A Machine Learning Approach to the "Bayes Ball" Hall of Fame

Simon Bertron Mitchell Brady John Rendleman Taylor Smith

The Recursive Bambino Team

ABSTRACT

This paper argues for an objective, machine-learning approach to prediction of Baseball Hall of Fame status based on the Sean Lahman baseball database and FanGraphs statistics. Using a logistic regression model and a neural network classifier, we built models that were able to predict many of the past Hall of Famers while projecting a few current players for future Hall status.

1. INTRODUCTION

Since 1936, the **Baseball Hall of Fame**, located in Cooperstown, New York, has sought to enshrine the players, managers, executives, umpires, and others who have contributed to the fabric of the history of baseball. As of 2019, 328 men and one woman have been inducted to the Hall[3]. Players are inducted into the Hall via election by the Baseball Writers' Association of America (BBWAA) or the Veterans' Committee[1]. Five years after his retirement, any player with at least 10 years of major league experience who passes a screening committee are eligible to be voted in by BBWAA members with more than 10 years' membership who have covered Major League Baseball (MLB) in the 10 years preceding the election. Each voter may name up to 10 candidates; any player who appears on greater than 75% of ballots is elected. Any player who appears on less than 5% of ballots is removed from future ballots but can be considered by the Veterans Committee. Players receiving between 5% and 75% of the votes can be reconsidered annually for a maximum of ten years. It falls upon the Veterans Committee to elect all other candidates, including long-retired players, non-playing personnel, and players who are not inducted by the BBWAA after 10 years of eligibility.

The complexity and elitist nature of this process leads to often endless debate among fans of the game over who is truly worthy of induction into the Hall of Fame[11]. The traditionalist nature of baseball (see appeals to "Unwritten Rules") and its scandals (see the Black Sox Scandal, Pete Rose, steroids, the BALCO Scandal, etc.) further complicate selection criteria. As Craig Edwards writes in his attempt to provide a statistical basis for Hall inclusion, "The Major League Baseball Hall of Fame is a lot like the game itself: wondrous, fascinating and great in scope. The voting process for the Hall of Fame, meanwhile, resembles the umpiring aspect of the game: even though the arbiters typically perform their job well enough, their failures receive considerable attention — nor is it particularly easy to determine who should be in charge of different aspects of gatekeeping"[5]. It is the purpose of this paper to investigate a machine learning, probabilistic approach that can reveal past biases as well as provide a more objective basis for induction. Other statisticians have attempted to contribute to this exact objective[4][7][5], but we believe that we can contribute a new, machine-learning based approach.

2. DATA COLLECTION

We collected data from the widely-used and cited Sean Lahman baseball database[9], which includes data on players from the years 1871-2018. This includes many common statistics like games, hits, runs batted in, pitching wins, and more. It also includes historical Hall of Fame voting data. However, it does not include advanced statistics like Wins Above Replacement (WAR), weighted on-base average (wOBA), fielding independent pitching (FIP), and more. For some of these advanced stats, we downloaded data from FanGraphs, a site which collects statistics on every player in MLB history, has its own calculations for WAR (often cited as fWAR in the literature), and calculates other advanced stats[6][2].

Using these data sources, we aggregated data on all players who are in the Hall of Fame currently (for whom we had statistics) or had at least 10 years of service time (calculated as length between debut and final game). We also threw out players who are currently playing or those who have retired during the last five years. Though these players may be great, they may bias our learning as they cannot be eligible for the Hall.[1] We focused on position players; pitchers are a source of consideration for future work. After taking special care to remove pitchers from the data set, we had full sets of data for $2,983$ players (only 151 of whom are Hall of Famers) to use in our models. After adding a binary variable for whether a player had been implicated in a major scandal[2] and converting batting and fielding hands into binary variables, we ended up with 86 features.[3] Before running the logistic regression and neural network models, we preprocessed the features by standardizing them (making

[1]Note that data was still collected on these players to use for prediction via our logistic regression and neural network models.

[2]These players were: Shoeless Joe Jackson, Pete Rose, Barry Bonds, Mark McGwire, Sammy Sosa, Manny Ramirez, Roger Clemens, Rafael Palmeiro, Gary Sheffield, and Alex Rodriguez.

[3]A full list of these features can be found in the Appendix.

them all 0 mean and unit variance).[4]

2.1 Shortcomings of Data

While we have the benefit of FanGraphs' extensive data set, we were unable to obtain certain advanced stats like FIP and others. Future work should be done to incorporate these as well as other stats that we may have overlooked.

Further, the years 1904 and 1994 are both omitted from our data set due to conflicts occurring in each respective year. In 1904, New York Giants manager John McGraw refused to allow his team to partake in the World Series versus the Boston Americans, a team from the newer, less legitimate American League. In 1994, the entire Major League season was canceled due to player strikes. The lack of this data could slightly skew results but not too much due to the vast amounts of other data.

3. FEATURE SELECTION

As a first step for examining our data before we constructed a logistic regression classifier, we used **scikit-learn**'s **SelectKBest** module for univariate feature selection to examine which 10 features were most correlated with selection to the Baseball Hall of Fame. For our first test, we used **Mutual Information**[5] for a discrete target variable (since our target is a binary classification).[10] This produced 10 features with the distribution according to **Figure 1**.[6] We also used the ANOVA f-value using scikit-learn's **f_classif** module to provide an alternate view of the features. The only difference between these two mechanisms was the exclusion of games fielding to the favor of career triples.[7]

The high correlation of these stats with Hall of Fame inclusion makes sense. A player does not rack up the career necessary to make the Hall if he does not play a lot of games and thus produce a lot of at-bats. A player would not remain in the major leagues if he could not produce offensively, so producing hits, doubles, triples, runs batted in, and walks means that he is getting on base and adding to his team's production. It is interesting that home runs does not appear on this list. However, it is only the modern game that is obsessed with home runs, so it is much more clear that other stats are historically more correlated with success.[8] Finally, the inclusion of Off and WAR are arguments in favor of these statistics; these calculations seek to explain a player's contribution to his team through a single number. That this number predicts Hall of Fame inclusion means that it performs its purpose well.[6]

4. LOGISTIC REGRESSION MODEL

We used scikit-learn's **LogisticRegressionCV** module to construct a logistic regression model for predicting which players will make the hall of fame. This module solves a L2 penalized logistic regression, which minimizes the following

[4]This was accomplished using scikit-learn's **StandardScaler** module.

[5]For this, we used scikit-learn's **mutual_info_classif** module.

[6]These features are: career games played, career at bats, career runs scored, career hits, career doubles, career runs batted in, career walks, Offensive Runs Above Average (Off)[12], WAR, and games fielding.

[7]Note that we could not use the χ^2 stats because our data set included some negative features, including WAR and Off.

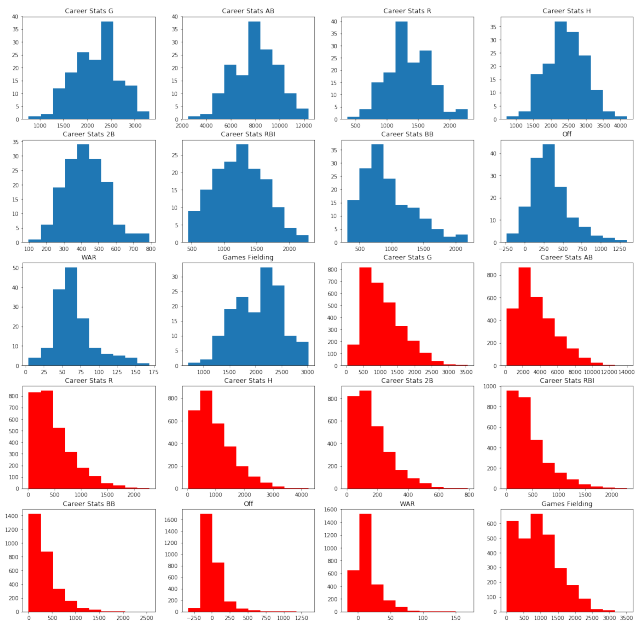

Figure 1: Distribution of 10 highest scoring features according to mutual information among Hall of Famers (Blue) and all players (Red).

cost function:

$$\min_{\beta_0,\beta} \sum_i y_i(\beta_0 + x_i^T\beta) - \log(1 + \exp(\beta_0 + x_i^T\beta)) - \frac{\lambda}{2}\beta^T\beta$$

We used 5-fold cross-validation to tune the hyperparameters and further train the model.

4.1 Model Results: Confusion Matrix

Our resulting model correctly predicted all of the Hall of Famers. However, it did produce 64 false positives[8], which seems like a large number compared to the number of Hall of Famers but in the context of the size of the data set is not as bad. All of these results lead to the confusion matrices in **Figure 2**.

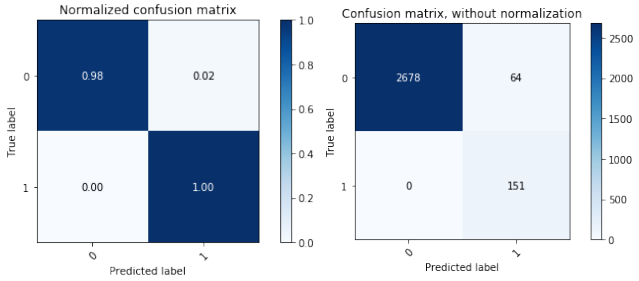

Figure 2: Normalized and non-normalized confusion matrices for L2 penalized logistic regression model.

4.2 Model Results: Scoring

[8]These false positives range from Barry Bonds and Pete Rose to Bob Johnson and Sid Gordon. See the Appendix for the full list of false positives.

Table 1: Classification Metrics for L2 penalized logistic regression model

Metric	Score
Average Precision Score	0.9262
Accuracy Score	0.9779
Balanced Accuracy Score	0.9883
Hinge Loss	0.9699
Matthews Correlation Coefficient	0.8282
ROC AUC	0.9960
F1 Score	0.8251
Hamming Loss	0.0221
Jaccard Similarity Coefficient Score	0.9779
Cross-Entropy Loss	0.7640
Zero-One Loss	0.0221

Table 2: Future Hall of Famers according to the logistic regression model

Name	Probability
Adrian Beltre	0.999906
Carlos Beltran	0.998823
Jason Giambi	0.889027
Derek Jeter	1.000000
Joe Mauer	0.996124
Brian McCann	0.964518
Yadier Molina	0.763639
Buster Posey	0.996227
Albert Pujols	0.980307
Alex Rodriguez	1.000000
Ichiro Suzuki	1.000000
Mike Trout	0.999999

We then explored various scoring metrics for the classification model. See **Table 1** for the results.

4.3 Soon-To-Be Hall of Famers

The logistic regression model anointed 12 players that have either recently retired or are currently playing to be tapped for the Hall of Fame. These players are summarized in **Table 2** together with the probability according to the model that they will make the Hall of Fame.

These names mostly make sense. The only players that are currently playing are Albert Pujols, Brian McCann, Yadier Molina, Buster Posey, and Mike Trout (though Ichiro, Adrian Beltre, and Joe Mauer concluded their careers within the last year). Pujols, McCann, Molina, and Posey are at the tail-end of illustrious careers, and so it is that much more remarkable that the 27 year-old Mike Trout is already tipped to make the Hall of Fame should he retire today. The high probability that Alex Rodriguez (A-Rod) makes the Hall is perhaps debatable due to his admission of performance enhancing drug use. Many speculate that Hall voters will punish A-Rod by making him wait until one of his last years of eligibility before finally letting him make it. The other names include hit merchant Ichiro (one of only 30 players ever to get $3,000$ hits) and Derek Jeter, who also had $3,000$ hits and was a key contributor to the 1990s success of the New York Yankees.

4.4 Discussion

With this model, it seems that there is a trade-off between the number of false positives and false negatives. The model correctly identifies all the Hall of Famers at the cost of producing 64 false positives.

Among these false positives are Barry Bonds and Pete Rose, who would be surefire picks for the Hall if not for their respective scandals.[9] Another advantage of this model is that it identifies Seattle Mariners slugger Edgar Martinez as a Hall of Fame pick; Martinez was selected for the Hall in his 10th and last year of eligibility in 2019. Given that there are so many players even in our reduced data set, it was likely that there were going to be some false positives, especially given the challenge of the induction process.

We believe that the classification metrics in **Table 1** argue in favor of a simple logistic regression model. However, we also explored a neural network classifier.

5. NEURAL NETWORK

We used scikit-learn's **MLPClassifier** module, which implements a Multi-Layer Perceptron classifier and optimizes the log loss, together with the **GridSearchCV** module to search different hyperparameters to train the best model. We explored logistic, ReLU, and hyperbolic tangent activation functions for our neurons, different numbers of hidden layers with different numbers of neurons, and different values of α, from 10^{-5} to 10^{2}, using 5-fold cross validation. Note that under the hood there is a final output layer with a logistic activation function, as our output is a binary variable.

The best model trained had one hidden layer with an amount of neurons twice the number of features, an alpha of 10^{-1}, and hyperbolic tangent activation functions.

5.1 Model Results: Confusion Matrix

Our model correctly predicted all of the Hall of Famers. Further, it produced only three false negatives: Heinie Groh, Stan Hack, and Buddy Myer. All these results lead to the confusion matrices in **Figure 3**.

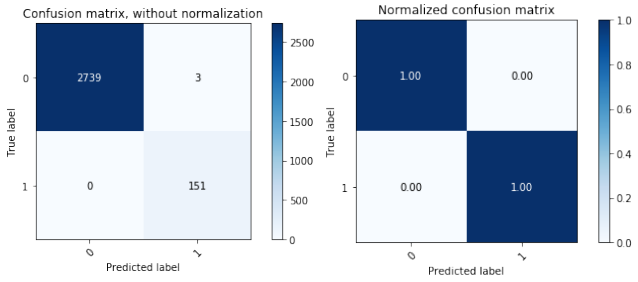

Figure 3: Normalized and non-normalized confusion matrices for L2 penalized logistic regression model.

5.2 Model Results: Scoring

We then explored various scoring metrics for the neural network model. See **Table 3** for the results along with the change from the logistic regression model.

[9]It has long been alleged that Bonds' use of PEDs led to his late-career burst to breaking the career home runs record. Pete Rose, MLB's all-time leader in hits, was banned from baseball after betting on games while he was employed as a player and manager of the Cincinnati Reds.

Table 3: Classification Metrics for neural network model.

Metric	Score	Change from Logistic Regression
Average Precision Score	1.0	↑
Accuracy Score	0.9990	↑
Balanced Accuracy Score	0.9995	↑
Hinge Loss	0.9488	↓
Matthews Correlation Coefficient	0.9897	↑
ROC AUC	1.0	↑
F1 Score	0.9902	↑
Hamming Loss	0.0010	↑
Jaccard Similarity Coefficient Score	0.9990	↑
Cross-Entropy Loss	0.0358	↑
Zero-One Loss	0.0010	↑

Table 4: Future Hall of Famers according to the neural network model

Name	Probability
Adrian Beltre	0.605095
Miguel Cabrera	0.989715
Derek Jeter	0.815430
Brian McCann	0.806608
Buster Posey	0.759265
Albert Pujols	0.952899
Alex Rodriguez	0.886148
Jimmy Rollins	0.994638
Ichiro Suzuki	0.998305
Mike Trout	0.899921
Chase Utley	0.520371

5.3 Soon-To-Be Hall of Famers

The neural network predicted that eleven recently retired or currently playing players would eventually be voted into the Hall of Fame. Interestingly, this list is quite different from the logistic regression list, with former Triple Crown winner Miguel Cabrera, speed-power superstar and former MVP Jimmy Rollins, and 2000s *Sports Illustrated* All-Decade team second-baseman Chase Utley making the Hall to the exclusion of Carlos Beltran, Jason Giambi, Joe Mauer, and Yadier Molina. Some of the probabilities are changed in this model as well. For example, Alex Rodriguez has gone from a surefire Hall of Famer to a less certain candidate. Further, that A-Rod continues to make the Hall is a mystery, as the neural network model excluded false positives with scandals like Pete Rose and Barry Bonds. Further, Mike Trout continues to show his status as a generational talent.

5.4 Discussion

The neural network model seemed to be much more sensitive to changes in the parameters than the logistic regression model. Frustratingly, several test runs of the neural network resulted in every player being mapped to a 0. Eventually, after iterating on different choices of αs, hidden layer sizes, and activation functions, we believe that we settled on choices that led to a better model.

The model is both as good at predicting Hall of Famers and has much less false positives compared to the logistic regression model. Interestingly, the three false positives dominated the early years of baseball, all playing before World War II. Perhaps this leads to the conclusion that future work should be done considering only baseball's modern era for predicting future Hall of Fame members, as these statistics would be much more suited to the modern game.

Though the model correctly identifies all the Hall of Famers, we should be sensitive to the fact that it might be overfitting. Certainly some players who have not made the Hall of Fame might be worthy of inclusion, which could be told to us by the model. Thus, more work should be done that establishes a greater balance between predicting past Hall of Famers and producing worthy criteria for inclusion for the Hall.

6. CONCLUSION AND DISCUSSION

A Hall of Fame model must balance correctly identifying past Hall of Famers, which could be interpreted as confirmation bias in an argument for already inducted players' inclusion, against its status as an objective predictor for future players' selection. It is not the basis of our work to argue for or against already selected players' Hall of Fame status. While perhaps some players whose plaques sit in Cooperstown do not deserve to be there, we believe that using their statistics is the most objective prior that we could find. Future work should be done to examine this hypothesis and examine other objective criteria on which to select players. Selection to the Hall is a challenging and arduous process, and our models identify a few men that may eventually be inducted purely on the basis of their playing statistics.

Our goal was to provide a more objective basis for Hall of Fame selection. Since our model was trained on the existing Hall, it perhaps includes the biases of the Hall as it exists today. Further, our feature for scandals may be an introduction of bias into the model. Future work should be done to examine any potential biases in our model and to further improve the model against bias.

Further, Hall of Fame selection is probably not based on the 50% probability cutoff that these models use. For example, Chase Utley's 52% likelihood of being a Hall of Famer based on the neural network model probably would not tell the average fan that he should be a definite candidate for the Hall. Future work should determine a better cutoff using ROC curves that balances false negatives against the model's use as a predictor for future Hall candidacy.

Other sources for future work building upon these models are work on pitchers, further examination of the differences among different eras of baseball, and exploration of better feature engineering.

APPENDIX

A. FEATURES

Our features are the following: games played, at bats, runs, hits, doubles, triples, home runs, runs batted in, stolen bases, caught stealing, walks, strikeouts, intentional walks, hits by pitch, sacrifice hits, sacrifice flies, times grounded into double plays, weight, height, scandal (defined above), walk percentage, strikeout percentage, Isolated Power (ISO), batting average on balls in play (BABIP), average (AVG),

on-base percentage (OBP), slugging percentage (SLG), wOBA, wRC+, Base Running (BsR), Off, Defensive Runs Above Average (Def), WAR, Triple Crown, MVPs, TSN Guide MVPs, TSN Major League Player of the Year awards, TSN Player of the Year, Rookie of the Year, Babe Ruth Awards, Lou Gehrig Memorial Awards, World Series MVPs, Gold Gloves, TSN Fireman of the Year Awards, All Star Game MVPs, Hutch Awards, Roberto Clemente Awards, NLCS MVPs, ALCS MVPs, Silver Slugger Awards, Branch Rickey Awards, Hank Aaron Awards, Comeback Player of the Year Awards, games fielding, games started fielding, inning outs fielding, putouts, assists, errors, double plays, postseason games fielding, postseason inning outs, postseason putouts, postseason assists, postseason errors, postseason double plays, postseason games, postseason at bats, postseason runs, postseason hits, postseason doubles, postseason triples, postseason home runs, postseason runs batted in, postseason stolen bases, postseason caught stealing, postseason walks, postseason strikeouts, postseason intentional walks, postseason hits by pitch, postseason sacrifice hits, postseason sacrifice flies, postseason times grounded into double plays, batting hand, and throwing hand.

B. LOGISTIC REGRESSION FALSE POSITIVES

The false positives of the logistic regression model are: Moises Alou, Dick Bartell, Buddy Bell, Wally Berger, Barry Bonds, Ken Boyer, Smoky Burgess, Ron Cey, Cupid Childs, Lave Cross, Bill Dahlen, Jake Daubert, Spud Davis, Jim Edmonds, Bob Elliott, Woody English, Art Fletcher, George Foster, Jack Fournier, Augie Galan, Steve Garvey, Jack Glasscock, Luis Gonzalez, Sid Gordon, Marquis Grissom, Heinie Groh, Stan Hack, Gil Hodges, Bob Johnson, Bob Johnson, Fielder Jones, Joe Judge, Billy Jurges, Chuck Knoblauch, Ed Konetchy, Kenny Lofton, Edgar Martinez, Frank McCormick, Dan McGann, John McGraw, Stuffy McInnis, Ed McKean, Bing Miller, Buddy Myer, Tony Oliva, Roger Peckinpaugh, Johnny Pesky, Scott Rolen, Al Rosen, Pete Rose, Wally Schang, Eddie Stanky, Riggs Stephenson, Vern Stephens, Gene Tenace, Bobby Thomson, Joe Torre, Omar Vizquel, Dixie Walker, Larry Walker, Lou Whitaker, Pinky Whitney, Ken Williams, and Maury Wills.

C. REFERENCES

[1] Bbwaa rules for election. https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election/, 2019. Accessed on 2019-05-05.

[2] Fangraphs career batting leaderboard. https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2018&month=0&season1=1871&ind=0&team=0&rost=0&players=0/, 2019. Accessed on 2019-05-05.

[3] Hall of famers. https://baseballhall.org/hall-of-famers/, 2019. Accessed on 2019-05-05.

[4] Hall of stats. http://www.hallofstats.com/, 2019.

[5] C. Edwards. An alternative hall of fame rating system. https://blogs.fangraphs.com/an-alternative-hall-of-fame-rating-system/, January 2016. Accessed on 2019-05-05.

[6] B. Harris. A sabermetric primer: Understanding advanced baseball metrics. https://theathletic.com/255898/2018/02/28/a-sabermetric-primer-understanding-advanced-baseball-metrics/, February 2018. Accessed on 2019-05-05.

[7] J. Jaffe. Jaffe war score system (jaws). https://www.baseball-reference.com/about/jaws.shtml/, 2012.

[8] E. Kelderman. What's with the home-run boom? major league baseball asked this professor to find out. https://www.chronicle.com/article/What-s-With-the-Home-Run/243518/, May 2018. Accessed on 2019-05-05.

[9] S. Lahman. Sean lahman baseball database. http://www.seanlahman.com/baseball-archive/statistics/, 2018. Accessed on 2019-05-05.

[10] E. G. Learned-Miller. Entropy and mutual information. https://people.cs.umass.edu/~elm/Teaching/Docs/mutInf.pdf/, September 2016. Accessed on 2019-05-05.

[11] T. Ringolsby. Ten for the hall. https://www.baseballamerica.com/stories/ringolsby-ten-for-the-hall/, July 2018. Accessed on 2019-05-05.

[12] N. Weinberg. Off. https://library.fangraphs.com/offense/off/, August 2014. Accessed on 2019-05-05.