ADDRESSING THE POSITIONAL ISSUE FOR BASEBALL HALL OF FAME CANDIDACY

Nicholas Torsky

Abstract

Baseball has long proven to be the first frontier for sports analytics. From Henry Chadwick's development of rudimentary statistics like batting average, runs scored, and runs allowed in the mid to late 1800s, to Hy Turkin's *Baseball Encyclopedia* in 1951, to Bill James's *Baseball Abstract*, and finally to the modern analytics systems employed by professional and amateur programs today like the Lahman database, the sport of baseball has been focused the concept of quantitative assessment of the game. Particularly, the source of much debate in the baseball community, particularly in the first few months of the Major League Baseball (MLB) off-season, is election to the Baseball Hall of Fame.

Every year, a list of former MLB players are placed onto a ballot for the Baseball Writers' Association of America (BBWAA) to vote on. Each writer is allowed to cast at most 10 votes (although many writers often cast fewer than this number) as they see fit among the players on the ballot. Players become eligible for the Hall of Fame ballot once they have been retired from the professional ranks for 5 years, provided that they played at least 10 seasons in Major League Baseball. Each player can be on the BBWAA ballot for a maximum of 10 years, with their name being removed only if they are elected into the Hall (>= 75% of the vote in a given year) or are dropped from the ballot (<5% of the vote in a given year). Voting results are usually released within the first few weeks of each new year. For example, although the BBWAA voting will take place in 2020 at this time of writing, the results will be released in early 2021, so it is referred to as the 2021 ballot.

For players who are not Hall of Fame "shoo-ins", hundreds and thousands of writers, fans, and members of the baseball community argue and debate over whether they deserve to be enshrined among baseball history's elite. An oft-repeated phrase used for this concept is "the size of the Hall", referencing how many people should be inducted. If one has a "large Hall", they are likely to cast votes for players that were very good players for their time, but not all-time record setters for various statistics. One with a "small Hall" views the Hall of Fame as a selection of truly elite and dominant players who made lasting impressions of the game. Modern "small Hall" proponents often fail to recognize the effect of a player's defensive position when considering a player's candidacy. Particularly, players at positions with demanding defensive responsibilities like shortstop or catcher often do not receive as much devotion as players at less defensive-minded positions like first base or the corner outfield positions, simply because their hitting numbers (especially for power statistics like home runs and slugging percentage) do not compare as favorably. The idea that players at these positions are not as potent offensive performers is simply a product of the way the game is structured; managers and team front offices have historically been willing to sacrifice premier offensive production at these positions in favor of a sure-handed defender and clubhouse leader. In the past, having players of this type at these positions has proven to be the best strategy to win games. In the entire history of Major League Baseball up until the last 15 years, shortstops and catchers with Hall of Fame-caliber hitting skills rarely have the defensive skills necessary to remain at these positions and are moved to others, and players that have both are exceedingly rare.

Many metrics determining potential Hall of Fame candidacy either neglect to make note of a player's position, or do so in a way that does not maximize the effect of a player's position. In this paper, I will address the positional issue by comparing players who have not earned induction into the Hall of Fame both to players at their positions and to all inductees.

Unfortunately, the availability and efficacy of evaluating a player's defensive performance statistically has not yet caught up to that of a player's offensive prowess. Current metrics fail to take many factors affecting a player's defense into account. For example, fielding percentage -- one of the oldest fielding metrics available and comparable to a hitter's batting average -- provides the ratio of defensive plays on which a player commits an error, where an error is a play in which a fielder fails to record an out on a play that "should have been made". Clearly, errors are subjective categorization, dependent on the judgement of a game's official scorer, and does not necessarily take into account a play's difficulty, the speed of a runner, or other circumstances relating to the ball put in play. Even the Major League Baseball Official Rules include plenty of stipulations for the ruling of an error (see Appendix 1). Today, player tracking software and long-term fielding data are collected and used for more in-depth metrics, but still present some bias, and their methodology can be difficult to find and replicate, not to mention their extremely limited and recent nature excluding seasons as recent as 15 years ago. For this reason, I will not include fielding metrics for determining Hall of Fame candidacy in this paper.

According to the Baseball Hall of Fame's official website (baseballhall.org), "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". Especially in today's statistically inclined baseball environment, criteria relating to a player's on-field contributions carry more weight than ever before, although, as the analysis will show, writers often invoke the "character clause" to justify voting against players with relevant off-field circumstances. In the past, the BBWAA had limited access to databases and analysis regarding eligible players, and even players already in the Hall of Fame to use as a reference. In this paper, I will estimate the probability of Hall of Fame inclusion for all Major League Baseball players with at least 10 years of service time, conditional on certain career statistics using a logistic regression model. The algorithm will compute a logistic regression on each statistic, giving a 0 to 1 probability of inclusion using leave-one-out cross validation. The 0 to 1 probability for each statistic will be averaged, giving a final 0 to 1 score that estimates Hall of Fame candidacy. Using the percentage of all players in the Hall of Fame, the algorithm compiles three tables: the first, a table of the top n% of scores amongst all qualified players, signifying players who have a high probability of inclusion (called Hall of Fame Worthy, or `HOF_worthy`); the second, a table of the top n% of scores at each position, signifying players who merit discussion about their Hall of Fame inclusion, if not a high probability of inclusion (colloquially, the Hall of Very Good, or HOVG); the third, a table of "false positives", or players inducted into the Hall whose offensive statistics do not fit the criteria of the logistic regression algorithm (`false_positives`). This third case is of particular interest, and is discussed at length in the "Results" section. The n percentages for the entire Hall of Fame, as well as at each position, are listed in Table 1.

Position	% of players in HOF
All	7.54
Catcher	3.90
First Base	10.90
Second Base	8.26
Third Base	6.09
Shortstop	8.76
Outfield	8.33

Table 1

Data

As the development of new statistics for baseball has progressed, so too has the accessibility and quality of baseball data. Information about any number of baseball related data is available from a variety of sources, chiefly through the work of volunteers. The data included here is the most recent incarnation of the Lahman Database -- a collection of tables containing hitting, pitching, fielding, and other statistics from 1871 to 2019, developed and compiled by Sean Lahman. Conveniently, the Lahman Database exists as an R package, which is where the calculations will take place. All data for the purposes of this paper will come from the Lahman database.

The tables in the Lahman package include hitting data from all players and all seasons from 1871 through 2019, and Hall of Fame election results as of the end of the 2019 season. As such, the players elected after the 2019 season (Derek Jeter, Ted Simmons, and Larry Walker), as well as players inducted during the 2019 season (Edgar Martinez and Harold Baines), are not included in the `HallOfFame` table. They are, however, clearly represented as potential Hall of Famers by the model.

The Lahman database provides season-by-season data for counting statistics like hits, home runs, and games played, and each row of the table corresponds to a unique player ID, year, team, league, and stint (number of times a player has played games for a given time during a distinct period of his career). The package also provides a function, `battingStats`, to calculate rate statistics like batting average, on-base percentage, and slugging percentage. For the purposes of both usefulness and data availability, I modify or completely omit the categories describing stolen bases, times caught stealing, strikeouts, intentional walks, times hit by pitch, times grounded into double plays, sacrifice bunts, sacrifice flies, triples, and batting average on balls in play (BABIP). Particularly, I found that triples tend to disproportionately favor players from before the so-called "Live Ball Era", named for the period in baseball after 1920 in which rule changes (particularly to how the baseball is treated) resulted in a marked increase in offensive statistics like home runs. The relationships between Hall of Fame inductees and non-Hall of Fame inductees for each of these statistics are represented graphically in Figure 1. Especially for players from the earlier periods of baseball, some of these statistics were not fully developed, and thus their values are either non-reliable or non-existent (i.e. caught steading, sacrifice hits/flies, and grounded into double play). Interestingly, players with excellent career stolen base totals (Rickey Henderson, for example) and players with few career strikeout totals relative to walk totals (George Sisler, Ted Williams, for example) often compiled exemplary career numbers in other statistics, so the exclusion of these statistics as predictors did not harm the model.

Another potential source of bias is in the nature of the statistics used. Many of the statistics in the `Batting` table are counting statistics, meaning that the longer a player plays, the better chance he has at accumulating the statistics necessary for Hall of Fame consideration under the model. While longevity is a crucial component for potential induction into the Hall, such statistics ignore relatively short-lived, yet dominant careers. For this reason, I considered using Lahman's `AwardsPlayers` table to favor players with shares of MVP voting and All-Star selections, but decided not to, as this would introduce a whole new bias. All-Star teams are decided mid-season rather than at season's end, so a player who performs better after the All-Star game may not be appropriately rewarded. Furthermore, All-Star and MVP voting are notorious for favoring players who play in larger media markets and tend to reflect a player's notoriety more so than his performance. Therefore, I decided to ignore the imbalance of counting statistics and rate statistics, instead addressing the Hall of Fame players whom the logistic regression model did not favor in the `false_positives` table.

I created a series of tables for filtering data dependent on useful information. `HOFers` was used to initially find Hall of Fame players elected by the BBWAA from `HallOfFame` (the table includes inductees from other election committees, as well as non-player inductees), `HOF_field` was created from `Fielding` and used for finding the primary position of Hall of Fame players, and information from each of these tables were combined with `Batting` into `HOF`. Similar processes were used for the creation and use of `all_field` and `all_batters`. `allHall` includes all inducted Hall of Famers, regardless of the method by which they were voted in.

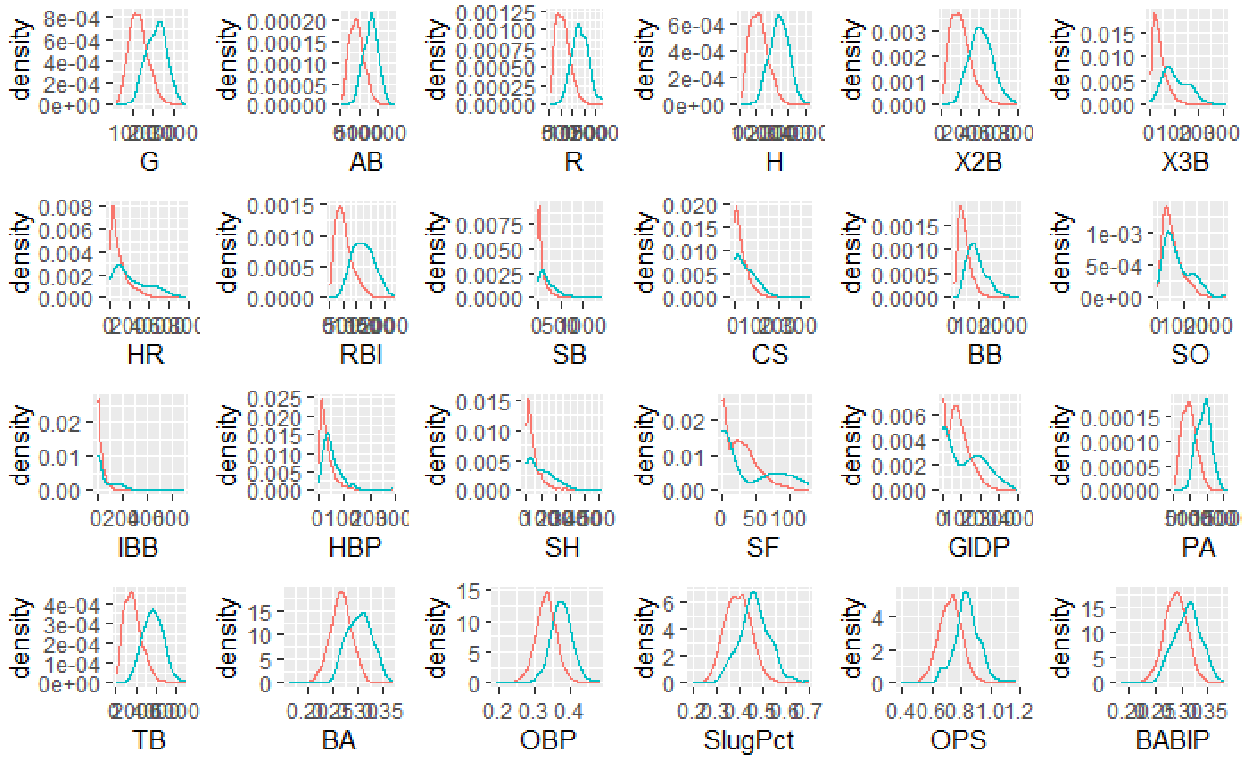

Figure 1

Methods

The methods involved in calculating the scores amongst all inductees and inductees at each position were slightly different. Comparing to all inductees, Hall of Fame candidacy was predicted using four generalized linear models (GLM) with two different sets of predictors. For the first, the GLM used statistics conventionally deemed significant for the BBWAA, including runs scored, hits, doubles, triples, home runs, runs batted in, batting average, and OPS. For the second, the GLM used all predictors deemed significant by the graphical analysis from Figure 1, with the exception of OPS (OPS stands for on-base percentage plus slugging percentage, and these two predictors are included in the model already). Both of these models are an initial GLM to obtain means and scales for the coefficients for each predictor, including the intercept, which were then fed into a Bayesian GLM, yet I used these models to create predictions as well (represented in the acc_LOOCV1 and acc_LOOCV2 columns of `all_batters`). Using leave-one-out cross validation and the `optimalCutoff()` function, the 0 to 1 predicted values for each of the four methods were cut off at the value given by `optimalCutoff()` (optimized for misclassification error), with values greater than the cutoff assigned a 1 (probable Hall of Famer) and values less than or equal to the cutoff assigned a 0 (not a probable Hall of Famer). For each position, the GLM performed a logistic regression on one predictor at a time, adding the predictor to a running sum vector containing the scores for each player only if the misclassification error was less than or equal to 0.1. The GLM used all predictors, as in the second model for all batters (the positional model included OPS, as the regression was performed one predictor at a time).

Due to many statistics spanning at least two orders of magnitude, some predictors were transformed onto a logarithmic scale. The predictors not transformed are listed in `keep_these`, and were rejoined to the transformed predictors in `log_all_batters`. This data frame was then split into subsets by position.

Ultimately, the decision to employ different models for the positional models and the overall model was driven by runtime. The four overall models using leave-one-out cross validation took a substantial amount of time already and implementing a logistic regression model on each predictor one by one for over 2000 entries, unfortunately, took too long to be effective.

There are a number of different metrics currently employed to determine Hall of Fame candidacy, and most are point-based systems developed by Bill James. Under these systems, players are awarded or deducted points for various career achievements. More information about these statistics can be found in the References section. These types of counting frameworks are designed less with the idea of measuring a player's greatness in mind, and more toward a player's probability of induction into the Hall. One advantage of these metrics is that it accounts for season-by-season performance – which serves to benefit players like Kirby Puckett, who only played 12 professional seasons, as well as other premier players with shorter careers – as well as playoff appearances and seasonal awards. The score developed in this paper differs from such methodologies in that it operates under the assumption that a player's career offensive greatness, especially given his primary defensive position, is the only factor affecting his chances of entering the Hall of Fame. As a result, of these so-called "Jamesian" metrics, the analysis presented here can be most likened to Similarity Scores. The Similarity Score compares one player to another, starting with 1000 points, and deducts points for differences between the two players' career stat totals at given intervals which are different for each statistic. One could apply a Similarity Score to any two players across baseball history to find the similarity in their career numbers, including a Hall of Fame inductee and a player in contention for Hall of Fame election. Rather than applying Similarity Scores pairwise to all qualified players in the Lahman database (needless to say, a computationally costly endeavor), the methods employed here take a distribution of players already deemed worthy of Hall of Fame induction and quantify the probability that a given player is better than the average Hall of Fame inductee at his position.

Simulations

The logistic regression model used leave-one-out cross validation (LOOCV) to generate Hall of Fame prediction probabilities, training the model on all entries in the data frame and testing the model on one entry while iterating through all n entries in the data frame. This method prevents overfitting data – that is, when the model fits the data to the response too closely, not taking noise into consideration. Overfitting is characterized by good accuracy, but poor out-of-sample results. LOOCV in particular can be contrasted with k-fold cross validation by looking at the bias-variance tradeoff: LOOCV results in higher variance and lower bias, while k-fold cross validation exhibits the opposite property.

Because the problem of predicting Hall of Fame candidacy is one of binary classification, the primary measure of fit is misclassification error. The tables representing the overall misclassification error rates can be found in Appendix 3. Equation 1 describes the method by which misclassification error is calculated.

	Predicted HOF	Predicted non-HOF
Actual HOF	a	b
Actual non-HOF	c	d

$$MCE = \frac{b+c}{a+b+c+d} = \frac{error}{total} \qquad \textit{Eqn. 1}$$

The `optimalCutoff()` function in the `InformationValue` R package was instrumental in finding the ideal cutoff values for each logistic regression model, with each model achieving better than 94.6% accuracy. Such miniscule out-of-sample prediction errors suggest, in general, the logistic regression models were extremely effective at classifying players while still allowing room for the premier non-inducted players (especially those active, not yet ballot eligible, or currently on the ballot). The nature of minimizing misclassification error prevented a large number of players who did not accumulate Hall of Fame-worthy career statistics from inclusion in the final data frames.

The effectiveness of the models can also be described using True Positive Rate (TPR) and False Positive Rate (FPR). These metrics describe the proportion of each class that were correctly and incorrectly identified, respectively. Their equations are described in Equations 2 and 3.

$$TPR = \frac{a}{a+b} = \frac{\text{Predicted HOF } AND \text{ Actual HOF}}{\text{Actual HOF}} \qquad \textit{Eqn. 2}$$

$$\mathrm{F}PR = \frac{\mathrm{c}}{\mathrm{c}+\mathrm{d}} = \frac{\text{Predicted HOF } AND \text{ Actual non-HOF}}{\text{Actual non-HOF}} \qquad \textit{Eqn. 3}$$

The FPR and TPR values for each of the four models are found in Appendices 4.1 and 4.2. The TPR values tended to be lower than one might expect: around 62.5%. This is largely due to the inclusion of pre-Live-Ball era players presenting outliers in various predictors, which is discussed at greater length in the Results section, as well as the relatively low percentage of players in the table who are classified as Hall of Fame inductees (see Table 1). The FPR values are each considerably small: near 2%. Again, this is largely due to the large number of players who have not been honored with Hall of Fame induction.

The final model performance metric is Lift, with its formula given in Equation 4, and is described as the ratio of target response to the average response, or the ratio of the percentage of correctly classified Hall of Famers to the ratio of correctly classified players in general. Lift is very similar is function to the power or receiver operating characteristic (ROC) curve. The Lift values for all four models for Hall of Fame prediction are shown in Appendix 5. Interpreting the high Lift values (over 9 in each model), players with logistic regression scores above the cutoff value for each model are nine times more likely to be inducted into the Hall of Fame as players drawn from `all_batters` at random.

$$\mathrm{Lift} = \frac{\mathrm{a}/(\mathrm{a}+\mathrm{b})}{(\mathrm{a}+\mathrm{c})/(\mathrm{a}+\mathrm{b}+\mathrm{c}+\mathrm{d})} = \frac{\text{TPR}}{\text{(Correctly Classified) / (All Players)}} \qquad \textit{Eqn. 4}$$

Results

The results of the logistic regression model were separated into three tables: `HOF_worthy` (containing the 44 players the algorithm misclassified as Hall of Fame inductees), `false_positives` (containing the 41 players the algorithm misclassified at NOT Hall of Fame inductees), and `pos_only` (containing the 31 players in the top n% of hitters at their position by Table 1, but correctly classified by the algorithm). All other players, while significant in their own rights, were correctly classified, and therefore do not merit further discussion here. For the sake of discussion, each of the three tables will be broken down into four groups: the active players (only applicable to `HOF_worthy` and `pos_only`), the old-timers (players whose careers were long enough ago that their relevance in today's voting procedures are largely forgotten), the steroid users, and the edge cases (players who do not fit into the three other categories). For each group, their career statistics and regression scores, along with the appropriate Hall of Fame statistical means, will be displayed in a table for context. Additionally, Figure 2 shows a smoothed regression line for each of the 15 statistics, with the points on the top of each graph (y=1) representing the corresponding x-axis coordinates of each Hall of Famer for the described statistic and the bottom points (y=0) to those not elected.

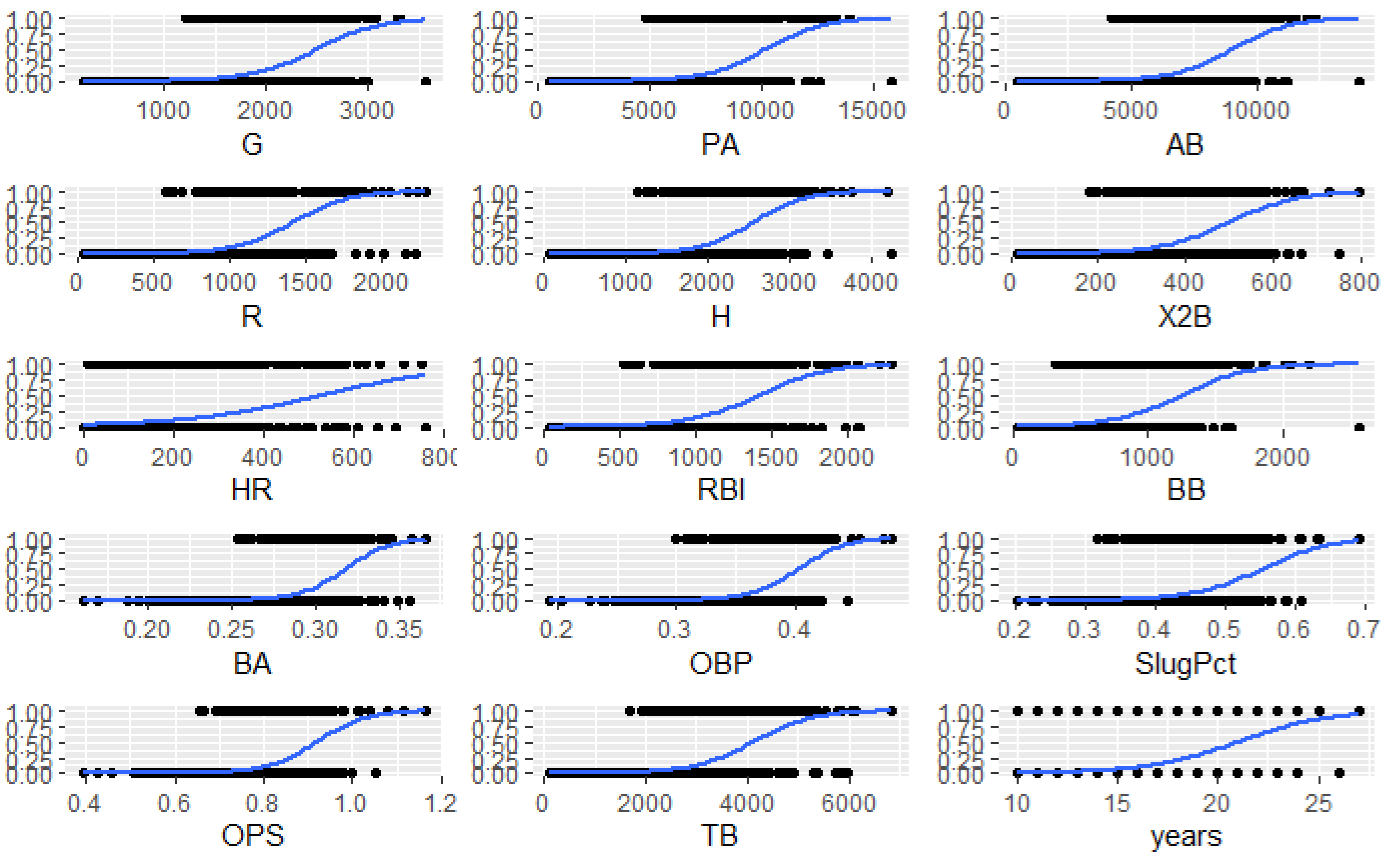

Figure 2

Each of the final four columns corresponds to the four logistic regression models used to compare all 2094 entries in the `all_batters` data frame on a 0 to 1 scale. These values can be interpreted as 0% to 100% probabilities of induction into the Hall of Fame. Recall that models 1 and 3 use predictors conventionally deemed significant by the BBWAA, while models 2 and 4 use all selected predictors. Also, models 3 and 4 are Bayesian Generalized Linear Models, using coefficients from models 1 and 2 as prior distributions.

As is the nature of complex systems like baseball, many of the players in each of these tables have reasons for their election or rejection from the Hall of Fame that go beyond their offensive production. Their candidacy will be discussed primarily in regard to their statistics and logistic regression probabilities, but with note of their extracurricular circumstances as well.

First, the "active Hall of Famers". While it is true that these players, as of the conclusion of the 2020 MLB season, are still demonstrating the ability to bolster their career statistics for at least one more season, the logistic regression algorithm has deemed their achievements to date worthy still of Hall of Fame induction. Miguel Cabrera and Albert Pujols earned Hall of Fame scores of around 80% and 87%, respectively. As evidenced by Table 2, these players are better than the Hall of Fame average in nearly every statistic considered (Cabrera has one season to eclipse the Hall of Fame mean career length, although we should recall that the Lahman database does not include the 2020 season at the time of publication). Given that the optimal cutoff values for each of the four models were just shy of 40%, these two players pass the eye test for induction (the cutoff values for each of the four models can be found in Appendix 2).

	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Cabrera	17	2400	8949	1429	2815	577	477	1694	1135	0.315	10236	4857	0.543	0.392	0.935	0.789	0.799	0.802	0.800
Pujols	19	2823	10687	1828	3202	661	656	2075	1322	0.300	12231	5863	0.549	0.379	0.928	0.871	0.866	0.882	0.866
Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 2

The "old-timers" in the `HOF_worthy` table include Jake Daubert (1910-1924), Jack Doyle (1889-1905), Joe Judge (1915-1934), Stuffy McInnis (1909-1927), Fred Tenney (1894-1911), Cupid Childs (1888-1901), Buddy Myer (1925-1941), Lave Cross (1887-1907), Larry Gardner (1908-1924, Steve Brodie (1890-1902), Doc Cramer (1929-1948), Patsy Donovan (1890-1907), Jimmy Ryan (1885-1903), George Van Haltren (1887-1903), Bobby Veach (1912-1925), Bill Dahlen (1891-1911), Ed McKean (1887-1899), Mike Tiernan (1887-1899), and "Shoeless" Joe Jackson (1908-1919). These players, as we can see from the years in which they played, are from a period of baseball that was very different than the one we see today, as most of them played the bulk of their careers prior to the dawn of the aforementioned "Live-Ball era". These players' regression scores were challenging ones to interpret, as they were premier players of their time, yet the style of the game in which they played resulted in very different career statistics compared to more recent comparisons, as shown in Table 3. Only two of the players (Ryan and Tiernan) eclipsed 100 career home runs, and many of these players' career statistics were close to or below the Hall of Fame averages. Their inclusion in the `HOF_worthy` data frame is a result of the logistic regression algorithm reading the similar statistics of Hall of Fame inductees from the same era (although the algorithm is unaware of the period in which any player played) and making prediction with that in mind. I toyed with the idea of omitting seasons prior to 1920 from consideration, but deemed it unfair to these players, and the idea of drawing an arbitrary line in the sand at the year 1920 worked to the detriment of even some modern players who employed a similar style of play. The inclusion of these players certainly affected the results of the algorithm, as these players and their Hall of Fame inducted counterparts skewed many of the regression coefficients for certain predictors (namely the power statistics like home runs and slugging percentage, as well as triples, as mentioned earlier). Particularly, the players in this section of the table tended to record fewer career hits, runs, and runs batted in than the average Hall of Famer, but with generally better-than-average career lengths, batting averages, and on-base percentages, as evidenced in Table 3. Many of these players' careers ended before the inception of the Hall of Fame in 1939, and several were therefore ineligible even for consideration on the BBWAA ballot. With this in mind, they may still earn baseball's highest honor via the Veterans' Committee – a series of committees, each devoted to a particular period in baseball history, specifically designed to consider players who are long since ineligible for the BBWAA ballot.

	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Steve Brodie	12	1438	5703	886	1728	191	25	900	420	0.303	6342	2172	0.381	0.365	0.746	0.471	0.409	0.463	0.408
Cupid Childs	13	1457	5622	1214	1721	205	20	743	991	0.306	6766	2188	0.389	0.416	0.805	0.647	0.691	0.644	0.691
Doc Cramer	20	2239	9140	1357	2705	396	37	842	572	0.296	9933	3430	0.375	0.340	0.715	0.543	0.489	0.552	0.489
Lave Cross	21	2277	9084	1338	2651	411	47	1378	466	0.292	9741	3475	0.383	0.329	0.712	0.775	0.739	0.784	0.739
Bill Dahlen	21	2444	9036	1590	2461	413	84	1234	1064	0.272	10405	3452	0.382	0.358	0.740	0.516	0.515	0.520	0.515
Jake Daubert	15	2014	7673	1117	2326	250	56	722	623	0.303	8742	3074	0.401	0.360	0.761	0.412	0.535	0.413	0.535
Patsy Donovan	17	1824	7505	1321	2256	208	16	738	457	0.301	8172	2662	0.355	0.348	0.703	0.767	0.724	0.772	0.724
Jack Doyle	17	1569	6055	977	1811	316	25	971	440	0.299	6589	2330	0.385	0.351	0.736	0.390	0.354	0.383	0.354

Larry Gardner	17	1923	6688	867	1931	301	27	934	654	0.289	7685	2571	0.384	0.355	0.739	0.333	0.485	0.323	0.485
Shoeless Joe Jackson	13	1332	4981	873	1772	307	54	785	519	0.356	5690	2577	0.517	0.423	0.940	0.733	0.874	0.736	0.874
Joe Judge	20	2171	7898	1184	2352	433	71	1034	965	0.298	9171	3316	0.420	0.378	0.798	0.514	0.593	0.517	0.593
Stuffy McInnis	19	2128	7822	872	2405	312	20	1063	380	0.307	8623	2979	0.381	0.343	0.724	0.795	0.811	0.798	0.811
Ed McKean	13	1655	6894	1227	2084	272	67	1124	636	0.302	7626	2873	0.417	0.365	0.782	0.577	0.591	0.581	0.592
Buddy Myer	17	1923	7038	1174	2131	353	38	850	965	0.303	8187	2858	0.406	0.389	0.795	0.500	0.565	0.499	0.565
Jimmy Ryan	18	2014	8172	1643	2513	451	118	1093	804	0.308	9124	3632	0.444	0.375	0.819	0.647	0.661	0.659	0.662
Fred Tenney	17	1994	7595	1278	2231	270	22	688	874	0.294	8809	2721	0.358	0.371	0.729	0.496	0.579	0.497	0.579
Mike Tiernan	13	1478	5915	1316	1838	257	106	853	748	0.311	6732	2737	0.463	0.392	0.855	0.432	0.467	0.431	0.467
George Van Haltren	17	1990	8043	1642	2544	286	69	1015	871	0.316	9017	3359	0.418	0.386	0.804	0.883	0.902	0.891	0.902
Bobby Veach	14	1821	6656	953	2063	393	64	1166	571	0.310	7557	2942	0.442	0.370	0.812	0.443	0.561	0.442	0.561
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 3

David Ortiz, Barry Bonds, Rafael Palmeiro, Sammy Sosa, Manny Ramirez, Gary Sheffield, and Alex Rodriguez constitute the "steroid users" portion of the `HOF_worthy` table. These players are either confirmed (Rodriguez, Bonds, Palmeiro, Sheffield, and Ramirez) or suspected (Ortiz, Sosa) to have tested positive for the use of anabolic steroids – a performance enhancing drug (PED) strictly prohibited by MLB's substance abuse policy – during their playing careers. The BBWAA, as well as public opinion, tends to negatively impact the Hall of Fame chances of players whose names are involved in PED rumors by using the Hall's "character clause" criterion (with a few notable exceptions[1]). Bonds in particular is the all-time career home run champion, as well as the all-time leader in walks and intentional walks, and was called the greatest hitter in baseball history by many during his career. With these statistics and his PED history in mind, however, Bonds's name appears on the Hall of Fame ballot for the 9th time in 2021, never having received greater than 60% of the vote in a given year. Without steroid allegations surrounding his name, his career statistics – and his logistic regression scores – in Tables 4 and 6 point to his obvious Hall of Fame candidacy. The results of his final two ballot eligible years will be followed closely by the baseball community. Palmeiro, Sosa, Ramirez, and Sheffield were all prolific sluggers of their time as well, landing at 13th, 9th, 15th, and 26th on the all-time home run leaderboard, respectively, with the complementary statistics and logistic regression scores to back them up. Specifically, Palmeiro is one of two players on this list to surpass the 3000 hit benchmark, although he failed to receive the minimum 5% of the vote to stay on the ballot after his 4th year in 2014, receiving a maximum vote share of only 12.2% in 2012. Sosa, Ramirez, and Sheffield appear on the ballot in 2021 in their 9th, 5th, and 7th years, respectively. Rodriguez and Ortiz have yet to appear on a Hall of Fame ballot, but both will become eligible for the first time on the 2022 ballot. These players are 4th and 17th in all-time home runs, respectively, with generally above-average Hall of Fame peripheral numbers to match. While this algorithm measures strictly offensive performance, it is important to note that Ortiz spent much of his career as a designated hitter rather than a first baseman, as he is listed in this analysis. Only one other primary DH has earned Hall of Fame induction (Edgar Martinez in 2019), largely due to the lack of defensive impact. Table 5 shows Martinez and Ortiz compared separately. Rodriguez, meanwhile, is considered to be among the greatest all-around hitters in baseball history, as the regression score shows (one of only 29 players to achieve greater than 90% probability under all four models), but his two positive PED tests and suspensions may prove to hurt his candidacy much in the same way as it has

[1] Mike Piazza, inducted 2016, admitted to PED use before ban; Ivan Rodriguez, inducted 2017, accused but not proven; Jeff Bagwell, inducted 2017, accused but not proven; Tim Raines, inducted 2017, admitted to cocaine use

Bonds. Table 6 compares Rodriguez and Bonds to other all-time best hitter candidates Willie Mays, Babe Ruth, Stan Musial, and Ted Williams.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Barry Bonds	22	2986	9847	2227	2935	601	762	1996	2558	0.298	12606	5976	0.607	0.444	1.051	0.973	0.962	0.976	0.962
David Ortiz	20	2408	8640	1419	2472	632	541	1768	1319	0.286	10091	4765	0.552	0.380	0.932	0.561	0.515	0.566	0.515
Rafael Palmeiro	20	2831	10472	1663	3020	585	569	1835	1353	0.288	12046	5388	0.515	0.371	0.886	0.730	0.716	0.743	0.716
Manny Ramirez	19	2302	8244	1544	2574	547	555	1831	1329	0.312	9774	4826	0.585	0.411	0.996	0.889	0.876	0.897	0.876
Alex Rodriguez	22	2784	10566	2021	3115	548	696	2086	1338	0.295	12207	5813	0.550	0.380	0.930	0.929	0.925	0.935	0.925
Gary Sheffield	22	2576	9217	1636	2689	467	509	1676	1475	0.292	10947	4737	0.514	0.393	0.907	0.807	0.813	0.816	0.814
Sammy Sosa	18	2354	8813	1475	2408	379	609	1667	929	0.273	9896	4704	0.534	0.344	0.878	0.501	0.515	0.505	0.515
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 4

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Edgar Martinez	18	2055	7213	1219	2247	514	309	1261	1283	0.312	8672	3718	0.515	0.418	0.933	0.527	0.538	0.531	0.538
David Ortiz	20	2408	8640	1419	2472	632	541	1768	1319	0.286	10091	4765	0.552	0.380	0.932	0.561	0.515	0.566	0.515
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 5

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Stan Musial	22	3026	10972	1949	3630	725	475	1951	1599	0.331	12712	6134	0.559	0.417	0.976	0.978	0.978	0.981	0.978
Babe Ruth	22	2503	8398	2174	2873	506	714	2217	2062	0.342	10616	5793	0.690	0.474	1.164	0.997	0.996	0.998	0.996
Ted Williams	19	2292	7706	1798	2654	525	521	1839	2021	0.344	9791	4884	0.634	0.482	1.116	0.988	0.982	0.989	0.983
Barry Bonds	22	2986	9847	2227	2935	601	762	1996	2558	0.298	12606	5976	0.607	0.444	1.051	0.973	0.962	0.976	0.962
Alex Rodriguez	22	2784	10566	2021	3115	548	696	2086	1338	0.295	12207	5813	0.550	0.380	0.930	0.929	0.925	0.935	0.925

Table 6

Harold Baines, Derek Jeter, Edgar Martinez, and Larry Walker have each been elected, either on the 2019 or 2020 ballots[2]. As previously mentioned, the Lahman database is not up to date at the time of this writing, but these players are Hall of Famers nonetheless.

Adrian Beltre, Ichiro Suzuki, and Carlos Beltran have each seen their playing careers end, but have not yet reached the minimum 5 years of retirement to be eligible for the Hall of Fame ballot. Each of these players is compared to the Hall averages in Table 7. Beltre, with his 3000+ hits and strong peripherals prove his candidacy (3000 hits is a common rule-of-thumb benchmark for induction), although his relatively low career batting average, on-base percentage, and walk total hurt his regression score. Suzuki (or Ichiro, as he is known both in the United States and his home nation of Japan) has compiled over 3000 career hits as well, albeit with very low power numbers (home runs, doubles, slugging percentage, runs batted in) and walk total, compared to Hall of Fame contemporaries. Ichiro is a prime example of a beneficiary of the inclusion of pre-live-ball era Hall of Famers, as his high-contact, low-power style of play is reminiscent of the much older style of baseball, and he is contrasted with the great hitters of that time (Ty Cobb, Billy Hamilton, Rogers Hornsby, and Ed Delahanty) in Table 8. Although he can be compares well to these players, Ichiro will ultimately be remembered for the relatively short period of dominance at the

[2] Baines was elected via the Veterans' Committee in 2019.

beginning of his career, during which he posted ten consecutive seasons of more than 200 hits, including one in which he logged 262 – both MLB records. Beltran and Bobby Abreu are very similar players, as Table 10 suggests (although, generally, Beltran was better), and both players will be discussed in the next paragraph.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Adrian Beltre	21	2933	11068	1524	3166	636	477	1707	848	0.286	12130	5309	0.480	0.339	0.819	0.549	0.507	0.565	0.508
Carlos Beltran	20	2586	9768	1582	2725	565	435	1587	1084	0.279	11031	4751	0.486	0.350	0.836	0.439	0.500	0.448	0.500
Ichiro Suzuki	19	2653	9934	1420	3089	362	117	780	647	0.311	10734	3994	0.402	0.355	0.757	0.573	0.480	0.589	0.481
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 7

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Ty Cobb	24	3035	11436	2247	4189	724	117	1944	1249	0.366	13071	5854	0.512	0.433	0.945	0.999	1.000	0.999	1.000
Ed Delahanty	16	1837	7510	1600	2597	522	101	1466	742	0.346	8400	3794	0.505	0.411	0.916	0.949	0.964	0.954	0.965
Billy Hamilton	14	1594	6283	1697	2164	242	40	742	1189	0.344	7608	2716	0.432	0.455	0.887	0.924	0.954	0.930	0.954
Rogers Hornsby	23	2259	8173	1579	2930	541	301	1584	1038	0.358	9475	4712	0.577	0.434	1.011	0.987	0.992	0.988	0.992
Ichiro Suzuki	19	2653	9934	1420	3089	362	117	780	647	0.311	10734	3994	0.402	0.355	0.757	0.573	0.480	0.589	0.481

Table 8

Todd Helton, Omar Vizquel, and Bobby Abreu are still on the ballot without any PED allegations, in their 3rd, 4th, and 2nd years in 2021, respectively. Their career statistics are on display in Table 9. Abreu barely survived his first ballot year in 2020, receiving only 5.5% of the vote. As mentioned, Abreu and Carlos Beltran posted very similar statistics across the life of their careers, with Beltran earning a slight advantage in nearly every category, as seen in Table 10. It is particularly interesting to see these players included in the `HOF_worthy` table, considering the fact that both players contributed significant value to their teams via stolen bases – a metric not considered in the logistic regression. Even without stolen bases, these players' career achievements earned them around 46% and 40% Hall of Fame probability from the models, respectively. Respectable, if unspectacular, hits, runs, and at bats totals with similarly mediocre (by Hall of Fame standards) rate statistics place these two players toward the bottom of this table in terms of Hall of Fame candidacy. Vizquel inches closer to election each year, earning 37.0%, 42.8%, and 52.6% of vote shares in 2018, 2019, and 2020, respectively. Aided by his high hit total and impressive longevity, Vizquel is conventionally known for his defensive prowess. The BBWAA models were far less favorable to his offensive statistics than the overall model, to the tune of 34% versus 43% – largely due to his lackluster on-base percentage and power numbers seen in Table 9. However, his exceptionally long career and solid career hit and walk totals were enough for at least two of the models to consider him an all-time great, and positional analysis would have placed him in the `pos_only` table were these statistics slightly less impressive. Helton faces the same barrier that Larry Walker overcame when Walker earned election to the Hall in 2020: both players played most or all of their careers in notoriously hitter-friendly Denver, Colorado. Denver baseball is played 4,000 feet higher in elevation than any other MLB stadium, leading the BBWAA (and the baseball community in general) to be wary of "inflated" offensive statistics from players who play the majority of their games there. Walker had the advantage in that he demonstrated premier offensive ability in eight of his 18 seasons away from Denver, but in Helton's case, the BBWAA may have a valid case. Figure 3 shows Helton's career splits at home (in Denver) versus on the road, and the results clearly portray Helton as a solid hitter over his career, but hardly a Hall of Fame worthy one considering he was a member of the Colorado Rockies for all of his 17 seasons.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Bobby Abreu	18	2425	8480	1453	2470	574	288	1363	1476	0.291	10081	4026	0.475	0.395	0.870	0.410	0.403	0.415	0.404
Todd Helton	17	2247	7962	1401	2519	592	369	1406	1335	0.316	9450	4292	0.539	0.414	0.953	0.681	0.678	0.692	0.679

Omar Vizquel	24	2968	10586	1445	2877	456	80	951	1028	0.272	12013	3727	0.352	0.336	0.688	0.338	0.430	0.341	0.430
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 9

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Bobby Abreu	18	2425	8480	1453	2470	574	288	1363	1476	0.291	10081	4026	0.475	0.395	0.870	0.410	0.403	0.415	0.404
Carlos Beltran	20	2586	9768	1582	2725	565	435	1587	1084	0.279	11031	4751	0.486	0.350	0.836	0.439	0.500	0.448	0.500
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 10

Split	G	GS	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	TB	GDP	HBP	SH	SF	IBB	ROE	BAbip	tOPS+
Home	1141	1084	4841	4038	874	1394	321	28	227	859	24	15	710	514	.345	.441	.607	1.048	2452	97	29	2	60	99	46	.348	119
Away	1106	1052	4612	3924	527	1125	271	9	142	547	13	14	625	661	.287	.386	.469	.855	1840	89	28	1	33	86	38	.312	80

Figure 3

Finally, the modern players whose lives on the Hall of Fame ballot came and passed: Julio Franco, Fred McGriff, Al Oliver, Dave Parker, and Rusty Staub. These five players are the epitome of baseball's colloquial "Hall of Very Good". Certainly, strong cases can be made for each of these five players, given how they match up against the Hall of Fame means in Table 11. Each player compiled above-Hall average career lengths, at-bats, hits, and plate appearances, with only Franco posting below-average doubles, runs batted in, and total bases – all by narrow margins. Franco, particularly, deserves more credit than his measly 1.1% of the vote share in 2013, his first and only year on the ballot, implies. Franco complied an impressive 23 seasons in Major League Baseball between the ages of 23 and 48 – even more impressive considering he missed the entire 1995 (labor strike), 1998 (Japan), and 2000 (Korea) seasons. His .365 career on-base percentage ranked 4th all-time amongst live-ball era shortstops, and his career hit totals and batting average are better than contemporaries Alan Trammell, Barry Larkin, and Ozzie Smith (Table 12). Each of the four logistic regression algorithms gave him around a 61% Hall of Fame probability, second only to Alex Rodriguez among shortstops not inducted. Considering the positional adjustment, Franco is likely deserving of Hall of Fame induction. Much like the pre-live-ball era players, these players all still have the opportunity to be inducted via the Veterans' Committee.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Julio Franco	23	2527	8677	1285	2586	407	173	1194	917	0.298	9731	3620	0.417	0.365	0.782	0.588	0.625	0.595	0.626
Fred McGriff	19	2460	8757	1349	2490	441	493	1550	1305	0.284	10174	4458	0.509	0.377	0.886	0.517	0.485	0.521	0.486
Al Oliver	18	2368	9049	1189	2743	529	219	1326	535	0.303	9778	4083	0.451	0.344	0.795	0.408	0.342	0.418	0.343
Dave Parker	19	2466	9358	1272	2712	526	339	1493	683	0.29	10184	4405	0.471	0.339	0.810	0.387	0.358	0.394	0.358
Rusty Staub	23	2951	9720	1189	2716	499	292	1466	1255	0.279	11229	4185	0.431	0.362	0.793	0.415	0.445	0.418	0.445
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 11

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Barry Larkin	19	2180	7937	1329	2340	441	198	960	939	0.295	9057	3527	0.444	0.371	0.815	0.238	0.260	0.238	0.260
Ozzie Smith	19	2573	9396	1257	2460	402	28	793	1072	0.262	10778	3084	0.328	0.337	0.665	0.143	0.163	0.139	0.162
Alan Trammell	20	2293	8288	1231	2365	412	185	1003	850	0.285	9375	3442	0.415	0.352	0.767	0.183	0.213	0.182	0.213
Julio Franco	23	2527	8677	1285	2586	407	173	1194	917	0.298	9731	3620	0.417	0.365	0.782	0.588	0.625	0.595	0.626

Table 12

The `false_positives` data frame is where one will find the "exceptions" among Hall of Fame inductees (at least, according to the logistic regression scores). Many of these players received low logistic regression scores largely due solely to the lack of reliable data available. Frank Chance, High Pockets Kelly, Home Run Baker, Johnny Evers, Roger Bresnahan, Buck Ewing, Ray Schalk, Max Carey, Earle Combs, Elmer Flick, Harry Hooper, King Kelly, Tommy McCarthy, Hack Wilson, Ross Youngs, Dave Bancroft, Travis Jackson, Joe Tinker, and John Ward all played most or all of their careers before 1920, and are featured as exceptions mainly due to the inclusion of the pre-1920 Hall of Fame players already discussed. Data from this time, especially pre-1900, can be incomplete, as records of games rely almost entirely on newspaper box scores and even then do not include some of today's commonplace statistics. Jackie Robinson (the first player to break MLB's color barrier in 1947), Larry Doby (the first American League player to break the color barrier, also in 1947), and Roy Campanella[3] each spent many years of their professional careers playing in the Negro Leagues, depriving them of their ability to accumulate the same professional statistics as their white contemporaries[4]. Statistics from these leagues are nearly non-existent, and the few records that do exist are not provided in the Lahman database. Some of the players in this data frame at defense-first positions (shortstop, second base, and catcher) were primarily inducted for just that reason – their defense – including Bill Mazeroski, Luis Aparicio, and others. Omission of these players may have been prudent for a strictly offensive algorithm like this one, but BBWAA voters rarely provide motive or justification for their votes, so determining why a player was elected is effectively impossible. Ralph Kiner (back) and Kirby Puckett (glaucoma) experienced unfortunate and untimely ends to their decidedly dominant baseball careers, with both retiring at relatively young ages (32 and 36, respectively). As each player's career statistics illustrate in Table 13, these players would likely not be included in this data frame (rather, correctly classified as Hall of Famers) had their bodies afforded them the luxury of more professional seasons.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Ralph Kiner	10	1472	5205	971	1451	216	369	1015	1011	0.279	6256	2852	0.548	0.398	0.946	0.084	0.055	0.077	0.055
Kirby Puckett	12	1783	7244	1071	2304	414	207	1085	450	0.318	7831	3453	0.477	0.360	0.837	0.251	0.205	0.254	0.205
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 13

In stark contrast to the PED argument, Orlando Cepeda and Ron Santo were aided rather than hurt by the Hall of Fame's "character clause". Cepeda's and Santo's statistics and logistic regression probabilities point to their likely inclusion in the infamous "Hall of Very Good", but eventually were elected based on their philanthropy, charity, and off-the-field contributions to society. Cepeda's and Santo's career statistics versus the Hall of Fame averages are shown in Table 14. As we can see, both players are extremely close to the Hall averages for all statistics, with neither having enough above-average numbers to sway the algorithm to correctly predict their inductions.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Orlando Cepeda	17	2124	7927	1131	2351	417	379	1365	588	0.297	8695	3959	0.499	0.350	0.849	0.331	0.285	0.333	0.285
Ron Santo	15	2243	8143	1138	2254	365	342	1331	1108	0.277	9396	3779	0.464	0.362	0.826	0.190	0.225	0.187	0.225
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 14

[3] Monte Irvin was initially included in the all_batters table and spent multiple seasons in the Negro Leagues as well, but was the only player removed after adding the 10 year career length filter to the HOF table.

[4] Willie Mays, Hank Aaron, and Ernie Banks each spent some time in the Negro Leagues as well, but the periods were shirt enough that they ultimately were each able to compile the requisite MLB statistics for the logistic regression to correctly classify them.

Still, there are players who have no extenuating circumstances, and simply compiled respectable careers that ultimately fail to match Hall of Fame standards. George Kell's statistics are listed in Table 15.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
George Kell	15	1795	6702	881	2054	385	78	870	621	0.306	7528	2773	0.414	0.367	0.781	0.145	0.150	0.141	0.150
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 15

The `pos_only` designation constitutes the very definition of the previously discussed "Hall of Very Good", but the primary purpose is to outline the players who were among the best offensive producers at their respective positions. As expected, players at traditional defensive-minded positions occupy the top of this table. One[5] of these players has earned election in via the Veterans' Committee, one[6] is still active, three[7] are featured on the 2021 ballot, four[8] have not yet reached the 5-year retirement requirement for ballot eligibility, and the remaining 22 players have seen their BBWAA ballot time come and go. Based on the logistic regression scores and difference between the players' scores and the mean scores at their positions, Table 16 includes the career statistics of Jeff Kent, Mark McGwire, Jason Giambi, Bernie Williams, Johnny Damon, Dwight Evans, Luis Gonzalez, Robinson Cano, Jimmy Rollins, Chase Utley, and Joe Mauer. Coupled with career statistics that are not quite Hall of Fame-caliber, McGwire, Giambi, Gonzalez, and Cano will likely never see election to the Hall due to PED use, assuming current voting trends continue. Jeff Kent is in his 8th season of ballot eligibility, earning his maximum 27.5% of vote shares in 2020 after hovering around 16% for his first 6 ballots. Williams, Damon, and Evans all failed to receive the necessary 5% of votes to stay on the ballot within their first three years of eligibility, owing largely to their generally below Hall of Fame average career statistics and high number of productive offensive players at the outfield position. Utley, Mauer, and Rollins each rank among the top offensive players in the history of their positions, but only Mauer managed to eclipse a .300 career batting average. Out of all 11 players in Table 16, overall career numbers and positional adjustments make Mauer and Kent the most likely Hall of Famers, with Kent having a positional regression score 47% better than the mean second baseman and Mauer 36% better than the mean catcher.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Robinson Cano	15	2185	8502	1234	2570	562	324	1272	607	0.302	9264	4170	0.490	0.352	0.842	0.256	0.190	0.261	0.190
Johnny Damon	18	2490	9736	1668	2769	522	235	1139	1003	0.284	10917	4214	0.433	0.352	0.785	0.299	0.304	0.307	0.304
Dwight Evans	20	2606	8996	1470	2446	483	385	1384	1391	0.272	10569	4230	0.470	0.370	0.840	0.312	0.341	0.312	0.341
Jason Giambi	20	2260	7267	1227	2010	405	440	1441	1366	0.277	8908	3753	0.516	0.399	0.915	0.375	0.343	0.368	0.343
Luis Gonzalez	19	2591	9157	1412	2591	596	354	1439	1155	0.283	10531	4385	0.479	0.367	0.846	0.331	0.318	0.335	0.318
Jeff Kent	17	2298	8498	1320	2461	560	377	1518	801	0.29	9537	4246	0.500	0.356	0.856	0.332	0.324	0.336	0.324
Joe Mauer	15	1858	6930	1018	2123	428	143	923	939	0.306	7960	3040	0.439	0.388	0.827	0.168	0.168	0.166	0.168
Mark McGwire	16	1874	6187	1167	1626	252	583	1414	1317	0.263	7660	3639	0.588	0.394	0.982	0.352	0.297	0.336	0.296
Jimmy Rollins	17	2275	9294	1421	2455	511	231	936	813	0.264	10240	3889	0.418	0.324	0.742	0.046	0.047	0.045	0.047
Chase Utley	16	1937	6857	1103	1885	411	259	1025	724	0.275	7863	3189	0.465	0.358	0.823	0.059	0.047	0.056	0.047
Bernie Williams	16	2076	7869	1366	2336	449	287	1257	1069	0.297	9053	3756	0.477	0.381	0.858	0.370	0.397	0.374	0.397

[5] Ted Simmons, C, inducted 2020
[6] Robinson Cano, 2B
[7] Jeff Kent, 2B, 8th ballot; Scott Rolen, 3B, 4th ballot; Aramis Ramirez, 3B
[8] Victor Martinez, C; Joe Mauer, C; Chase Utley, 2B; Jimmy Rollins, SS

HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 16

Finally, we come to Pete Rose. Rose's career statistics and logistic regression scores in Table 17 paint the picture of a top-10 hitter across baseball history, especially bolstered by his 4,256 career hits, 14,053 career at-bats, and 3,562 career games played – all MLB records – alongside 24 professional seasons and stellar peripheral statistics. He played from 1963 to 1986 so he was well within the live-ball era. He was never suspended or even accused of taking PEDs, either. For each of the features included in the regression model, he ranks close to or above the Hall of Fame average. It was not his play, the time of his career, nor his drug use that disqualifies him for Hall of Fame candidacy, but rather a permanent ban from professional baseball issued to him in 1989 surrounding allegations that he gambled on MLB games during his tenure as manager of the Cincinnati Reds. This conflict of interest may cost one of the greatest hitters in baseball history his opportunity at induction into the Hall of Fame.

names	years	G	AB	R	H	X2B	HR	RBI	BB	BA	PA	TB	SLG	OBP	OPS	m1	m2	m3	m4
Pete Rose	24	3562	14053	2165	4256	746	160	1314	1566	0.303	15861	5752	0.409	0.375	0.784	0.905	0.897	0.917	0.897
HOF Average	18	2161	8008	1338	2419	417	224	1232	913	0.302	9110	3720	0.464	0.376	0.840	0.517	0.515	0.520	0.515

Table 17

Discussion

Ultimately, the logistic regression model is imperfect, based on voting history and public perception. Were a survey to be published, asking a population to judge the classification of the algorithm, some players would almost certainly be deemed by many baseball fans to be under the wrong classification. A potentially significant predictor not included in the model could be a factor describing the status of a player's PED allegations and/or suspensions, especially given that in recent memory, the BBWAA tends to penalize players whose integrity in this regard is in doubt. Additionally, it would be interesting to see the consideration and effect on the model of players with statistical outliers – players who dominated a particular statistic not included in the model, like Sam Crawford's triples (309; an advantage of 118 over the Paul Waner's 191, the highest total of a player who spent his entire career in the live-ball era) or Rickey Henderson's stolen bases (1406; an advantage of 468 over second place Lou Brock, a record that may never be broken).

Wins Above Replacement (WAR; a single composite statistical measure of a player's on-field value), 7-year peak WAR (the greatest 7-year sum of single season WAR totals in a player's career), JAWS (the average of a player's career WAR and 7-year peak), and other Jamesian Hall of Fame metrics negate the need for such linear regression analysis to varying degrees, as Hall of Famers are generally considered using a single threshold value for these numbers. However, the WAR statistics in particular present as intriguing candidates for predictors in a similar algorithm. Although Jamesian statistics can be notoriously difficult to obtain in bulk, the comparison of these metrics to the logistic regression scores shown here may indicate significant relationships, and may indicate differences or overlooked features as well.

A clustering algorithm on a wider set of predictors may also be an interesting and fruitful approach that mirrors Ball James's Similarity Scores. Especially taking the positional issue into account, a clustering approach to grouping players may be a conceptually sensible one, and would allow a direct and straightforward analysis, both visually and computationally.

While imperfect, the logistic regression Hall of Fame scoring algorithm provides a reliably accurate baseline metric for assessing induction candidacy to professional baseball's highest honor, at least from an offensive perspective. Most importantly, positional considerations show that Hall of Fame voting trends at less premier offensive positions like catcher, shortstop, and third base fail to consider candidates' offensive contributions relative to players at their shared position, especially in recent memory. The `HOF_`worthy data frame clearly and effectively lists Hall of Fame-caliber players' offensive careers relative to all hitters who qualify for the Hall of Fame, while `pos_only` does the same with those who were in the top echelon of Hall of Fame inductees at each position – with the unique added feature of each player's performance above the mean player at their position. These final data frames contribute scalable, easy to parse scores and comparisons for players across baseball, giving the necessary positional consideration where it is due. Whether voters consider themselves to be "small Hall" or "large Hall" proponents, the regression scores both overall and positionally provide valid arguments for a large number of players from a wide range of playing styles and eras. I implore all BBWAA voters to make use of these considerations as they continue to hold the fate of the pinnacle of a player's baseball career in their hands.

References

Freindly, Michael, et al. "Lahman." 6 June 2020.
https://cran.r-project.org/web/packages/Lahman/Lahman.pdf

"Leaderboard Glossary." *Baseball*, www.baseball-reference.com/about/leader_glossary.shtml.

"Similarity Scores." *Baseball*, www.baseball-reference.com/about/similarity.shtml.

Alderson, Sandy, et al. "2019 Official Baseball Rules." The Office of the Commissioner of Baseball, 2019.

Cawley, Gavin & Talbot, Nicola. (2008). Efficient approximate leave-one-out cross-validation for kernel logistic regression. Machine Learning. 71. 243-264. 10.1007/s10994-008-5055-9.

Appendix

Appendix 1

The reproduction or transmission of the Official Baseball Rules is prohibited without the express written consent of the Office of the Commissioner of Baseball. However, the Rules are publicly available on Major League Baseball's website here.

Appendix 2

Model	Optimal Cutoff
m1 (BBWAA predictors)	0.37868
m2 (all predictors)	0.39950
m3 (Bayesian, BBWAA)	0.37896
m4 (all predictors)	0.39951

Appendix 3

Model	Misclassification Error
m1	0.94747
m2	0.94651
m3	0.94747
m4	0.94651

Appendix 4

Appendix 4.1

Model	TPR
m1	0.64331
m2	0.61446
m3	0.64331
m4	0.61446

Appendix 4.2

Model	FPR
m1	0.02168
m2	0.02117
m3	0.02168
m4	0.02117

Appendix 5

Model	Lift
m1	9.42025
m2	9.34604
m3	9.42025
m4	9.34604