Predicting Major League Baseball Hall of Fame Status for Pitchers

Casey Delaney

04/16/2023

Table of Contents

Abstract

The objective of this paper is to build an accurate classification model that can predict whether or not a Major League Baseball pitcher was inducted into the Hall of Fame based on career statistics including games played, wins, and career accolades. Classification methods used include logistic regression, random forests, k-nearest numbers, and Naive Bayes. Features and parameters will be manipulated to find optimal outcomes. The goal of the research was to determine which statistics can influence the Hall of Fame status of a pitcher, and whether or not career awards and All Star Game appearances have an effect on this outcome.

From the research, it was concluded that statistics such as wins, innings pitched, and shutouts have greater significance to a pitcher’s Hall of Fame status than statistics including career earned runs allowed, home runs given up, and hits allowed. While players in the Hall of Fame have more awards and All Star Game appearances on average, these were not indicative of induction status. The random forest model proved to be the most accurate model created, whereas Naive Bayes was the least effective.

1. Introduction

The Major League Baseball Hall of Fame is a storied and exclusive collection of the greatest, honored, and infamous individuals to ever have represented Major League Baseball. Located in Cooperstown, New York, the museum represents the game's finest players, managers, executives, and umpires throughout its history. Players can only be voted in five years after they retire from the sport, and it is considered to be the greatest achievement of one's career. This study will consist of an analysis that attempts to predict whether or not a player will be voted in based on their career statistics. Statistics include both in game results as well as awards received, such as Most Valuable Player and All Star Games played in.

Questions that will be researched are whether there are significant statistics that cause a player to be voted into the Hall of Fame, if there is a direct correlation between All Star Game appearances and Hall of Fame status, and if there is a threshold of a certain amount of career accolades a player must receive to be voted in. As a side note, every year the MLB hosts an All Star Game in the middle of the season at a different stadium, players earn the right to play in the All Star Game via a fan vote. The majority of the time, the best players thus far that year are voted in. However, sometimes a struggling player who has a history of prominence is voted in. These variables will be tested through a feature selection process, and the models used to analyze the data will be logistic regression, random forest, k-nearest numbers, and Naive Bayes. Cross-validation will be used to avoid overfitting the data, as only ~1.15% of MLB pitchers have been voted into the Hall of Fame. Models will be fitted to a subset of training data, parameters and features will be selected and tuned, and then the final models will be tested on a testing subset to gauge the final accuracy.

The research will conclude with an explanation of the findings, and what was discovered in relation to the primary research questions. An analysis of how the information obtained can assist modern day baseball players and researchers will also be discussed.

2. Data Source and Explanations

The data used for this study comes from the Lahman Baseball Database, which is a baseball database that was created and is updated annually by a baseball journalist/researcher named Sean Lahman (Lahman, 2022). Lahman pioneered an effort to make baseball statistics freely available to the general public. A lengthy and highly detailed description of the database and its contents is featured on his site as a "read me" text file (Lahman, 2022). Descriptions of the variables used in this study will be given later on.

Data consists of Hall of Fame selections from 1936 until 2022, as well other baseball related statistics dating back to the late 1800s. The analysis will be driven primarily by players who were inducted as pitchers; the data used will be pitcher specific data (the pitcher is the player who "pitches", or throws, the ball to the hitter; think of a bowler in cricket). Only regular season statistics will be analyzed, playoff statistics will not be included.

3. Data Cleaning

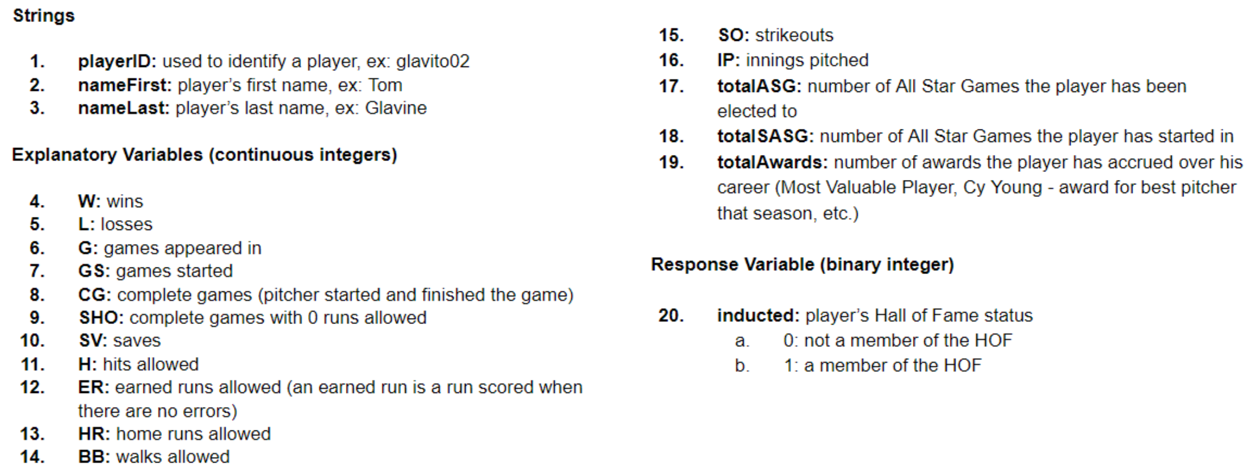
The initial data came in the form of CSV files. Five files were chosen that contained the correct information needed to perform this experiment. Each CSV was converted to a table, and these tables were cleaned before performing the analysis.

Various columns in each table were removed due to the variables being insignificant for the analysis. Variables such as lgID (league name) and teamID (team name) were removed because teams, divisions, and league names have changed many times throughout MLB history, and they are unimportant to a statistical analysis of a player. Furthermore, certain pitching statistics such as BK (balks) and HBP (hit by pitch), are generally trivial stats that didn't start being recorded until the late 1900s, so many players have “zero”. In the Hall of Fame dataframe, inductees were sorted so that it just contained players, and not managers, umpires, and pioneers/executives.

Each dataframe lists each player's statistics by year. For example, if someone played for 10 straight seasons, there would be 10 rows of data for him. To fix this and get the totals, the dataframes were grouped by the playerID so that the columns could be totalled and later be merged with others.

The “inducted” variable was manipulated so that it would show a 0 if a player has not been inducted, and a 1 if the player was inducted. Lastly, pitchers who appeared in less than 10 games were removed. This is because many players who are primarily hitters can be called upon to pitch once or twice throughout their careers in a lopsided game so that the team can rest their pitchers. This would lead to the data being skewed, so these players were removed (Figure 1).

Below is a sample of the variables that were used after these steps were taken. Strings were used just to identify players, and will not be included in the models:



Strings

1. **playerID:** used to identify a player, ex: glavito02
2. **nameFirst:** player's first name, ex: Tom
3. **nameLast:** player's last name, ex: Glavine

Explanatory Variables (continuous integers)

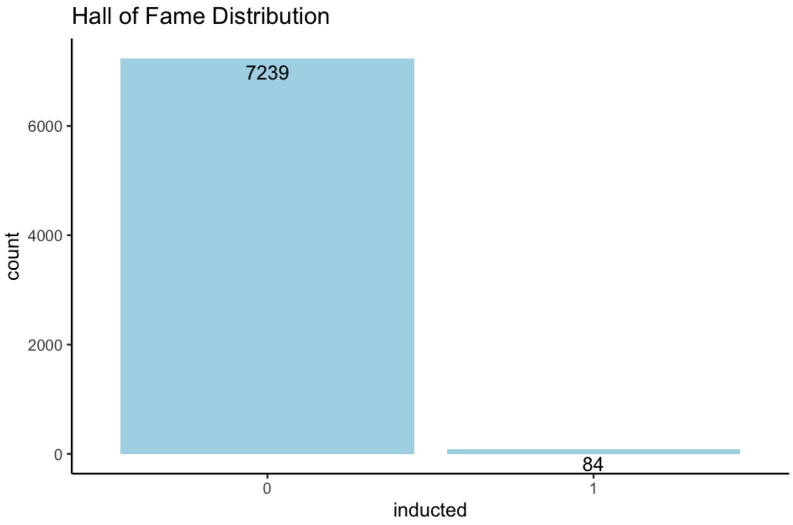
4. **W:** wins
5. **L:** losses
6. **G:** games appeared in
7. **GS:** games started
8. **CG:** complete games (pitcher started and finished the game)
9. **SHO:** complete games with 0 runs allowed
10. **SV:** saves
11. **H:** hits allowed
12. **ER:** earned runs allowed (an earned run is a run scored when there are no errors)
13. **HR:** home runs allowed
14. **BB:** walks allowed
15. **SO:** strikeouts
16. **IP:** innings pitched
17. **totalASG:** number of All Star Games the player has been elected to
18. **totalSASG:** number of All Star Games the player has started in
19. **totalAwards:** number of awards the player has accrued over his career (Most Valuable Player, Cy Young - award for best pitcher that season, etc.)

Response Variable (binary integer)

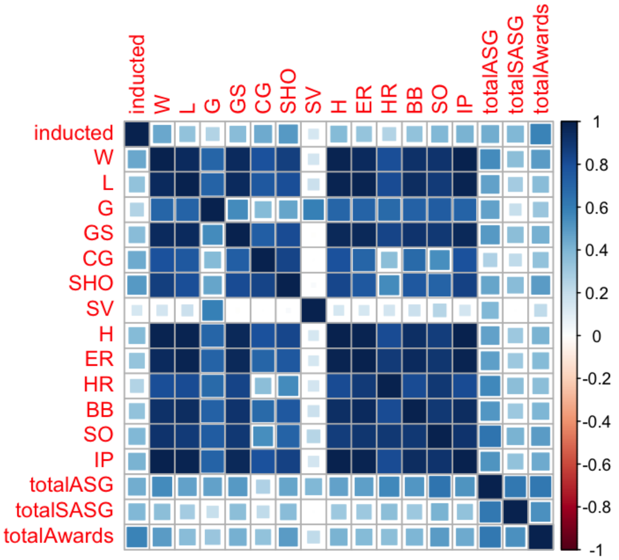
20. **inducted:** player's Hall of Fame status
    a. 0: not a member of the HOF
    b. 1: a member of the HOF

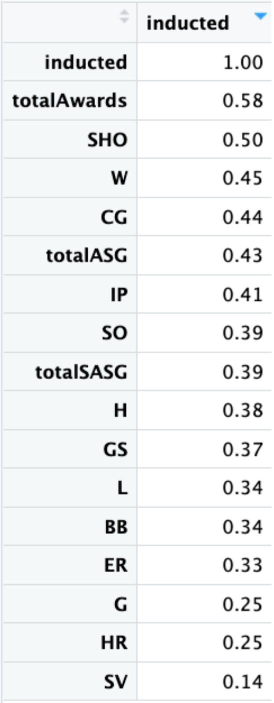
4. Exploratory Data Analysis

After cleaning the data, it was examined in order to determine a few key statistics. To start, it was discovered that of the selected data, only 1.15% of observations were Hall of Fame inductees:



Correlations were computed, mapped out on a heatmap, and also plotted on a table. It was discovered that the five variables with the highest correlations to Hall of Fame status were total awards, shutouts, wins, complete games, and All Star Game appearances:





	inducted
inducted	1.00
totalAwards	0.58
SHO	0.50
W	0.45
CG	0.44
totalASG	0.43
IP	0.41
SO	0.39
totalSASG	0.39
H	0.38
GS	0.37
L	0.34
BB	0.34
ER	0.33
G	0.25
HR	0.25
SV	0.14

Next, a summary of each category (inducted into the Hall of Fame vs. not inducted into the Hall of Fame) based on total awards and All Star Games was computed. This was to note if there were any distinctions between the two categories. Tables representing these summaries can be seen below:

<table>
<tr><th colspan="6"><b>Awards Earned Over a Player’s Career
(Pitchers Inducted into the Hall of Fame)</b></th></tr>
<tr><td>Minimum</td><td>1st Quarter</td><td>Median</td><td>Mean</td><td>3rd Quarter</td><td>Maximum</td></tr>
<tr><td>0.00</td><td>2.00</td><td>5.00</td><td>8.17</td><td>12.25</td><td>33.00</td></tr>
</table>

<table>
<tr><th colspan="6"><b>All Star Game Participations Over a Player’s Career
(Pitchers Inducted into the Hall of Fame)</b></th></tr>
<tr><td>Minimum</td><td>1st Quarter</td><td>Median</td><td>Mean</td><td>3rd Quarter</td><td>Maximum</td></tr>
<tr><td>0.00</td><td>0.00</td><td>3.50</td><td>4.07</td><td>8.00</td><td>16.00</td></tr>
</table>

The average Hall of Fame member who was inducted as a pitcher played in about 4 All Star Games and earned more than 8 total awards over his career.

<table>
<tr><th colspan="6"><b>Awards Earned Over a Player’s Career
(Pitchers Not Inducted into the Hall of Fame)</b></th></tr>
<tr><td>Minimum</td><td>1st Quarter</td><td>Median</td><td>Mean</td><td>3rd Quarter</td><td>Maximum</td></tr>
<tr><td>0.00</td><td>0.00</td><td>0.00</td><td>0.15</td><td>0.00</td><td>22.00</td></tr>
</table>

<table>
<tr><th colspan="6"><b>All Star Game Participations Over a Player’s Career
(Pitchers Not Inducted into the Hall of Fame)</b></th></tr>
<tr><td>Minimum</td><td>1st Quarter</td><td>Median</td><td>Mean</td><td>3rd Quarter</td><td>Maximum</td></tr>
<tr><td>0.00</td><td>0.00</td><td>0.00</td><td>0.20</td><td>0.00</td><td>11.00</td></tr>
</table>

For pitchers who did not make the Hall of Fame, they average less than one All Star Game appearance and zero awards. The maximum values are high due to outliers, such as Roger Clemens, who will never be voted into the Hall as it was discovered he played while using steroids.

These four tables illustrate that the typical Hall of Fame inductee who earns their spot as a pitcher will likely have more All Star Game appearances and career accolades than those players who do not.

5. Methodology

5.1. Splitting the Data

Before the data can be analyzed, it must be split into a training set and testing set. The training set contains 80% of the total data, and the testing set contains 20%. To determine which observations would be included in each subset, a random sample of 1,464 integers, uniformly distributed, was created; 20% of the 7,323 total observations is 1,464.6. The row numbers for the testing set were taken from this distribution, and the ones not included were used for the training set. The seed 179 was set to ensure the experiment could be tweaked and run in the future, while having the same training and testing sets.

5.2. Models Selected, Cross-Validation, Feature Selection

Feature selection was done after splitting the data into training and testing subsets. It was done in this order to avoid the leaking of information from the testing set into the training pipeline. Furthermore, the feature selection was done during each fold of cross-validation because if it was done prior to cross-validation, significant bias and overfitting can occur. This was very important due to the small number of observations in the “1” class of the dependent variable, inducted.

5.2.1. Logistic Regression

A logistic regression model returns the probability of a label and compares it to a predefined threshold (Edgar, 2017). The comparison of the probability and the threshold determines the classification. This model was chosen due to it being one of the more simple machine learning models to deploy on a data set, making it a reasonable choice to start with. The logistic regression model deployed will be cross-validated with 10 folds, and features will be selected by analyzing the significance of each predictor during training. A final model to be used on the testing data will be created using just those variables that were significant.

5.2.2. Random Forest

A random forest model is a collection of decision trees where each tree is trained on a random sampling of the data and a majority vote is taken in order to determine the number of variables used that gives the highest model accuracy when classifying (Liberman, 2017). The random forest model deployed will be cross-validated with 5 folds, and features will be selected by analyzing the variable importance. A final model to be used on the testing data will be created using just those variables that were significant.

5.2.3. K-Nearest Numbers (KNN)

The k-nearest numbers (KNN) algorithm uses proximity of observations to others in order to make predictions about the classification of an observation (Harrison, 2017). It is a simple and versatile method of both regression and classification. The final number of k groups will be determined by using a cross-validation technique with 10 folds, in order to determine which value of k gives the highest accuracy. The optimal k will be used on the testing set.

5.2.4. Naive Bayes

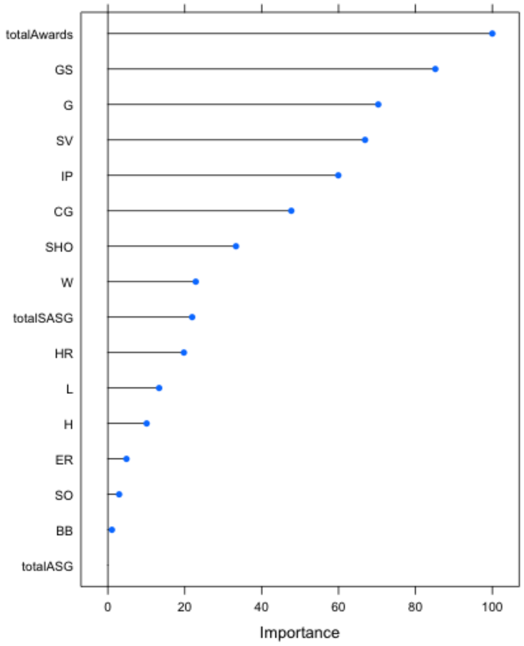
The Naives Bayes classifier calculates the conditional probability of a class based on prior knowledge gained during training. However, a key note is that the Naive Bayes classifier assumes that each observation is independent of one another (Brownlee, 2019). Naive Bayes was chosen because it would be interesting to see if there was a noticeable change in accuracy due to the fact that each observation would be deemed as independent. The Naive Bayes model deployed will be cross-validated with 10 folds, and features will be selected by analyzing the variable importance. A final model to be used on the testing data will be created using just those variables that were significant.

6. Results

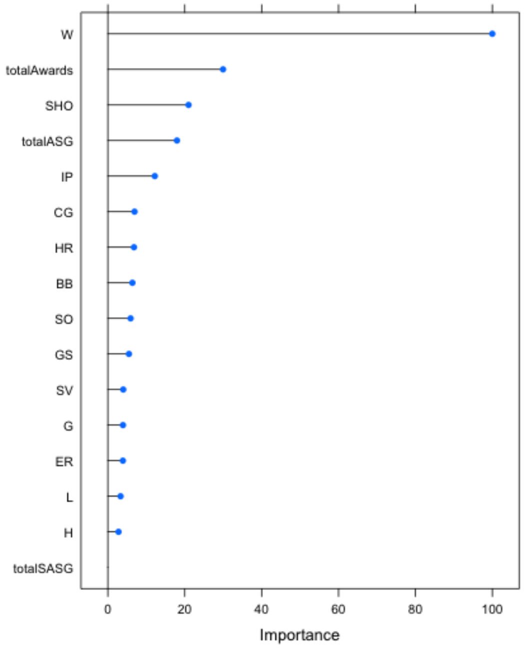
Before analyzing the training and testing errors, as well as answering the proposed research questions, feature selection and optimal parameters must be discussed.

6.1. Feature Selection and Optimal Parameters

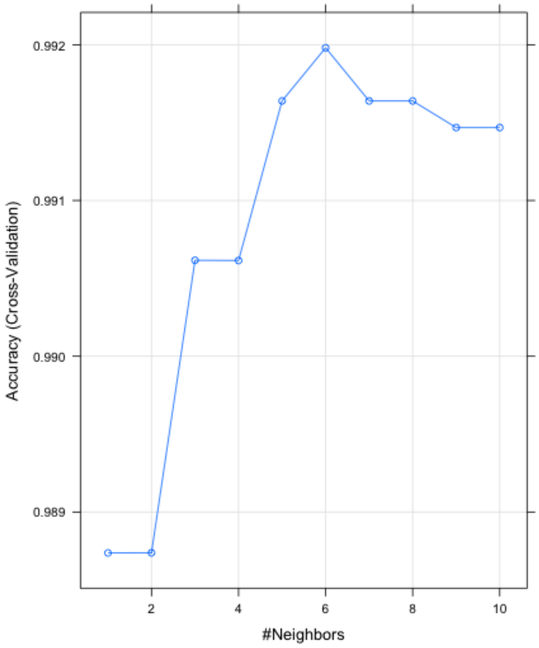
The predictors that were most significant and furthermore selected for the final logistic regression model were G (games), GS (games started), CG (complete games), SV (saves), IP (innings pitched), and totalAwards. This makes sense, as the first five selected indicate that a pitcher has had a long career, and the only way a pitcher has a long career is if he is effective and excelling. The last predictor selected, totalAwards, can be validated as well. The more awards a pitcher receives over his career indicates that not only did he have a long career, but he was recognized for his outstanding achievements, meaning there is a higher probability that he would be voted into the Hall of Fame. Figure 2 in the appendix gives the numerical values.



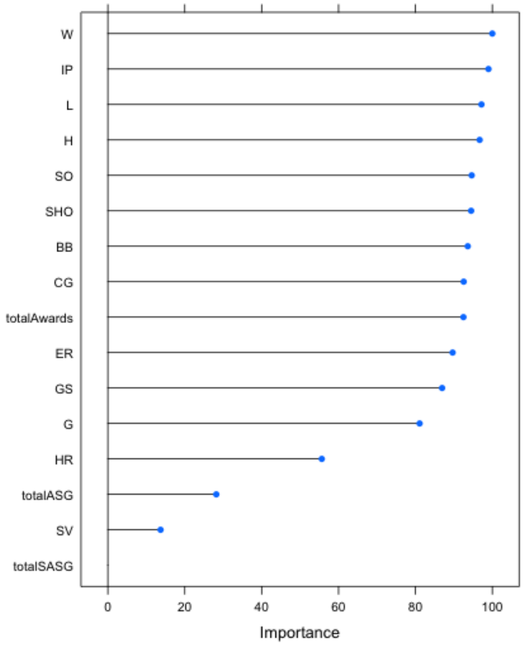
The 5 fold cross validation for the random forest model determined that in terms of variable importance, W (wins), totalAwards, SHO (shutouts), totalASG, and IP (innings pitched) were the 5 most important variables (Figure 3.



The final k value used for the KNN model was k = 6. K values of 1 through 10 were tested, and k = 6 provided the highest accuracy and Kappa (Figure 4).



The only predictors that were unimportant to the Naive Bayes model were totalSASG, SV, totalASG, and HR (home runs allowed). This was interesting as the initial hypothesis that All Star Games were important does not seem to ring true. The most important variable was W (wins). This model seemed to rank playing statistics as being more important than accolades (Figure 5).



6.2. Training Errors Using Optimal Parameters and Predictors

Logistic Regression	Random Forest	KNN k = 6	Naive Bayes
0.00597	**0.00051**	0.00802	0.04694

6.3. Testing Errors Using Optimal Parameters and Predictors

Logistic Regression	Random Forest	KNN k = 6	Naive Bayes
0.00615	**0.00478**	0.01366	0.04850

7. Conclusion

Upon completion of the analysis, it was discovered that the random forest model was the most accurate of the models tested for predicting a Major League Baseball pitcher's Hall of Fame status, when compared to logistic regression, k-nearest number, and Naive Bayes. The random forest model is the most complex of the four models, however this does not necessarily mean that it will be the most accurate classification method on any random data set. The Naive Bayes model was the least accurate model on both the training and the testing subsets. This could be due to the fact that it deems all observations as being independent. The logistic regression model had the smallest difference in accuracy between the training and testing set, indicating that it was the least prone to overfitting. This is interesting because random forest is known as being highly effective in eliminating overfitting, so one could predict that it would have had the smallest difference.

7.1. Scientific Research Questions Answered

The first question proposed in this study was to learn which pitcher specific statistics were most vital in predicting Hall of Fame status. Total wins, shutouts, and innings pitched were three predictors that consistently showed up when checking variable importance for each model. A larger amount of innings pitched illustrates the longevity of a pitcher's career. Pitchers who are able to stay in the Major Leagues for longer periods of time are ones who are dominant. Wins and shutouts are an example of a pitcher's success, particularly shutouts. A pitcher can receive a win even if he doesn't pitch well, if his team scores a lot of runs he gets the win anyway. This doesn't mean it should be downplayed, it just highlights the shutouts category more. A shutout is earned if a pitcher completes the entire game without surrendering any runs, another example of a pitcher's great performance. It can be stated that these three variables were significant in predicting Hall of Fame status.

The next question was whether All Star Game appearances had a correlation between Hall of Fame status. It was determined initially that there was a 0.43 positive correlation. Additionally, when comparing pitchers who were voted into the Hall of Fame and those who weren't, pitchers who were voted in had on average 3.87 more All Star Game appearances. However, logistic regression and Naive Bayes ranked totalASG low in predictor importance, which was interesting to see given the previous observations.

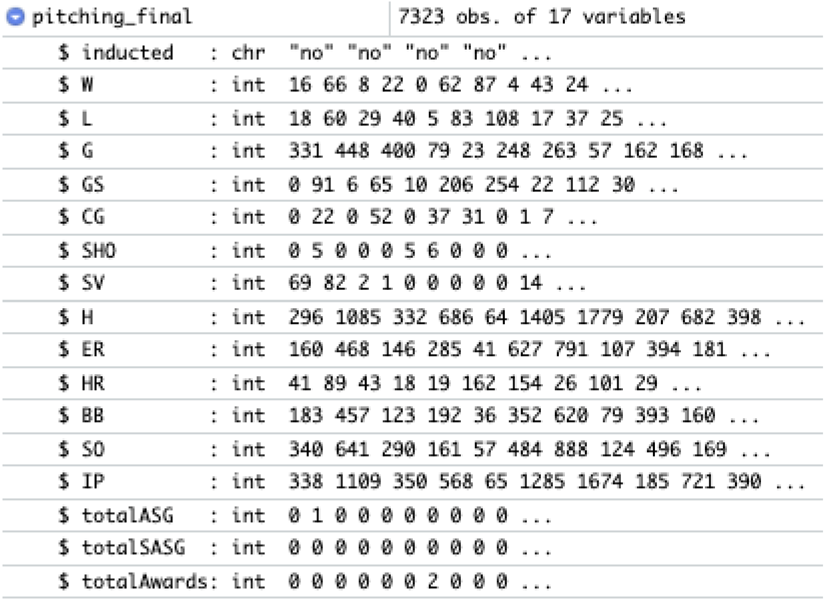
Lastly, the relationship between Hall of Fame status and career accolades was explored. The objective was to see if there was a certain threshold of awards a player must receive in order to be voted in. Through the analysis, there wasn't a number that could be determined as a threshold. Various players inducted did not have a single award throughout their entire careers, which was surprising to see. However, this obviously does not mean that they should not have been voted in, their career statistics backed up their Hall of Fame status. One thing to note is that certain awards were not available to older players, so they never had a chance to even win them, such as the Cy Young Award which started in 1956. This award is given to the best pitcher in baseball for that season. Another example is Rookie of the Year, which can obviously only be won during a player's rookie season.

7.2. Real World Implications

This study showed that for a pitcher to be voted into the Hall of Fame, he typically must have a longer career. A further bit of research could have introduced the longevity of a pitcher’s career as another variable to test. It also shows that while awards, accolades, and All Star Games can build a pitcher’s resume, they are not the final determinants of Hall of Fame status. Pitchers who strive to receive the MLB’s greatest honor, being inducted into the Hall of Fame, should focus on their health as much as possible, as there are countless pitchers who are at the top of their game but cannot stay healthy in order to prolong their careers.

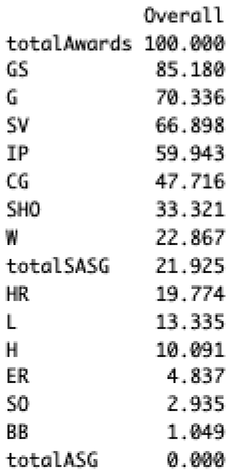
8. Appendix

Figure 1. Cleaned Data Frame (inducted is “no” and “yes” per a further data type change to factor)



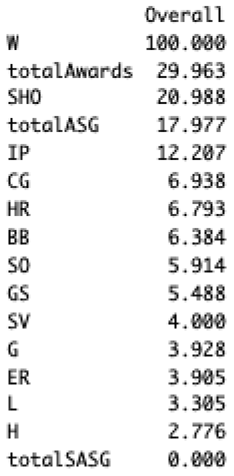
```
pitching_final             7323 obs. of 17 variables
    $ inducted   : chr  "no" "no" "no" "no" ...
    $ W          : int  16 66 8 22 0 62 87 4 43 24 ...
    $ L          : int  18 60 29 40 5 83 108 17 37 25 ...
    $ G          : int  331 448 400 79 23 248 263 57 162 168 ...
    $ GS         : int  0 91 6 65 10 206 254 22 112 30 ...
    $ CG         : int  0 22 0 52 0 37 31 0 1 7 ...
    $ SHO        : int  0 5 0 0 0 5 6 0 0 0 ...
    $ SV         : int  69 82 2 1 0 0 0 0 0 14 ...
    $ H          : int  296 1085 332 686 64 1405 1779 207 682 398 ...
    $ ER         : int  160 468 146 285 41 627 791 107 394 181 ...
    $ HR         : int  41 89 43 18 19 162 154 26 101 29 ...
    $ BB         : int  183 457 123 192 36 352 620 79 393 160 ...
    $ SO         : int  340 641 290 161 57 484 888 124 496 169 ...
    $ IP         : int  338 1109 350 568 65 1285 1674 185 721 390 ...
    $ totalASG   : int  0 1 0 0 0 0 0 0 0 0 ...
    $ totalSASG  : int  0 0 0 0 0 0 0 0 0 0 ...
    $ totalAwards: int  0 0 0 0 0 0 2 0 0 0 ...
```

Figure 2. Logistic Regression Variable Importance:



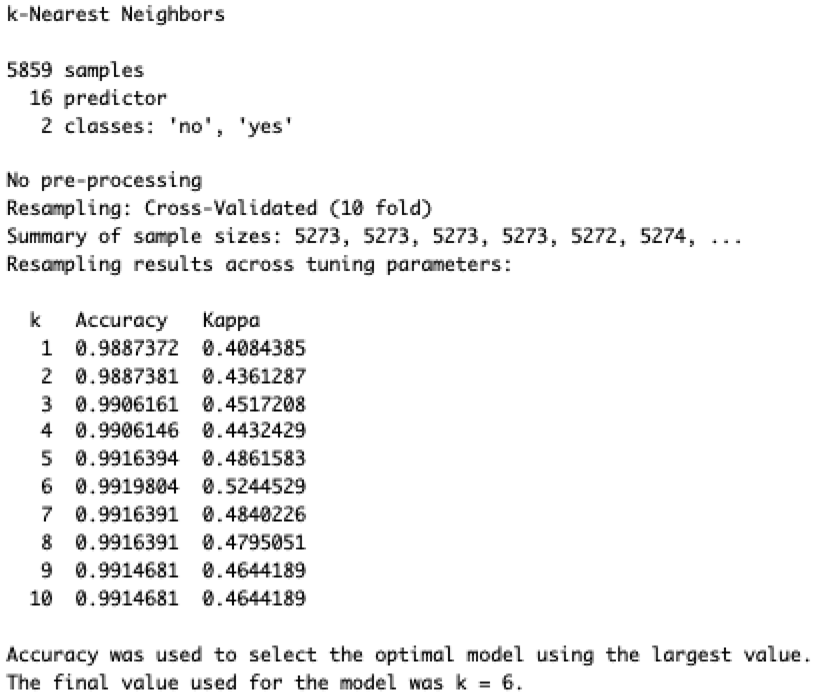
```
            Overall
totalAwards 100.000
GS           85.180
G            70.336
SV           66.898
IP           59.943
CG           47.716
SHO          33.321
W            22.867
totalSASG    21.925
HR           19.774
L            13.335
H            10.091
ER            4.837
SO            2.935
BB            1.049
totalASG      0.000
```

Figure 3. Random Forest Variable Importance:



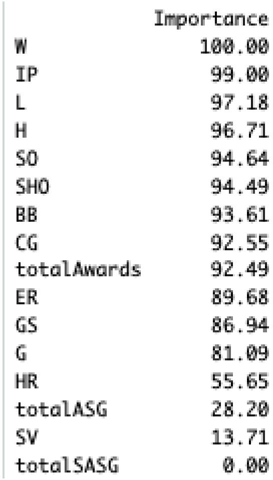
```
             Overall
W            100.000
totalAwards   29.963
SHO           20.988
totalASG      17.977
IP            12.207
CG             6.938
HR             6.793
BB             6.384
SO             5.914
GS             5.488
SV             4.000
G              3.928
ER             3.905
L              3.305
H              2.776
totalSASG      0.000
```

Figure 4. KNN Analysis:



```
k-Nearest Neighbors

5859 samples
  16 predictor
   2 classes: 'no', 'yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 5273, 5273, 5273, 5273, 5272, 5274, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9887372  0.4084385
   2  0.9887381  0.4361287
   3  0.9906161  0.4517208
   4  0.9906146  0.4432429
   5  0.9916394  0.4861583
   6  0.9919804  0.5244529
   7  0.9916391  0.4840226
   8  0.9916391  0.4795051
   9  0.9914681  0.4644189
  10  0.9914681  0.4644189

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 6.
```

Figure 5. Naive Bayes Variable Importance:



```
            Importance
W               100.00
IP               99.00
L                97.18
H                96.71
SO               94.64
SHO              94.49
BB               93.61
CG               92.55
totalAwards      92.49
ER               89.68
GS               86.94
G                81.09
HR               55.65
totalASG         28.20
SV               13.71
totalSASG         0.00
```

9. Bibliography

Brownlee, J. (2019, October 7). *Machine Learning Mastery.* https://machinelearningmastery.com/classification-as-conditional-probability-and-the-naive-bayes-algorithm/

Edgar, T. (2017). *Science Direct.* Logistic Regression. https://www.sciencedirect.com/topics/computer-science/logistic-regression

Harrison, O. (2018, September 10). *Towards Data Science.* Machine Learning Basics with the K-Nearest Neighbors Algorithm. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Lahman, S. (2022, March 8). *Download Lahman's Baseball Database*. https://www.seanlahman.com/baseball-archive/statistics/

Lahman, S. (2022, March 8). *The Lahman Baseball Database 1871-2021.* https://www.seanlahman.com/files/database/readme2021.txt

Liberman, N. (2017, January 26). *Towards Data Science.* Decision Trees and Random Forests. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991