Predicting MLB Hall of Famers

Name: RJ Burjek, netID: rburjek2

STAT 432 Summer 2024

Introduction

Each year, Major League Baseball recognizes some of the game's greatest players with their induction into the Baseball Hall of Fame. Between recorded statistics and accolades earned, players who excelled above and beyond during their time as a professional baseball player are recognized with the greatest honor: being named a Hall of Famer. In order to be eligible for induction, players must meet a few requirements. The two main requirements are that players must be at least five years removed from their retirement from the league, and they must have played in at least ten Major League Baseball seasons (*BBWAA ELECTION RULES | Baseball Hall of Fame*, n.d.). Once on the ballot, honorary members of the Baseball Writers' Association of America have a yearly vote on who they believe to be worthy of Hall of Fame status (*BBWAA ELECTION RULES | Baseball Hall of Fame*, n.d.). Players who receive votes from at least 75% of voters within a maximum of ten years being on the ballot will ultimately be inducted into the Hall of Fame (*BBWAA ELECTION RULES | Baseball Hall of Fame*, n.d.).

Goal

The purpose of this study is classify which players currently on the Hall of Fame ballot should be inducted and which players should be left out. Looking at players who have completed their Major League careers and comparing their statistics with players who are currently on the ballot, the goal is to predict who may be the next group of players receiving baseball's greatest honor.

Data

The data used in this study is provided by Stathead Baseball and Sports Reference ("Player Batting Season & Career Stats Finder - Baseball," n.d.). It consists of the career statistics of retired Major League Baseball position players and their Hall of Fame status. Therefore, pitchers are not included in this study. A player's Hall of Fame status is denoted by a binary response variable, with "Yes" denoting they are currently in the Hall of Fame and "No" denoting that they did not get inducted after their retirement. The rest of the predictors in the data set are numeric variables that are both totals, like hits and home runs accumulated, or rates, like batting average or on base percentage.

Players included in the data are restricted to those who played in at least 1,250 games throughout their career. The data is also restricted to players who debuted in or after the 1939 Major League Baseball season, as many advanced baseball statistics did not begin to be recorded until around this time.

Methods

Before beginning the model selection and prediction analysis, the data was read into R Studio and any necessary modifications were made, like omitting any N/A values present in the data. The variable "HoF"

was also transformed into a factor variable instead of a character variable. Lastly, the players that are currently on the ballot were removed from our training and testing sets. Because those players have not received a definitive yes or no in terms of their Hall of Fame status, it would not be fair to have their data considered in our model calculations as if they have been denied to the Hall of Fame. Those player's statistics may be worthy of Hall of Fame selection, so having their numbers count towards the "No" category could lead to incorrect prediction. That set of players currently on the ballot, however, will be used at the end of the study to predict their future status as Hall of Famers.

While the 2025 Baseball Hall of Fame ballot is not yet finalized, all of the names used in the "onBallot" list are eligible for the ballot and have the opportunity to appear on it for the 2025 election. The list of players is as follows ("2025 Potential Hall of Fame Ballot," n.d.):

```
#>  [1] "Andruw Jones"      "Carlos Beltrán"    "Álex Rodríguez"
#>  [4] "Manny Ramírez"     "Chase Utley"       "Omar Vizquel"
#>  [7] "Bobby Abreu"       "Jimmy Rollins"     "Torii Hunter"
#> [10] "David Wright"      "Ichiro Suzuki"     "Dustin Pedroia"
#> [13] "Ian Kinsler"       "Troy Tulowitzki"   "Ben Zobrist"
#> [16] "Curtis Granderson" "Hanley Ramírez"    "Russell Martin"
#> [19] "Adam Jones"        "Brian McCann"      "Martín Prado"
#> [22] "Carlos González"   "Melky Cabrera"     "Ian Desmond"
#> [25] "Kendrys Morales"   "Mark Reynolds"
```

Logistic Regression

The first method of choice was to build a logistic regression model to predict the probability of a player getting voted into the Hall of Fame. When it comes to Hall of Fame baseball players, they are not all built the same. Some are inducted for their elite performance on offense, and some are recognized for their world-class defense. In all, different players excel in different areas, and logistic regression is a good way to quantify those numeric statistics and be able to estimate a player's induction status when some of their statistics are at or below league average, yet others are some of the best numbers the game of baseball has ever seen.

To begin, both a null model will be built along with a full model using all the numeric predictors to estimate our response variable. Once again, the response variable in question is the binomial factor variable "HoF" containing a categorical response of either "Yes" or "No". The original data was randomly split into training and testing data sets with 80% of the data points being put into the training set and 20% being put into the testing set, and then another 80%-20% split was made on the training set into estimation and validation data sets. The null and full models were built using the estimation data set.

Next, a logistic regression model was built using backward selection to select the feature variables that would be included in the model.

```
#>
#> Call:
#> glm(formula = HoF ~ YearsPlayed + AgeStart + G + PA + AB + R +
#>     H + `1B` + `2B` + RBI + SB + CS + SO + BA + SLG + OPS + `OPS+` +
#>     HBP + SH + SF + IBB + WAR + WAA + oWAR + dWAR + Rbat + TOBwe +
#>     RC + BAbip + BtWins + ISO + `AB/SO` + `SB%` + SBatt, family = "binomial",
#>     data = offBallot.est %>% select(-Player, -From, -To))
#>
#> Coefficients:
#>               Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -420892.54 1297527.38  -0.324    0.746
#> YearsPlayed     685.23    2127.80   0.322    0.747
```

```
#> AgeStart        279.02     865.16   0.323    0.747
#> G               -76.04     236.45  -0.322    0.748
#> PA             -230.15     709.72  -0.324    0.746
#> AB              242.73     748.23   0.324    0.746
#> R              -108.72     337.84  -0.322    0.748
#> H              -221.46     682.36  -0.325    0.746
#> `1B`            -68.05     211.92  -0.321    0.748
#> `2B`            -74.27     229.94  -0.323    0.747
#> RBI              11.58      36.01   0.322    0.748
#> SB              647.57    6698.36   0.097    0.923
#> CS              455.03    6546.09   0.070    0.945
#> SO               36.53     113.35   0.322    0.747
#> BA          2075626.14 6379050.66   0.325    0.745
#> SLG         -932884.94 2851389.22  -0.327    0.744
#> OPS           81029.91  260176.97   0.311    0.755
#> `OPS+`          916.91    2835.19   0.323    0.746
#> HBP              42.86     132.72   0.323    0.747
#> SH              288.21     889.62   0.324    0.746
#> SF              202.04     622.74   0.324    0.746
#> IBB             -64.76     200.20  -0.323    0.746
#> WAR            1555.41    4804.16   0.324    0.746
#> WAA           -1271.65    3904.19  -0.326    0.745
#> oWAR            166.76     515.96   0.323    0.747
#> dWAR            241.44     758.02   0.319    0.750
#> Rbat           -383.70    1190.30  -0.322    0.747
#> TOBwe           286.27     885.31   0.323    0.746
#> RC               42.30     129.97   0.325    0.745
#> BAbip       -402957.46 1251012.36  -0.322    0.747
#> BtWins         1716.41    5305.46   0.324    0.746
#> ISO          803272.68 2462831.13   0.326    0.744
#> `AB/SO`        1698.08    5283.22   0.321    0.748
#> `SB%`            47.86     151.26   0.316    0.752
#> SBatt          -585.56    6643.79  -0.088    0.930
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 3.7063e+02  on 498  degrees of freedom
#> Residual deviance: 6.9841e-04  on 464  degrees of freedom
#> AIC: 70.001
#>
#> Number of Fisher Scoring iterations: 25
```

After running the backwards selections, the feature variables that were included in the logistic regression model as are listed above, and the model features an AIC value of 70.001.

Now, using the null model, forward selection is used to build another logisistic model.

```
#>
#> Call:
#> glm(formula = HoF ~ WAR + `AB/SO` + WAA + SLG + XBH + BtRuns +
#>     `OPS+` + OBP + RBI + BAbip + BtWins + AgeStart, family = "binomial",
#>     data = offBallot.est %>% select(-Player, -From, -To))
#>
#> Coefficients:
```

```
#>               Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -1.204e+02  2.328e+01  -5.174 2.29e-07 ***
#> WAR          6.098e-01  1.439e-01   4.238 2.25e-05 ***
#> `AB/SO`      2.662e-01  1.086e-01   2.451 0.014246 *
#> WAA         -4.657e-01  1.365e-01  -3.412 0.000644 ***
#> SLG          3.482e+01  2.450e+01   1.421 0.155249
#> XBH         -9.311e-03  6.919e-03  -1.346 0.178402
#> BtRuns      -1.079e-01  3.087e-02  -3.494 0.000475 ***
#> `OPS+`       3.512e-01  1.350e-01   2.603 0.009252 **
#> OBP          1.040e+02  3.115e+01   3.339 0.000841 ***
#> RBI          8.986e-03  3.210e-03   2.800 0.005116 **
#> BAbip        4.937e+01  1.849e+01   2.670 0.007579 **
#> BtWins       6.078e-01  2.974e-01   2.044 0.040972 *
#> AgeStart    -3.379e-01  1.936e-01  -1.746 0.080893 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 370.631  on 498  degrees of freedom
#> Residual deviance:  95.593  on 486  degrees of freedom
#> AIC: 121.59
#>
#> Number of Fisher Scoring iterations: 9
```

The forward model that is built results in a model with slightly different feature variables, including the variables WAR, AB/SO, WAA, SLG, XBH, BtRuns, OPS+, OBP, RBI, BAbip, BtWins, and AgeStart. Despite having more significant variables than the backwards model, the forward model displays a higher AIC with a value of 121.59

Now that both the models are built, predictions will be made on them using the validation data set and the misclassification error will be calculated for both.

```
#>   StepDirection      AIC MisclassificationRate
#> 1      Backward  70.0007            0.08064516
#> 2       Forward 121.5932            0.03225806
```

From the table, it can be seen the backwards selection resulted in a model with a lower AIC value, however the forward model resulted in a lower misclassification rate.

Lastly, the model sensitivity, specificity, and accuracy will be calculated for each model.

```
#>          actual
#> predicted  No Yes
#>       No  112   4
#>       Yes   6   2
```

```
#>          actual
#> predicted  No Yes
#>       No  115   1
#>       Yes   3   5
```

```
#>      Model Sensitivity Specificity  Accuracy
#> 1 Backward   0.3333333   0.9491525 0.9193548
#> 2  Forward   0.8333333   0.9745763 0.9677419
```

For these values, the forward selection model also displayed better results, having a higher sensitivity, specificity, and accuracy than the backwards model.

k-NN

After building two logistic models with forward and backward feature variable selection, the next method of choice was to build a k-nearest neighbors model. This kind of modeling is great for comparative classification, which is what a bulk of voters take into consideration when voting for Hall of Famers. If a player performed very similarly to another Hall of Famer throughout their career, then it would make sense that they be inducted into the Hall of Fame as well.

For this method, the data in our model was initially normalized, and a k-NN model was trained using the estimation data set. Using 10-fold cross-validation, a best k value of k=7 was calculated, and from there, the k-NN model was retrained using the entire training data set.

```
#>      Model Sensitivity Specificity  Accuracy
#> 1 Backward   0.3333333   0.9491525 0.9193548
#> 2  Forward   0.8333333   0.9745763 0.9677419
#> 3     k-NN   0.6667000   0.9478000 0.9097000
```

Similarly to the logistic models, the model sensitivity, specificity, and accuracy for the k-NN model was calculated. However, the k-NN model displayed lower values in all three of these calculations, and therefore failed to perform as well as the forward selection model.

Final Prediction

After deciding that the logistic regression using backward selection led to the best and most accurate classification, the model was refit to the entire training data set, and our final predictions of future Hall of Famers was made.

```
#>            Player HoF
#> 2  Álex Rodríguez Yes
#> 3    Andruw Jones Yes
#> 5     Bobby Abreu Yes
#> 7  Carlos Beltrán Yes
#> 16  Ichiro Suzuki Yes
#> 23   Omar Vizquel Yes
```

With a final model misclassification rate of 0.032, Álex Rodríguez, Andruw Jones, Bobby Abreu, Carlos Beltrán, Ichiro Suzuki, and Omar Vizquel are the six players currently on ballot that are classified as being worthy of Baseball Hall of Fame status.

Conclusion

To recap, using career statistics of former Major League Baseball players both in the Hall of Fame and not in the Hall of Fame, two logistic regression models along with a k-nearest neighbors model were all built with the goal of classifying which former players would be worth of Hall of Fame status. After determining which model performed the best, predictions were made on that model using the players currently on the Hall of Fame ballot, and Álex Rodríguez, Andruw Jones, Bobby Abreu, Carlos Beltrán, Ichiro Suzuki, and Omar Vizquel were all classified as Hall of Famers.

References

2025 Potential Hall of Fame Ballot. (n.d.). In *Baseball-Reference.com*. Retrieved July 30, 2024, from https://www.baseball-reference.com/awards/hof_2025.shtml

BBWAA ELECTION RULES | Baseball Hall of Fame. (n.d.). Retrieved July 30, 2024, from https://baseballhall.org/hall-of-fame/election-rules/bbwaa-rules

Player Batting Season & Career Stats Finder - Baseball. (n.d.). In *Stathead.com*. Retrieved July 30, 2024, from https://stathead.com/baseball/player-batting-season-finder.cgi