What Distinguishes Hall of Fame Batters Statistically?

Calvin Makelky, A01872013

November 13, 2016

Since 1898, almost every season has had 154 games or more according to http://research.sabr.org/journals/schedule-changes-since-1876. 10 seasons are needed in order to be considered for the Baseball Hall of Fame. Unfortunately baseball-reference will not let you search by seasons played. To low ball the minimum number of games played to play for 10 seasons, I estimated (154*10)/2 = 770 games. Because a player can't be considered for the HOF until 5 years after retirment and can be considered for 10 years, I determined to only look at players whose last year was 2001 (2016-15).Here is the query I used to find these non-Hall of Fame batters who played for ten or more seasons.

Spanning Multiple Seasons or entire Careers, From 1898 to 2001, not a Hall Of Fame Member (as mlb players), Played 50% of games at C, 1B, 2B, 3B, SS, LF, CF, RF, DH or PHR, (requiring year_max<=2001 and At least 770 games).

Because baseball-reference will not let you randomly select or see more than 10 players in the results without a subscription, I used a random criteria to sort by that wouldn't bias the data.

I sorted by greatest first season, birth month, birth, name, final season and sorted by lowest First season, birth month, birth, name, and final season.

I copy and pasted the csv of this data into a text file.

For the hall of famers, I downloaded a .csv file of the career stats of non-pitchers. I excluded two players who predominatly played in the negoro leaugues since this only contains MLB data. There is over a 100 hall of fame batters and over a 100 batters not in the hall of fame.

```
hof = read.csv("hof_batting.csv", header=TRUE)
hof$HOF = 1
#create new column, hof, coded as 1 that will be the response variable

nhof = read.csv("Not_HOF_Batters.txt", header=TRUE)
nhof$HOF = 0
#create new column, hof, coded as 0 that will be the response variable
nhof$Yrs = nhof$To - nhof$From + 1
#calculate how many seasons a player has played in their career
nhof= subset(nhof, Yrs>9)
#removing players with less than 10 years

nhof = subset(nhof, select = -c(Rk, Player, X, Age, CS, Pos, Tm, IBB, HBP, SF, From, To, GDP, PA))
hof = subset(hof, select = -c(Rk, Name, Inducted, X, X.1, ASG, WAR.pos, X.2, X.3, X.4, X.5, PA))
#getting rid of unmatching columns and useless columns
hof = rbind(hof, nhof)
#combining hall of fame batter dataset with non-hall of fame batter dataset


hof$HOF <- factor(hof$HOF,
  levels = 0:1, labels = c("no", "yes"))
#better to have response variable factor type than numeric

library(party)
```

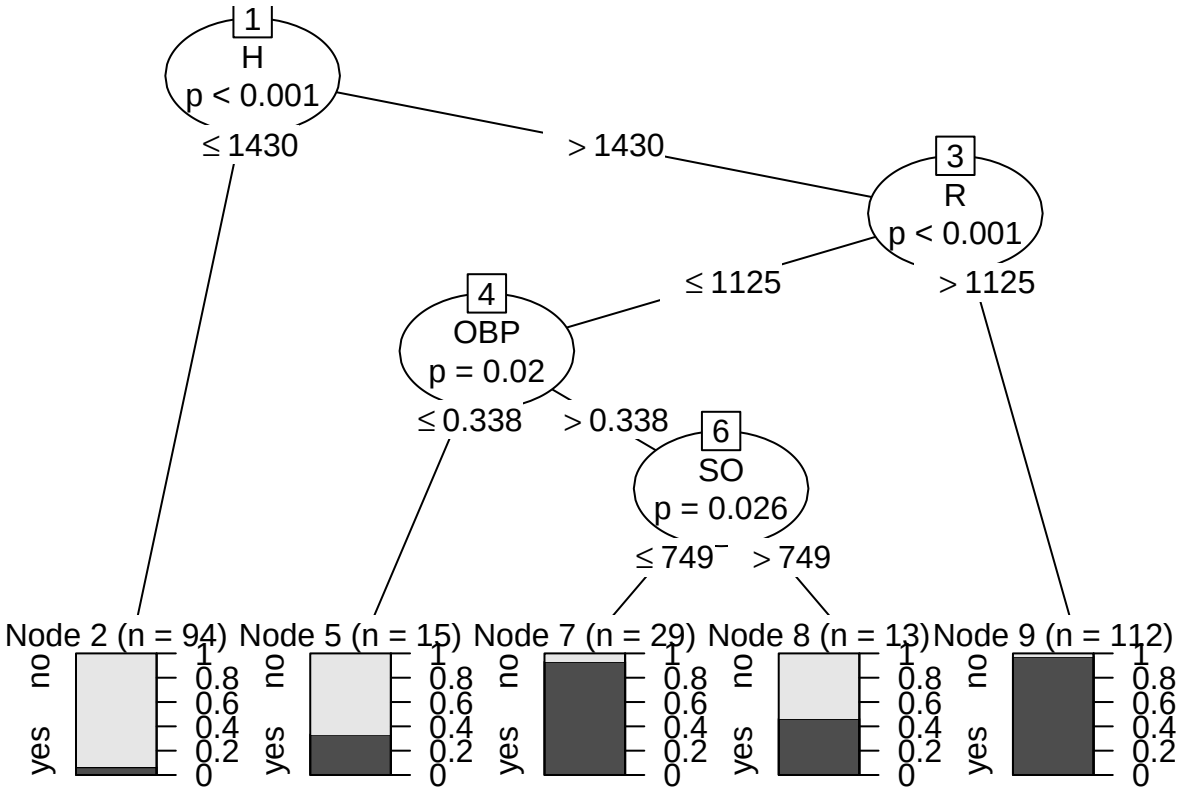
```
## Loading required package: grid
```

```
## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
hof.ct = ctree(HOF ~ Yrs + G  + AB + R + H + X2B + X3B +
HR + RBI + SB + BB + SO + SH + BA + OBP + SLG + OPS, data = hof)
#constructing classification tree model
hof.ct
```

```
##
##   Conditional inference tree with 5 terminal nodes
##
## Response:  HOF
## Inputs:  Yrs, G, AB, R, H, X2B, X3B, HR, RBI, SB, BB, SO, SH, BA, OBP, SLG, OPS
## Number of observations:  263
##
## 1) H <= 1430; criterion = 1, statistic = 146.531
##   2)*  weights = 94
## 1) H > 1430
##   3) R <= 1125; criterion = 1, statistic = 27.164
##     4) OBP <= 0.338; criterion = 0.98, statistic = 10.473
##       5)*  weights = 15
##     4) OBP > 0.338
##       6) SO <= 749; criterion = 0.974, statistic = 9.989
##         7)*  weights = 29
##       6) SO > 749
##         8)*  weights = 13
##   3) R > 1125
##     9)*  weights = 112
```

```
plot(hof.ct)
```



The classification tree is very interesting. Almost none of the hall of fame batters had less than 1430 hits. So the criteria in order to be in the hall of fame as position player is as follows. Batters must have more than 1430 career hits and more than 1125 career runs. Or must have more than 1430 career hits, an OBP above 0.338, and less than 750 strikeouts. Those are the two scenarios the classification tree predicts batters to be a hall of famer.