Machine Learning Course Project

Temi A. Sorungbe

September 20, 2016

Summary

This project analyzes the wearable computing dataset- a large amount of data quantifying how well people execute their personal fitness activity . Specifically, data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants are recorded while they perform barbell lifts correctly and incorrectly in 5 different ways.

The cross validated random forest model best predicts how well each participants executes the barbell lifts and was selected as the final predictive model for the wearable computing dataset.

This model was selected after preprocessing the train dataset to select key features and comparing prediction accuracy of linear discriminant analysis (lda), quadratic discriminant analysis (qda) and classification and regression tree (CART) methods- rpart and random forest.

The validation data (out of sample) errors were equal to or higher than the train data (in sample) errors for all prediction models.

Pre-processing

The following steps were carried out on the train, validation and test dataset:

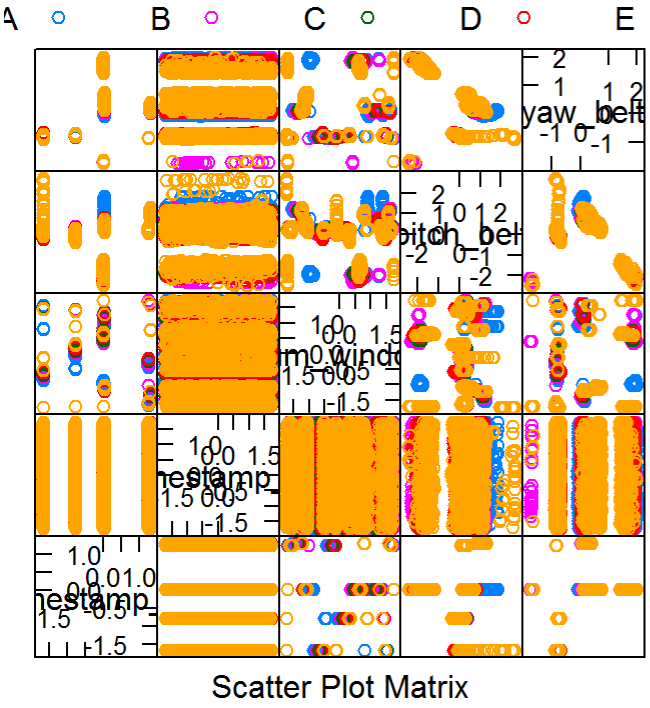
1. Selecting features from the training dataset to be used for predictions by:

i. removing variables with high missingness (> 95%)
ii. removing near zero variance variables
iii. removing highly correlated variables (ie. correlations > 0.75)

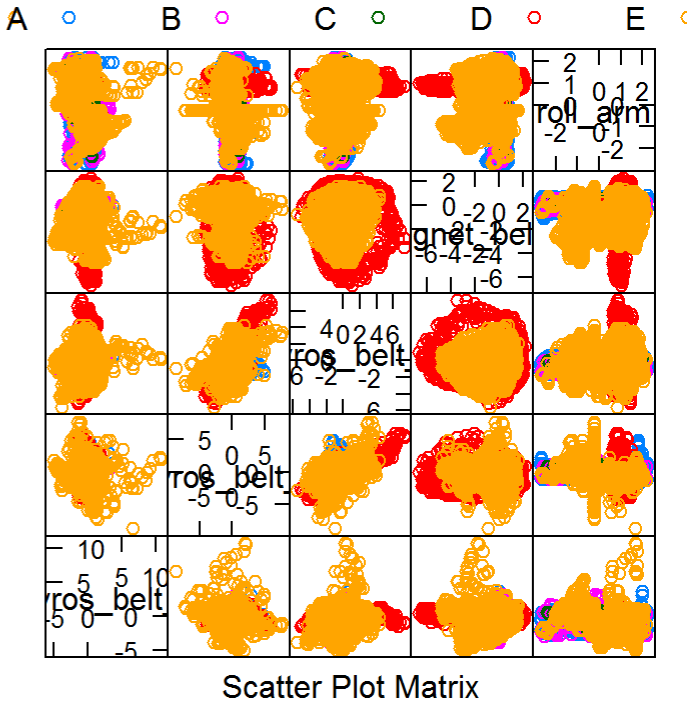
2. Centering and scaling all variables except “classe”

Explore the data

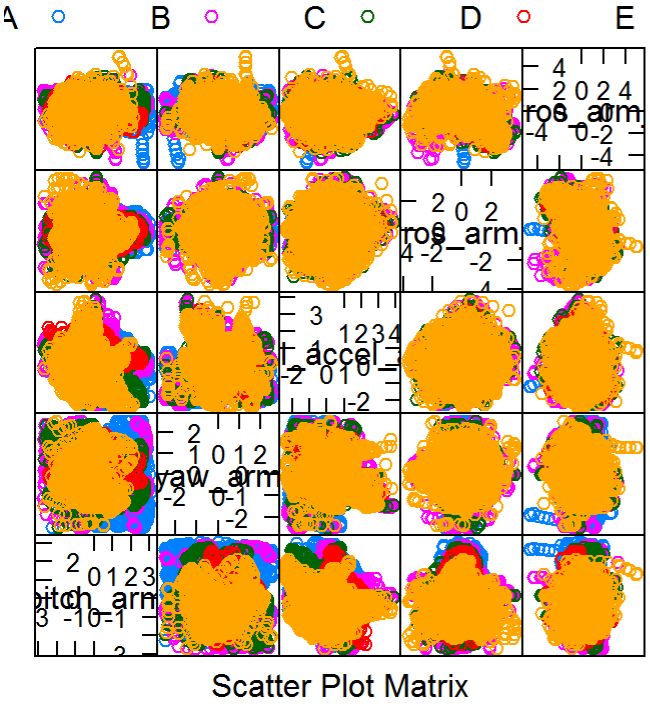
The following plots display the relationships between selected features and the classe variables in the train dataset.



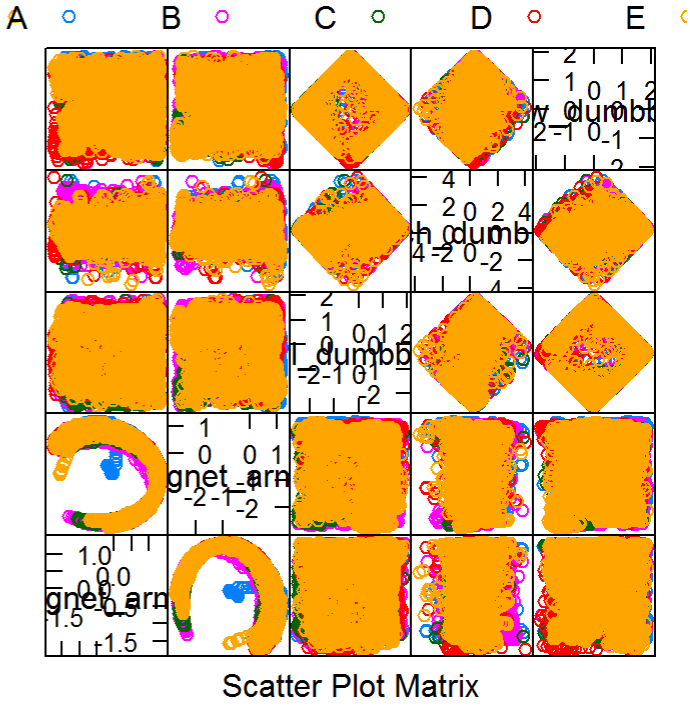
Scatter Plot Matrix



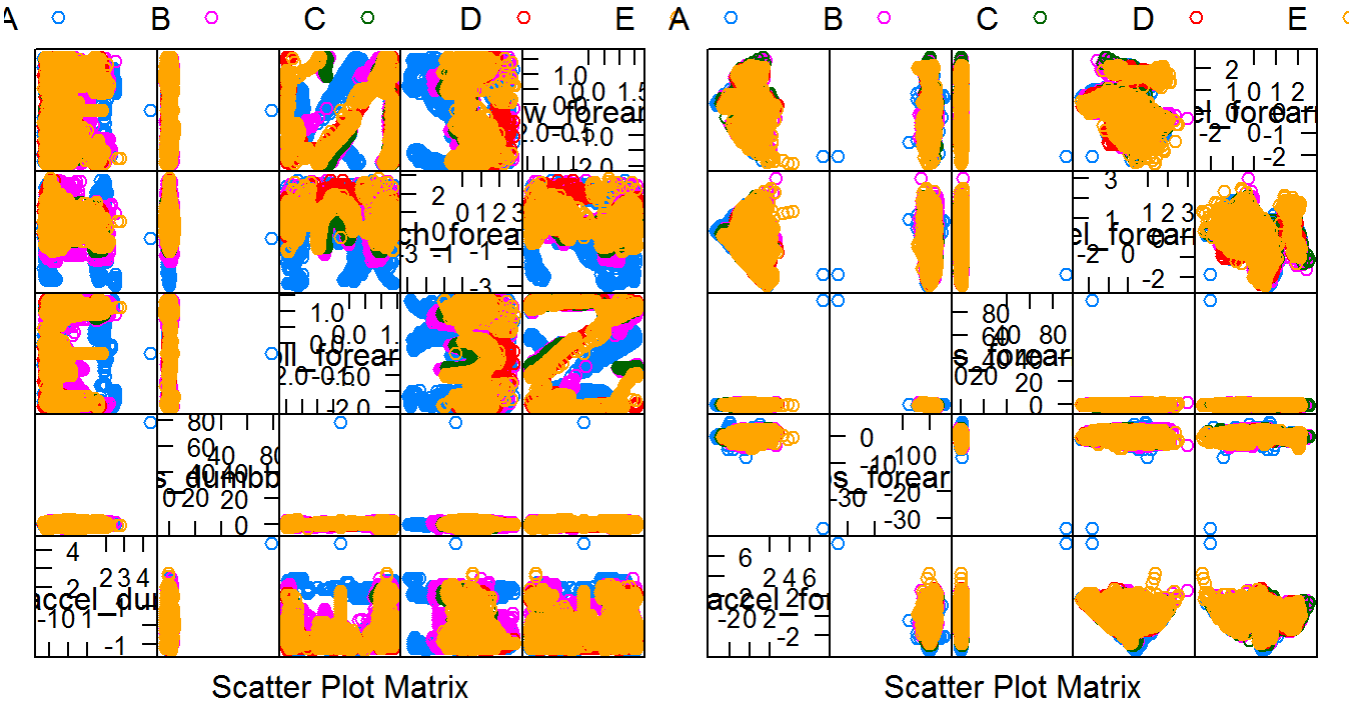
Scatter Plot Matrix



Scatter Plot Matrix



Scatter Plot Matrix



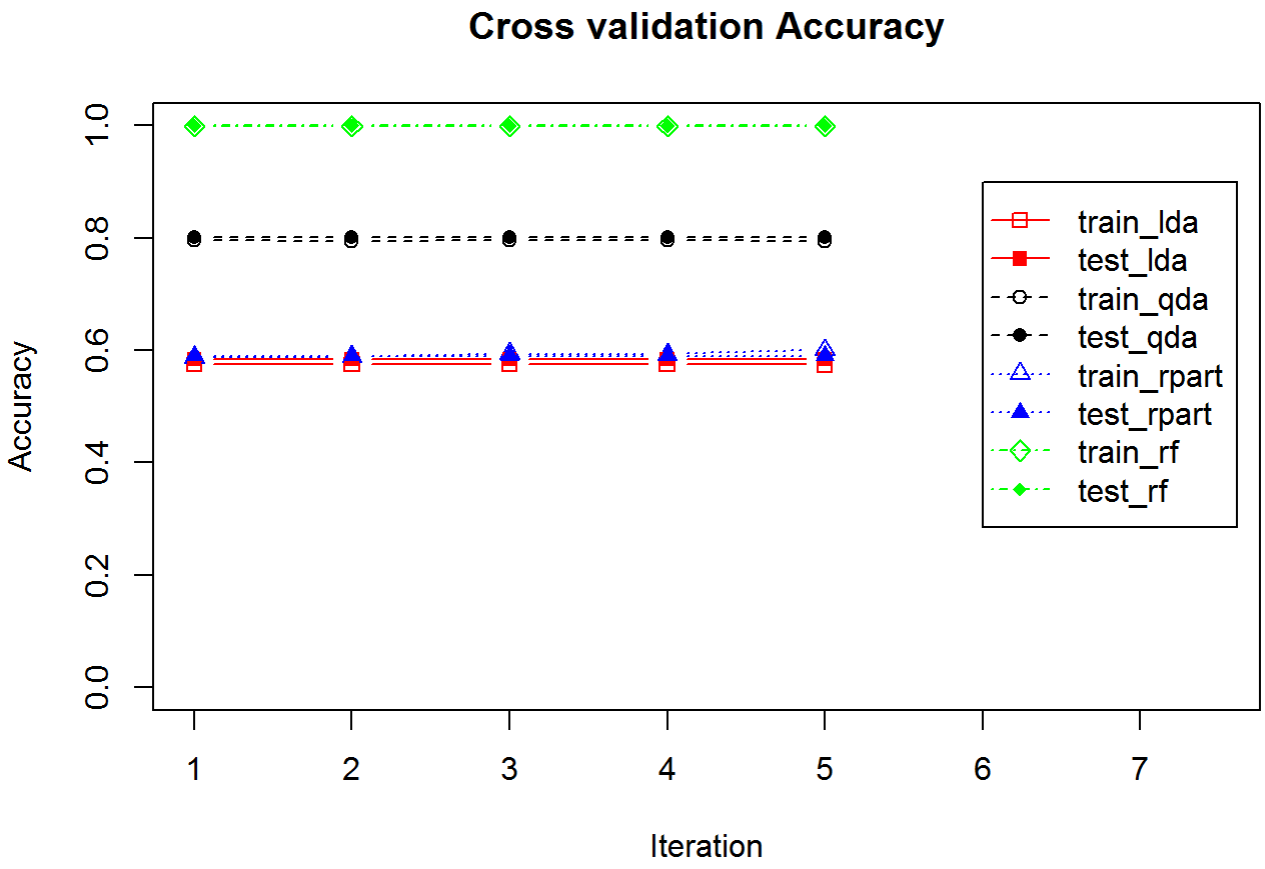
Modelling, model assessment and model selection

The “classe” variable will be predicted using the following classification methods:

1. linear discriminat analysis (assuming common variance across classes)
2. quadratic discriminat analysis (assuming class specific variance)
3. classification and regression trees (rpart and random forests)

In addition a 10 fold cross validation method is used and out of sample error is calculated for final model selection. I expect the “out of sample”" error to be greater than the “in sample error” in all models.

Results



Final model

The random forest model gives the most accurate prediction of the classe variable in the validation dataset with an average prediction accuracy (across 5 iterations) of 1 versus 0.5838164 from the linear discriminant analysis model, 0.8007246 from the quadratic discriminant analysis model, and 0.5903382 from the classification and regression trees model (rpart).

Predictions

Random forest model predictions on the test dataset are: B, A, B, A, A, E, D, B, A, A, B, C, B, A, E, E, A, B, B, B