Predictive Modeling for the National Baseball Hall of Fame

Jackson Bandow

Advisor: Ethan Meyers

I. Abstract

Every year elections are held to vote for the new members of baseball's National Hall of Fame, but there is often controversy in the players that are selected. Is there a set of criteria in which players are typically inducted into the Hall? And if so, what factors of a player's career matter the most? The following paper builds a predictive logistic model using five-fold cross-validation on player data from 1924 to 2004. By no surprise, the most important factor is the player's career abilities, which is often expressed in terms of Wins-Above-Replacement (WAR). Other important factors include single-season achievements, team performance bias, career longevity, and off-field penalties, like gambling or steroids. The resulting model can predict induction accurately with an 8.8% and 7.5% misclassification rate for batters and pitchers on only the most qualified players and a 0.959 and 0.958 AUC on the ROC curve for the same groups. After observing misclassified outliers, three conclusions can be drawn. First, there is a bias in Veterans Committee selections. Second, sometimes players are inducted that probably would not be today. Third, the statistical reasoning for induction is constantly evolving.

II. Introduction

In 2017, 2.4 million kids played Little League baseball (Zminda, 2010). The figure is consistent with yearly expectations. They dream of playing professionally one day. All of them have a hometown team and a favorite player that they follow and look up to. Unfortunately, the reality is that less than 1% make it to the Major League Baseball stage (*What Are Your Odds of Making the Pros?,* n.d.). The journey is beyond difficult. Starting from Little League, a player must make it through middle school travel ball leagues, find a spot on the high school varsity team, get recruited to play in college, get drafted into professional baseball, and earn promotions through the 5+ levels of the minor league system before he plays his first inning in "The Show." But the tiers do not stop there. At the MLB level, there is an even more elite group of players: The National Baseball Hall of Fame. Perhaps just as selective, only around 1% of MLB players are inducted into the organization (James, 2019). On average, just under 4 players are inducted annually, yet millions of kids start their journey every year.

The National Baseball Hall of Fame sets forth a complete set of regulations on how an individual is inducted. An association of active baseball writers with at least 10 years of experience compose the electorate (*BBWAA Election Rules,* 2014). A committee among them determines the group of players to appear on the ballot, all of whom have been retired from their 10+ year careers for at least 5 years, but no more than 15. According to the National Baseball Hall of Fame website, "voting shall be based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played." Each elector can vote up to 10 times. After the ballots have been counted, any player that received at least 75% of the vote is awarded his place in the most prestigious group in all of Major League Baseball (Lindbergh, 2020). When a player appears on less than 5% of ballots or his 10 years on

the ballot are up, he is no longer eligible to be voted into the Hall of Fame by the BBWAA. The only remaining path is via the Veterans Committee, who vote on past players that they believe the BBWAA incorrectly passed on.

Many sabermetrics, or baseball statistics, experts use career WAR, or wins above replacement, as a predictor for Hall of Fame membership probability (James, 2019). WAR is a measure of a player's individual contributions to his team compared to an average free agent or top minor league player. In other words, WAR measures the additional games that a team won because of the given player. For example, Clayton Kershaw had a WAR of 8.1 in 2013 (*Clayton Kershaw,* 2020). If he had not played, the Dodgers would have won approximately 8.1 games less. Other baseball statisticians have derived similar measures. Bill James published a statistic in his book, "Win-Shares," that describes, in thirds, the number of the wins attributed to a given player. Clearly these measures are very similar, and James uses a combination of the two called "Hall of Fame Value" to calculate his list of likely candidates. However, one's contribution to his team is only one factor for induction into the Hall of Fame, according to the organization's own website. Moreover, baseball is a sport that accumulates individual performances, where a player on a losing team can have a great performance, but not overcome the total performance of the opposition. For example, Felix Hernandez is well regarded as a top pitcher in his generation. He played for the Seattle Mariners for 14 years, but they did not make the playoffs once during his time with them. So, does WAR and Win Shares really capture an individual's performance unconditional on the team's performance? And what are the driving statistics behind WAR and Hall of Fame induction?

The following analysis will build a five-fold cross-validated predictive logistic model to both determine the driving factors in voters' decisions and predict the modern players' induction

probabilities. I will start with a basic model using only WAR, representing a player's abilities to predict induction, then add in features that decrease the misclassification rate and increase the AUC. The rest of the paper details the process and lays out the findings of the model.

The remaining analysis is organized as follows. Section III discusses prior research in induction prediction. Section IV explains relevant statistics and how they are calculated. Section V cites the data sources that are used in the analysis. Section VI explains the logistic modeling process. Section VII lays out the results of the modeling in its entirety. Section VIII concludes the findings. Section IX cites the references used in the paper.

III. Prior Research

In 1977, Bill James published his first book, the *Baseball Abstract*. Before his time, baseball analysts reacted off "gut feel," but James attempted to quantify that feeling. He created obscure new baseball statistics, such as player performance by month or stolen bases allowed by a starting pitcher. Just five years later, he had a national publisher for his annual *Baseball Abstract* and a small following, but still lacked recognition from the professional teams. Old-school coaches tended to strongly dislike the shift to a statistics lens in baseball. In 1984 James started Project Scoresheet, a network of volunteers that collected every major-league game's play-by-play data. The organization inspired many other early statistic collectors in the baseball world. In 2002, as baseball organizations began to incorporate sabermetrics, James was hired by the Boston Red Sox. They won the World Series two years later (Zminda, 2010).

In the 1980 issue of his *Baseball Abstract,* James first introduced what he called the Value Approximation Method. At the time, it was the only attempt to quantify the value of a player, using a quasi-replacement level benchmark. The replacement level is the class of performance that could be replaced by the average free-agent or top minor league player. Four

years later, Pete Palmer's *The Hidden Game of Baseball* presented a similar metric that expressed value in terms of wins, but used the average player as a benchmark. Over the years, James and other early baseball statisticians cemented the replacement level as the benchmark but kept Palmer's use of wins as a unit (Lindbergh, 2020). Among the next evolution of player valuation metrics was Bill James's Win Shares. He published a book, *Win Shares*, in 2002 with Jim Henzler that detailed the statistic. Essentially, it distributed the wins on a team during a given season to the players based on their performance. This particular metric did not use any benchmark (James, 2019). Then in 2008, FanGraphs presented the most-used player valuation statistic today: Wins-Above-Replacement or WAR. Baseball-Reference also published their own WAR statistic in 2010 (Lindbergh, 2020). The main difference between WAR and Win Shares is that Win Shares attributes wins to players for their performance, while WAR attributes wins above the replacement level. A good analogy is raw return vs. excess return of a portfolio or asset in finance, where Win Share is the raw return and WAR is the excess return over some benchmark.

Recently, baseball analysts have used these valuation statistics to compare the best players of all time. In his 2019 annual publication, now called *The Bill James Handbook*, James applies Win Shares to Hall of Fame status. He claims that players with more than 325 Win Shares are usually inducted into the Hall of Fame, while those with less than 275 are usually not. WAR is also a good predictor of Hall of Fame status. James does not argue that one is better than the other, rather that they are different. Win Shares typically favors players that have a short but great career, while WAR typically favors a longer, steady career. To combat the bias in their distribution, James created Hall of Fame Value as a combination statistic to unbiasedly predict

Hall of Fame induction (James, 2019). I will discuss these metrics in greater detail in the following section.

IV. Relevant Statistic Calculations

Before considering new predictors of Hall of Fame status, it is important to be familiar with the generally accepted value measures for baseball players. The most common metric is WAR. WAR is the number of wins above replacement, where a replacement is an average free agent of minor league player in that year. Essentially, the term describes the number of excess wins that a player produced for his team that the team would not have won if he had not played. Of course, the actual calculation of WAR is much more complicated. At a high level, the formula for WAR is given by

$$WAR = \frac{BR + BRR + FR + PA + LA + RR}{Runs/Win}$$

where BR is batting runs, BRR is baserunning runs, FR is fielding runs, PA is a positional adjustment, LA is a league adjustment, RR is the replacement runs, and $Runs/Win$ is the number of runs per win (chalsey15, n.d.). However, each component is a very technical computation. For example, batting runs is based off of $wRAA$, or weighted runs above average given by

$$wRAA = \frac{wOBA - lgwOBA}{wOBAScale}\ PA$$

where $wOBA$ is a player's weighted on base average, $lgwOBA$ is the league average $wOBA$, $wOBAScale$ is an yearly adjustment based on $wOBA$, and PA is plate appearances. Finally, to get batting runs,

$$BR = wRAA - (PF)lgR - wRC$$

where PF is a park-factor adjustment, lgR is regular season league runs, and wRC is the weighted runs created. The technical calculation of WAR is beyond the scope of this paper, not to mention that the complete calculation is done behind closed doors and is not publicly shared. The prior example shows the complexity of finding just one part of the equation. The important takeaway is that WAR is a very complex evaluation metric for a baseball player that is composed of the runs that a player generates over a benchmark in each component of the game, adjusted for factors, such as the stadium and the league, and then divided by the average number of runs per win. To add more complexity, the main contributors to WAR are Baseball Reference and FanGraphs, who calculate it with slightly different methods. For example, Baseball Reference regresses park factors on the prior three years of competition, while FanGraphs uses five. FanGraphs uses pitchers in batting league averages, but Baseball Reference does not (*WAR Comparison Chart*, n.d.). There are several small differences between the methods, but they create a significant change in the metrics.

The philosophy behind Bill James's Win Shares is similar to WAR. The number of Win Shares that a team has in a given season was triple their wins, so in 2018, when the Boston Red Sox won 108 games, they would have 324 Win Shares to divide among their players (James, 2019). From there, the total number of Win Shares is divided and distributed to members of the team based on their batting, pitching, and fielding performances. James claims that although it seems like there would be a bias towards players on winning teams, winning teams win because they have more higher valued players. Thus, Win Shares is more or less unconditional on a team's performance.

Both Win Shares and WAR are solid predictors for one's induction into the Hall of Fame. James claims that Win Shares favors the longevity of a career, while WAR favors the shorter,

more explosive career, but he is not sure which is a better measure (James, 2019). By combining the two metrics into one, called Hall of Fame Value, he reduced the individual biases. Since the variance of WAR is almost exactly 25% of the variance of Win Shares, James derived Hall of Fame Value by

$$HoFV = WinShares + (4)WAR$$

The statistic may measure the Hall of Fame likelihood with accuracy, but when removed so far from the actions of players on the field, it is easy to lose sight of the actual impact of a play on a player's value. In other words, how valuable are certain accomplishments relative to others? And what are the benchmarks that sends a Hall of Fame candidate over the top?

V. Data Sources

My primary source of data is Sean Lahman's dataset, found on his website and automatically installed in R under the Lahman library (Lahman, n.d.). The files include extensive yearly data of all MLB players from 1871 to 2019. I took the relevant data from the season-by-season statistics, condensed them to career statistics and compiled them to a master data frame for the following data sets: pitching performance, batting performance, fielding performance, Hall of Fame voting results, player appearances, team records, player awards, all-star appearances, and postseason results. Additionally, I scraped WAR data off the Baseball-Reference website. WAR will be an intermediary for predicting player value and I will compare the differences between it and a linear combination of standard statistics, such as hits, runs, and runs-batted-in (RBI).

The Hall of Fame Induction process has evolved constantly since the first election in 1936 (*Hall of Fame Ballot History*, n.d.). One important note is that Major League Baseball

debuted 65 years earlier in 1871. That means that the early elections were not only responsible for inducting recently retired players like they are today, but also for looking back through over 60 years of players and picking out players that should have been inducted had there been a hall of fame in the earlier era. As a result, these elections will naturally select different players than an ordinary year. They may even pass on a player that would normally be inducted if the selection pool was not so competitive. For this reason, I will exclude the earliest era of players from the data. On the other end, there is an issue with keeping the most recent era of players. Those who had not yet retired prior to the 2004 season may still be eligible for induction through the BBWAA in the most recent year of the data. I will use my model to predict the results of the recent era of players, but will not use them in the training set, as the outcome of their careers (inducted or not inducted) is not yet certain for those not currently inducted. Before starting analysis, I will make one more limitation on the data. The data will exclude players that played in the Negro Leagues for two main reasons. The first is that the Negro Leagues did not track statistics as well as Major League Baseball, so there is little to no non-anecdotal record of performance. Players that played both in the Negro Leagues and the Major League will be undervalued by statistics because only part of their careers will be officially recorded. The second reason to exclude these players is that the purpose of this model is to predict future induction. The Negro Leagues are an important part of baseball history, but their records will not apply to future generations of players. The results section that follows will begin with the universe of players that retired after the 1923 season but before the 2004 season and did not play in the Negro Leagues.

VI. Logistic Modeling

To predict a binary result from a mix of continuous and categorical variables, I will use a logistic regression. A linear regression could also apply to the BBWAA results in terms of percentage of vote, which require at least a 75% vote for induction, but the veterans committee does not always release specific ballot results, beyond the names of any inductees. It follows that a logistic regression is more appropriate. Specifically, my analysis will begin by regressing induction status of a player by the player's statistical value. Then, I will add in features to make the most accurate model possible. To protect against overfitting, I will use a five-fold cross-validation, where I begin by splitting the data into randomly selected 20% folds. Then, I train the model on 80% of the data and test on the remaining 20%. I repeat this process five times, testing on each fold once. To evaluate the performance of my model, I will use both the misclassification rate and the Area-Under-the-Curve (AUC) metric for the Receiver Operating Characteristic Curve (ROC), which graphs the true positive rate against the false positive rate. The misclassification rate is simply the percent of observations incorrectly classified, found by averaging this rate for each fold. The AUC is found by taking the area under the ROC curve and again averaging over each fold. While the goal is to minimize the misclassification rate, the best models maximize the AUC because they maximize the true positive rate relative to the false positive rate.

One assumption of logistic regression is for there to be no multicollinearity among independent variables. This assumption will not be met for the following analysis because most baseball statistics are highly correlated. For example, a player that has more at-bats, typically has more hits, home runs, runs, RBIs, etc. Under normal logistic regression assumptions, only one of these variables would be able to be used as a predictor in the model. By using a linear

combination of correlated variables in a logistic regression, the predictive ability of the model is still effective, but the model loses some interpretive ability. For example, under normal assumptions, the coefficients of the independent variables in the regression give the direct effect of the independent variable on the dependent variable. However, with multicollinearity, the effects of an individual independent variable are diluted, so the coefficients cannot be interpreted in the same manner. I will have to take further steps to provide interpretability in the model.

VII. Results

Section A. Selecting Eligible Players

The current universe of players includes any player that played his last game in the 80 years between the 1924 and 2003 seasons, inclusive, and did not play in the Negro Leagues. For simplicity, I will first look at non-pitchers (hereafter referred to as batters), where position is defined by the highest percentage of defensive appearances in relation to the other eight defensive positions. I removed all players that are currently on the list of banned players. These players have been banned by the MLB Commissioner and are not allowed to be inducted in the Hall of Fame. Most notable on the list is Pete Rose, the all-time hits leader, who was banned from Major League Baseball when he was caught gambling on games, while serving as the manager of the Reds. In the current data set, one important note is that the number of Hall of Fame members is much smaller than non-members. Specifically, Hall of Fame batters compose approximately 2.2% of the current data, which is problematic. For example, a model for Hall of Fame induction that predicts false every time would be correct 97.8% of the time. Further steps should be taken to limit the dataset to the most relevant observations.

I took two steps to limit the data to appropriate observations. The first issue that I identified was that there were many players in the data with very few at-bats. For a model predicting Hall of Fame induction, these players are not relevant because they will clearly be ruled out. The second issue is similar in that even if a player has a significant number of at-bats, he may be an obviously bad player and have no chance to be inducted to the Hall of Fame. Including him in the data does not add any value to the model. To address the issues that I identified, I first removed all players with career WAR under 10, so that all players left would at least have some value. The minimum WAR of Hall of Fame batters during this period is 25.3, so this limitation includes all Hall of Fame members in the previous data with a cushion for any future Hall of Fame candidates on the lower end of player value. Next, I took the top 60 batters in terms of at-bats from each retirement decade. Retirement decade is defined as the ten-year period during which a player played his final game in Major League Baseball with a four-year offset, such that the first decade is 1924 to 1933. The lowest ranking Hall of Fame member in at-bats in a decade had 51st most at-bats during his retirement decade, which occurred during the decade with the most inductions (22). Taking the top 60 players during a given decade again leaves room for a possible lower-end Hall of Fame candidate in the future. In the adjusted data, all Hall of Fame members from the previous data are included and make up 24.0% of observations. By limiting data by retirement decade, I ensured that the number of observations was consistent over time.

Section B. Initial Modeling

Undoubtedly, the factor contributing most to a given player's induction should be his value as a player. This section will explore the best logistic models that predict Hall of Fame induction by a linear combination of statistics that show a player's value and abilities. First, the

most universal statistic for evaluating a player's value is WAR, as discussed extensively thus far. The first model will regress induction by WAR. Then, the second uses a linear combination of the following standard statistics: games played, at-bats, walks, hits, home runs, runs, RBIs, stolen bases, fielding percentage, on-base percentage, slugging percentage, and batting average. Finally, the third model will combine the features of the prior two. Table 1 displays the misclassification rate and AUC of the first three models. The combined model has the best AUC and misclassification rate, but still has an AUC of only 0.926. Let's continue to search for a better identifier of player value.

First Player Value Logistic Modeling Results

Regressors	Misclass Rate	AUC
WAR	14.0%	0.915
Standard Statistics	12.3%	0.899
Combined	10.4%	0.926

Table 1. Logistic regression results for the first attempt at modeling induction status by player value.

It is important to note that player value can change over time. For example, Hall of Fame candidates in the 1920s may not have been as good as modern-day candidates but were inducted because of their relative performance. Players can only be judged relative to the standards of their time. To address this point, I will standardize the previously used statistics by retirement decade so that each decade has roughly normal distribution for each statistic with mean zero and standard deviation one. All features are roughly normal when log transformed, as confirmed by normal quantile plots, so the transformations are described by

$$Transformation = \frac{log(Statistic) - \mu\big(log(Statistic)\big)}{\sigma\big(log(Statistic)\big)}$$

where μ is the mean function and σ is the standard deviation function. Then, the previous models were calculated using the newly transformed data, and the results are reported in Table 2. Interestingly, WAR is now a much better predictor. The misclassification rate of the model using

only WAR is the same as the combined model, but the WAR model has a slightly higher AUC. The differences are so small that there is likely no significant difference between the models, although the model using only WAR can produce almost identical results with only one feature.

Second Player Value Logistic Modeling Results
Standardized by Retirement Decade

Regressors	Misclass Rate	AUC
WAR	11.5%	0.935
Standard Statistics	13.1%	0.892
Combined	11.5%	0.933

Table 2. Logistic regression results for the second attempt at modeling induction status by player value.

The final models using only player value as a predictor will explore how the importance of player value itself has changed over time. Since the standardized models improved the performance from the initial models, the final models will add retirement decade interactions to all the predictors in the model to pick up on trends over time. For example, home runs rare in early baseball, but are now a big part of the game. By adding time interactions, the importance of home runs in the model could increase. Table 3 shows the results of the standardized models with retirement decade interactions. In the third round of initial models, the model regressing on only WAR outperformed all models with a misclassification rate of 11.0% and an AUC of 0.943. Although it didn't have the lowest misclassification rate of all the models, its AUC was the highest. Its performance statistics and simplicity make it an appealing initial model. I will proceed to the next section with the logistic model regressed on WAR using standardized data and retirement decade interactions.

Third Player Value Logistic Modeling Results
Standardized with Retirement Decade Interactions

Regressors	Misclass Rate	AUC
WAR	11.0%	0.941
Standard Statistics	13.5%	0.895
Combined	12.3%	0.937

Table 3. Logistic regression results for the third attempt at modeling induction status by player value.

Section C. Improving the Model

A closer look at the given model reveals that it performs relatively well but has room for improvement. Figure 1 shows a scatterplot of WAR against induction status and colored by model accuracy. Using data with 24% positive results, the baseline model, which would predict "false" every time, is 76% accurate or has a 24% misclassification rate. A starting point of 11.0% misclassification is good but would ideally reach well into single digits after modifications. There are a handful of large outliers that could be addressed. Specifically, there appear to be a few players with low career WARs relative to other Hall of Fame members that were inducted. In general, there are a few possible reasons for misclassification. The first is that the model is missing a feature that could be patched so that the outlier is not misclassified in the next model.

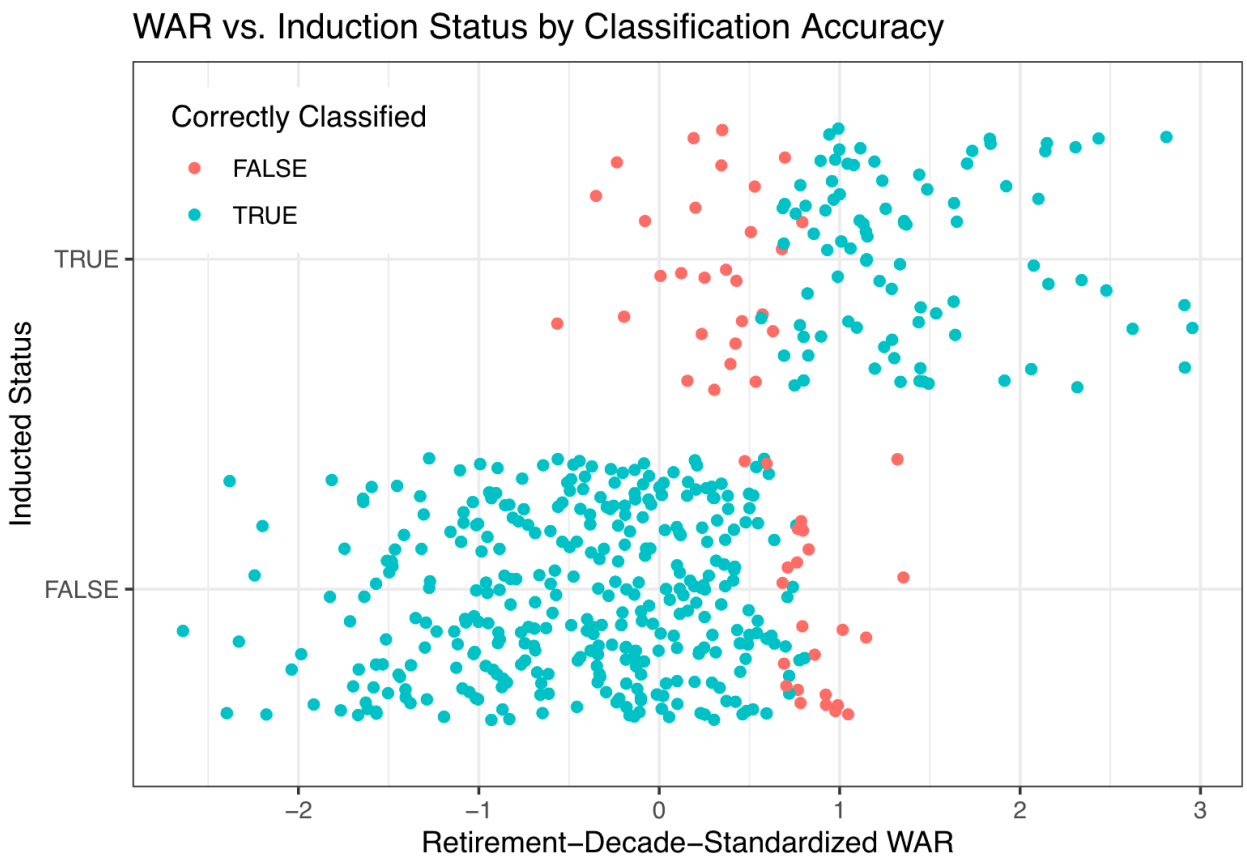

Figure 1. Initial modeling results shown on a scatterplot of induction stats by retirement-standardized WAR. There is a clear WAR threshold where the model switches from classifying "not inducted" to "inducted".

Another reason is that the outlier has certain qualities that are difficult to quantify. These qualities are more likely among older players, as there was not much access to statistics in the early era of baseball, so the voters had to rely on more qualitative abilities. Finally, the last plausible reason that I will discuss is that the selection committee could have simply made a mistake. It is very possible that there are players in the Hall of Fame that would not be there if the election was redone today. The remainder of the section will investigate specific outliers and determine possible factors to improve the model.

To determine other important factors in a player's chances of induction into the Hall of Fame, it is important to remember the official criteria published by the Hall of Fame itself. According to its website, "voting shall be based upon the player's record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which the player played." After analyzing the errors in the initial model and referencing the voting guidelines for the Hall of Fame, I found features in the following categories improved the model: single season achievements, longevity of career, team performance, and extracurricular activities. First, single season achievements include recognitions for high performance in a given season in the form of the MVP award, Triple Crown, Silver Slugger, and All-Star appearances. Some of these awards are voted on by the BBWAA, so there is no surprise that they translated to Hall of Fame induction. Second, longevity of career is simply measured by the date of his final game minus the date of his first game in years. When an athlete plays longer, he typically has a higher chance of induction. Third, a player's contributions to his team are a direct factor of induction according to the Baseball Hall of Fame, which can directly be measured by the team's win percentage. I measured a given player's win-percentage by using the team's single season win-percentage in a weighted average by the player's total games played. The voters seemed to favor players on

winning teams above their value as a player. Finally, the last category of features I added to the model was extracurricular activities. More specifically, I attempted to add steroid use to the model by indicating all players that were either implicated by the Mitchell Report or suspended at some point for steroid use. There was not a huge impact on the training set, because the steroid era in baseball is from 1993 to 2002 and my data is limited by players that retired before 2004. However, there is a very negative view of steroids by the voters, so it was important to include steroid use in the model for its future predictive purposes.

Improved Logistic Modeling Results

Model	Misclass Rate	AUC
Initial	11.0%	0.941
Modified	8.8%	0.959

Table 4. Performance statistics of the original and improved logistic models on the batters.

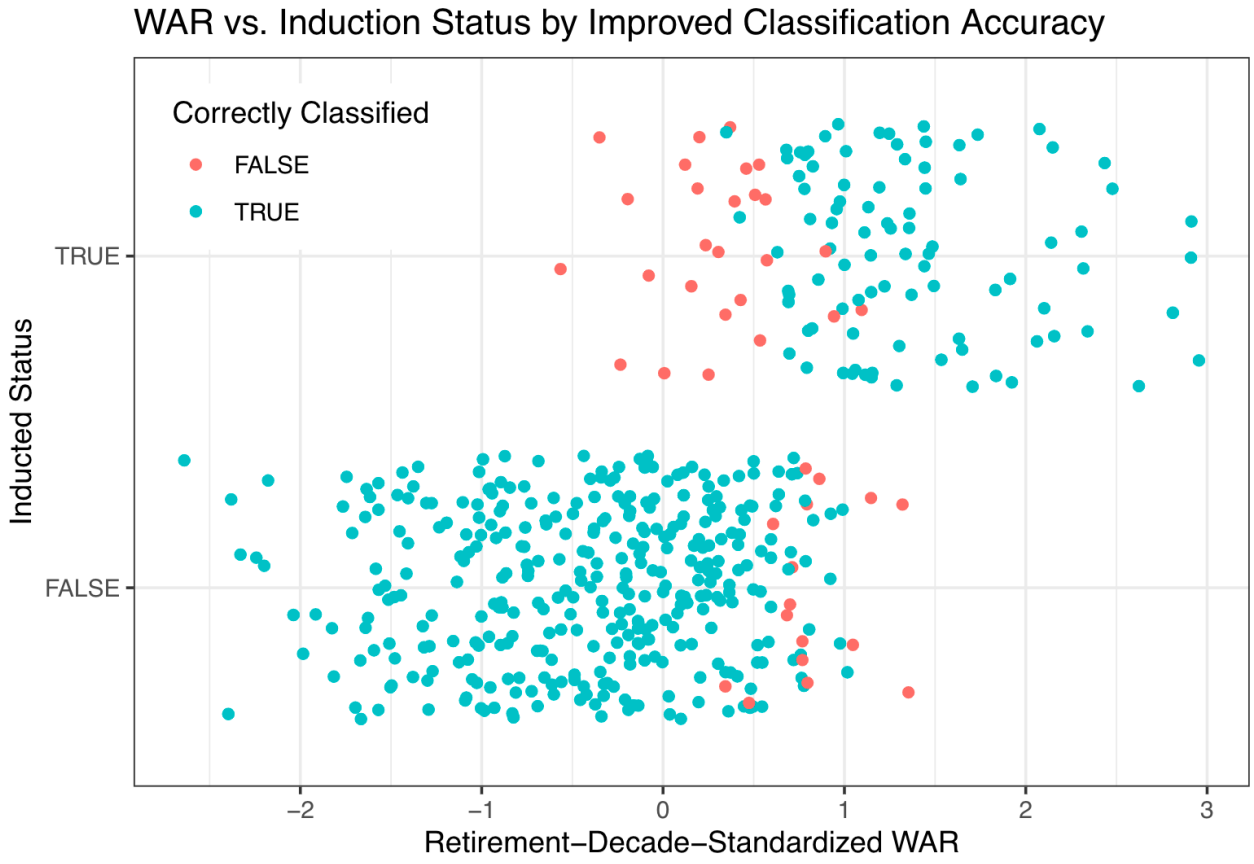

Figure 2. Improved classification results from the new logistic model shown as a scatterplot of induction status by retirement-standardized WAR.

The addition of the previously described features added significant value to the model. Table 4 displays the new performance statistics. After modifications, the model's AUC increased from 0.941 to 0.959, and its misclassification rate decreased from 11.0% to 8.8%. Additionally, Figure 2 shows WAR graphed against induction status and colored by predictive accuracy. Compared to Figure 1, the new model better captures observations that are below the typical WAR threshold for Hall of Fame members or that are above the threshold for non-members.

To determine the importance of features in the model, I used a dominance analysis. In a dominance analysis, every feature is added one at a time to every possible subset of features to calculate its additional predictive ability, measured in pseudo-R-squared. A normal linear regression uses R-squared to measure the amount of variability explained by the model, but the same metric does not apply to logistic regression, so pseudo-R-squared is used. The following analysis will use the Nagelkerke index, which corrects the Cox and Snell index to fill the range zero to one, like the traditional R-squared would. Figure 3 shows the average additional pseudo-R-squared across each level of subset for each feature, where level one corresponds to a model with one initial predictor. As expected, WAR is the most dominant feature, followed closely by the interaction between WAR and retirement decade. A player's abilities should be the leading contributor to his status as a Hall of Fame member. WAR is followed by the various awards given to players for a single season, such as All-Star appearances. This secondary feature is not surprising either because the writers of the BBWAA are involved in some of the voting for these awards, such as the Most Valuable Player (MVP) award, so these awards should cover some of the voting bias. Although not as large of a factor, a player's winning percentage adds as much value to the model as a mid-level single-season award would. Its significance in the model is a bit surprising because a team's performance helps the player's Hall of Fame candidacy. I would

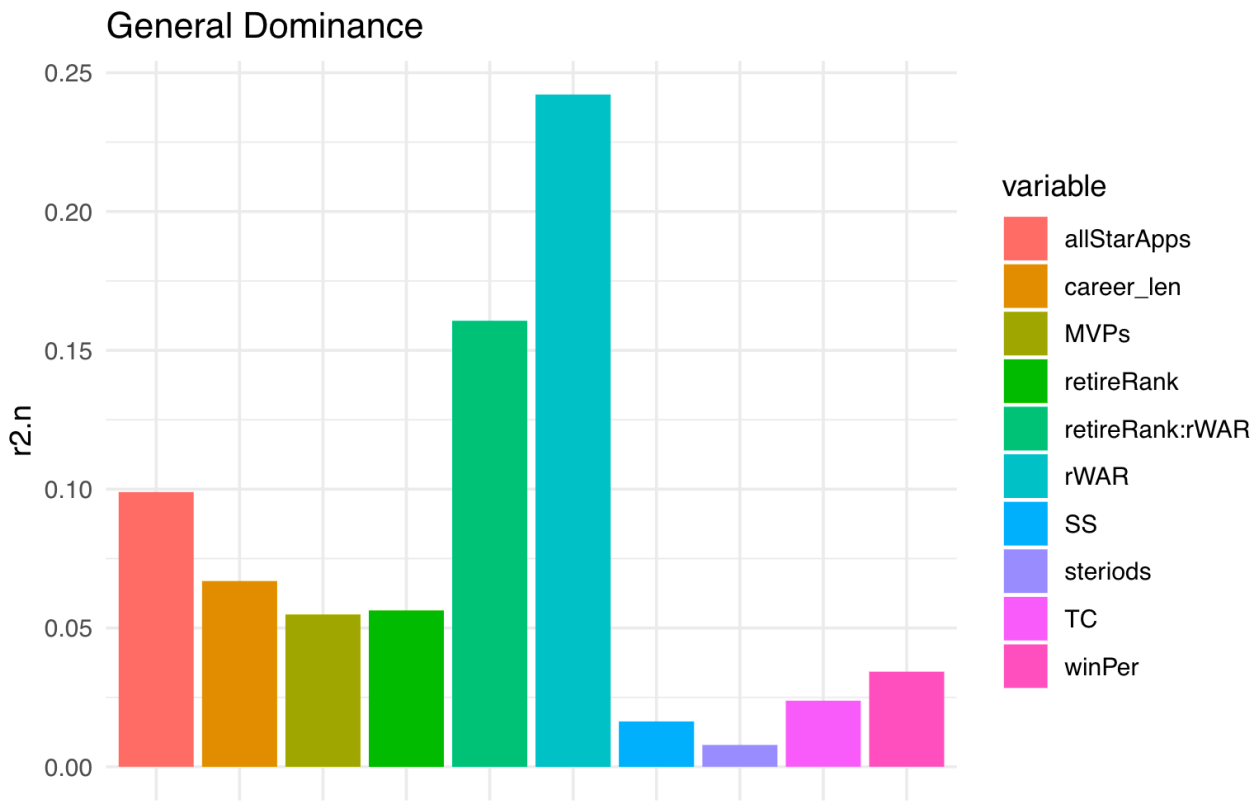

Figure 3. General dominance analysis results. Bar graph shows the average additional pseudo-R-squared added to the model by each feature.

argue that it is more difficult to perform at an elite level on a team that is struggling than a team that is winning. However, the model shows that the BBWAA and the Veterans Committee have a bias in the opposite direction. Finally, the steroids feature adds the least amount of value to the model, which is expected, given the starting date of the steroid era. The steroid feature was included in the model because the current BBWAA shows extreme bias against steroid use. For example, Barry Bonds has the second highest WAR of all time, behind only Babe Ruth but has been rejected from the Hall of Fame nine times and is expected to note be voted in again on his final year on the ballot because of his steroid use allegations. While the steroid feature may not be very relevant in the earlier years of baseball, I would expect most of its relevance to come in the recent years. In summary, the factors that determine a Hall of Fame candidate's likelihood of induction are, from highest to lowest importance: his career accomplishments, his single-season

accomplishments, his team accomplishments, and extra-curricular activities that may hurt his candidacy.

Section D. Application of the Model to Pitchers

The previously improved model can be applied to best predict pitchers' induction probabilities. The same categories from the batters' model are essential in evaluating the careers of the pitchers: player's ability, single season achievements, longevity of career, team performance, and extracurricular activities. Each category was modified slightly due to both positional differences between pitchers and batters and biases from the Hall of Fame voters. To limit the data set to the most relevant pitchers, I took the top 30 players for each retirement decade with the highest sum of wins and saves. The combined metric was the best method to unbiasedly select both starting pitcher and relieving pitchers. I found that if I used innings pitched, the algorithm would select too many starters, but if I used games played, it would select too many relievers.

When modifying the features that account for a player's abilities, I first changed the WAR predictor in the previous model to WAR from pitching. Essentially, the modification removes all batting aspects of WAR from pitchers. A batter's total WAR closely resembles his WAR from batting because batters rarely pitch. If they do, then they add value to their team that should be accounted for in Hall of Fame considerations. However, pitchers in the modern American League typically do not bat because they are replaced in the lineup by the designated hitter (DH). Then, a pitcher's total WAR would partially be a function of the league that he plays in and would not account as well for his abilities as a pitcher as his WAR from pitching would. Additionally, I removed the interaction between WAR and a player's retirement decade because

the additional feature reduced the accuracy in the cross-validated model. It appears that the contribution of a player's ability to his induction likelihood has not changed much over time.

In terms of single season achievements, no significant changes were made to the model. In the previous model, single season achievements were defined by All-Star appearances, MVP awards, Silver Slugger awards, and Triple Crowns. While pitchers are eligible for All-Star appearances and MVPs, the Silver Slugger award and Triple Crown title are both specific to batters. I replaced these features with the Pitching Triple Crown and the Cy Young award. The Triple Crown title is given to a player that leads a league in batting average, home runs, and RBIs, while the Pitching Triple Crown title is given to a pitcher that leads a league in wins, strikeouts, and ERA. The Silver Slugger is awarded to the best offensive player at every position. As mentioned, a pitcher's hitting ability is not very highly valued when evaluating the career of a pitcher, so the Silver Slugger award was not included in the pitcher's model. Instead, I included the Cy Young award, which is given to the best player in each league every season. After these simple modifications, the single season achievement features were easily applied to the pitchers' data.

The longevity of career category was the only category that did not need modification. The only feature in this category is the duration in years from the start to the end of a player's career. While the feature itself did not change, its coefficient moved from significantly positive in the batters' model to significantly negative in the pitchers' model. This change is evidence that a batter's chances of induction increase the more he plays, but a pitcher's career is more evaluated by its efficiency. The pitcher still needs to earn the approximate Hall of Fame benchmark statistics, but his chances of induction decrease for every additional year that it takes him.

Team performance was arguably the most notable change in the model. The batters' model used the weighted average of team performance by games played. However, the same metric did not help the pitchers' model prediction accuracy. In fact, when I removed the statistic from the model, the misclassification rate dropped around 4%. Instead, I replaced team win percentage with wins and saves, which do not appear as team performance statistics at first glance. Wins and saves are both statistics that award the pitcher's performance relative to his team's. A win is awarded to a pitcher if he completes most of the game, leaves the game while the team is winning, and the team wins the game without giving up the lead after the pitcher leaves. For example, if a pitcher throws six innings and leaves the game with a 2-1 lead before the team wins 2-1, then the pitcher is awarded the win. However, if the pitcher leaves the game, having let up the same one run, but his team does not score, losing 1-0, then the pitcher receives a loss for the exact same performance. The save is scored in a similar manner. The team's performance only seems to matter when evaluating a pitcher's career during games that he plays in. While wins and saves appears to isolate the pitcher's performance more than the team win percentage feature in the batters' model, pitchers also play less often than batters do. In other words, both features account for the team's performance in games played by a player. Team win percentage is not a significant feature in the pitchers' model because there are many games that the pitchers do not contribute to, unlike the batters.

Finally, the last category of consideration is extracurricular activities, specifically steroid usage. Unfortunately, there were no pitcher steroid users in the training set, so that feature could not be included in the model. However, there is no reason to believe that steroid use would be viewed differently between batter and pitchers. For that reason, I will include steroid use in the pitchers' model moving forward, using the coefficient derived in the batters' model.

Additionally, from a strictly qualitative perspective, the BBWAA strongly opposes the induction of steroid users to the Hall of Fame. To optimize predictive ability, I will continue with the high negative coefficient on steroid use, operating under the assumption that involvement with steroids will effectively end Hall of Fame induction chances.

Logistic Modeling Results for Pitchers

Model	Misclass Rate	AUC
Batters	8.8%	0.959
Pitchers	7.5%	0.958

Table 5. Performance statistics for both the previous batters model and the application of that model to the pitchers data.

The resulting model performs very similarly to the batters' model. Table 5 displays the two performance metrics for both models. The pitchers' model's misclassification rate is only 0.6% higher than the batters' and its AUC is only 0.009 lower. Additionally, Figure 4 shows the general dominance resulting from the same dominance analysis performed on the previous

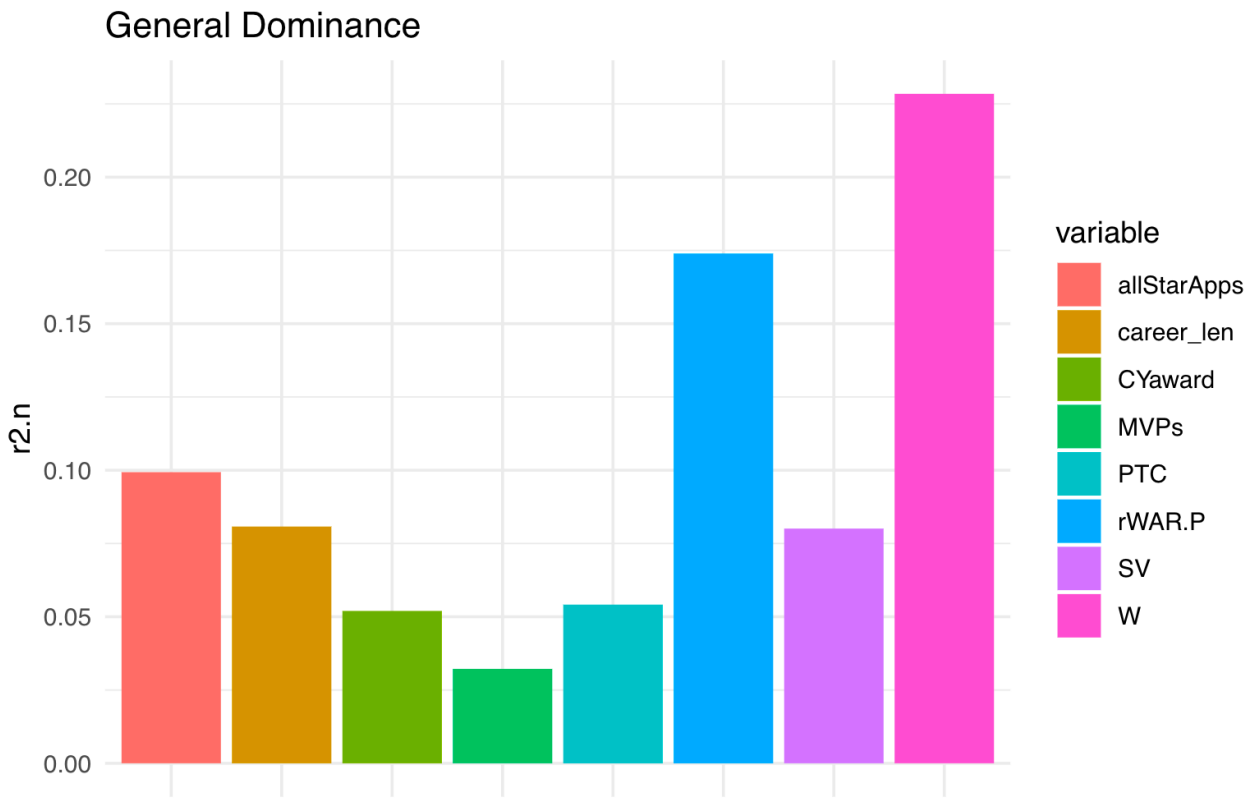

Figure 4. General dominance analysis results for the pitchers model, displaying a bar graph of the average added pseudo-R-squared for each feature.

model. Interestingly, WAR does not have as large of a role in the pitchers' model as it did in the batters'. A pitcher's contributions to his team's performance seems to be a much higher factor in this model, which makes sense because the pitcher has a much higher impact on an individual game that one of nine batters does. Further, wins and saves do capture a player's abilities, as well as his team's performance. Otherwise, the feature importance is consistent with the previous model. The two models perform strongly and are very good candidates moving forward to both predict the voting behavior on future generations and locate statistical inconsistencies of these voting patterns.

Section E. Unexplained Voting Behavior

Like any good model, not all observations can be clearly and accurately defined. There exist a handful of players whose induction status was incorrectly classified. Patterns in these errors hint at possible aspects of Hall of Fame candidacy that were not explained by the models above and could be improved in future models. The majority of the significant outliers fall into the following camps: bias from the Veteran's committee, evolution of baseball statistics, and legitimate errors by of the voters.

Veteran's committee bias is a significant force for helping former teammates sneak into the Hall of Fame. The main example of this bias, often dubbed "cronyism," occurred in the 1970s when Frankie Frisch led the committee, backed by former teammate Bill Terry and two sportswriters that covered his teams to induct five former teammates (*Committee on Baseball Veterans,* 2019). Additionally, after Frisch's death, his influence remained, leading to two more teammates being inducted shortly after. While Frisch could theoretically have played with outstanding Hall of Fame candidates that legitimately deserved induction, this was not the case. The models predicted less than 12.5% probability of induction for six of the seven players,

making these players among the worst outliers in Hall. To combat the bias, the Veteran's committee expanded membership to fifteen and limited elections to no more than two players per year. However, there is still evidence of bias among the group. In 2019, Harold Baines was elected to the Hall of Fame via the Veteran's committee. While Baines was a solid player, many question his induction. Over his 22-year-long career, Baines was only named to six all-star games and won one silver slugger. He did not win an MVP or lead the league in any important offensive category during any season. What he did have is four friends out of the fifteen total members on the Veterans committee. Baines received the minimum 12 votes for induction. While I am not saying that Baines was voted into the Hall solely because of his connections, it did not hurt his chances. When deciding on voting between two comparable candidates, a voter would reasonably be biased toward a personal connection. Therefore, the Veteran's committee will always host some bias as long as all teams are not equally represented.

Another factor in the difficulty of predicting Hall of Fame inductions is the evolution of baseball statistics since the start of baseball in 1871. While advanced statistics like WAR are available for historical players now, they were never seen by Hall of Fame voters in the past. Because of their lack of information, past voters relied on basic statistics like batting average, or career landmarks, like 200 wins for a starting pitcher. While they are often good indicators of success, sometimes they include biases of the era and do not reflect a player's abilities as well as they appear to. For example, batters in the "Live Ball" era performed very well relative to other decades, often hitting above .300. A career .300+ hitter in the Live Ball era, like Hack Wilson was viewed optimistically when in reality, he only kept up with the competition and did not stand out. Over his career, he put up a solid, but not particularly Hall-of-Fame-worthy 38.2 WAR before he was inducted by the Veteran's committee in 1979. Other examples of odd statistical

reasonings include career longevity and defensive ability, like Tony Perez and Bill Mazeroski, respectively. Essentially, the Hall of Fame tends to consist of normally, highly valued players and those that aren't necessarily in the same class in the traditional sense but in some other way are the absolute best at what they do.

Finally, there are those that simply deserve to be in the Hall of Fame but are not and those that do not deserve to be in the Hall of Fame but are. In other words, sometimes the voters make the wrong choice. It is important to keep perspective during this section of the paper and remember that outliers in general do not occur very often. Even less often are outliers that are hard to even justify from the perspective of the voters. However, false negatives can be corrected by the Veterans committee. On the other hand, false positives are permanent, as no one can be removed from the Hall post induction. One important example of an unjustifiable inconsistency is the difference between Ken Boyer and Ron Santo. Both players were third basemen in the 1960s with identical statistics and postseason awards. Third base has notoriously few inductees, but that does not explain why Ron Santo was inducted in 2012 and Ken Boyer remains on the outside. In my opinion, because of their similarities, I would expect the Veterans committee to return for Ken Boyer in the future. As of now, some players that have not been inducted have directly comparable careers to current Hall of Famers, which makes the induction process slightly confusing.

Section F. Predicting Future Generations

The final section of analysis will explore the future generations of players, which includes players that have already been inducted, those still on the ballot, and those that have not yet retired. The following results assume that current players were to retire after the 2019 season. They can and probably will build their Hall resumes further, but the analysis compares all

players at their current states. The process to predict was the same as above, but this time the model was trained on the entire training set. Then, the future data was limited by the same cutoffs and standardized like before.

First, the model predicted the induction probability of the batters. Table 6 displays all batters with a Hall of Fame prediction probability of at least 15%. At the top of the list sit Miguel Cabrera and Albert Pujols, who are still playing today. The fact that these players are not retired is even more impressive because they have more time to build on their already unbelievable careers. The model did very well with all players above 50%, correctly classifying every player that has been eligible thus far. One of the two remaining players above 50% that has not yet been on the ballot, Adrian Beltre, is widely considered a lock for induction on the first ballot ().

Hall of Fame Batter Predicitons

Player	Probability	Inducted	Eligible	HOF Class
Miguel Cabrera	100.0%			
Albert Pujols	99.4%			
Ken Griffey	99.4%	X	2016	2016
Barry Larkin	99.1%	X	2010	2012
Ivan Rodriguez	98.8%	X	2017	2017
Derek Jeter	97.2%	X	2020	2020
Mike Piazza	96.0%	X	2013	2016
Chipper Jones	79.1%	X	2018	2018
Vladimir Guerrero	79.0%	X	2017	2018
David Ortiz	74.3%		2022	
Roberto Alomar	68.9%	X	2010	2011
Frank Thomas	59.3%	X	2014	2014
Adrian Beltre	58.8%		2024	
Craig Biggio	55.1%	X	2013	2015
Edgar Martinez	46.4%	X	2010	2019
Sammy Sosa	40.8%		2013	
Carlos Beltran	24.4%		2023	
Ichiro Suzuki	23.6%		2025	
Jim Thome	19.6%	X	2018	2018
Jeff Bagwell	19.4%	X	2011	2017
Larry Walker	19.4%	X	2011	2020
Barry Bonds	15.2%		2013	

Table 6. Application of the previously derived batters model to the modern generation of players. Some players have already been inducted, some are still in the process, and some are still playing, making them not yet eligible.

Towards the bottom of the list, there appear to be a few errors in the model. Jim Thome, Jeff Bagwell, and Larry Walker are either undervalued by my model or inconsistently overvalued by the voters. I will throw Ichiro Suzuki in the same group, too, as he is widely considered a lock for the Hall of Fame. A couple of these players

were or will be inducted because of their outlier status in a few obscure categories, like Jim Thome's 22-year-long career, or Ichiro's defensive abilities. Note that much of Ichiro's career was spent in Japan before he joined the MLB, so he is also undervalued for that reason. Regarding the few false negatives, I will offer two possible explanations. The first is that the negative coefficient on the retirement decade carries too strong of a penalty for future application. This error should be corrected as Hall of Fame elections become more regularly supported with advanced statistics and inductions become more consistent in the future. The second explanation is that many of these players played during the steroid era of the MLB, so there is no precedent to the actions of the voters. Many voters look very negatively toward steroid use, which is why no one named in the Mitchell Report or suspended for steroids at any time has ever been elected to the Hall of Fame. The voters will be tested on how they handle this generation of candidates. Never has a player as good as Barry Bonds been denied induction; if he is denied, then how do they handle candidates that were never suspended, but accused, like Sammy Sosa or David Ortiz? These unanswered questions will set the precedent for voting moving forward. As far as the false negatives go, if the voters were to deny the best players because of their involvement with steroids, do they replace them with the best players of the era that didn't do steroids? Surely, these players would be of lesser value. Obviously, the actions of the voters are very important at this time.

As far as pitchers, the model seemed to predict future generations slightly more accurately than the batters. Table 7 displays the pitchers with a predicted induction probability greater than 15%. In the same class as Miguel Cabrera and Albert Pujols, the model predicts Justin Verlander as a safe pick for the Hall of Fame even as his career continues. In general, it seems that the probabilities are fairly accurate. For now, John Franco is a false positive, as he

Hall of Fame Pitcher Predicitons

Player	Probability	Inducted	Eligible	HOF Class
Mariano Rivera	100.0%	X	2019	2019
Greg Maddux	100.0%	X	2014	2014
Randy Johnson	100.0%	X	2015	2015
Trevor Hoffman	100.0%	X	2016	2018
Justin Verlander	99.9%			
Tom Glavine	99.7%	X	2014	2014
John Smoltz	97.0%	X	2015	2015
Roger Clemens	96.3%		2013	
Mike Mussina	96.1%	X	2014	2019
CC Sabathia	89.4%		2025	
Francisco Rodriguez	84.4%		2023	
John Franco	75.1%		2011	
Joe Nathan	59.4%		2022	
Billy Wagner	58.5%		2016	
Jonathan Papelbon	48.2%		2022	
Mark Buehrle	38.7%		2021	
Tim Hudson	38.2%		2021	
Craig Kimbrel	24.4%			
David Wells	15.9%		2013	

Table 7. Application of the previously derived pitchers model to the modern generation of players. Some players have already been inducted, some are still in the process, and some are still playing, making them not yet eligible.

was removed from the BBWAA ballot after failing to receive 5% of the vote. However, his induction status could change as he is now eligible to be voted in via the Veteran's Committee. The one big question mark in the table is Roger Clemens. Even with the same steroid coefficient penalty as Barry Bonds, Clemens' probability drops only to 96.3%. He was truly one of the best to ever play, posting a pitching WAR of 138.7, 354 career wins, one MVP, seven Cy Young awards (more than any other pitcher), two Pitching Triple Crowns, and 11 All-Star appearances. The BBWAA has never rejected a pitcher as good as Clemens, but with only one year left on the ballot, his numbers do not trend favorably for induction. Like with the batters, the voters have an important precedent to set during this election era.

VIII. Conclusion

The induction process for the National Baseball Hall of Fame is constantly evolving, making predictions more difficult. During its first elections in 1936, voters had to make up for the lack of elections over the prior 65 years by both voting in modern players and looking back

to find the elite of the past. Early baseball relied on count statistics like home runs, hits, and strike outs to judge players, which while legitimate statistics, did not always tell the story well. Corruption also played a part in election results, as seen by Freddie Frisch's Veteran's Committee in the 1970s. Finally, the push towards sabermetrics led to new ways of evaluating players.

In general, the logistic models were able to identify the five most important categories of a Hall of Fame candidacy: career performance, single season achievements, longevity of career, team performance, and extracurricular penalties. A dominance analysis revealed that typically, the career performance was the most important factor in a candidacy, followed by single season achievements. I used the extracurricular penalties section to essentially exclude candidates that had an off-the-field issue that could potentially bar them from the Hall of Fame, like gambling or steroid use. After this era of voting is completed, the factors could be fine-tuned to predict how much involvement in steroid activities could penalize induction probabilities. The most surprising factor was a significant positive coefficient for team performance for batters. Since the model already accounts for player career and single season achievements, this factor corresponds to the performance of the team above a player's contributions to it. It is evidence that the voters are biased towards players that played for better teams over the abilities that they actually demonstrate.

While the final performance of my model was very accurate, there is always room for improvement. After performing a five-fold cross-validation logistic regression on the set of data including players that retired between 1924 and 2004, the misclassification rates were 8.8% and 7.5% and the AUC metric was 0.959 and 0.958 for the batters and pitchers, respectively. In terms of model improvement, some factors would be very difficult to incorporate, like committee bias,

but others could be worked into the model, like unique statistic outliers. Moving forward, I would expect a higher prediction ability, as sabermetrics give future voters consistent and mostly complete information about the players they are considering. That being said, the voters will always face new and unique challenges that set the precedent for future voting.

IX. References

BBWAA Election Rules. (2014). National Baseball Hall of Fame.

chalsey15. (n.d.). *Calculating the WAR statistic*. Instructables.

Clayton Kershaw. (2020). Baseball Reference.

Committee on Baseball Veterans. (2019, September 9). Baseball Reference.

Hall of Fame Ballot History. (n.d.). Baseball Reference.

James, B. (2019). *The Bill James Handbook: The complete up-to-date statistics on every major league player, team, and manager through last season*. Acta Sports.

Lahman, S. (n.d.). *SeanLahman.com*. Minimalist Blog.

Lindbergh, B. (2020, September 24). *Long Before WAR, Nobody Knew What MLB Players Were Worth*. The Ringer.

WAR Comparison Chart. (n.d.). Baseball Reference.

What Are Your Odds of Making the Pros? (n.d.).

Zminda, D. (2010). Henry Chadwick Award: Bill James. *Summer 2010 Baseball Research Journal*.