Machine Learning Report

Brian Hallberg

6/21/2018

Applying Machine Learning to Project

How well can all-star appearances and other baseball player statistics predict the induction into the hall of fame for a player?

This is a supervised problem because we are attempting to predict a specific dependent variable using a model based on a set of independent variables.

This is a classification problem.

Based on some model testing, the independent variables I will use are as follows:

- All Star Game Appearances
- Non Pitcher Statistics
  - Career Hits
  - Career Home Runs
  - Career Batting Average
  - Career Runs Batted In
  - Career Runs Scored
- Pitcher Statistics
  - Career Wins
  - Career Strike Outs Thrown
  - Career Games Pitched

I'm going to start with logistic regression and use results from that to help identify the best independent variables. I'll use the confusion matrix to see how well the model works. If this does not provide good results, I'll look at the random forest model to see if it works better.

I will be testing the model both using a train/test split in the original data. There is also the ability to use a few more players that are not currently in the hall and have not retired to see if they would be inducted. There are a few years of inductions not in the database so that could be useful.