# PML-Proj

Anand

6/17/2020

## Practical Machine Learning

## Course Project

DATA PREPROCESSING

```
# required libraries
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
# read the training data into training data frame while excluding columns with NA, blank or division by zero
training <- read.csv("c:\\Users\\Dell\\Documents\\pml-training.csv", na.strings = c("NA","","#DIV/0!"))

# remove blank columns
na_cols <- which(colSums(is.na(training)) > 0)
training <- training[,-na_cols]

# consider only those rows containing new_windows = 'no'
training <- training[training$new_window=="no",]


# read the test data into test_set data frame
test_set <- read.csv("c:\\Users\\Dell\\Documents\\pml-testing.csv")

# remove blank columns
na_cols <- which(colSums(is.na(test_set)) > 0)
test_set <- test_set[,-na_cols]

# remove a few variables that do not seem to be predictors
training <- training[,-c(1:7)]
test_set <- test_set[,-c(1:7)]

# size of training set
dim(training)
```

```
## [1] 19216    53
```

```
# size of test set
dim(test_set)
```

```
## [1] 20 53
```

```
# END OF DATA PREPROCESSING
```

DATA ANALYSIS USING MACHINE LEARNING TECHNIQUES

It is mentioned in the course that random forest and AdaBoost are the most popular techniques. So, the random forest approach is used for classification in this project.

```r
# set the seed value for replicability
set.seed(455)

# carate training and test sets from the training set given by coursera
inTrain <- createDataPartition(training$classe, p=0.7, list=FALSE)
new_train <- training[inTrain,]
new_test  <- training[-inTrain,]

modelRF <- train( classe ~ .,  method = "rf", ntree = 100, data = new_train )

# try the model on the new_test set to check the accuracy
predictions <- predict( modelRF, newdata = new_test )

# confusion matrix
confusionMatrix( predictions, new_test$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1632   11    0    0    0
##          B    6 1102    2    0    1
##          C    2    2 1000    9    3
##          D    0    0    3  934    1
##          E    1    0    0    1 1053
##
## Overall Statistics
##
##                Accuracy : 0.9927
##                  95% CI : (0.9902, 0.9947)
##     No Information Rate : 0.2847
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9908
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9945   0.9883   0.9950   0.9894   0.9953
## Specificity            0.9973   0.9981   0.9966   0.9992   0.9996
## Pos Pred Value         0.9933   0.9919   0.9843   0.9957   0.9981
## Neg Pred Value         0.9978   0.9972   0.9989   0.9979   0.9989
## Prevalence             0.2847   0.1935   0.1744   0.1638   0.1836
## Detection Rate         0.2832   0.1912   0.1735   0.1621   0.1827
## Detection Prevalence   0.2851   0.1928   0.1763   0.1628   0.1831
## Balanced Accuracy      0.9959   0.9932   0.9958   0.9943   0.9974
```

```r
# error in classification.
classif_error <- (1 - mean(predictions == new_test$classe)) * 100

# classification erro in percentage
classif_error
```

```
## [1] 0.7287871
```

```r
# END OF MODEL BUILDING
```

# PREDICTION ON THE TEST SET

```r
# Prediction on the actual testing dataset provided by coursera
predictions_test_set <- predict(modelRF, newdata=test_set)

# print the predictions for each problem ID in the test set
predictions_test_set
```

```
##  [1] B A B A A E D B A A B C B A E E A B B B
## Levels: A B C D E
```

CONCLUSION

Using a random forest with 100 trees, the classification was attempted. An accuracy of 0.9929 was noticed. It was a lot higher than the “No Information Rate” of 0.2847. So, it could be considered a good model on the training set. A classification error of 0.7 % was noticed on the test set built out of training data. It could be considered as a validation set.

On the test set provided by coursera, the following classifications were obtained: B A B A A E D B A A B C B A E E A B B B

If the number of trees in the forest are increased and the mtry parameter is tweaked, the accuracy might improve further.