Machine Learning Project

Project Introduction

Summary

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. In this project, our goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants and try to predict the manner in which they did the exercise. This is the “classe” variable in the training set. We will compare the accuracy of the models and choose the best model to apply on the given test set; the obtained results will be chosen as answers to the multiple choice quiz.

Data

The training data for this project are available here: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
The test data are available here: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

```r
#set library path
.libPaths("D:/R/R-3.5.0/library")

#Load Packages
library(caret)
library(rpart)
library(rpart.plot)
library(RColorBrewer)
library(rattle)
library(randomForest)
library(knitr)
```

Load the training and testing data.

```r
#LOAD DATA
pml_training_url <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
pml_testing_url <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

#read data and replace the empty/non existing cells with NA
pml_training <- read.csv(url(pml_training_url), na.strings=c("NA","#DIV/0!",""))
pml_testing <- read.csv(url(pml_testing_url), na.strings=c("NA","#DIV/0!",""))

rm(pml_training_url,pml_testing_url)
```

Split the pml_training data into train data & test data; our outcome is the “classe” variable.

pml_training: we will allocate 60% of to be our training data and the rest; 40% to be our test data
pml_testing: we will apply our selected best prediction model on this data and obtain the results we need to insert later in the multiple choice quiz.

```r
inTrain <- createDataPartition(pml_training$classe, p=0.6, list=FALSE) # 60% training and
#40% testing
my_training <- pml_training[inTrain, ]
```

```r
my_testing <- pml_training[-inTrain, ]

dim(my_training);
```

```
## [1] 11776   160
```

```r
dim(my_testing)
```

```
## [1] 7846  160
```

```r
rm(inTrain)
```

Cleaning data

```r
# Remove the extra first column; equivalent to row names.
my_training <- my_training[c(-1)]


#Remove predictors that have one unique value (i.e: zero variance predictors) or those that
#satisfy the following:
#1. Very few unique values relative to the number of samples
#2. Ratio of the frequency of the most common value to the frequency of the second most
#common value is large
nzv <- nearZeroVar(my_training, saveMetrics=TRUE)
my_training <- my_training[,nzv$nzv==FALSE]


nzv<- nearZeroVar(my_testing,saveMetrics=TRUE)
my_testing <- my_testing[,nzv$nzv==FALSE]



# Clean variables with more than 70% NA
index_to_rm <- NULL
for (i in 1:ncol(my_training))
{
if  ( length(which(is.na(my_training[,i])==T)) / nrow(my_training) >= 0.7 )
index_to_rm <- append(index_to_rm, i)
}

my_training <- my_training[,-index_to_rm]

rm(i, index_to_rm,nzv)




# Specify which columns to include in each of pml_testing & my_testing data frames for consistency with
#my_training.


# we will use classe variable to build the confusion matrix (together with our prediction)
columns_to_include_in_mytesting <- colnames(my_training)
```

```r
# since the classe variable doesn't exist in pml_testing
columns_to_include_in_pmltesting <- colnames(my_training[, -58])

# allow only variables in mytesting that are also in mytraining (this includes classe variable)
my_testing <- my_testing[columns_to_include_in_mytesting]

 # allow only variables in pml_testing that are also in mytraining (this doesn't include classe variabl
pml_testing <- pml_testing[columns_to_include_in_pmltesting]


rm(columns_to_include_in_mytesting,columns_to_include_in_pmltesting)
```

```r
#  print the dimensions of my_testing and pml_testing
dim(my_testing)
```

```
## [1] 7846   58
```

```r
dim(pml_testing)
```

```
## [1] 20 57
```

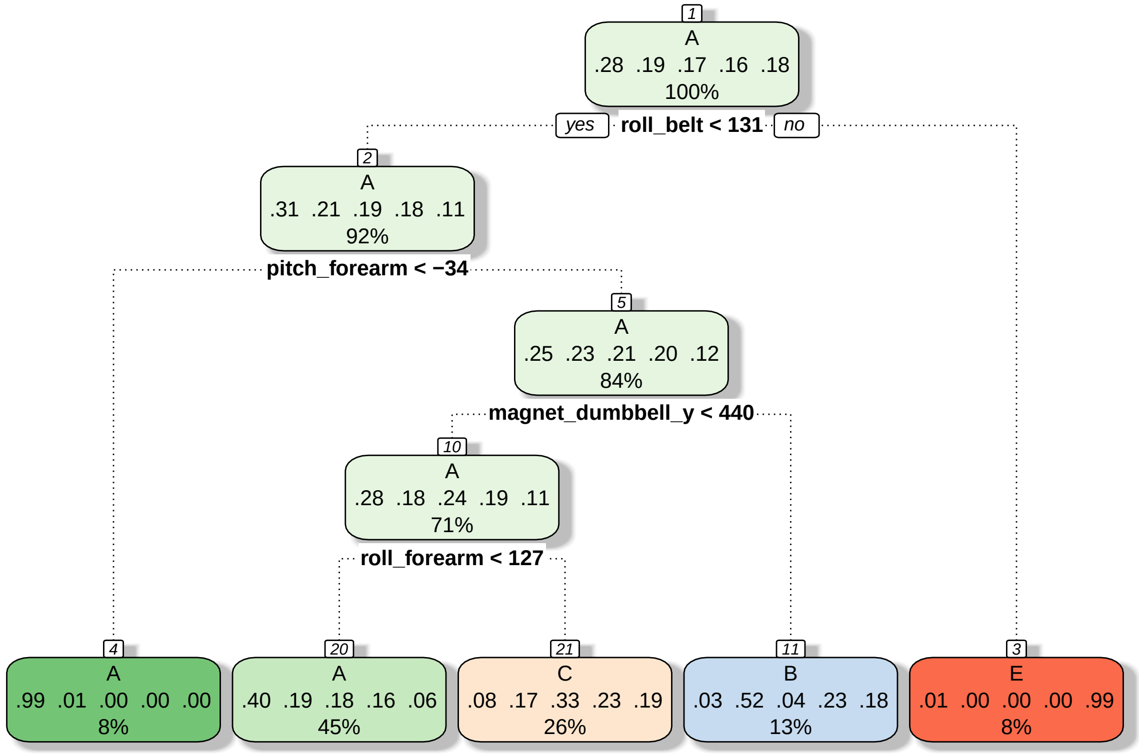
```r
#  Unifying data structures between my_training, my_testing and pml_testing by converting all
#integer variable types into numerics (for the commonly existing columns)



# To get the same class between testing and mytraining
my_testing <- rbind(my_training[2, ] , my_testing)
my_testing <- my_testing[-1,]



# To get the same class between testing and mytraining
pml_testing <- rbind(my_training[2, -58] , pml_testing)
pml_testing <- pml_testing[-1,]
```

Prediction

We will experiment with 3 prediction models and choose the one that gives us the highest accuracy and apply it on plm_testing in order to obtain our results for the quiz section.

Decision Trees

```r
set.seed(1)
model_1<- train(classe ~ ., data=my_training, method="rpart") #use all variables to predict the
#classe outcome.
fancyRpartPlot(model_1$finalModel)
```



Rattle 2019−Jan−16 16:38:24 RDajani

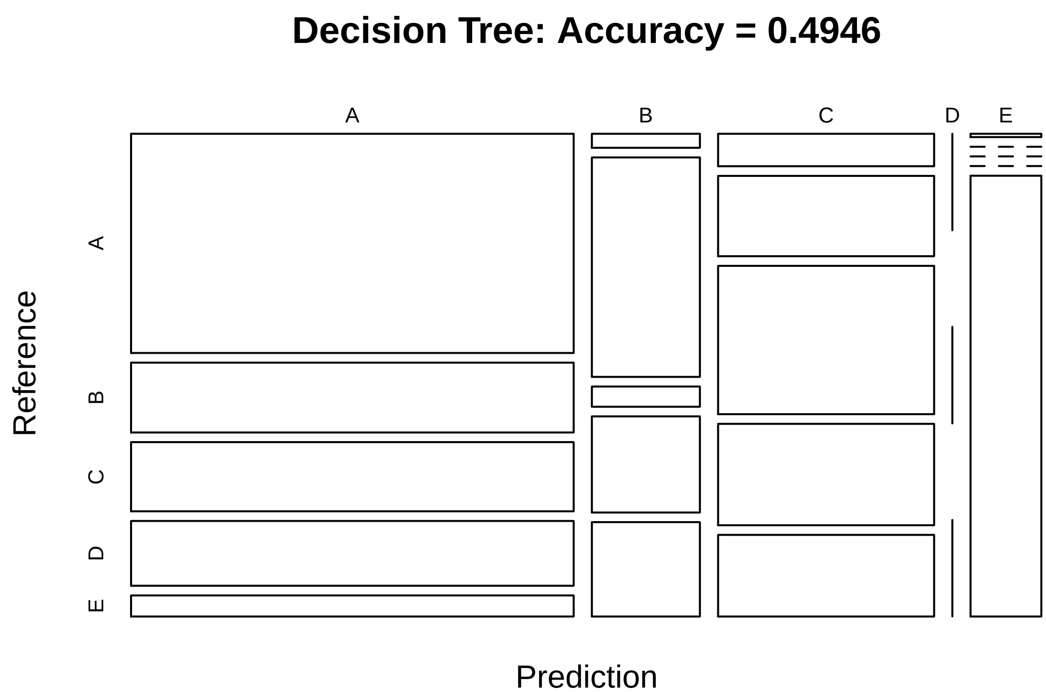
In order to test accuracy apply the model on the test set: my_testing

```
predictions1 <- predict(model_1, my_testing) #apply model on my_testing

#insepect accuracy and other statistics via confusionMatrix by
#comparing with the original my_testing data
cmtree <- confusionMatrix(predictions1, my_testing$classe)
cmtree
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2047  652  646  605  197
##          B   32  500   46  219  215
##          C  148  366  676  462  372
##          D    0    0    0    0    0
##          E    5    0    0    0  658
##
## Overall Statistics
##
##                Accuracy : 0.4946
##                  95% CI : (0.4835, 0.5058)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3387
```

```
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9171  0.32938  0.49415   0.0000  0.45631
## Specificity            0.6259  0.91909  0.79191   1.0000  0.99922
## Pos Pred Value         0.4936  0.49407  0.33399      NaN  0.99246
## Neg Pred Value         0.9500  0.85104  0.88114   0.8361  0.89085
## Prevalence             0.2845  0.19347  0.17436   0.1639  0.18379
## Detection Rate         0.2609  0.06373  0.08616   0.0000  0.08386
## Detection Prevalence   0.5285  0.12898  0.25797   0.0000  0.08450
## Balanced Accuracy      0.7715  0.62424  0.64303   0.5000  0.72776
```

```r
plot(cmtree$table, col = cmtree$byClass, main = paste("Decision Tree: Accuracy =",
                                                       round(cmtree$overall['Accuracy'], 4)))
```

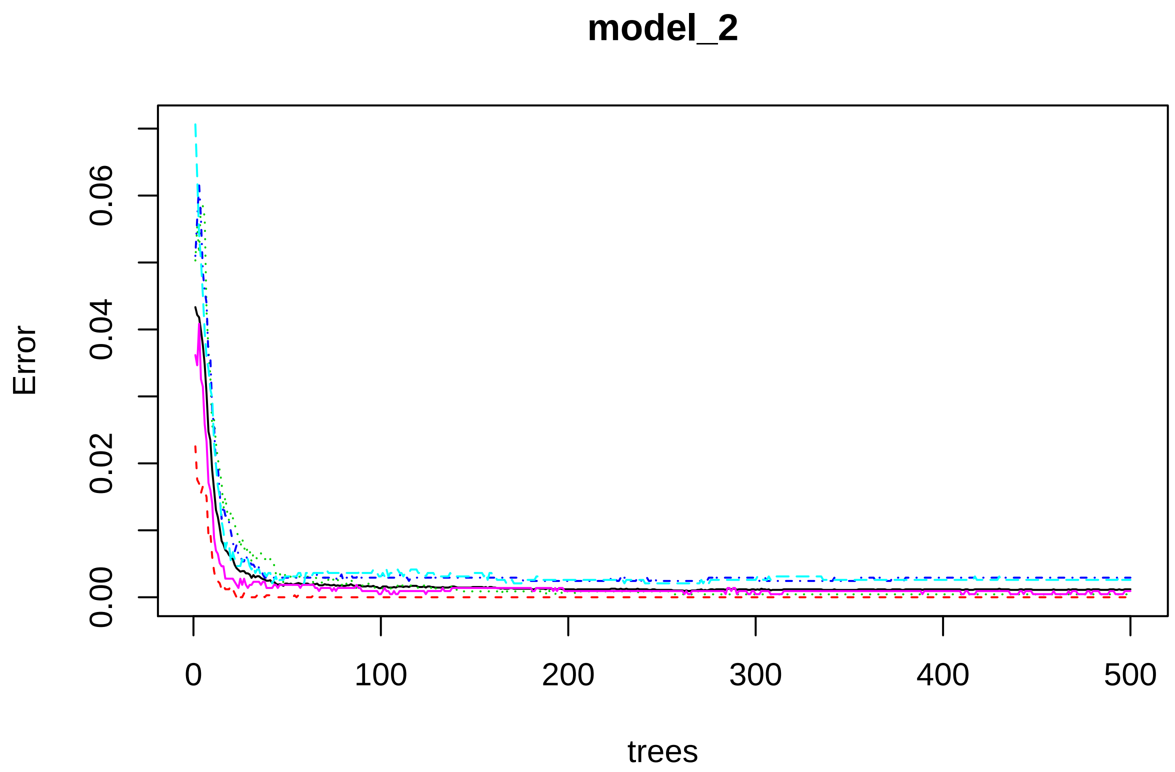


Random Forests

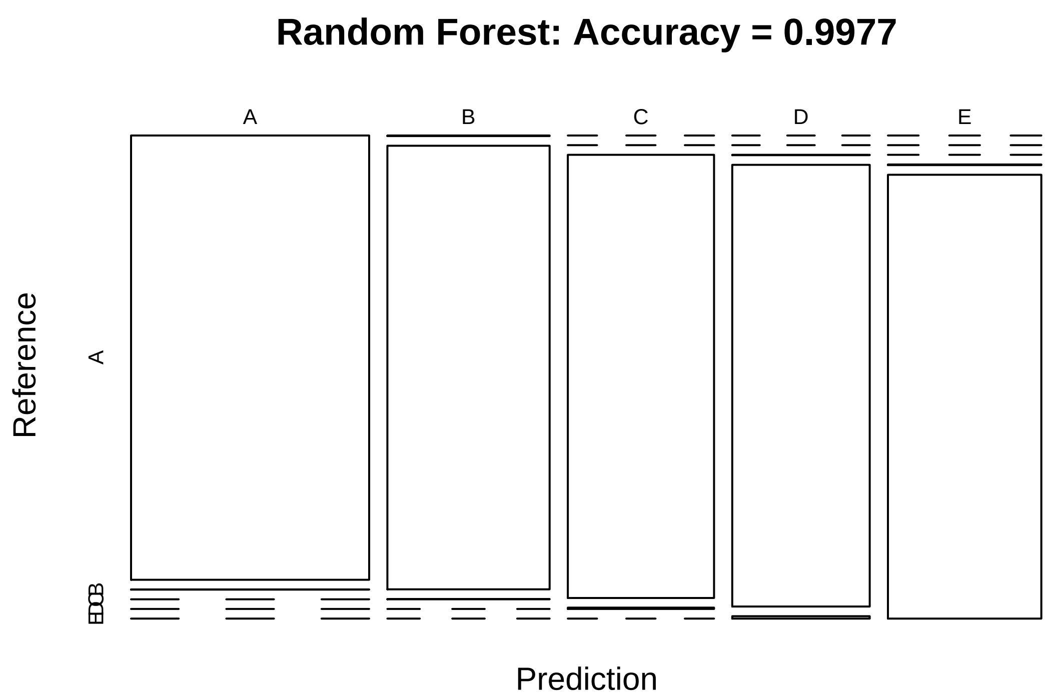
```r
set.seed(1)

model_2 <- randomForest(classe ~ ., data=my_training)
prediction2 <- predict(model_2, my_testing, type = "class")
cmrf <- confusionMatrix(prediction2, my_testing$classe)
cmrf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2230    1    0    0    0
##          B    2 1517    1    0    0
##          C    0    0 1366    4    0
##          D    0    0    1 1280    7
##          E    0    0    0    2 1435
##
## Overall Statistics
##
##                Accuracy : 0.9977
##                  95% CI : (0.9964, 0.9986)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9971
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9991   0.9993   0.9985   0.9953   0.9951
## Specificity            0.9998   0.9995   0.9994   0.9988   0.9997
## Pos Pred Value         0.9996   0.9980   0.9971   0.9938   0.9986
## Neg Pred Value         0.9996   0.9998   0.9997   0.9991   0.9989
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2842   0.1933   0.1741   0.1631   0.1829
## Detection Prevalence   0.2843   0.1937   0.1746   0.1642   0.1832
## Balanced Accuracy      0.9995   0.9994   0.9990   0.9971   0.9974
```

```
plot(model_2)
```




```
plot(cmrf$table, col = cmtree$byClass, main = paste("Random Forest: Accuracy =", round(cmrf$overall['Acc
```



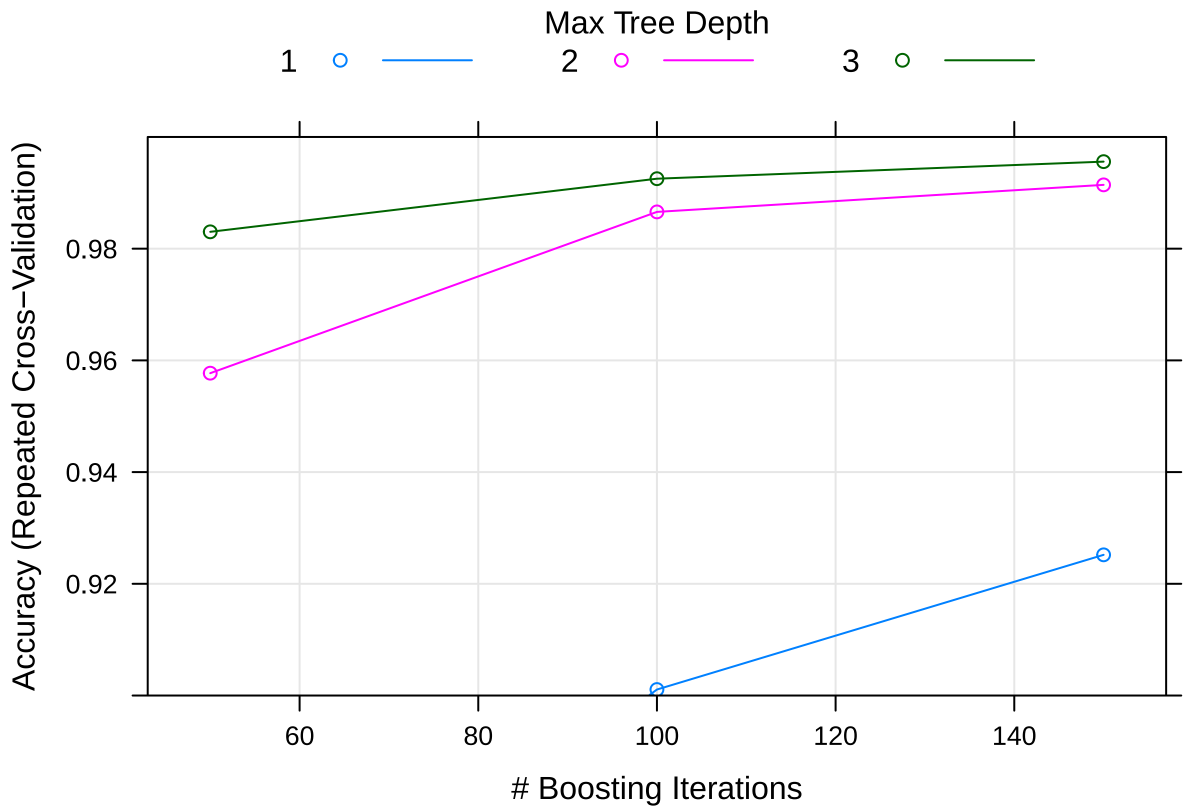
Boosting with Trees

```
set.seed(1)


fitControl <- trainControl(method = "repeatedcv",
                           number = 5,
                           repeats = 1)

model_3 <- train(classe ~ ., data=my_training, method = "gbm",
                 trControl = fitControl,
                 verbose = FALSE)



prediction3 <- predict(model_3, newdata=my_testing)
cmGBM <- confusionMatrix(prediction3, my_testing$classe)
cmGBM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2231    3    0    0    0
##          B    1 1514    3    0    0
```

```
##          C    0    1 1355    2    0
##          D    0    0   10 1281   11
##          E    0    0    0    3 1431
##
## Overall Statistics
##
##                Accuracy : 0.9957
##                  95% CI : (0.9939, 0.997)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9945
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9996   0.9974   0.9905   0.9961   0.9924
## Specificity            0.9995   0.9994   0.9995   0.9968   0.9995
## Pos Pred Value         0.9987   0.9974   0.9978   0.9839   0.9979
## Neg Pred Value         0.9998   0.9994   0.9980   0.9992   0.9983
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2843   0.1930   0.1727   0.1633   0.1824
## Detection Prevalence   0.2847   0.1935   0.1731   0.1659   0.1828
## Balanced Accuracy      0.9995   0.9984   0.9950   0.9965   0.9960
```

```
plot(model_3, ylim=c(0.9, 1))
```



Selecting the best model to apply on the Test Data

By comparing the accuracy among the 3 models, we notice that the Random Forests model gave us the highest accuracy 99.81%; thus the expected out-of-sample error is 100-99.81 = 0.19%. We choose RF to apply on the test data: plm_testing. The obtained results are the ones to be selected as answers to the multiple choice quiz.

```
prediction_model_2 <- predict(model_2, pml_testing, type = "class")
prediction_model_2
```

```
##  1  2  3  4  5  6  7 81  9 10 11 12 13 14 15 16 17 18 19 20
##  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
```