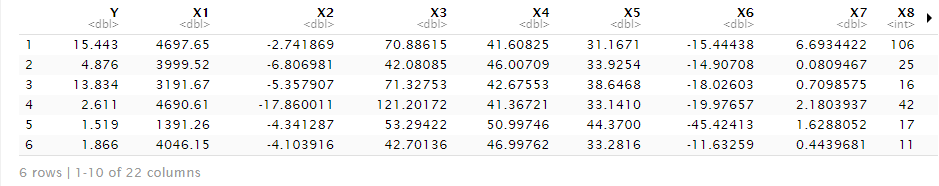
STAT 452 Project 1

Introduction & Set-up

In this project, a most-suitable model is used to predict a set of response variables. The training data consists of 20 explanatory variables and 10000 data points. All variables are given numerically, and no additional context has been provided.



A set of fitting techniques are used, and one of the models will be chosen to predict the response values. The model chosen is based on the lowest mean-square predicted error (MSPE). The fitting techniques, in sequence, are:

* Ridge
* LASSO
* Least squares (LS)
* Stepwise
* Partial least squares (PLS)
* Regression tree (RT)
* Random forest (RF)
* Boosting

Initially, all 20 variables will be used, and no interactive terms are considered. Once the model is picked, the pairwise interaction terms are used, if necessary, to attempt to improve the prediction error.

A 100-fold CV is used (each fold has a validation set of 100 data rows).

The details of each method is explained in each subsection. This report does not include the description of the methods themselves.

Ridge

The λ values used for testing range from 0 to 100, with increments of 0.05. In total, there would be 2001 values of λ. The λ that gives the least GCV in each fold is used for prediction.

The MSPE is .

LASSO

Both the λ­min­­ and λ­1SE are taken into consideration.

The MSPE for λ­min is .

The MSPE for λ­1SE is .

Least Squares

The MSPE for the simple linear regression model is .

Hybrid Stepwise

The stepwise method begins with the null model, and adds variables depending on the importance.

The MSPE for the selected stepwise model is .

Partial Least Squares

The MSPE for the PLS model is .

Regression Tree (with bootstrap)

Once a default regression tree is fit, prune it using the minimum complexity error.

The MSPE for the pruned regression tree is .

Random Forest

Boosting