**Coursera machine learning exercise**

**Title:** Can accelerometer measures predict how well individuals exercise?

##Disclosure: I encountered a number of problems with the “caret” and “rmarkdown” library. Libraries are either corrupted or do not work well with my operating system. I let “coursera” know. For this reason I was unable to generate the html file.##

The goal of this exercise was to use data from accelerometers on the belt, forearm, arm and dumbbell of 6 participants to predict their performance on barbell lifting. The participants were instructed to lift barbells correctly and incorrectly in 5 different ways. The variable “classe” is the categorical variable we want to predict and includes information on how well individuals exercised.

I downloaded the training and testing datasets as instructed and then partitioned the training dataset into two sets: a training (mytraining; 60% of the initial dataset) and a testing (mytesting) sets. I excluded variables with more than 60% of missing values, and zero variance variables, e.g. variables containing “0” observations and/or useless predictors. I removed the ID column from mytraining as it was of no use for further prediction purposes. I then made sure that both mytraining, mytesting, and testing would contain the same set of variables and adjusted the three datasets (coerce) so that they are of the same data type.

Then I decided to create 2 models of the data to estimate their accuracy on unseen data. To start with I set up the test harness to use a 3-fold cross validation. This split the dataset into 3 parts, train in 2 and test on 1, and release for all combinations of train-test split in an effort to get a more accurate estimate. I used the metric of “FitControl” to evaluate the models. We will be using the metric variable when we run build and evaluate each model next.

To build the final model I evaluated 2 algorithms: “generalized boosted regression” (gbr) and “Random Forest” (rf). Gbr produces a prediction model in the form of an ensemble of prediction models, typically decision trees.it builds a model in a stage-wise fashion using boosting algorithms. These algorithms optimize a cost function over function space by teaching a model to predict values by minimizing the mean squared error averaged over some training set of actual values of the output variable. I then generated a probability matrix for each of the observations of the mytesting dataset. rf uses multiple models for better performance and adjusts to a number of settings (e.g. large number of variables, can provide a measure of variable importance).

Based on an assessment of these 2 models fits and out-of sample results, and confusion matrix it looks like the random forest model outperform the rpart model.In mytesting dataset random forests gave an accuracy of 99.87% , which is greater than that observed with recursive partitioning (86.26%). The expected out-of-sample error is 100-99.89 = 0.11%.

I then applied to the random forest model to predict observations in the testing dataset.In other words I will use the testing dataset to predict classe results for each of the 20 observations based on the information collected in the validation sample.