STAT 636, Fall 2015 - Assignment 8
Due Friday, December 4, 3:00pm Central
Online Students: Submit your assignment through WebAssign.
On-Campus Students: Email your assignment to the TA.

The data file "HOF_tr.csv" contains data on 677 retired Major League Baseball (MLB) players who are already in or are eligible for induction to the Hall of Fame (HOF). Only players who finished their career after 1960 are included. Pitchers and designated hitters have been excluded, as have any players known or alleged to have used performance enhancing drugs, as has Pete Rose (since he has been banned from HOF consideration for gambling on MLB games). For each player, you have the following variables and career statistics:

- `HOF`: Hall of fame status. "Y" = in the HOF; "N" = not in the HOF.
- `FY`: The first season in which the player appeared.
- `LY`: The last season in which a player appeared.
- `SP`: Number of seasons played. This may be different from `LY - FY + 1`, e.g., if a player missed a season due to injury.
- `POS`: A code for primary position played. "1B" = first base, "2B" = second base, "3B" = third base, "C" = catcher, "OF" = outfielder, "SS" = shortstop. If an individual played multiple positions over the course of his career, the position at which he played the most games was selected.
- `ASG`: The proportion of seasons played in which the player was selected for the All Star Game.
- `G`: Number of games played.
- `AB`: Number of at-bats.
- `R`: Number of runs.
- `H`: Number of hits.
- `DB`: Number of doubles.
- `TP`: Number of triples.
- `HR`: Number of home runs.
- `RBI`: Number of runs batted in.
- `SB`: Number of stolen bases.
- `CS`: Number of times caught stealing.
- `BB`: Number of base on balls (walks).
- `SO`: Number of strikeouts.

- `AVG`: Batting average (number of hits divided by number of at-bats). Has been multiplied by 1000.
- `SLG`: Slugging percentage (total bases divided by number of at-bats). Has been multiplied by 1000.
- `OBP`: On-base percentage. Equal to (`H` + `BB` + `HBP`) / (`AB` + `BB` + `HBP` + `SF`), where `H`, `BB`, and `AB` are as described above, `HBP` is the number of times the player was hit by a pitch, and `SF` is the number of sacrifice flies. Has been multiplied by 1000.

Your task is to build a model for classifying players as being in the HOF or not, given data on the above variables. You can use any classification method you want. You can use any subset and / or transformation of the variables you want. Specific expectations:

1. Use cross validation (CV) to guide the selection of your model and report a CV-based estimate of your model's sensitivity and specificity. Turn in all `R` code.
2. Provide text defining an R function that will use your model to classify new observations. There are an additional 339 players for whom I am holding back data. For each of these held-out players, I have the exact same variables as you (including HOF status). Your function needs to take a single argument, with which we will pass these data as a 339×21 data frame. The data frame will have identical column names and types as those in "HOF_tr.csv". Your function should (only) return 339 "Y"s and "N"s, depending on whether each held-out player is classified as being in the HOF or not.

Grading criteria will be as follows: (i) correct implementation of the model(s) used, (ii) appropriate use of CV for model selection and accuracy estimation, and (iii) functionality of your classification `R` function (i.e., it should work when we run it). In addition, each student's submission will be assessed an overall (weighted) accuracy defined by (`sensitivity` + 3× `specificity`) / 4, where `sensitivity` and `specificity` are those values achieved by your method when applied to the held-out data. The student with the highest weighted accuracy on the held-out data will be awarded a $50 gift card to Amazon.com. In the event of a tie, Dr. Dabney will make the final decision on which one student gets the award, based on the approach taken and clarity / efficiency of code. To be clear, your grade will <u>not</u> be dependent on your model's accuracy.