Practical Machine Learning Project

## Performing Data Cleaning

* Read the pml training data
* Substitute empty cell value, "#DIV/0!" and "NA"" cell value to R NA value

library(caret)

## Loading required package: lattice  
## Loading required package: ggplot2

trainData = read.csv("pml-training.csv",header=TRUE,na.strings=c("","#DIV/0!","NA"))

* Get a logical matrix of the data frame where cell value is NA

NAData = data.frame(lapply(trainData,function(x){is.na(x)}),stringsAsFactors=FALSE)

* Dropped the column where it has more than 50% of missing data

ratio = colSums(NAData)/nrow(NAData)  
filterTrainData = trainData[,names(ratio[ratio <= 0.5])]

* Drop the first 5 column data that should not impact on the model. The X observation index column, user name column and 3 timestamp columns

filterTrainData = subset(filterTrainData,select=-c(X,user\_name,raw\_timestamp\_part\_1,raw\_timestamp\_part\_2,cvtd\_timestamp))

* Create a cross validation dataset with 70%:30% ratio

train = createDataPartition(y=filterTrainData$classe,p=0.7,list=FALSE)  
cvTrain = filterTrainData[train,]  
cvTest = filterTrainData[-train,]  
set.seed(647)

## Model Building

* Use random forest to model the data with column "classe" as the label data
* Use out of bag (oob) resampling method to estimate the out of sample error to give an unbiased estimate of the test set error.
* Search for the best mtry parameter value that give lowest oob error for 10 different setting

trControl = trainControl(method = "oob",allowParallel = TRUE)  
oob = train(classe ~.,data=cvTrain,method="rf",trControl=trControl,tuneLength=10,verbose=TRUE)

Final model parameters:

* mtry value:

## [1] 19

* Accuracy:

## [1] 0.998

* Out of bag/sample error:

## [1] 0.001965

As the model is trained with OOB resampling method, the accuracy of the model is calculated using (1- oob error)

Cross Validation:

* In random forests, there is no need for cross-validation or a separate test set as the out of bag error is estimated using the sample not used in the tree construction.
* The purpose of this exercise is to prove it using the initialy sampled 30% training data
* The prediction accuracy on the cross validation data should be close to the final model reported accuracy

## Loading required package: randomForest  
## randomForest 4.6-7  
## Type rfNews() to see new features/changes/bug fixes.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 1674 0 0 0 0  
## B 0 1139 0 0 0  
## C 0 0 1026 1 0  
## D 0 0 0 963 2  
## E 0 0 0 0 1080  
##   
## Overall Statistics  
##   
## Accuracy : 0.999   
## 95% CI : (0.999, 1)  
## No Information Rate : 0.284   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.999   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 1.000 1.000 1.000 0.999 0.998  
## Specificity 1.000 1.000 1.000 1.000 1.000  
## Pos Pred Value 1.000 1.000 0.999 0.998 1.000  
## Neg Pred Value 1.000 1.000 1.000 1.000 1.000  
## Prevalence 0.284 0.194 0.174 0.164 0.184  
## Detection Rate 0.284 0.194 0.174 0.164 0.184  
## Detection Prevalence 0.284 0.194 0.175 0.164 0.184  
## Balanced Accuracy 1.000 1.000 1.000 0.999 0.999