I. Abstract

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?

- Add paragraph about methods/results
- Add paragraph about the following structure of 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 known yet. 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 a part of history and are not a factor that will 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. Figure 1 displays the final game of the current universe of batters by their total career at-bats. The total number of Hall of Fame inductees seems somewhat consistent over time. However, it appears that the number of at-bats that a typical Hall of Fame member accumulates increases over time. One important note is 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. Figure 2 displays the date of the final game of new universe of relevant players by their number of at-bats. 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 incorrectly classifies 11.5% of observations. Let's continue to search for a better identifier of 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 lower than the combined model, but the combined model has a slightly higher AUC. The differences are so small that there is likely no significant difference in the model, although the model using only WAR can produce almost identical results with only one feature.

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 10.8% and an AUC of 0.943. Its high-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.

Section C. Improving the Model

A closer look at the given model reveals that it performs relatively well but has room for improvement. Figure 3 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 10.8% 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. 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 1994. 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.

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 0.943 to 0.961, and its misclassification rate decreased from 10.8% to 7.3%. Additionally, Figure 4 shows WAR graphed against induction status and colored by predictive accuracy. Compared to Figure 3, 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 4 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

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

Qualitative discussion of unexplained outliers

Predictions to future generations

Conclusion

References

BBWAA Election Rules. (2014). National Baseball Hall of Fame.
chalsey15. (n.d.). *Calculating the WAR statistic*. Instructables.

Clayton Kershaw. (2020). 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*.