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.

We'll be using R version 4.0.x throughout this project, so please update to at least R version 4.0 to be able to replicate our process.

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

TODO: Add more EDA details

Steroids Clustering

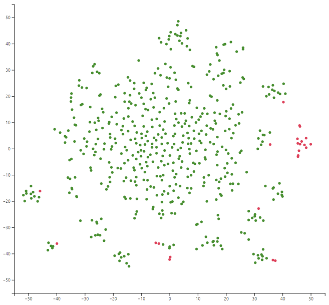


Figure 1: Clustering with Steroids in Red

Methods

TODO: improve format

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 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 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.

These models were trained on the same corresponding training dataset, which consists of `80%` of the full dataset are using. Thus, we used a 80/20 split, to get our training and testing data. This allows us to further evaluate the performance of our models as the model will be predicting on unseen data.

Confusion Matrix

TODO: Explain why we're using our training data, fiddle with plot – text size too big

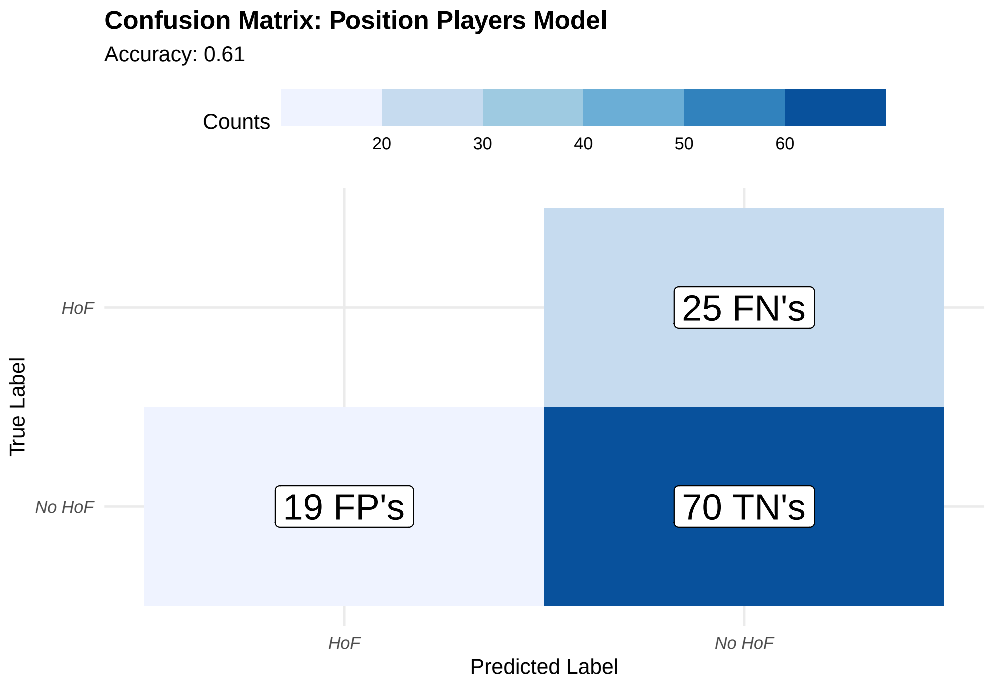


Figure 2: Position Player Confusion Matrix

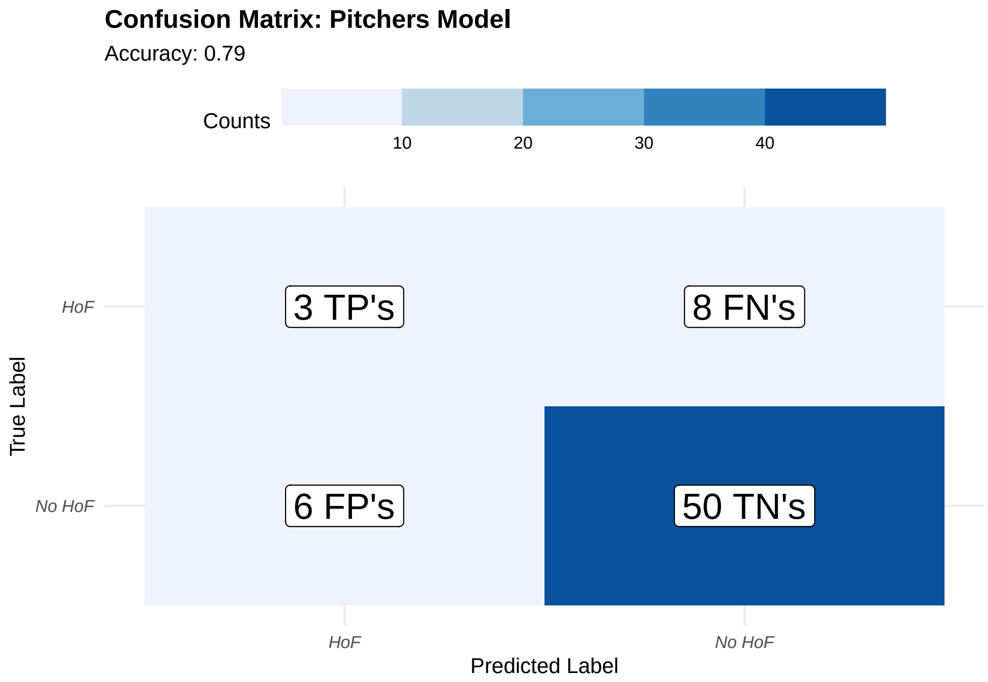


Figure 3: Pitchers Confusion Matrix

Our position player model seems to be identifying the non-hof players 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 dataset as there are many more non-hof players compared to hof players. Thus, with only around 20 of the testing data consisting of hall of famers, we could consider other possibilities like downsampling to balance out the dataset.

Same goes with the pitchers mode, however, it is predicting a bit better than our position model.

Results

TODO: format table so that text wrap over and under the table, more detail necessary

Table 1: Pitchers Model(s)

	Dependent variable:		
	wein		
	Iteration 1	Iteration 2	Iteration 3
	(1)	(2)	(3)
W	0.117*** (0.043)	0.172*** (0.064)	0.168*** (0.061)
L	0.002 (0.039)	0.016 (0.042)	−0.079* (0.041)
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)	1.190*** (0.442)
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)	0.028 (0.019)
cy_young_award	1.188 (0.981)	0.784 (0.993)	0.775 (1.267)
Steroids		−19.433 (2,477.702)	−18.329 (2,720.043)
most_valuable_player	2.200 (1.792)	2.287 (1.729)	4.710** (2.328)
pitching_triple_crown	2.273 (1.679)	1.841 (1.640)	
rolaids_relief_man_award	1.112* (0.666)	1.389* (0.782)	1.844* (1.039)
nice_guy_awards	−1.377* (0.761)	−1.581* (0.821)	
votedByVeterans			12.888*** (4.188)
StarterRelieverStarter			−6.464 (5.575)
Steroids:most_valuable_player			
Constant	−10.682 (9.017)	−12.588 (9.418)	−28.239*** (10.536)
Observations	271	271	320
Log Likelihood	−24.206	−22.742	−12.310
Akaike Inf. Crit.	80.412	79.483	48.620
Note:			*p<0.1; **p<0.05; ***p<0.01

The former table is the summary information of the models we obtained from many iterations of trying different variables for both pitchers. 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.

As you can observe, the AIC for the model 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. The same process was applied to the position players model.

Our final models are as follows:

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

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

$$
\begin{aligned}
\log\left[\frac{P(\widehat{wein = 1})}{1 - P(\widehat{wein = 1})}\right] = &-28.24 + 0.17(W) - 0.08(L) + \\
&0(SO) + 0.03(SV) - 18.33(Steroids) + \\
&4.71(most\_valuable\_player) + 1.19(all\_star) + 0.77(cy\_young\_award) + \\
&1.84(rolaids\_relief\_man\_award) + 12.89(votedBy_{Veterans}) - 6.46(StarterReliever_{Starter}) + \\
&NA(Steroids \times most\_valuable\_player)
\end{aligned}
\tag{2}
$$

Coefficients

TODO: label plots with figure # and change font of plots

In this section, 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.

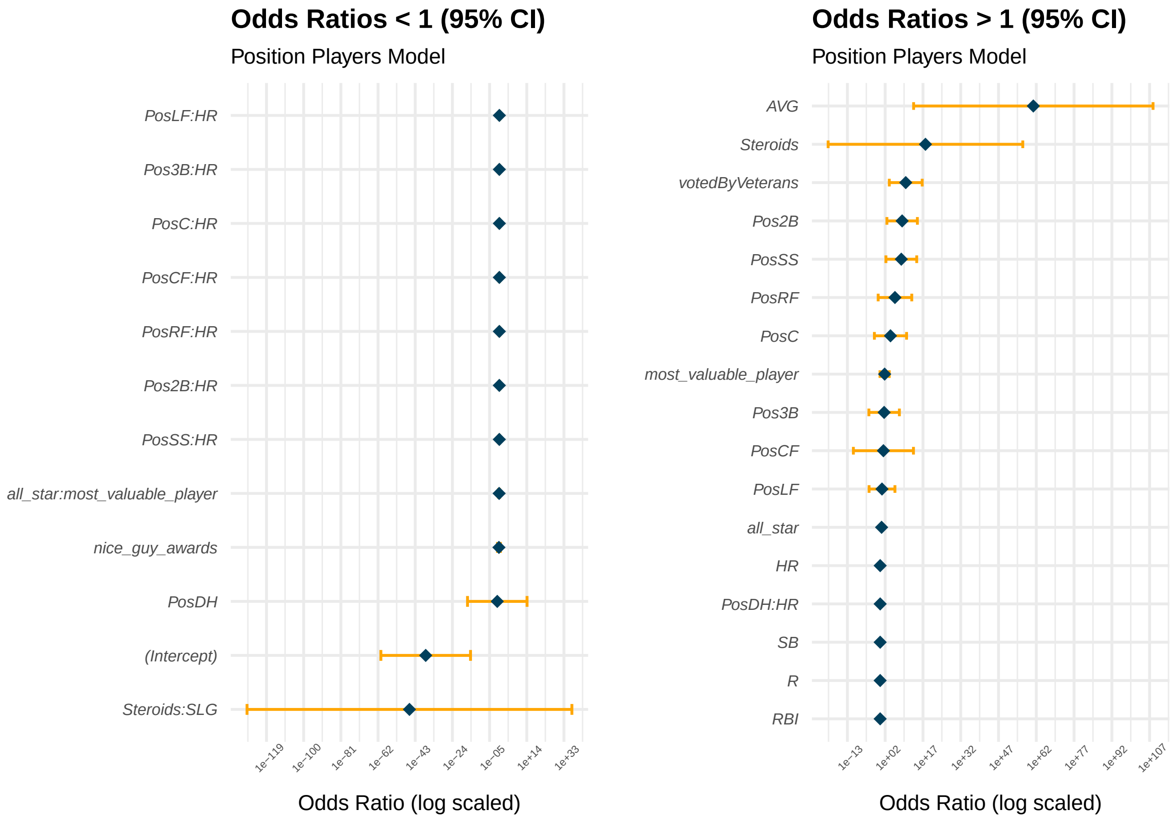


Figure 4: Position Players Odds Ratio CI

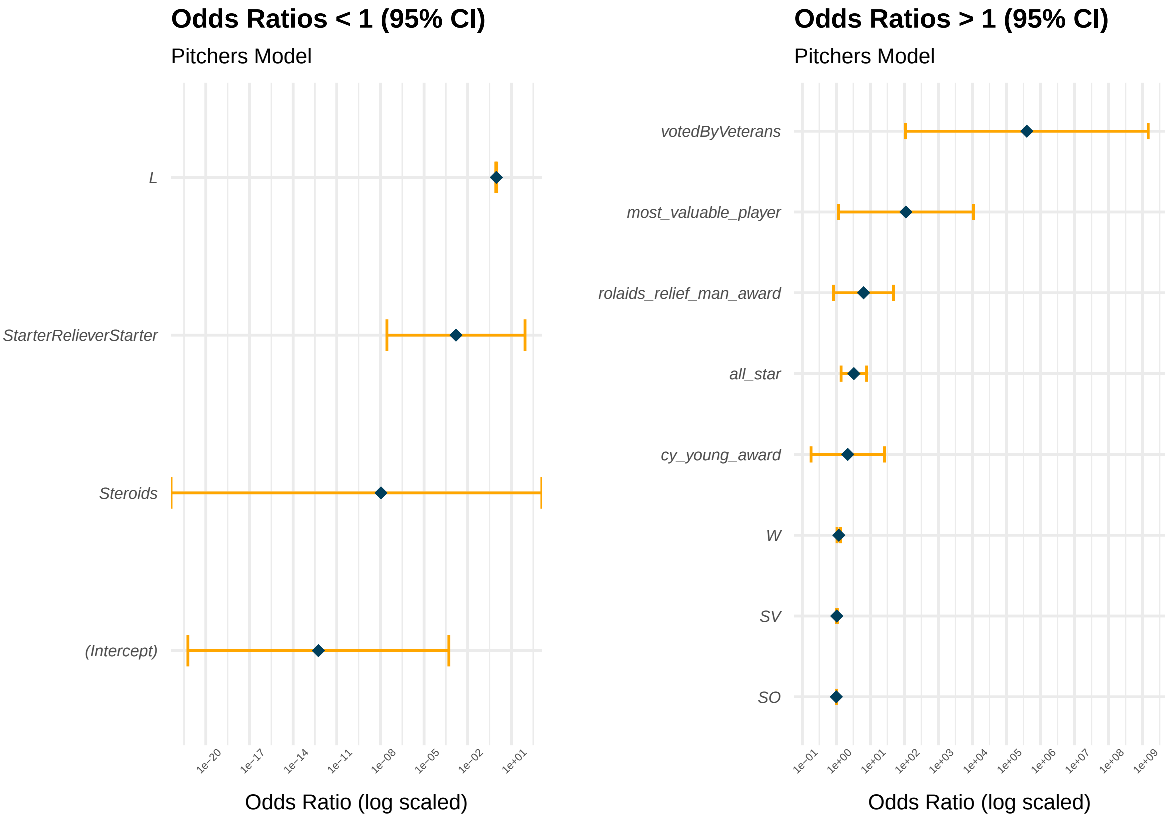
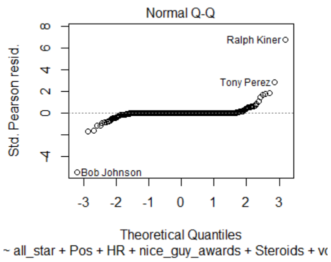
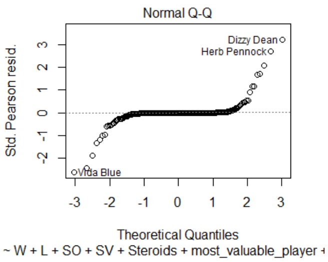


Figure 5: Pitchers Odd Ratio CI

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. 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.

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

- 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