuable_player	2.200 (1.792)	2.287 (1.729)	3.405* (1.784)
pitching_triple_crown	2.273 (1.679)	1.841 (1.640)	
rolaids_relief_man_award	1.112* (0.666)	1.389* (0.782)	1.704** (0.853)
nice_guy_awards	−1.377* (0.761)	−1.581* (0.821)	
votedByVeterans			9.630*** (2.485)
StarterRelieverStarter			−5.151 (3.845)
Steroids:most_valuable_player			−31.906 (7,125.941)
Constant	−10.682 (9.017)	−12.588 (9.418)	−24.465*** (6.872)
Observations	271	271	401
Log Likelihood	−24.206	−22.742	−18.617
Akaike Inf. Crit.	80.412	79.483	63.233
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: 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

The former 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.

Our final models are as follows:

$$\begin{aligned}\log\left[\frac{P(\widehat{wein}=1)}{1-P(\widehat{wein}=1)}\right] &= -86.69 + 1.32(all\_star) + 20.09(Pos_{2B}) + \\ &\quad 3.63(Pos_{3B}) + 9.44(Pos_{C}) + 2.98(Pos_{CF}) - \\ &\quad 2.48(Pos_{DH}) + 1.69(Pos_{LF}) + 13.54(Pos_{RF}) + \\ &\quad 19.34(Pos_{SS}) + 0.03(HR) - 0.53(nice\_guy\_awards) + \\ &\quad 41.43(Steroids) + 23.4(votedBy_{Veterans}) + 0.01(RBI) + \\ &\quad 140.04(AVG) + 4.09(most\_valuable\_player) + 0.01(R) + \\ &\quad 0.01(SB) - 0.06(Pos_{2B} \times HR) - 0.02(Pos_{3B} \times HR) - \\ &\quad 0.03(Pos_{C} \times HR) - 0.03(Pos_{CF} \times HR) + 0.02(Pos_{DH} \times HR) + \\ &\quad 0(Pos_{LF} \times HR) - 0.05(Pos_{RF} \times HR) - 0.08(Pos_{SS} \times HR) - \\ &\quad 0.3(all\_star \times most\_valuable\_player) - 105.8(Steroids \times Steroids_{SLG})\end{aligned} \tag{1}$$

$$\begin{aligned}\log\left[\frac{P(\widehat{wein}=1)}{1-P(\widehat{wein}=1)}\right] &= -24.46 + 0.13(W) - 0.05(L) + \\ &\quad 0(SO) + 0.02(SV) - 17.59(Steroids) + \\ &\quad 3.4(most\_valuable\_player) + 1.17(all\_star) + 0.65(cy\_young\_award) + \\ &\quad 1.7(rolaids\_relief\_man\_award) + 9.63(votedBy_{Veterans}) - 5.15(StarterReliever_{Starter}) - \\ &\quad 31.91(Steroids \times most\_valuable\_player)\end{aligned} \tag{2}$$

Coefficients

Now, we take a look at the coefficients that were computed from our model, specifically we look at the odds ratio and the corresponding 95% confidence interval to gain insight on what kind of variables may be important in our final models. Note, only the parameters with $< .05$ p-value have plotted odds ratios below.

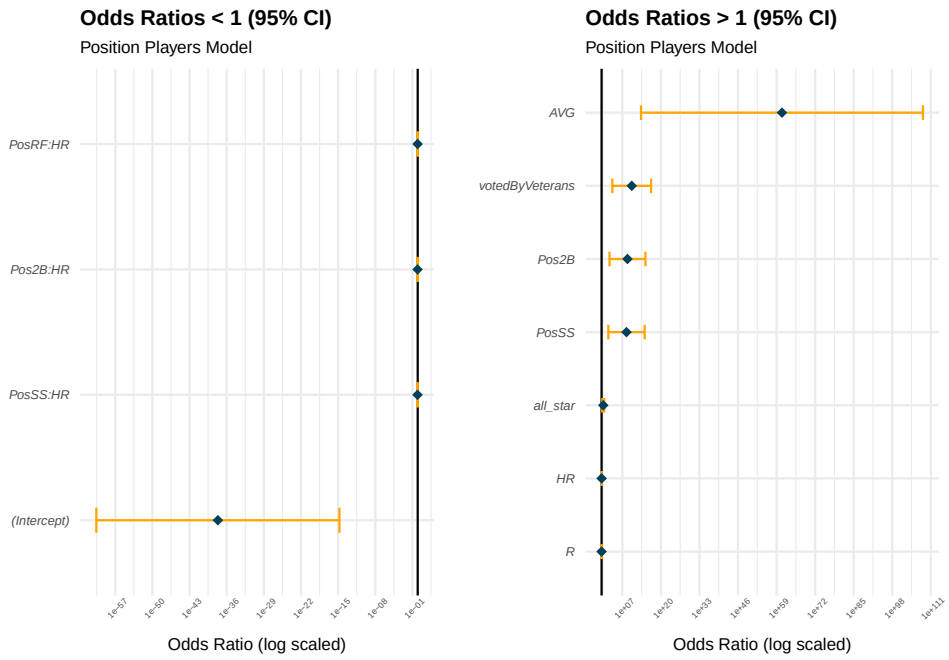


Figure 1: Position Players Odds Ratio CI

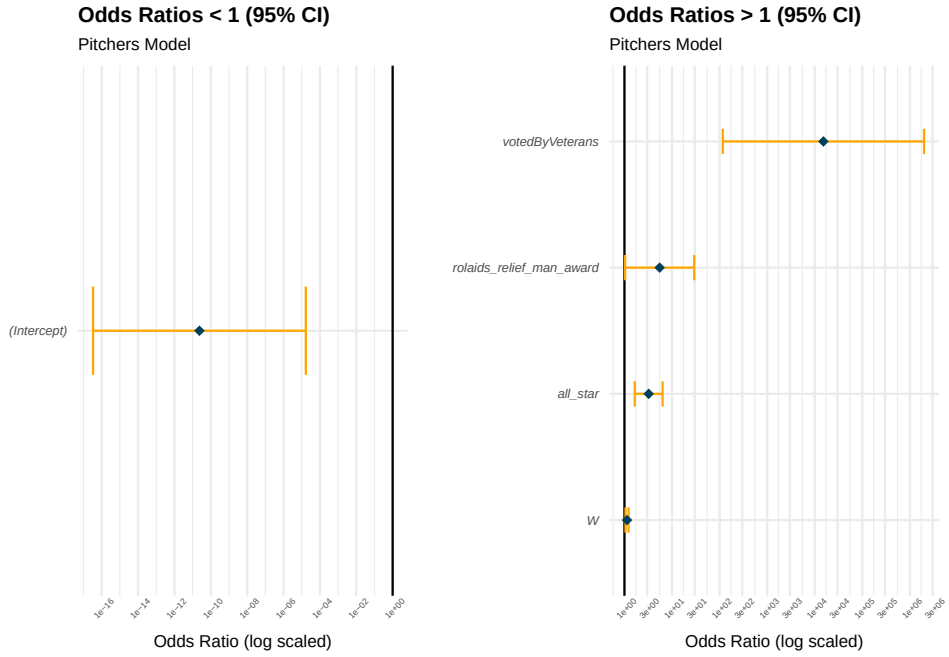
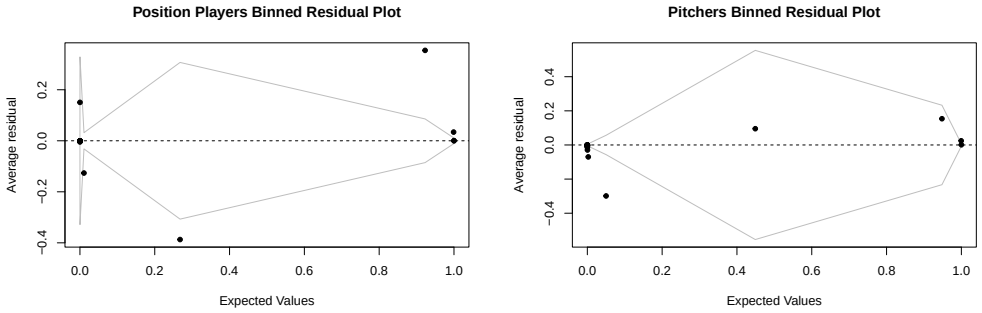


Figure 2: Pitchers Odd Ratio CI

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. While not significant, and therefore not included above in the plot, one thing of note is that the `Steroids` coefficient is positive, but that does not include

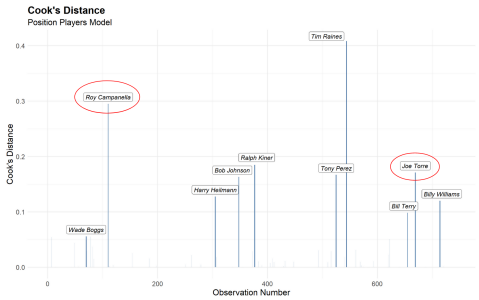
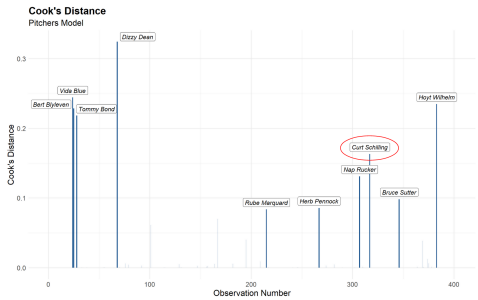
the interaction of `Steroids` with `SLG`, which means that the overall effect `Steroids` has on the chance a player has to get in the hall of fame is still negative.

Binned Residual Plot Diagnostics



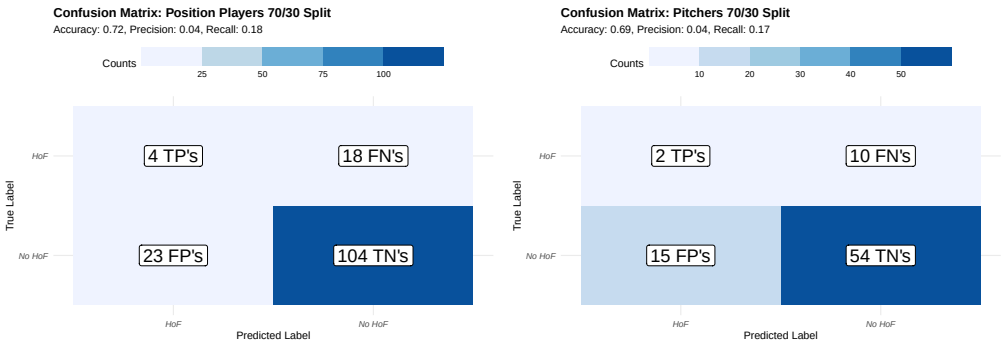
For both models, we can tell that many of the residuals are outside or near the standard-error bounds. This may mean that our models deviate some from true model. As such, even though our models have a low AIC relative to other models we have tried, they still are nowhere close to perfect models

Cook's Distance Diagnostics

Position Players	Pitchers
	

These Cook's Distance plots pass the eye test for weird observations. Joe Torre was inducted more for his contributions to the game as baseball's Commissioner, and not so much for his prowess as a player. Roy Campanella's Kansas City Monarch data is not included in his MLB counting stats, so our model is a bit confused to why he is inducted even though it's entirely deserving. Curt Schilling's character off the field has been widely attributed to why he is not elected despite his successful career.

Training/Test Set



We can tell above that neither model particularly performed very well on the test set. While the overall accuracy is high, both models clearly struggle with correctly identifying a hall of fame player. The precision, which is the number of true positives out of total positives, is below 50% for both models. The same is true for the recall, which is the number of true positives out of total hall of fame players in the test set.

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: **https://klobby19.shinyapps.io/explorable/**

In terms of our our predictions, we would say that they are, for the most part, accurate. 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, both of whom have low predicted chance of getting in. 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.

Discussions and Conclusions

Interpretation

The goal of our research question has always been prediction, and thus, the AIC value which we used to assess our model has left us with many more parameters than if we used the BIC. Below are selected coefficients from our final position player and pitching glm objects. We will interpret them based on the fact that they met our family wise signifiance level, and not necissarly because they have a sign or magnitude which makes sense contextually.

- From the exponentiated `all_star` parameter $e^{\beta_{allstar}} = e^{1.321} = 3.75$, we can say for every all-star selection of a position player, the odds of being elected into the Baseball Hall of Fame increase by 3.75 times when holding

all other predictors constant. From the 95% C.I. [1.95, 9.38], we are 95% confident that for every all-star appearance, the odds of being elected into the Hall of Fame increases by 1.95 to 9.38 times. This makes sense intuitively, as consistent all-star selections is synonymous with consistent player performance. The positive feedback loop of baseball writers awards players accolades and then using those accolades to justify Hall of Fame election is shown through many parameters in both the batter and pitcher models.

- From the exponentiated `votedByVeterans` parameter $e^{\beta_{votedByVeterans}} = e^{12.89} = 396329$, we can say the odds of a pitcher being elected into the Baseball Hall of Fame increase by 396329 times if done by the Veterans Committee when holding all other predictors constant. From the 95% C.I. [1062, 39143030000], we are 95% confident that for every all-star appearance, the odds of being elected into the Hall of Fame increases by 1062 to 39143030000 times. This seems like an insane result, but when you consider how few players are considered by the Veterans Committee and how they are almost always elected into the Hall of Fame from their consideration, then it makes sense why our number seems so ludacris based on our lack of statistical power.

We had several fringe parameters for our position player and pitcher models which did not quite meet our bonferroni correction significance levels, but were close. We listed them below on a selection criteria of being close to significant and interesting. They follow the same interpretation logic as above.

- Position Players
  - Pos2B, 530855280 w/ 95% C.I. [6199, 2.13×10^{16}]
  - VotedbyVeterans, 1.46×10^{10} w/95% C.I.[3.69×10^{5}, 8.47×10^{19}]
- Pitchers
  - Wins, 1.183 w/ 95% C.I. [1.081, 1.388]
  - all_star, 3.286 w/ 95% C.I. [1.72, 10.63]

Limitations

We acknowledge some limitations of our processes. Firstly, our binomial random variable response of being elected into the Baseball Hall of Fame (`wein`) has a naturally low probability of success Therefore, many cross-validation methods of model selection and assessment are unavailable to us. Additionally, our research question surrounding Hall of Fame voting is not a perfect science by any means. Any election process harbors some roots in favoritism, superficial perception, and dynamic measures of success Baseball Hall of Fame voting is no different. Lastly, the poor record keeping of baseball statistics prior to the mid 20th century has imparted many `NA`s in our data frames. This greatly reduces the power in which we can effectively measure which covariates are associated with Hall of Fame election.

Future

In future research, we would like to solve our issues of inappropriate predictors and the lack of power they cause. It would interesting to use time series analysis with election into the Hall of Fame as the terminating event, while also making use of cohort-predictor-time variables. It would be interesting to see if we can somehow quantify popularity of players by possibly looking at their social media pages (though it might be hard to gauge popularity before the social media age).

References

- Fangraphs. (n.d.). Positional Adjustment. Retrieved from https://library.fangraphs.com/misc/war/positional-adjustment/
- Baseball Reference. (n.d.). Baseball Statistics and History. Retrieved from https://www.baseball-reference.com/
- Lahman, S. (n.d.). The Baseball Archive. Retrieved from https://www.seanlahman.com/baseball-archive/statistics/
- Wikipedia. (2021, December 14). List of Major League Baseball players suspended for performance-enhancing drugs. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_Major_League_Baseball_players_suspended_for_performance-enhancing_drugs
- ESPN. (2009, July 31). A-Rod admits using PEDs from 2001-03. Retrieved from https://www.espn.com/mlb/news/story?id=4366683
- RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Appendix

- All: General data tydying including the joining, mutation, and filtering necessary for analysis. Developed presentation slides and report. Quality control of final report.
- Andy:
- Dante: Proposed research question and gathered initial datasets. Deligated tasks based on project needs. Assumptions and issues. Awards EDA, Odd ratios interpretation, Cook's Distances, impact of our results.
- Kai:
- Kobe: