Assignment 3

Hall of Fame Standards

UMGC Data 630

Professor Ami Gates

**Introduction**:

**Objective**:

The objective of this analysis is to create a model that showcases what specific attributes and statistics are the most important in determining whether a baseball player is worthy of being enshrined in the Hall of Fame. The purpose of this model is so that those voting on future inductees may have a baseline with which to compare current players with those already enshrined.

**Problem Domain**:

In 1939, the Baseball Hall of Fame and Museum was opened in Cooperstown, New York since Cooperstown was hailed as the birthplace of the sport. The inaugural class of inductees included legends like Babe Ruth, Honus Wagner, and Ty Cobb. The museum continues to grow each year, now housing over 40,000 artifacts spread around 5 buildings and accommodating an average of 3,000 visitors per day.

As of today, 333 players have been inducted which averages to just over four players being inducted each year. In order to be eligible for induction, a player must have played at least 10 seasons in Major League Baseball and have retired at least 5 years prior to induction. Any baseball writer who has maintained a beat for 10 years is eligible to vote for players to be enshrined and receives a mail-in ballot for the eligible players each year, which has caused the vote count to range from 150 to 581 over the course of the museum’s history. A voter may vote for a maximum of 10 players each year. If a player receives 75% of the vote or higher, then they are elected to the Hall of Fame. If a player receives greater than 5% of the vote, they are eligible to appear on the next year’s ballot for another chance at election, but if they fail to receive 5%, they are taken off the ballot and are ineligible for induction into the Hall of Fame.

The words of the voting criteria specifically state “Voting shall be based upon the player’s record, playing ability, integrity, sportsmanship, character, and contributions to the team(s) on which they played” (baseballhall.org). There is also a stipulation that states there will be no automatic election based on a single statistic. For example, batting a .400 average over the course of one’s career does not guarantee admission into the Hall of Fame. This has been seen with players such as Barry Bonds and Mark McGwire who are numbers 1 and 11 respectively on the all-time home run list but had their integrity tarnished by being involved with performance enhancing drugs.

**Method Rationale**:

While the Hall of Fame denies any automatic election for players based on a singular statistic, having a model that can compare current eligible candidates to those already enshrined in the Hall of Fame may be helpful for the baseball writers who are determining whether a player is deserving of enshrinement. By intaking a variety of career statistics, a decision tree analysis (specifically ctree) will be taken that can split players based on attributes such as batting average, runs scored, home runs, etc. and determine probabilities for being elected into the Hall of Fame based on what the player accomplished over the course of their career. The final output is intended to be shared with the Hall of Fame voters so they may compare all players on whom they are voting with those already in the Hall of Fame.

**Analysis**:

**Data**:

The dataset utilized for this analysis is a census that covers all MLB players who retired prior to the 1993 MLB season and were eligible for the Hall of Fame (played at least 10 years). The dependent variable is Hall of Fame Membership. This is marked with one of three factor levels. 0 means that the player was not inducted. 1 means the player was inducted within 5-15 years after their playing career ended. 2 means the player was inducted on a “legacy” rule meaning they were added to the ballot well after their playing career for various reasons and were then inducted. The numeric independent variables are Number of Seasons, Number of Games, Official At-Bats, Runs Scored, Hits, Doubles, Triples, Home Runs, RBIs, Walks, Strikeouts, Batting Average, On Base Percentage, Slugging, Adjusted Production, Batting Runs, Adjusted Batting Runs, Runs Created, Stolen Bases, Caught Stealing, Stolen Base Runs, Fielding Average, Fielding Runs, and Total Player Rating. The one independent factor variable is the players’ fielding position, marked as a C for Catcher, 1 for First Base, 2 for Second Base, S for Shortstop, 3 for Third Base, and O for Outfield. It should be noted that due to the nature of this data being batting-statistic centric, pitchers were omitted from the dataset.

**Exploratory Analysis**:

Using the str function on the dataset, it is shown that almost all variables are integers or numeric (due to the decimal nature of many baseball statistics). The Name and Position variables are the only character variables present. Using the summary function, it is shown that many of the variables have some immense spreads. Season played ranges from 10 to 26, hits range from 48 to 4256, triples range from 0 to 309, and stolen bases range from 0 to 938. This function also shows a number of missing variables that will be addressed in the Preprocessing.

Figure 1 shows a boxplot of the players’ batting averages broken down by whether they made the Hall of Fame. Induction Type 1 or 2 means that the player made the Hall of Fame while Induction Type of 0 means they did not. As expected, players who made the Hall of Fame tended to have higher batting averages as the majority of data in the second and third boxplots are higher than that of the first boxplot. However, it is worth noting that the 1st quartile mark for the second and third boxplots (the Hall of Famers) is just below a .300 batting average while that is also about where the 3rd quartile mark is for the non-Hall of Famers as well. This means that in terms of batting average, the top 25% of non-Hall of Famers had better batting averages than the bottom 25% of Hall of Famers. It is also evident that there are some very high outliers for the non-Hall of Famers as well. Based on this visualization, Hall of Fame enshrinement is clearly based on more than just average since many non-Hall of Famers have outperformed Hall of Famers in this statistic.

A screenshot of a cell phone

Description automatically generated

Figure

Creating similar visualizations for Home Runs yields the distribution shown in Figure 2. In this visualization, it appears that hitting home runs improves a player’s chance of getting into the Hall of Fame, but really it shows a similar alignment with the top 25% of non-Hall of Famers outperforming the bottom 25% of Hall of Famers. There is such a major spread in the upper 75% of Hall of Famers which causes this. The boxplot does show that any player who hit over 450 runs at this point in baseball history was enshrined in the Hall of Fame.

A screenshot of a cell phone

Description automatically generated

Figure

**Preprocessing**:

In preprocessing this data to prepare for the creation of the model, the first step was to eliminate the variable of “Name”. While it is very interesting to see the names of some legendary players line up with their stats, for the purpose of a ctree analysis, the names are irrelevant. This removal was done using the null command. Next, to address the character variables, the variable of Primary Position Played was converted to a factor variable with each possible position represented by a factor level. The same was done for the variable of Hall of Fame Membership and also a new variable was created that mirrors the Hall of Fame Membership variable but combines factor levels 1 and 2 so that all Hall of Famers are grouped together.

The summary function completed in the exploratory analysis showed several variables in which the data was missing a value. This issue will be addressed per variable. For the variable of strikeouts, there were 20 NA’s. This variable does have a large range of values from 0 to 2597 total strikeouts which may make it difficult to approximate the missing values. Figure 3 was A screenshot of a cell phone

Description automatically generatedcreated to show that the reason for this immense spread is the existence of high outliers. With this knowledge, it was decided to replace the value of the missing data with the mean of 445.7 strikeouts since that is a strong measure of center for this particular variable.

Figure

Similar reasoning was used to replace 2 missing values within the variable of stolen bases as well. The spread on this variable was 0 to 938 with the mean and median within 50 stolen bases of each other showing only moderate skew. Thus, these two NA’s for stolen bases were replaced with the mean of 104.4 stolen bases.

The variables of Caught Stealing and Stolen Base Runs are particularly problematic as they have 264 and 622 missing values respectively. This is because baseball statisticians have varied throughout history in their desire to actually record that statistic so it was recorded some years and not others and thus may even be inaccurate even for observations that are not actually missing the value. For this reason, both of these variables were eliminated by being set to null. Utilizing the summary command after these steps were taken yields no missing values.

Figures 1, 2, and 3 all showcase that the variables in this dataset all have some very strong outliers. Since this analysis’s objective is, in a way, quantifying greatness, the decision was made to keep all outliers in the data. For the positive stats such as hits or home runs, it makes sense that a player with such a quantity of each statistic such that they are considered an outlier would also be inducted into the Hall of Fame. It also makes sense for a negative statistic such as strikeouts, since being an outlier in that variable may be bad for the overall reputation and play a factor in not getting elected to the Hall of Fame.

**Algorithm Intuition**:

In order to have a large enough sample from the data to create an accurate model, as well as a separate and large enough sample with which to test the accuracy of the model, the dataset of 1340 player records was split 70/30, placing 952 players into the training dataset and 388 players into the testing dataset. Using the training dataset, the ctree command will be used, with its output stored as one of two models. This function requires the “party” package within R and takes the input of a dataset and a formula based on a dependent variable and however many independent variables are desired. The dependent variable for this model will be Hall of Fame Membership and the independent variable will be scripted as the wildcard “.” telling the function to output the best model using all possible independent variables. Model 1 will be created keeping all three levels of Hall of Fame Membership while Model 2 will be created using the combined Membership variable that has all Hall of Famers grouped together. The models will be compared for both the homogeneity of their terminal nodes and the accuracy of their confusion matrices and the better performing model will be utilized.

**Modeling Fitting**:

Model 1 yielded a decision tree with 17 total nodes that split by 7 different variables. Creating a confusion matrix for this based on the training data yields the results shown in Figure 4. This shows an accurate prediction for 896 of the 952 observations, giving this model a 94.12% accuracy. Model 2 yielded a decision tree with 11 nodes splitting by 4 unique variables. Since the variable of Hall of Fame Membership was converted from 3 factor levels to binary, this also changed the visualization of each terminal node which will be discussed in the results. The confusion matrix for this model is also shown in Figure 4. It shows a correct prediction of 909 of 952 observations, showing a 95.48% accuracy on the training data. For this purpose, Model 2 will be discussed in the results section.

A screenshot of a cell phone

Description automatically generated**A screenshot of a cell phone

Description automatically generated**

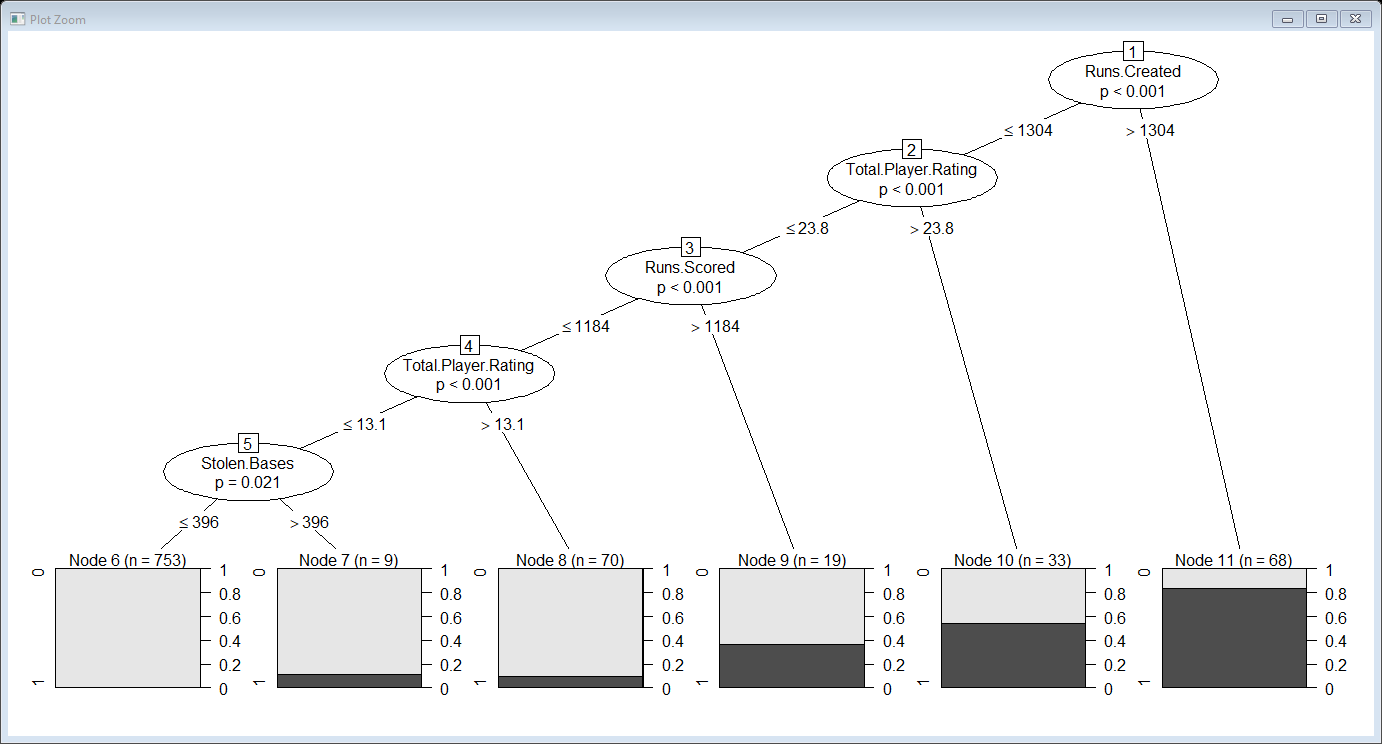
Figure

**Result**:

**Output and Properties:**

Figure 5 showcases the visual plot of the decision tree created by Model 2, the model based on combining both types of Hall of Famers into one category in order to compare them to non-Hall of Famers. Since this combination of factor levels simplified the variable of Hall of Fame Membership to be binary, the terminal nodes can show direct percentages rather than histograms. The first splitting variable is Runs Created, which is a calculated measure based on hits, walks, total bases, and at-bats. The split shows that having over 1,304 runs created in a career yields an 83.8% chance of getting into the Hall of Fame, calculated based on 68 players who fit this description. The second splitting variable, for players who created 1,304 or less runs, is Total Player Rating, a metric used to compare player value against what an average player is expected to do. Having a total player rating greater than 23.8 yields a 54.5% chance of getting inducted to the Hall of Fame based on 33 players who fit into this node. The third splitting variable, for players who created less than 1,304 runs and have a Total Player Rating less than or equal to 23.8, is Runs Scored. For the observations in this node are split at 1184 runs scored. Players who scored more than 1,184 runs showed a 36.8% chance of being inducted, based on 19 players. The fourth splitting variable was once again Total Player Rating, this time split at 13.1. Players in this node who had a total player rating greater than 13.1 showed a 10% chance of induction, based on 70 players. The final splitting variable is stolen bases. Players who stole more than 396 bases in this node showed an 11.1% chance of induction based on 9 players. For the players who did not fit any of the criteria to follow a rightward branch out of a node, meaning the their Runs Created was less than 1,304, their TPR was less than 13.1, their runs Scored was less than or equal to 1,184, and their stolen bases were less than or equal to 396, the probability of induction into the Hall of Fame was 3%.

This model meets the objective set of creating a model in which current and future Hall of Fame hopefuls may be placed to see how they compare to those already enshrined. Two of the variables used (Runs Created and Total Player Rating) are actually combined metrics which factor in several of the other variables such as hits, extra base hits, at bats, etc. to compare players. These statistics are available for any eligible retired player and thus the model can be used to determine a player’s likelihood of being inducted, which may help inform the voters of which players are deserving.



Figure

**Evaluation**:

A picture containing clock

Description automatically generatedEmploying a confusion matrix on the data that was set aside for testing purposes yields the matrix in Figure 6. It shows that of the 388 observations reserved for testing, 362 were predicted correctly and 26 were predicted incorrectly. This yields an accuracy rating of 93.3% which is just slightly lower than the results of the confusion matrix created on the training data. This is to be expected as the model was based on the training data while the testing data is all new to the model until tested with the confusion matrix.

Figure

**Conclusion**:

**Summary**:

While being great in one statistical category is not enough for a player to be inducted into the Baseball Hall of Fame, the created model and analysis shows key variables/statistics that correlate to a player’s chances of being enshrined. While this may be informative for players in trying to pad their stats to ensure their induction, the real power of this model comes from its usage by those who elect players into the Hall of Fame. By using this model, voters can be sure that they are fair to all players by being able to compare possible inductees to past inductees to see if they achieved the same level of statistical greatness.

**Limitations**:

The greatest limitation to this model is the stipulation clearly stated in the rules of the election process that a player’s character must also be considered. Due to several scandals, mainly surrounding steroids and performance enhancing drugs, many players in recent history would be considered highly likely to be inducted just based on their numbers, but things like failed drug tests may keep them out of the Hall of Fame. For example, Barry Bonds hit 762 career homeruns which helped him achieve the highest ranking for the Runs Created metric: 2,892. According to this model, having 1,304 Runs Created is enough to give somebody an 83% chance of induction, but he has double that and will not be inducted. It is because of this emphasis on character and playing the game in a clean way that may cause this model to start failing as more and more eligible players will be remembered for things like failed drug tests.

**Improvement Areas**:

In order to improve this model, some further experimentation with the independent variables may be done. By refining the dependent variable from three factor levels to binary, the accuracy was improved by an estimated 2%. Editing the possible input independent variables may also be able to improve the accuracy. As seen in some of the visualizations, there were levels for some statistics that almost guaranteed induction into the Hall of Fame and perhaps experimenting with this may have led to more homogenous terminal nodes and a higher accuracy based on the confusion matrix. Also, as previously discussed, induction into the Hall of Fame has become about more than just the numbers. It may be possible to account for things such as failed steroid tests with a simple binary variable that indicates whether a player has ever failed a test to account for the players who should statistically be inducted but never will be due to their reputation.

**References:**

History of the Museum. (n.d.). Retrieved July 01, 2020, from https://baseballhall.org/about-the-hall

Svrluga, B. (2018, January 23). Perspective | Cut the nonsense: Roger Clemens and Barry Bonds should be in the Hall of Fame. Retrieved July 01, 2020, from https://www.washingtonpost.com/sports/cut-the-nonsense-roger-clemens-and-barry-bonds-should-be-in-the-hall-of-fame/2018/01/23/89c6a5ea-006c-11e8-9d31-d72cf78dbeee\_story.html