1. Built the model

I built two models using this dataset. One is Random Forest, another is Decision Tree. The dependent variable is “classe”, which is a factor of five levels (A, B, C, D and E). I used all variables to build the model, except these with nothing but missing values. After cleaning data, I did cross validation by splitting the training data into two parts: one part is to fit the model, another part is to predict the results and test the accuracy of prediction. Then compare the accuracy of two models, the model with higher accuracy is chosen to be the best model.

2. Cross-validation

Cross-validation is done by splitting training data set randomly without replacement into 2 parts: subTrain data (70% of the original training dataset) and subTest data (30% of the original training dataset). SubTrain is used to fit the model, and subTest is for prediction and comparsion.

3. Expected out-of-sample error

The expected out-of-sample error is (1 – accuracy) in the subTest dataset.
After cross validation, I got subTrain and subTest datasets. I fitted the model using subTrain dataset, and then I can use the model to predict the results using subTest dataset. Then compare the results from the model and from the original subTest dataset, I know the expected accuracy of this model. The expected value of out-of-sample error will correspond to the expected number of missclassified observations in the subTest dataset. Therefore the expected out-of-sample error is (1 – accuracy) in the subTest data.

4. Reasons for my choices

The dependent variable “classe” is a factor variable so I can build confusion matrix to compare whether the predicted is the same as the original in the test dataset. The amount of the same is accuracy then I can also get the out-of-sample error.
When building the model, the huge dataset is the base for cross validation, which can help avoid overfitting. It is important to predict the results using subTest dataset first and compare accuracy.