Predicting Baseball Hall Of Fame Inductions

Michael Hirsch

ILLC, University of Amsterdam

michaelahirsch@gmail.com

Abstract

Every year, the Baseball Writers Association of America votes on a new Hall of Fame class. Each ballot consists of 10 votes, and players need to appear on 75% of ballots in order to be inducted into the hall. Players have 15 years to get inducted, and are no longer eligible if that time period has passed. Here we attempt to classify hall of fame batters based on their career statistics using decision trees and an ensemble method, random forests.

1. Introduction

Major League Baseball (MLB) has been keeping thorough records of batting, picthing, and fielding statistics since its innagural season in 1869. Recently, with the advent of sabermetrics by the Society for American Baseball Research (SABR), many new metrics building on traditional statistics were created. Such metrics caught the eye of stasticians, as these new metrics allowed for the creation of even more powerful predictive models.

In this paper, we are investigating what makes a Hall Of Fame (HOF) batter. There have been over 20,000 Major League baseball players in the history of the organization, and only 211 of them have been inducted into the HOF. There are several ways in which a player can be inducted, and we only concern ourselves with players inducted from Baseball Writers Association of America (BBWAA) ballots, as this process is regulated. Often, fans and players feel that a worthy candidate is unfairly denied entry into the HOF because of voter bias, stacked ballots, or negative associations with performance enhancing drugs. The model proposed in this paper should be able account for these cases.

2. Data

Data was collected from the Lahman Baseball Database, a freely available database that has been contiously updated since 1994 with the help of SABR and many individual researchers. There are many data tables available for download, but the ones that we are focusing on are *Batting.dat* and *HallOfFame.dat*. Batting statistics were supplied on a by-year basis, so a player's career statistics were computed by aggregating his yearly results. The batters' data was then merged with hall of fame voting results using the player's ID string. All data preparation was done within R, the environment where we also build our learning models.

Attention is paid to 18 different batting statistics over the course of the player's career:

Below is a list of features used

1. G: games played
2. AB: at-bats
3. R: runs
4. H: hits
5. X2B: doubles
6. X3B: triples
7. HR: homeruns
8. RBI: runs batted in
9. SB: stolen bases
10. CS: caught stealing
11. BB: walks
12. SO: strike outs
13. IBB: intentional walks
14. HBP: hit by pitches
15. SH: sacrifice hits
16. SF: sacrifice flys
17. GIDP: grounded into double plays
18. Inducted: classifier for HOF induction

We needed to further filter our data by not including pitchers in our set of observations. This was done by removing those observations that indicated that the player recorded less than 300 RBI in his career. This number was not arbitrary, but rather chosen by cross referencing my data with HOF data from Baseball-Reference.

3. Tree-Based Methods for Classification

In this section, we will give an account of two learning methods that we will use for the basis of prediction: classification trees, and an ensemble method, random forests.

3.1. Classification Trees

Classification trees are a fairly effective predictive tool that are lauded for their high degree of interpretability. When constructing a classification tree, we begin with the full set of observations and begin dividing the predictor space into non-overlapping regions by means of binary splitting. In prediction, we assign each observation in a given region of the predictor space to the most commonly occuring class of the training observations in that region. When splitting, we are concerned with the textitInformation Gain at each split, that is, we want to choose the feature to split on based on which feature split will give us the most information gain. This defined as:

$$IG(T, a) = H(T) - H(T|a)$$

where T denotes a set of training examples, each of the form $(\mathbf{x}, y) = (x_1, \ldots, x_k, y)$ where each $x_a \in vals(a)$ is the value of the a^{th} attribute of $\mathbf{x}$ and y is the classification label. So, the information gain for a split on attribute a is definted in terms of entropy as:

$$IG(T, a) = H(T) - \sum_{v \in vals(a)} \frac{|\{\mathbf{x} \in T : x_a = v\}|}{|T|} \cdot H(\{\mathbf{x} \in T : x_a = v\})$$

which is a measure of the variance over the K classes. $\hat{p}_{mk}$ represents the proportion of training observations in the m^{th} region that are from the k^{th} class. This is taken to be a measure of each node's textitpurity, with a low value for G indicating that the node contains predominatly observations from a single class.

Another measure that we are concerned with in building classification trees is the *information gain*, defined as:

3.2. Ensemble Methods

3.2.1. Bagging

3.2.2. Random Forests