Modeling Players' Chances of Entering the Baseball Hall of Fame

Ryan Anderson

December 9, 2022

This report describes an attempt to model using different statistical techniques the selection process for the Baseball Hall of Fame (HOF). We describe the history and election procedures of the Hall of Fame, and also the appropriateness of kernel estimation and machine learning methods to the problem of predicting which players will make it in.

This report investigates two distinct questions relating to HOF selection: (1) how best to model and estimate the max share of HOF votes received given a player's vector of career batting data, and (2) whether machine learning models can identify whether a player will be elected to the HOF, once more given a vector of their career batting data.

1 Introduction

Election to the Baseball Hall of Fame is the great dream of professional baseball players all over the world. Part of what makes election so valuable is how rare it is – out of some 20,000 professional baseball players for which there are records, only about 200 have entered the Hall.

The breadth of data compiled in the course of professional baseball at the player level suggests that we may be able to use common statistical techniques to predict which players will someday make it in.

1.1 The Baseball Hall of Fame

Professional baseball is the oldest of the major American sports – the National League, one of the two halves of Major League Baseball as currently constructed, began game operations in 1876 with direct predecessors of today's Atlanta Braves and Chicago Cubs as member clubs. With this depth of history, it is no surprise that professional baseball is also the sport most focused on its history and tradition.

The success of and interest in the Baseball Hall of Fame is one of the main causes for such rich sports history. Founded in 1936 in Cooperstown, NY by Stephen Carlton Clark (heir to the Singer Sewing Machine fortune), the HOF elected five men to its first class: Ty Cobb, Walter Johnson, Christy Mathewson, Babe Ruth, and Honus Wagner. Since then, many of the game's greatest players have been enshrined in an annual election process. Though there

is more than one avenue towards inclusion in the HOF, we'll focus on one particular process known as the BBWAA election.

In the BBWAA election, a voting body made up of hundreds of long-standing baseball writers spread out across the country receive ballots once a year and may vote for up to ten eligible players. Any player receiving at least 75% of the votes in a given year is elected to the HOF. Players are eligible for the BBWAA election upon the fulfillment of two requirements: (1) that they played in the MLB for at least ten years and (2) that five years have passed since they retired from playing professional baseball.

As per the Hall of Fame website, the BBWAA writers are encouraged to vote "based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played." As integrity, sportsmanship and character are less easily quantifiable components of a player's career, we will build our models that follow with an understanding that BBWAA voting decisions cannot be entirely explained by career batting data; indeed, BBWAA voters are not accountable to anyone else and may decide arbitrarily how to vote.

One illustration of the embedded arbitrariness in BBWAA voting is the lack of unanimous electees – all time legends of the game such as Willie Mays or Ken Griffey, Jr. have frequently received only about 95% of the vote. It was not until the 2019 election of Yankees relief pitcher Mariano Rivera that any player was unanimously selected.

1.2 Data Description

Baseball has always been the sport most concerned with statistics – when Joe DiMaggio was having his breakout seasons with the Yankees in the mid-1930s, the legend goes that his father, a poor illiterate fisherman who cast off from San Francisco's Fisherman's Wharf, taught himself division to follow his son's batting average.

Classically, batters were evaluated based on a few simple metrics, among which chiefly their batting average. This figure, the number of hits divided by the number of at-bats, varies across players in any given season between 0.150 and 0.400. Over an entire career, achieving a 0.300 batting average is a remarkable feat.

In the last quarter-century or so, commentators have invested energy into compiling further metrics derived from the simple metrics. Bill James, the sportswriter and statistician who spearheaded much of this push, was essentially a nobody writing into the wind all the way back in the 1970s. Then Aaron Sorkin wrote and Brad Pitt starred in a movie about the early-aughts Oakland Athletics' successful use of his philosophy, which privileges metrics like on-base percentage and batting average on balls in play over hits, runs and runs batted in.

Since then, the so-called sabermetrics community has come to dominate baseball discourse, seeking ever more accurate predictors of future success. The writer Dan Szymborski, who works at FanGraphs, is known for a model called ZiPS developed to assess the future impact of players and thereby the teams they play on, out to time horizons extending a decade or more.

Others in the baseball statistics community work on recording and reporting in great detail historical metrics. The team at Baseball Databank maintains several data sets which were used for this project. In particular, three of their data sets were of concern here: HOF

Voting, Batting Data, and People.

The HOF Voting set contains records on which players were eligible in every year of voting going back to 1936, how many votes they received in that year, and whether they ever were inducted into the HOF. This per-player-year data provides our main response variable in Section X, *max_pct*, which we calculate as

$$max\_pct_j = \max_{i \in Y_j}(\frac{V_{i,j}}{B_i}),$$

where *max_pct* for the j^{th} player is given by taking the max across all the player's eligible years, Y_j, of the ratio of votes received by that player in year i, $V_{i,j}$, to the total ballots available in that same year, B_i. Players who have successfully been elected into the HOF are precisely those who have $max\_pct > 0.75$.

The Batting Data set contains season-wide totals for each player on a set of metrics, which we subset here: Games Played, At-Bats, Hits, and Runs. We group and total these at the player level to end up with a table we call Career Totals. Career-wide data is as described above more natural when considering HOF classification.

The final data set of note is the People file, which allows us to recover full names of players detailed in the former two data sets. We join on the unique identifier *playerID* and can then pull strings of first and last name. This is also useful in Section Y, where we predict the results for the 2023 first-time eligible players.

1.3 Problem Statement

With the above data described, we will attempt to examine two questions.

The first relates to taking in a given vector of career batting totals and estimating the *max_pct* a player whose career produced those totals would earn. We will investigate this question with a linear model and kernel regression on the covariates.

The second question relates again to taking in a given vector of career batting totals, but now we attempt to classify the player who produced those totals into yes/no for the HOF. This boils down to predicting whether that player's *max_pct* was above or below the 75% cutoff noted above. We will investigate this question by recourse to machine learning algorithms and will attempt to compare which among a popular few algos produces the best accuracy.

2 Data Exploration

Here we briefly describe some features of our response variable *max_pct*.

Figs. 1a and 1b show histograms and kernel density estimates for *max_pct*. Note that kernel density estimates are computed with a Gaussian kernel and bandwidth set to the value given by Silverman's rule of thumb,

$$bw(X) = \frac{0.9}{n^{\frac{1}{5}}} \min(\sqrt{Var(x)}, \frac{IQR(X)}{1.349})$$

The charts match our intuition – the majority of players receive close to no votes for the HOF, and interestingly even fewer players ever receive max votes in the 60-70% range.

After 75% there appears more mass in the distribution, corresponding to the players who do successfully get in.

In Fig. 5 we zoom in on the distribution of successful inductees' *max_pct*. It's most common for successful inductees to earn about 80% of the vote. We can visualize in the kernel density estimate the fact noted above that only one person ever has been elected unanimously, i.e., with 100% of the vote (a few others have received >98%!).

3 Predicting *max_pct*

With the above in mind, we fit a linear model for *max_pct* against our career batting data as covariates. The model summary can be found in Fig. 3. All the batting variables are found to be significant to the 0.0001 level. The equation to estimate $\hat{max\_pct}$ is given by

$$\hat{max\_pct} = -0.3079 - 0.00012G - 0.0001AB + 0.0003R + 0.0004H + 0.0349career\_years$$

However, we note that the residuals of this model exhibit quite differently to what we would expect under standard homoscedasticity assumptions. The residuals vs fitted and normal Q-Q plots are depicted in Fig 4. With such uneven residual plots, we conclude that linear models may not be well suited to prediction of the *max_pct* variable. We attempt another regression analysis by performing kernel regressions against two of the career batting covariates. The results are shown in Figs. 5a and 5b. We also include 95% confidence intervals for the point estimates produced by the kernel regression.

Idiosyncrasies of the history of baseball may contribute to the wide CIs seen towards the upper end of the range for both covariates analyzed. As a case in point, there is one player who had on the order of 4000 hits represented in the Career Batting data set – this is Pete Rose, an outfielder who helped lead several Cincinnati Reds teams to successful seasons in the 1970s. He holds MLB records in many other batting statistics, including at-bats and games played. In 1989, Rose was discovered to have been gambling on the outcome of games he played in and managed, earning as a result a lifetime ban from professional baseball activity. He has never been elected to the Hall of Fame and likely will never be.

4 Classifying HOF Induction Status

We turn now to the second problem posed in the introduction, that of using machine learning algorithms to predict HOF induction status based on a given vector of career batting totals.

We generate train and test sets using the Career Batting dataset and evaluate four different approaches: quadratic discriminant analysis, logistic regression, and regression trees with and without cross-validation. Accuracy results for the four approaches are presented in Fig. 6.

Overall, machine learning approaches are highly effective in the HOF induction problem, earning about 80% accuracy across the methods. For illustration's sake, we depict in Figs. 7a and 7b the branching points of the regression tree analyses. The utility of the cross-validated regression tree is in the simple rules given for decision making – any batter with more than 2,210 hits and 1,366 runs has a good case for inclusion in the Baseball Hall of Fame.

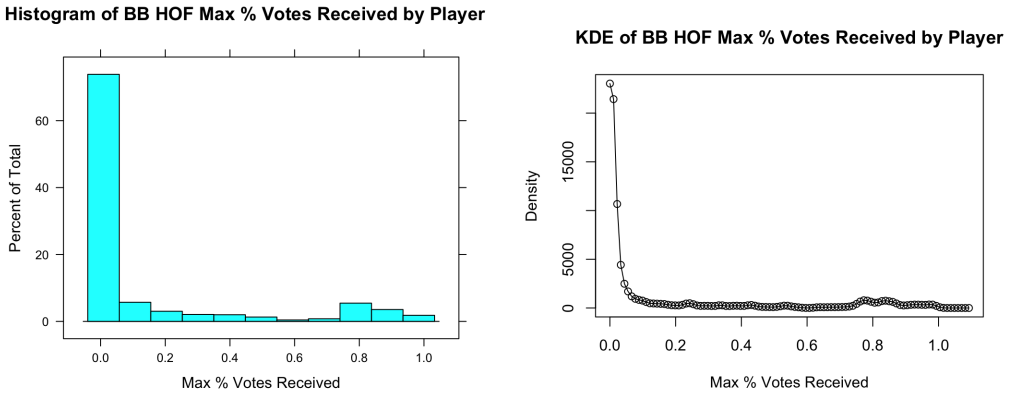
5 Challenges and Limitations

In the above report, we explore the application of statistical computing techniques to the question of analyzing data relating to induction into the Baseball Hall of Fame. While linear and kernel regression techniques come up short, we find that machine learning methods are highly applicable to the classification problem and can achieve 80% accuracy without much fine-tuning.

There were of course limitations to our approach. By only using Career Batting data, we ignore the fact that both batters and pitchers are included in the Hall of Fame. Any future work should seek to fold pitching data, which is comprised of variables distinct from those important in batting data, into the models used to predict or classify.

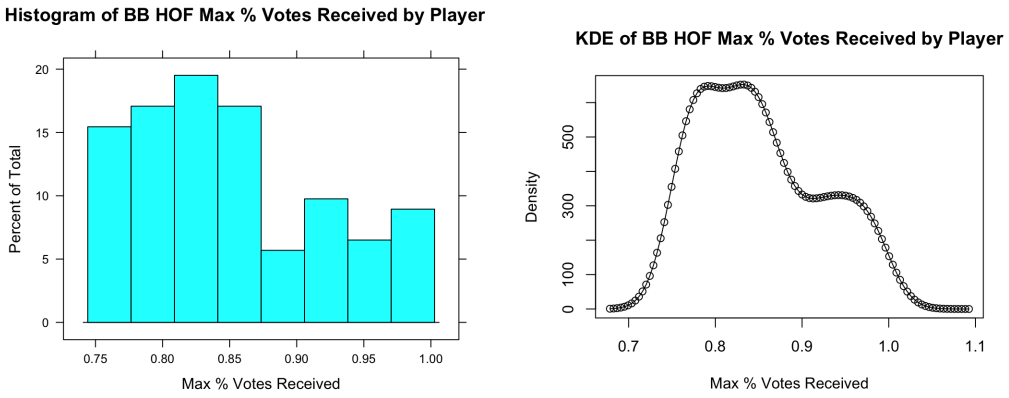
Another limitation is historical. Some of the greatest baseball players from a statistical standpoint were caught up in the steroid scandal of the late 1990s and early 2000s. Figures such as Barry Bonds, Mark McGwire, and Sammy Sosa rewrote batting statistics records playing under the influence of performance enhancing drugs and, as such, have effectively been shadowbanned from consideration for inclusion in the HOF. This is difficult for techniques which only have insight into the batting totals to identify *a priori*.

6 Figures



(a) (b)

Figure 1: Histogram and KDE for All Players



(a) (b)

Figure 2: Histogram and KDE for Successful Inductees

```
> pct_votes_model <- lm(max_pct~.-playerID,data=batting_pct_votes_data)
> pct_votes_model

Call:
lm(formula = max_pct ~ . - playerID, data = batting_pct_votes_data)

Coefficients:
 (Intercept)             G            AB             R             H  career_years
  -0.3079062    -0.0001888    -0.0001057     0.0003092     0.0003650     0.0348828

> anova(pct_votes_model)
Analysis of Variance Table

Response: max_pct
               Df Sum Sq Mean Sq  F value    Pr(>F)
G               1  8.020  8.0199 150.1871 < 2.2e-16 ***
AB              1  0.352  0.3522   6.5951   0.01035 *
R               1  5.154  5.1535  96.5093 < 2.2e-16 ***
H               1  1.554  1.5542  29.1056 8.319e-08 ***
career_years    1 10.591 10.5906 198.3275 < 2.2e-16 ***
Residuals    1148 61.302  0.0534
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Linear model for *max_pct*

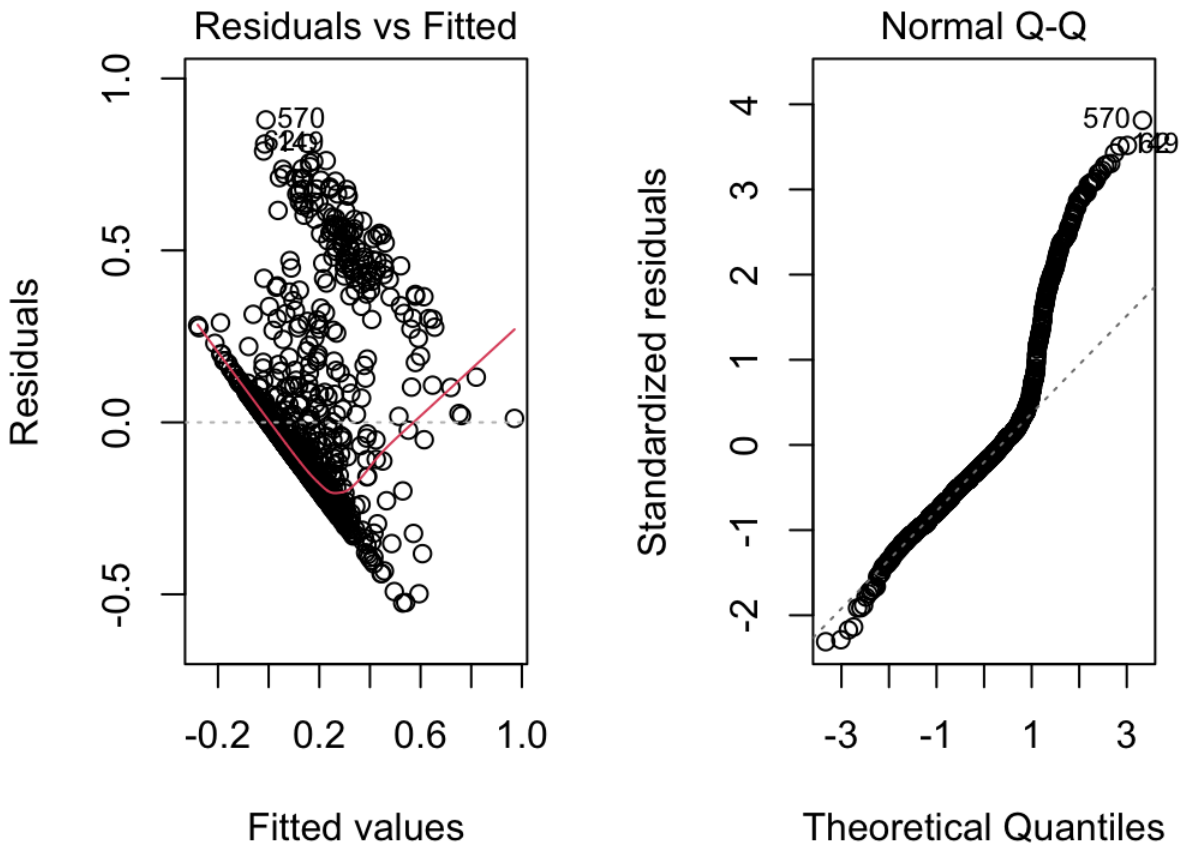
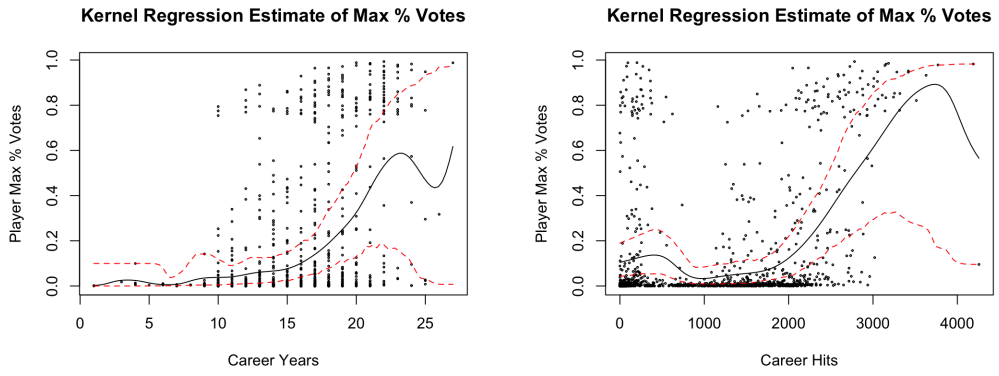


Figure 4: Linear model for *max_pct*

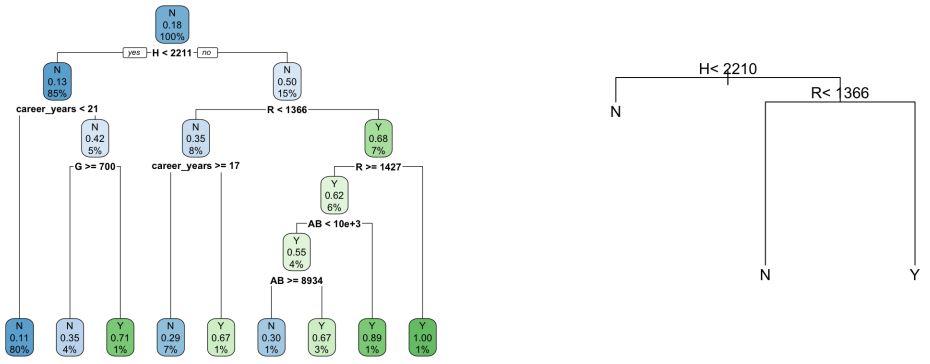


(a) (b)

Figure 5: Kernel Regression of *max_pct* Against Career Years and Hits

Method	Accuracy
Quadratic Discriminant	0.814
Logistic Regression	0.822
Non-CV Regression Tree	0.789
CV Regression Tree	0.815

Figure 6: Accuracy from Confusion Matrices for Different ML Approaches



(a) (b)

Figure 7: Regression Tree Branching Points