

# Practical Machine Learning Course

## Proejct

RS

13 9 2020

## R Markdown

### Background

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: <http://web.archive.org/web/20161224072740/http://groupware.les.inf.puc-rio.br/har> (<http://web.archive.org/web/20161224072740/http://groupware.les.inf.puc-rio.br/har>) (see the section on the Weight Lifting Exercise Dataset).

### Data Preparation

Load the packages and the training and test data. Create partition with the training data set

```
#Load libraries  
library(caret)
```

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

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

```
library(corrplot)
```

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

```
## corrplot 0.84 loaded
```

```
library(rpart)  
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.3
```

```
library(rattle)
```

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

```
## Loading required package: tibble
```

```
## Loading required package: bitops
```

```
## Rattle: A free graphical interface for data science with R.  
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.  
## Geben Sie 'rattle()' ein, um Ihre Daten mischen.
```

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##  
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:rattle':  
##  
##     importance
```

```
## The following object is masked from 'package:ggplot2':  
##  
##     margin
```

```
library(gbm)
```

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

```
## Loaded gbm 2.1.8
```

```
# Load training and test data
myTrain <- read.csv("pml-training.csv")
myTest <- read.csv("pml-testing.csv")

#Create partition with the training data set
inTrain <- createDataPartition(myTrain$classe, p=0.7, list=FALSE)
myTrainSet <- myTrain[inTrain, ]
myTestSet <- myTrain[-inTrain, ]

dim(myTrainSet)
```

```
## [1] 13737 160
```

```
dim(myTestSet)
```

```
## [1] 5885 160
```

```
#Remove variables with nearly zero variance AND ariables that are almost always NA

myNZV <- nearZeroVar(myTrainSet)
myTrainSet <- myTrainSet[, -myNZV]
myTestSet <- myTestSet[, -myNZV]

myMNA <- sapply(myTrainSet, function(x) mean(is.na(x))) > 0.95
myTrainSet <- myTrainSet[, myMNA==F]
myTestSet <- myTestSet[, myMNA==F]

# remove ID only variables: columns 1 - 5
myTrainSet <- myTrainSet[, -(1:5)]
myTestSet <- myTestSet[, -(1:5)]

dim(myTrainSet)
```

```
## [1] 13737 54
```

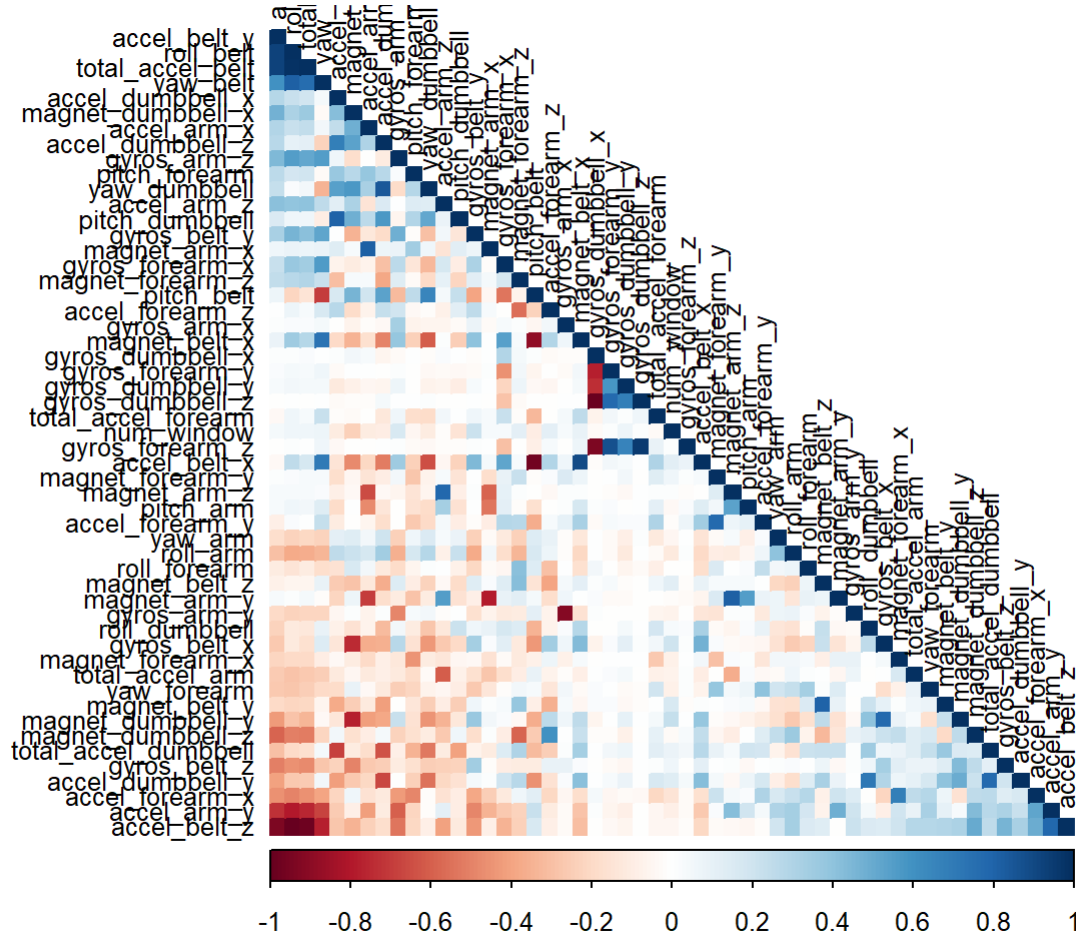
```
dim(myTestSet)
```

```
## [1] 5885 54
```

## Correlation Analysis

Perform a correlation analysis among variables before proceeding with the modeling procedure. Correlated variables/predictors are shown in dark colors

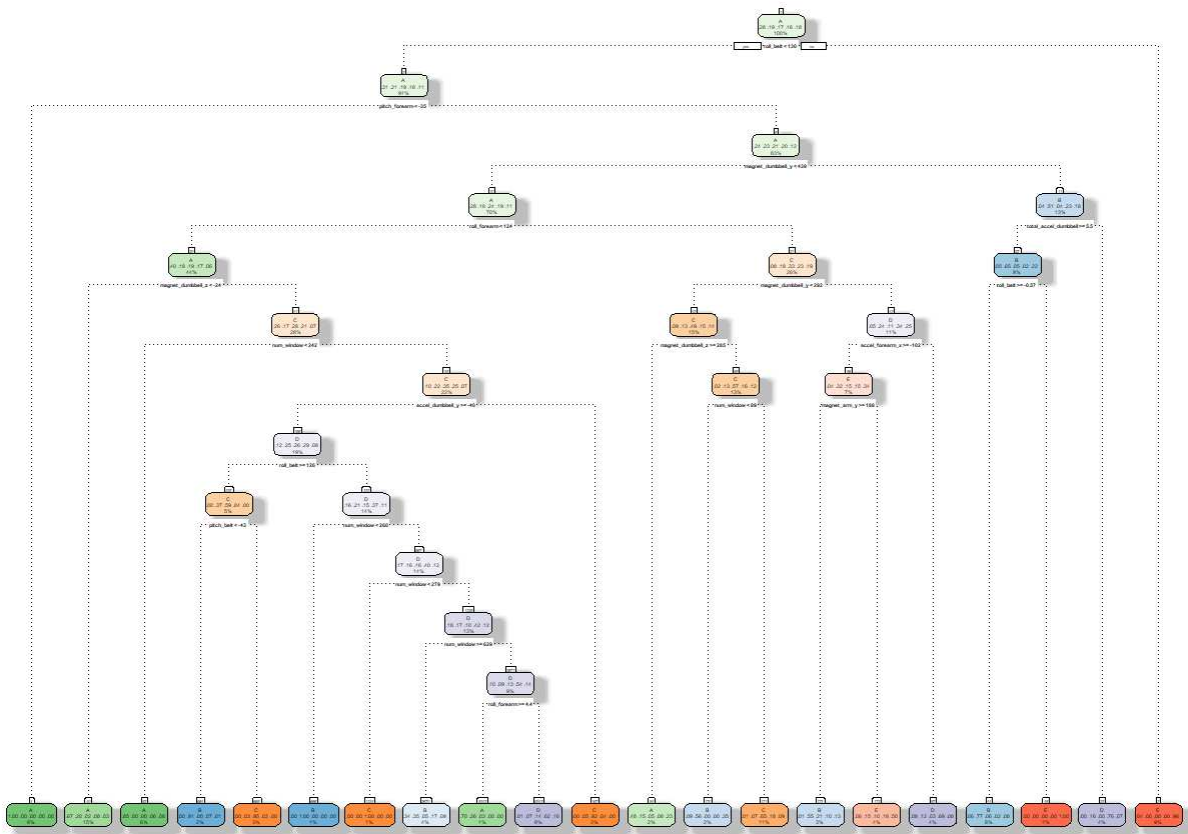
```
myCorMatrix <- cor(myTrainSet[, -54])
corrplot(myCorMatrix, order = "FPC", method = "color", type = "lower",
         tl.cex = 0.8, tl.col = rgb(0, 0, 0))
```



## Model Evaluation / Building

### I : Decision Trees

```
# Fit the model
set.seed(12345)
myDecTreeModel <- rpart(classe ~ ., data=myTrainSet, method="class")
fancyRpartPlot(myDecTreeModel)
```



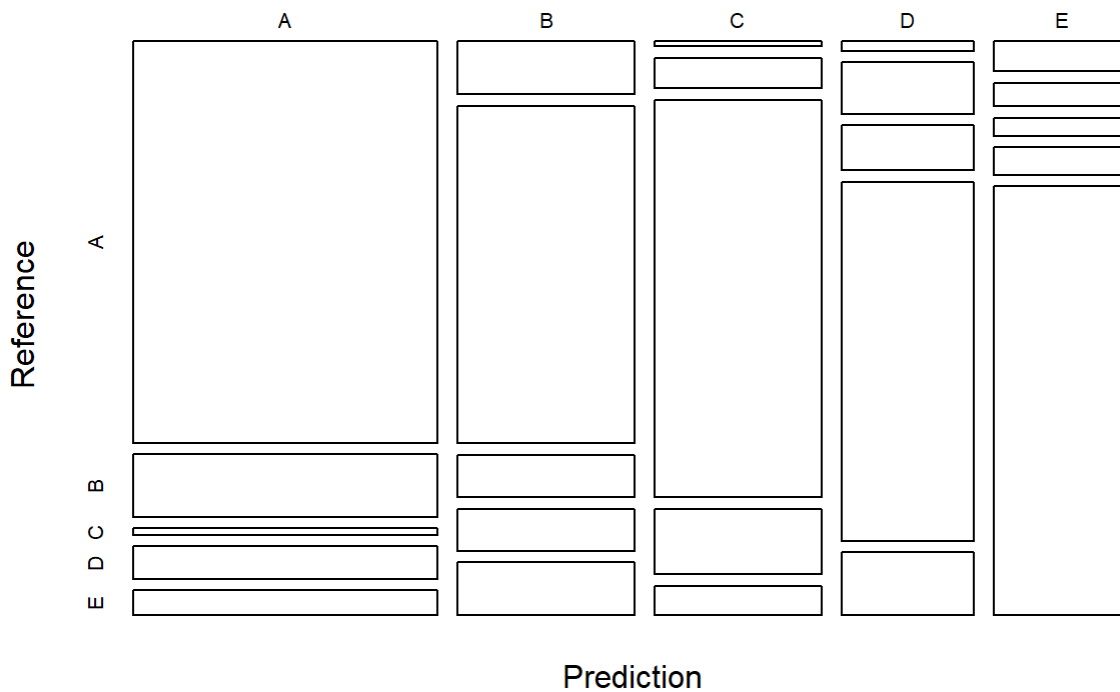
Rattle 2020-Sep-13 22:18:55 Ronnz

```
# Perform prediction on test data set
myDecTreePredict <- predict(myDecTreeModel, newdata=myTestSet, type="class")
myDecTreeConf <- confusionMatrix(myDecTreePredict, myTestSet$classe)
myDecTreeConf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   A     B     C     D     E
##           A 1485  230    25  119    93
##           B  114  725    92   90   114
##           C   10   62   807  133    60
##           D   15   83    72  576   101
##           E    50   39    30   46   714
##
## Overall Statistics
##
##           Accuracy : 0.7319
##           95% CI : (0.7203, 0.7431)
##           No Information Rate : 0.2845
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.6587
##
## Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity      0.8871  0.6365  0.7865  0.59751  0.6599
## Specificity      0.8891  0.9136  0.9455  0.94493  0.9656
## Pos Pred Value   0.7608  0.6388  0.7528  0.68005  0.8123
## Neg Pred Value   0.9519  0.9128  0.9545  0.92299  0.9265
## Prevalence       0.2845  0.1935  0.1743  0.16381  0.1839
## Detection Rate   0.2523  0.1232  0.1371  0.09788  0.1213
## Detection Prevalence 0.3317  0.1929  0.1822  0.14393  0.1494
## Balanced Accuracy 0.8881  0.7751  0.8660  0.77122  0.8128
```

```
# Plot the matrix results
plot(myDecTreeConf$table, col = myDecTreeConf$byClass,
     main = paste("Decision Tree Accuracy =",
                  round(myDecTreeConf$overall['Accuracy'], 3)))
```

## Decision Tree Accuracy = 0.732



: Random Forest

```
#Fit the model

set.seed(12345)
myControlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)
myRFModFit <- train(classe ~ ., data=myTrainSet, method="rf",
                    trControl=myControlRF)
myRFModFit$finalModel
```

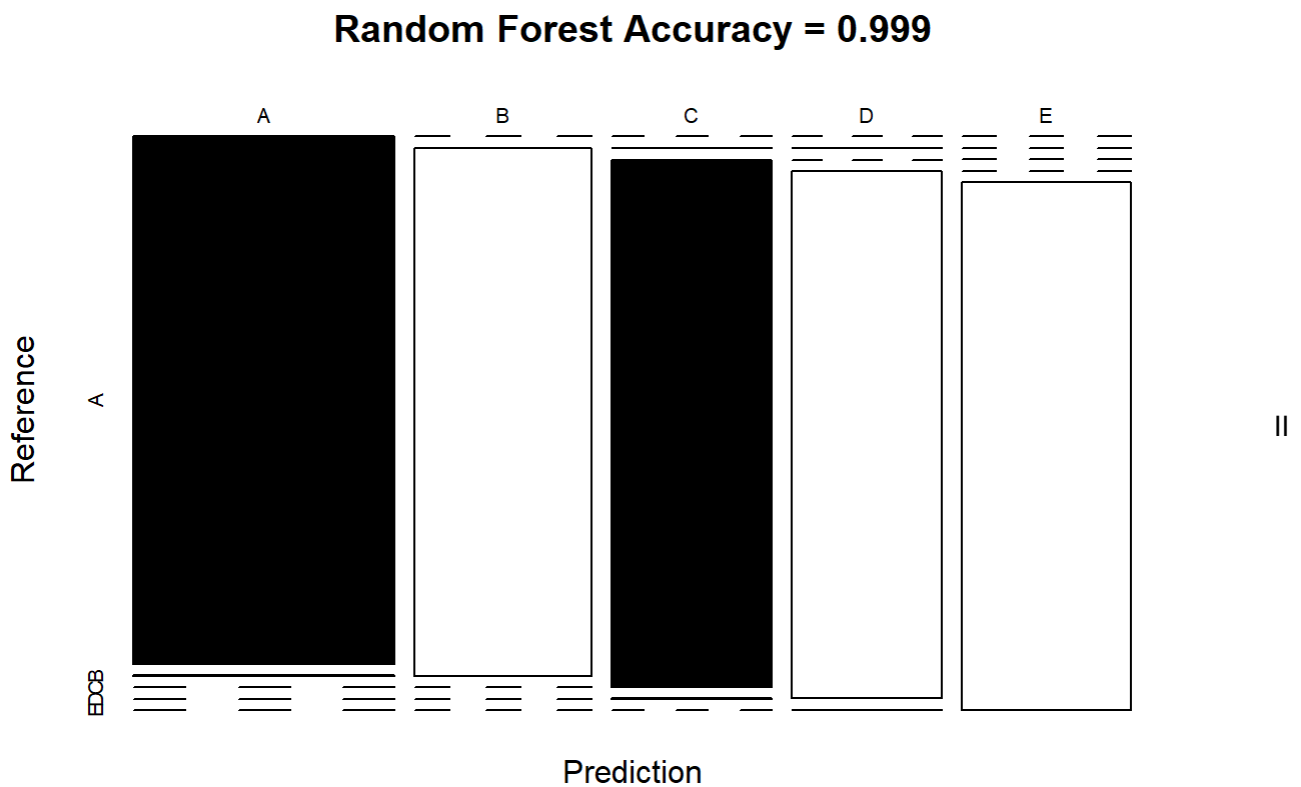
```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: classification
##                Number of trees: 500
## No. of variables tried at each split: 27
##
##                OOB estimate of  error rate: 0.23%
## Confusion matrix:
##      A      B      C      D      E  class.error
## A 3904      2      0      0      0 0.0005120328
## B      5 2651      2      0      0 0.0026335591
## C      0      5 2391      0      0 0.0020868114
## D      0      0  11 2240      1 0.0053285968
## E      0      0      0      5 2520 0.0019801980
```

```
# Perform prediction on test data set
myRFPredict <- predict(myRFModFit, newdata=myTestSet)
myRFConf <- confusionMatrix(myRFPredict, myTestSet$classe)
myRFConf
```



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   A     B     C     D     E
##           A 1674     1     0     0     0
##           B     0 1136     0     0     0
##           C     0     1 1026     1     0
##           D     0     1     0  963     1
##           E     0     0     0     0 1081
##
## Overall Statistics
##
##           Accuracy : 0.9992
##           95% CI : (0.998, 0.9997)
##           No Information Rate : 0.2845
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.9989
##
## McNemar's Test P-Value : NA
##
## Statistics by Class:
##
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity       1.0000   0.9974   1.0000   0.9990   0.9991
## Specificity       0.9998   1.0000   0.9996   0.9996   1.0000
## Pos Pred Value    0.9994   1.0000   0.9981   0.9979   1.0000
## Neg Pred Value    1.0000   0.9994   1.0000   0.9998   0.9998
## Prevalence        0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate    0.2845   0.1930   0.1743   0.1636   0.1837
## Detection Prevalence 0.2846   0.1930   0.1747   0.1640   0.1837
## Balanced Accuracy  0.9999   0.9987   0.9998   0.9993   0.9995
```

```
# Plot the matrix results
plot(myRFConf$table, col = myRFConf$byClass,
     main = paste("Random Forest Accuracy =",
                  round(myRFConf$overall['Accuracy'], 3)))
```



: Generalized Boosted Model

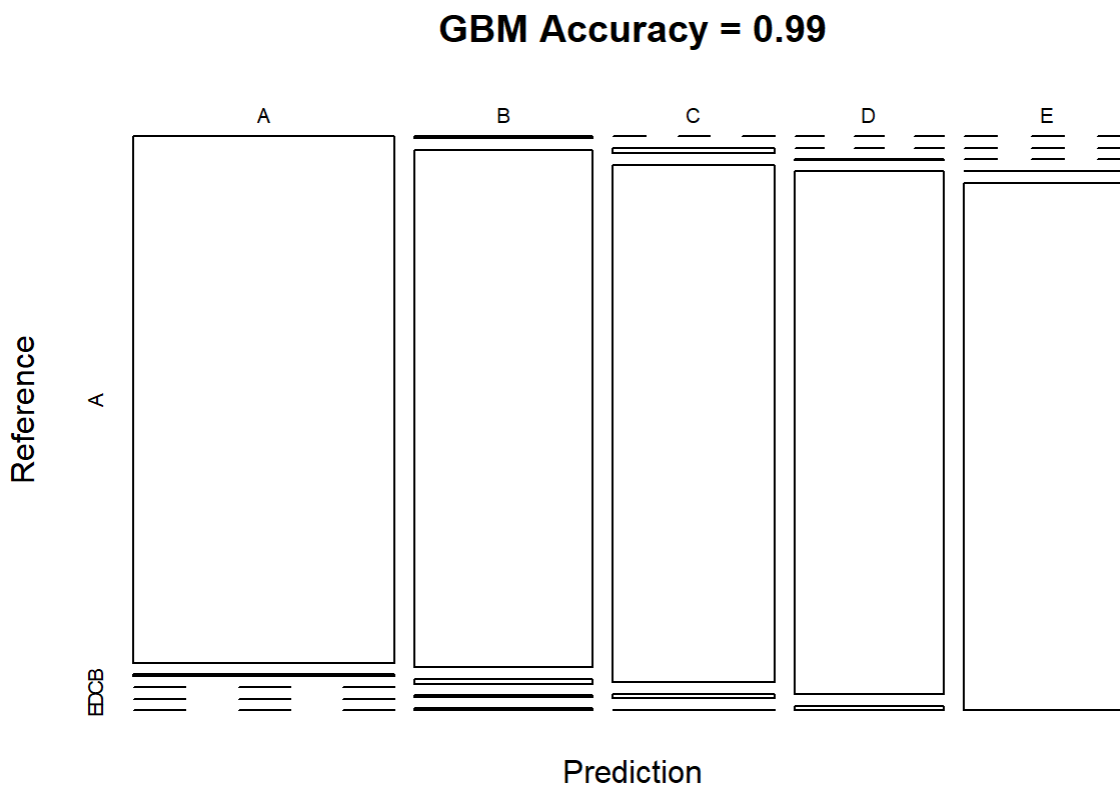
```
#Fit the model
set.seed(12345)
myControlGBM <- trainControl(method = "repeatedcv", number = 5, repeats = 1)
myGBMmodfit <- train(classe ~ ., data=myTrainSet, method = "gbm",
                     trControl = myControlGBM, verbose = FALSE)
myGBMmodfit$finalModel
```

```
## A gradient boosted model with multinomial loss function.
## 150 iterations were performed.
## There were 53 predictors of which 53 had non-zero influence.
```

```
#Perform prediction on test data set
myGBMPrediction <- predict(myGBMmodfit, newdata=myTestSet)
myGBMconfm <- confusionMatrix(myGBMPrediction, myTestSet$classe)
myGBMconfm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   A      B      C      D      E
##           A 1670      5      0      0      0
##           B   4 1123     11     4     4
##           C   0   11 1014      9     1
##           D   0    0      1   950     8
##           E   0    0      0    1 1069
##
## Overall Statistics
##
##           Accuracy : 0.99
##           95% CI : (0.9871, 0.9924)
##           No Information Rate : 0.2845
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.9873
##
## Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity      0.9976   0.9860   0.9883   0.9855   0.9880
## Specificity      0.9988   0.9952   0.9957   0.9982   0.9998
## Pos Pred Value   0.9970   0.9799   0.9797   0.9906   0.9991
## Neg Pred Value   0.9990   0.9966   0.9975   0.9972   0.9973
## Prevalence       0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate   0.2838   0.1908   0.1723   0.1614   0.1816
## Detection Prevalence 0.2846   0.1947   0.1759   0.1630   0.1818
## Balanced Accuracy 0.9982   0.9906   0.9920   0.9918   0.9939
```

```
# Plot the matrix results
plot(myGBMconfm$table, col = myGBMconfm$byClass,
     main = paste("GBM Accuracy =", round(myGBMconfm$overall['Accuracy'], 3)))
```



Applying the Selected Model to the Test Data. Since Random forest has best prediction use Random forest

```
#Random Forest model will be applied
myResults <- predict(myRFModFit, newdata=myTestSet)
myResults
```

[illegible]

[illegible]

[illegible]

[illegible]