JHU_PracticalMachineLearning_Assignment4

Danilo Steckelberg

6/2/2020

Project description

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 dumbell 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 (see the section on the Weight Lifting Exercise Dataset).

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

Method

The analysis made in this project considers the following steps: 1) Downloading, loading and cleaning the data; 2) Testing Machine Learning algorithms, understanding the best predictor and using it to predict the test data set. The algorithms used in this analysis is Decision Trees and Random Forests.

These models are chosen since they are the ones presented in course that are most suitable for predicting classification objects.

Downloading, loading and cleaning the data

```
library(caret)
```

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

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

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

```
## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
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
```

```
#### Download the training and validation data files ####
dataset.url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
validation.url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

dataset.filename <- "pml-training.csv"
validation.filename <- "pml-testing.csv"

download.file(dataset.url, dataset.filename)
download.file(validation.url, validation.filename)

dataset <- read.csv(dataset.filename, na.strings = c("NA","#DIV/0!",""))
dataset.variables <- dim(dataset)[2]
dataset.entries <- dim(dataset)[1]

validation <- read.csv(validation.filename, na.strings = c("NA","#DIV/0!",""))
validation.variables <- dim(validation)[2]
validation.entries <- dim(validation)[1]

# Checking if variables are the same in both dataset and validation set
numOfVarsCheck <- ifelse(validation.variables == dataset.variables,TRUE,FALSE)
numOfMatchinVariables <- sum(ifelse(colnames(validation)==colnames(dataset),1,0))
if(numOfVarsCheck == TRUE & numOfMatchinVariables == dataset.variables-1){
  print("All variables equal except for output variable - We are good to go :)")
}else{warning("Variables Mismatch - please check variables")}
```

```
## [1] "All variables equal except for output variable - We are good to go :)"
```

```r
dim(dataset); dim(validation)
```

```
## [1] 19622   160
```

```
## [1]  20 160
```

```r
head(dataset)
```

```
##   X user_name raw_timestamp_part_1 raw_timestamp_part_2   cvtd_timestamp
## 1 1  carlitos           1323084231               788290 05/12/2011 11:23
## 2 2  carlitos           1323084231               808298 05/12/2011 11:23
## 3 3  carlitos           1323084231               820366 05/12/2011 11:23
## 4 4  carlitos           1323084232               120339 05/12/2011 11:23
## 5 5  carlitos           1323084232               196328 05/12/2011 11:23
## 6 6  carlitos           1323084232               304277 05/12/2011 11:23
##   new_window num_window roll_belt pitch_belt yaw_belt total_accel_belt
## 1         no         11      1.41       8.07    -94.4                3
## 2         no         11      1.41       8.07    -94.4                3
## 3         no         11      1.42       8.07    -94.4                3
## 4         no         12      1.48       8.05    -94.4                3
## 5         no         12      1.48       8.07    -94.4                3
## 6         no         12      1.45       8.06    -94.4                3
##   kurtosis_roll_belt kurtosis_picth_belt kurtosis_yaw_belt
## 1                 NA                  NA                NA
## 2                 NA                  NA                NA
## 3                 NA                  NA                NA
## 4                 NA                  NA                NA
## 5                 NA                  NA                NA
## 6                 NA                  NA                NA
##   skewness_roll_belt skewness_roll_belt.1 skewness_yaw_belt max_roll_belt
## 1                 NA                   NA                NA            NA
## 2                 NA                   NA                NA            NA
## 3                 NA                   NA                NA            NA
## 4                 NA                   NA                NA            NA
## 5                 NA                   NA                NA            NA
## 6                 NA                   NA                NA            NA
##   max_picth_belt max_yaw_belt min_roll_belt min_pitch_belt min_yaw_belt
## 1             NA           NA            NA             NA           NA
## 2             NA           NA            NA             NA           NA
## 3             NA           NA            NA             NA           NA
## 4             NA           NA            NA             NA           NA
## 5             NA           NA            NA             NA           NA
## 6             NA           NA            NA             NA           NA
##   amplitude_roll_belt amplitude_pitch_belt amplitude_yaw_belt
## 1                  NA                   NA                 NA
## 2                  NA                   NA                 NA
## 3                  NA                   NA                 NA
## 4                  NA                   NA                 NA
## 5                  NA                   NA                 NA
## 6                  NA                   NA                 NA
##   var_total_accel_belt avg_roll_belt stddev_roll_belt var_roll_belt
## 1                   NA            NA                NA            NA
```

```
## 2                   NA            NA               NA            NA
## 3                   NA            NA               NA            NA
## 4                   NA            NA               NA            NA
## 5                   NA            NA               NA            NA
## 6                   NA            NA               NA            NA
##   avg_pitch_belt stddev_pitch_belt var_pitch_belt avg_yaw_belt
## 1             NA                NA             NA           NA
## 2             NA                NA             NA           NA
## 3             NA                NA             NA           NA
## 4             NA                NA             NA           NA
## 5             NA                NA             NA           NA
## 6             NA                NA             NA           NA
##   stddev_yaw_belt var_yaw_belt gyros_belt_x gyros_belt_y gyros_belt_z
## 1              NA           NA         0.00         0.00        -0.02
## 2              NA           NA         0.02         0.00        -0.02
## 3              NA           NA         0.00         0.00        -0.02
## 4              NA           NA         0.02         0.00        -0.03
## 5              NA           NA         0.02         0.02        -0.02
## 6              NA           NA         0.02         0.00        -0.02
##   accel_belt_x accel_belt_y accel_belt_z magnet_belt_x magnet_belt_y
## 1          -21            4           22            -3           599
## 2          -22            4           22            -7           608
## 3          -20            5           23            -2           600
## 4          -22            3           21            -6           604
## 5          -21            2           24            -6           600
## 6          -21            4           21             0           603
##   magnet_belt_z roll_arm pitch_arm yaw_arm total_accel_arm var_accel_arm
## 1          -313     -128      22.5    -161              34            NA
## 2          -311     -128      22.5    -161              34            NA
## 3          -305     -128      22.5    -161              34            NA
## 4          -310     -128      22.1    -161              34            NA
## 5          -302     -128      22.1    -161              34            NA
## 6          -312     -128      22.0    -161              34            NA
##   avg_roll_arm stddev_roll_arm var_roll_arm avg_pitch_arm stddev_pitch_arm
## 1           NA              NA           NA            NA               NA
## 2           NA              NA           NA            NA               NA
## 3           NA              NA           NA            NA               NA
## 4           NA              NA           NA            NA               NA
## 5           NA              NA           NA            NA               NA
## 6           NA              NA           NA            NA               NA
##   var_pitch_arm avg_yaw_arm stddev_yaw_arm var_yaw_arm gyros_arm_x
## 1            NA          NA             NA          NA        0.00
## 2            NA          NA             NA          NA        0.02
## 3            NA          NA             NA          NA        0.02
## 4            NA          NA             NA          NA        0.02
## 5            NA          NA             NA          NA        0.00
## 6            NA          NA             NA          NA        0.02
##   gyros_arm_y gyros_arm_z accel_arm_x accel_arm_y accel_arm_z magnet_arm_x
## 1        0.00       -0.02        -288         109        -123         -368
## 2       -0.02       -0.02        -290         110        -125         -369
## 3       -0.02       -0.02        -289         110        -126         -368
## 4       -0.03        0.02        -289         111        -123         -372
## 5       -0.03        0.00        -289         111        -123         -374
## 6       -0.03        0.00        -289         111        -122         -369
```

```
##   magnet_arm_y magnet_arm_z kurtosis_roll_arm kurtosis_picth_arm
## 1          337          516                NA                 NA
## 2          337          513                NA                 NA
## 3          344          513                NA                 NA
## 4          344          512                NA                 NA
## 5          337          506                NA                 NA
## 6          342          513                NA                 NA
##   kurtosis_yaw_arm skewness_roll_arm skewness_pitch_arm skewness_yaw_arm
## 1               NA                NA                 NA               NA
## 2               NA                NA                 NA               NA
## 3               NA                NA                 NA               NA
## 4               NA                NA                 NA               NA
## 5               NA                NA                 NA               NA
## 6               NA                NA                 NA               NA
##   max_roll_arm max_picth_arm max_yaw_arm min_roll_arm min_pitch_arm
## 1           NA            NA          NA           NA            NA
## 2           NA            NA          NA           NA            NA
## 3           NA            NA          NA           NA            NA
## 4           NA            NA          NA           NA            NA
## 5           NA            NA          NA           NA            NA
## 6           NA            NA          NA           NA            NA
##   min_yaw_arm amplitude_roll_arm amplitude_pitch_arm amplitude_yaw_arm
## 1          NA                NA                  NA                NA
## 2          NA                NA                  NA                NA
## 3          NA                NA                  NA                NA
## 4          NA                NA                  NA                NA
## 5          NA                NA                  NA                NA
## 6          NA                NA                  NA                NA
##   roll_dumbbell pitch_dumbbell yaw_dumbbell kurtosis_roll_dumbbell
## 1      13.05217      -70.49400    -84.87394                     NA
## 2      13.13074      -70.63751    -84.71065                     NA
## 3      12.85075      -70.27812    -85.14078                     NA
## 4      13.43120      -70.39379    -84.87363                     NA
## 5      13.37872      -70.42856    -84.85306                     NA
## 6      13.38246      -70.81759    -84.46500                     NA
##   kurtosis_picth_dumbbell kurtosis_yaw_dumbbell skewness_roll_dumbbell
## 1                      NA                    NA                     NA
## 2                      NA                    NA                     NA
## 3                      NA                    NA                     NA
## 4                      NA                    NA                     NA
## 5                      NA                    NA                     NA
## 6                      NA                    NA                     NA
##   skewness_pitch_dumbbell skewness_yaw_dumbbell max_roll_dumbbell
## 1                      NA                    NA                NA
## 2                      NA                    NA                NA
## 3                      NA                    NA                NA
## 4                      NA                    NA                NA
## 5                      NA                    NA                NA
## 6                      NA                    NA                NA
##   max_picth_dumbbell max_yaw_dumbbell min_roll_dumbbell min_pitch_dumbbell
## 1                 NA               NA                NA                 NA
## 2                 NA               NA                NA                 NA
## 3                 NA               NA                NA                 NA
## 4                 NA               NA                NA                 NA
```

```
## 5                 NA               NA                NA                 NA
## 6                 NA               NA                NA                 NA
##   min_yaw_dumbbell amplitude_roll_dumbbell amplitude_pitch_dumbbell
## 1               NA                      NA                       NA
## 2               NA                      NA                       NA
## 3               NA                      NA                       NA
## 4               NA                      NA                       NA
## 5               NA                      NA                       NA
## 6               NA                      NA                       NA
##   amplitude_yaw_dumbbell total_accel_dumbbell var_accel_dumbbell
## 1                     NA                   37                 NA
## 2                     NA                   37                 NA
## 3                     NA                   37                 NA
## 4                     NA                   37                 NA
## 5                     NA                   37                 NA
## 6                     NA                   37                 NA
##   avg_roll_dumbbell stddev_roll_dumbbell var_roll_dumbbell
## 1                NA                   NA                NA
## 2                NA                   NA                NA
## 3                NA                   NA                NA
## 4                NA                   NA                NA
## 5                NA                   NA                NA
## 6                NA                   NA                NA
##   avg_pitch_dumbbell stddev_pitch_dumbbell var_pitch_dumbbell
## 1                 NA                    NA                 NA
## 2                 NA                    NA                 NA
## 3                 NA                    NA                 NA
## 4                 NA                    NA                 NA
## 5                 NA                    NA                 NA
## 6                 NA                    NA                 NA
##   avg_yaw_dumbbell stddev_yaw_dumbbell var_yaw_dumbbell gyros_dumbbell_x
## 1               NA                  NA               NA                0
## 2               NA                  NA               NA                0
## 3               NA                  NA               NA                0
## 4               NA                  NA               NA                0
## 5               NA                  NA               NA                0
## 6               NA                  NA               NA                0
##   gyros_dumbbell_y gyros_dumbbell_z accel_dumbbell_x accel_dumbbell_y
## 1            -0.02             0.00             -234               47
## 2            -0.02             0.00             -233               47
## 3            -0.02             0.00             -232               46
## 4            -0.02            -0.02             -232               48
## 5            -0.02             0.00             -233               48
## 6            -0.02             0.00             -234               48
##   accel_dumbbell_z magnet_dumbbell_x magnet_dumbbell_y magnet_dumbbell_z
## 1             -271              -559               293               -65
## 2             -269              -555               296               -64
## 3             -270              -561               298               -63
## 4             -269              -552               303               -60
## 5             -270              -554               292               -68
## 6             -269              -558               294               -66
##   roll_forearm pitch_forearm yaw_forearm kurtosis_roll_forearm
## 1         28.4         -63.9        -153                    NA
## 2         28.3         -63.9        -153                    NA
```

```
## 3         28.3         -63.9        -152                    NA
## 4         28.1         -63.9        -152                    NA
## 5         28.0         -63.9        -152                    NA
## 6         27.9         -63.9        -152                    NA
##   kurtosis_picth_forearm kurtosis_yaw_forearm skewness_roll_forearm
## 1                     NA                   NA                     NA
## 2                     NA                   NA                     NA
## 3                     NA                   NA                     NA
## 4                     NA                   NA                     NA
## 5                     NA                   NA                     NA
## 6                     NA                   NA                     NA
##   skewness_pitch_forearm skewness_yaw_forearm max_roll_forearm
## 1                     NA                   NA               NA
## 2                     NA                   NA               NA
## 3                     NA                   NA               NA
## 4                     NA                   NA               NA
## 5                     NA                   NA               NA
## 6                     NA                   NA               NA
##   max_picth_forearm max_yaw_forearm min_roll_forearm min_pitch_forearm
## 1                NA              NA               NA                NA
## 2                NA              NA               NA                NA
## 3                NA              NA               NA                NA
## 4                NA              NA               NA                NA
## 5                NA              NA               NA                NA
## 6                NA              NA               NA                NA
##   min_yaw_forearm amplitude_roll_forearm amplitude_pitch_forearm
## 1              NA                     NA                      NA
## 2              NA                     NA                      NA
## 3              NA                     NA                      NA
## 4              NA                     NA                      NA
## 5              NA                     NA                      NA
## 6              NA                     NA                      NA
##   amplitude_yaw_forearm total_accel_forearm var_accel_forearm
## 1                    NA                  36               NA
## 2                    NA                  36               NA
## 3                    NA                  36               NA
## 4                    NA                  36               NA
## 5                    NA                  36               NA
## 6                    NA                  36               NA
##   avg_roll_forearm stddev_roll_forearm var_roll_forearm avg_pitch_forearm
## 1               NA                  NA               NA                NA
## 2               NA                  NA               NA                NA
## 3               NA                  NA               NA                NA
## 4               NA                  NA               NA                NA
## 5               NA                  NA               NA                NA
## 6               NA                  NA               NA                NA
##   stddev_pitch_forearm var_pitch_forearm avg_yaw_forearm
## 1                   NA                NA              NA
## 2                   NA                NA              NA
## 3                   NA                NA              NA
## 4                   NA                NA              NA
## 5                   NA                NA              NA
## 6                   NA                NA              NA
##   stddev_yaw_forearm var_yaw_forearm gyros_forearm_x gyros_forearm_y
```

```
## 1                 NA              NA            0.03            0.00
## 2                 NA              NA            0.02            0.00
## 3                 NA              NA            0.03           -0.02
## 4                 NA              NA            0.02           -0.02
## 5                 NA              NA            0.02            0.00
## 6                 NA              NA            0.02           -0.02
##   gyros_forearm_z accel_forearm_x accel_forearm_y accel_forearm_z
## 1           -0.02             192             203            -215
## 2           -0.02             192             203            -216
## 3            0.00             196             204            -213
## 4            0.00             189             206            -214
## 5           -0.02             189             206            -214
## 6           -0.03             193             203            -215
##   magnet_forearm_x magnet_forearm_y magnet_forearm_z classe
## 1              -17              654              476      A
## 2              -18              661              473      A
## 3              -18              658              469      A
## 4              -16              658              469      A
## 5              -17              655              473      A
## 6               -9              660              478      A
```

We have a data set with 19622 observations of 160 variables, and a validation set with 20 observations and 160 variables. We checked if all variables are the same, and they match. No need to modify anything so far.

Taking a look at the dataset, we see a lot of NAs. For the sake of simplicity, we will not use those variables. If we run the models and the predicting variables are not sufficient to explain the classes reasonably, we will have to revisit and process theses variables with NAs so they are useful.

We also understand that the quality of exercises should be time independent. So we won't use any timestamps either.

We will check if there are any variables that seem to have no significant effect in predicting the results by doing a Near Zero Variables check.

Follows the routine to clean these variables:

```
# Remove the timestamp columns, ID, window flags and names - assuming the quality is independent of tim
dataset.clean <- dataset[,-(1:6)]
validation.clean <- validation[,-(1:6)]

# Remove data with excessive NAs (threshold > 50% of NAs)
n <- length(dataset.clean[,1])
remove.cols = sapply(dataset.clean, function(x) (sum(is.na(x))/n > 0.5))
dataset.clean = dataset.clean[!remove.cols]

n <- length(validation.clean[,1])
remove.cols = sapply(validation.clean, function(x) (sum(is.na(x))/n > 0.1))
validation.clean = validation.clean[!remove.cols]

# Near Zero Variables check
checkNearZeroVariables <- nearZeroVar(dataset.clean, saveMetrics=TRUE)
if(length(checkNearZeroVariables$nzv[checkNearZeroVariables$nzv == TRUE]) == 0){
  print("No Near Zero Variables - all of them can be relevant to prediction")
}else{
  warning(paste("\nThe following variable appear to be irrelevant:",colnames(dataset.clean)[checkNearZe
}
```

```
## [1] "No Near Zero Variables - all of them can be relevant to prediction"
```

```
dim(dataset.clean);dim(validation.clean)
```

```
## [1] 19622    54
```

```
## [1] 20 54
```

We have no Near Zero Variables, so all variables seem to be relevant. Now, our cleaned data and validation sets have 54 variables.

We now create a partition for the training and test sets. We will split in 60% of the observations for training and 40% for testing.

```
#### Creating partition with test and training sets ####
inTrain <- createDataPartition(y=dataset.clean$classe, p=0.6, list=FALSE)
training <- dataset.clean[inTrain,]; testing <- dataset.clean[-inTrain,]
dim(training); dim(testing)
```

```
## [1] 11776    54
```

```
## [1] 7846   54
```

There are, after the partition, 11776 observations for training set and 7846 for testing set. Time to get our hands on the algorithms :)

Machine Learning algorithms testing

The first algorithm to be tested is the Decision Tree.

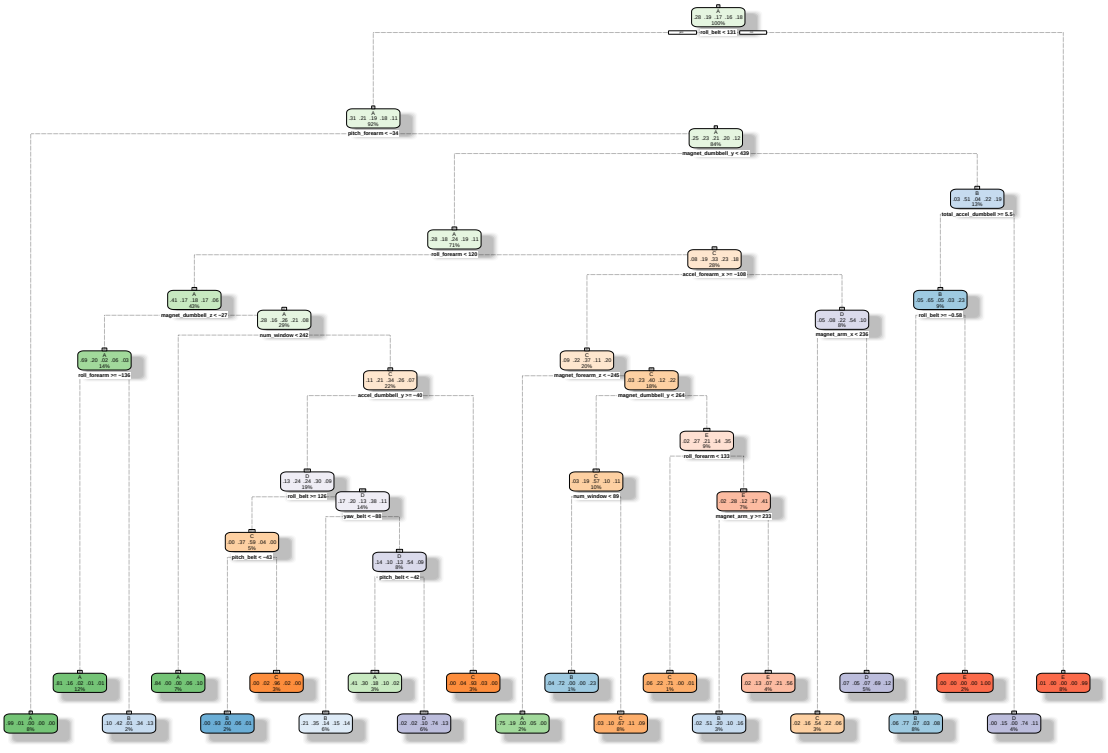
```
modFit.decisionTree <- rpart(classe~.,method = "class", data = training)
pred.decisionTree <- predict(modFit.decisionTree, testing, type="class")
confusionMatrix(pred.decisionTree, testing$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2003  249   51   58   61
##          B  144  986  198  187  204
##          C   31  146 1027  111   69
##          D   46   88   66  847  151
##          E    8   49   26   83  957
##
## Overall Statistics
##
##                Accuracy : 0.7418
##                  95% CI : (0.7319, 0.7514)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

```
##                   Kappa : 0.6722
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.8974   0.6495   0.7507   0.6586   0.6637
## Specificity            0.9254   0.8842   0.9449   0.9465   0.9741
## Pos Pred Value         0.8270   0.5736   0.7421   0.7070   0.8522
## Neg Pred Value         0.9578   0.9132   0.9472   0.9340   0.9279
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2553   0.1257   0.1309   0.1080   0.1220
## Detection Prevalence   0.3087   0.2191   0.1764   0.1527   0.1431
## Balanced Accuracy      0.9114   0.7669   0.8478   0.8026   0.8189
```

```
fancyRpartPlot(modFit.decisionTree)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



Rattle 2020−jun−03 20:51:26 danilo

We can see that our Decision Tree model has an overall accuracy of 0.7483. A visual plot of the decision tree is also embedded, which help in understanding how predict a new case.

Let's check if Random Forests will do better:

```
modFit.randomForest <- randomForest(classe ~. , data=training)
pred.randomForest <- predict(modFit.randomForest, testing, type="class")
confusionMatrix(pred.randomForest, testing$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2232    2    0    0    0
##          B    0 1516    4    0    0
##          C    0    0 1364    9    0
##          D    0    0    0 1276    5
##          E    0    0    0    1 1437
##
## Overall Statistics
##
##                Accuracy : 0.9973
##                  95% CI : (0.9959, 0.9983)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9966
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            1.0000   0.9987   0.9971   0.9922   0.9965
## Specificity            0.9996   0.9994   0.9986   0.9992   0.9998
## Pos Pred Value         0.9991   0.9974   0.9934   0.9961   0.9993
## Neg Pred Value         1.0000   0.9997   0.9994   0.9985   0.9992
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2845   0.1932   0.1738   0.1626   0.1832
## Detection Prevalence   0.2847   0.1937   0.1750   0.1633   0.1833
## Balanced Accuracy      0.9998   0.9990   0.9978   0.9957   0.9982
```

We can see that our accuracy has greatly improved. We have now 99.45% accuracy! We won't need to use the variables that were removed in the beggining of this project.

Now we apply this model to the validation set and write the results in the desired output:

```
validation.randomForest <- predict(modFit.randomForest, validation.clean, type="class")

ouputFiles <- function(x){
  n = length(x)
  for(i in 1:n){
    filename = paste0("problem_id_",i,".txt")
    write.table(x[i],file=filename,quote=FALSE,row.names=FALSE,col.names=FALSE)
  }
}

ouputFiles(validation.randomForest)
```