Stats202: Homework 8

Chris Walker

November 27, 2015

Name of Kaggle team:

cdubs

Members:

Chris Walker

Public leaderboard RMSE score:

0.57703

Process

The first step of my process was to clean the data. I imputed missing values using a simple median application. I also added indicators of missingness to the dataset in case the missing values were interesting in and of themselves.

For the modeling I first tried some parametric approaches to the problem using both the ridge and lasso regressions. I cross-validated the regularisation parameter lambda and chose its value such that it minimised the test set RMSE. The Lasso significantly outperformed the ridge, indicating a large degree of redundancy in the data. I then attempted tree-based methods and settled on tree boosting. Again I cross-validated the tuning parameter and the number of trees, settling on a final model with 1000 trees and a tuning parameter (lambda) of 0.006. Finally I fit the model to the aggregated training and test set and predicted on the leaderboard set.