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

Because the independent variables are a mix of discrete/categorical variables and continuous variables, the Classification and Regression Trees (CART), specifically Random Forests, type model is a good choice.

I will be testing the random forest model both using a train/test split in the original data, and also using a new test data set.