Practical-Machine-Learning-JHU project

Danlu Z

12/31/2019

Overview

This document analyses how well the different activities hasve been performed by sports monitor users. Based on the data collected on the performance (ABCD) vs various activities, various predictive models have been built. A well-fitting model was picked based on the highest accuracy and low out-of sample error rate.

Data Processing

```
download.file("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv",d
estfile="training.csv")
# pay attention here, the missing values, NA have been treated as na.strings
file1<-read.csv("training.csv",na.strings=c("NA","#DIV/0!", ""))
summary(file1$classe)
```

```
##    A    B    C    D    E
## 5580 3797 3422 3216 3607
```

```
dim(file1)
```

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

Library Loading

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.2
```

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

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

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

```
## Warning: package 'rpart.plot' was built under R version 3.6.2
```

```
## Loading required package: rpart
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 -
-
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() -
-
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(tidyr)
library(rpart)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.2
```

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

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

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

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

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

Data Cleaning

```
#Delete columns with all missing values
file2<-file1[,colSums(is.na(file1))==0]
#New dataframe dimensions
dim(file2)
```

```
## [1] 19622    60
```

```
# Rplace NA in each column with median of that column.
for(i in 1:ncol(file2))
   {
  file2[is.na(file2[,i]), i] <- median(as.numeric(file2[,i]), na.rm = TRUE)
}
# Do some exploratory study on the data set
summary(file2)
```

```
##        X            user_name    raw_timestamp_part_1 raw_timestamp_part_2
##  Min.   :    1   adelmo  :3892   Min.   :1.322e+09    Min.   :   294
##  1st Qu.: 4906   carlitos:3112   1st Qu.:1.323e+09    1st Qu.:252912
##  Median : 9812   charles :3536   Median :1.323e+09    Median :496380
##  Mean   : 9812   eurico  :3070   Mean   :1.323e+09    Mean   :500656
##  3rd Qu.:14717   jeremy  :3402   3rd Qu.:1.323e+09    3rd Qu.:751891
##  Max.   :19622   pedro   :2610   Max.   :1.323e+09    Max.   :998801
##
##           cvtd_timestamp  new_window    num_window      roll_belt
##  28/11/2011 14:14: 1498   no :19216   Min.   :  1.0   Min.   :-28.90
##  05/12/2011 11:24: 1497   yes:  406   1st Qu.:222.0   1st Qu.:  1.10
##  30/11/2011 17:11: 1440               Median :424.0   Median :113.00
##  05/12/2011 11:25: 1425               Mean   :430.6   Mean   : 64.41
##  02/12/2011 14:57: 1380               3rd Qu.:644.0   3rd Qu.:123.00
##  02/12/2011 13:34: 1375               Max.   :864.0   Max.   :162.00
##  (Other)         :11007
##    pitch_belt          yaw_belt       total_accel_belt  gyros_belt_x
##  Min.   :-55.8000   Min.   :-180.00   Min.   : 0.00    Min.   :-1.040000
##  1st Qu.:  1.7600   1st Qu.: -88.30   1st Qu.: 3.00    1st Qu.:-0.030000
##  Median :  5.2800   Median : -13.00   Median :17.00    Median : 0.030000
##  Mean   :  0.3053   Mean   : -11.21   Mean   :11.31    Mean   :-0.005592
##  3rd Qu.: 14.9000   3rd Qu.:  12.90   3rd Qu.:18.00    3rd Qu.: 0.110000
##  Max.   : 60.3000   Max.   : 179.00   Max.   :29.00    Max.   : 2.220000
##
##   gyros_belt_y       gyros_belt_z      accel_belt_x       accel_belt_y
##  Min.   :-0.64000   Min.   :-1.4600   Min.   :-120.000   Min.   :-69.00
##  1st Qu.: 0.00000   1st Qu.:-0.2000   1st Qu.: -21.000   1st Qu.:  3.00
##  Median : 0.02000   Median :-0.1000   Median : -15.000   Median : 35.00
##  Mean   : 0.03959   Mean   :-0.1305   Mean   :  -5.595   Mean   : 30.15
##  3rd Qu.: 0.11000   3rd Qu.:-0.0200   3rd Qu.:  -5.000   3rd Qu.: 61.00
##  Max.   : 0.64000   Max.   : 1.6200   Max.   :  85.000   Max.   :164.00
##
##   accel_belt_z     magnet_belt_x   magnet_belt_y   magnet_belt_z
##  Min.   :-275.00   Min.   :-52.0   Min.   :354.0   Min.   :-623.0
##  1st Qu.:-162.00   1st Qu.:  9.0   1st Qu.:581.0   1st Qu.:-375.0
##  Median :-152.00   Median : 35.0   Median :601.0   Median :-320.0
##  Mean   : -72.59   Mean   : 55.6   Mean   :593.7   Mean   :-345.5
##  3rd Qu.:  27.00   3rd Qu.: 59.0   3rd Qu.:610.0   3rd Qu.:-306.0
##  Max.   : 105.00   Max.   :485.0   Max.   :673.0   Max.   : 293.0
##
##     roll_arm         pitch_arm          yaw_arm          total_accel_arm
##  Min.   :-180.00   Min.   :-88.800   Min.   :-180.0000   Min.   : 1.00
##  1st Qu.: -31.77   1st Qu.:-25.900   1st Qu.: -43.1000   1st Qu.:17.00
##  Median :   0.00   Median :  0.000   Median :   0.0000   Median :27.00
##  Mean   :  17.83   Mean   : -4.612   Mean   :  -0.6188   Mean   :25.51
##  3rd Qu.:  77.30   3rd Qu.: 11.200   3rd Qu.:  45.8750   3rd Qu.:33.00
##  Max.   : 180.00   Max.   : 88.500   Max.   : 180.0000   Max.   :66.00
##
```

```
##   gyros_arm_x        gyros_arm_y       gyros_arm_z       accel_arm_x
##  Min.   :-6.37000   Min.   :-3.4400   Min.   :-2.3300   Min.   :-404.00
##  1st Qu.:-1.33000   1st Qu.:-0.8000   1st Qu.:-0.0700   1st Qu.:-242.00
##  Median : 0.08000   Median :-0.2400   Median : 0.2300   Median : -44.00
##  Mean   : 0.04277   Mean   :-0.2571   Mean   : 0.2695   Mean   : -60.24
##  3rd Qu.: 1.57000   3rd Qu.: 0.1400   3rd Qu.: 0.7200   3rd Qu.:  84.00
##  Max.   : 4.87000   Max.   : 2.8400   Max.   : 3.0200   Max.   : 437.00
##
##   accel_arm_y      accel_arm_z       magnet_arm_x     magnet_arm_y
##  Min.   :-318.0   Min.   :-636.00   Min.   :-584.0   Min.   :-392.0
##  1st Qu.: -54.0   1st Qu.:-143.00   1st Qu.:-300.0   1st Qu.:  -9.0
##  Median :  14.0   Median : -47.00   Median : 289.0   Median : 202.0
##  Mean   :  32.6   Mean   : -71.25   Mean   : 191.7   Mean   : 156.6
##  3rd Qu.: 139.0   3rd Qu.:  23.00   3rd Qu.: 637.0   3rd Qu.: 323.0
##  Max.   : 308.0   Max.   : 292.00   Max.   : 782.0   Max.   : 583.0
##
##   magnet_arm_z    roll_dumbbell     pitch_dumbbell     yaw_dumbbell
##  Min.   :-597.0   Min.   :-153.71   Min.   :-149.59   Min.   :-150.871
##  1st Qu.: 131.2   1st Qu.: -18.49   1st Qu.: -40.89   1st Qu.: -77.644
##  Median : 444.0   Median :  48.17   Median : -20.96   Median :  -3.324
##  Mean   : 306.5   Mean   :  23.84   Mean   : -10.78   Mean   :   1.674
##  3rd Qu.: 545.0   3rd Qu.:  67.61   3rd Qu.:  17.50   3rd Qu.:  79.643
##  Max.   : 694.0   Max.   : 153.55   Max.   : 149.40   Max.   : 154.952
##
##  total_accel_dumbbell gyros_dumbbell_x    gyros_dumbbell_y   gyros_dumbbell_z
##  Min.   : 0.00        Min.   :-204.0000   Min.   :-2.10000   Min.   : -2.380
##  1st Qu.: 4.00        1st Qu.:  -0.0300   1st Qu.:-0.14000   1st Qu.: -0.310
##  Median :10.00        Median :   0.1300   Median : 0.03000   Median : -0.130
##  Mean   :13.72        Mean   :   0.1611   Mean   : 0.04606   Mean   : -0.129
##  3rd Qu.:19.00        3rd Qu.:   0.3500   3rd Qu.: 0.21000   3rd Qu.:  0.030
##  Max.   :58.00        Max.   :   2.2200   Max.   :52.00000   Max.   :317.000
##
##  accel_dumbbell_x  accel_dumbbell_y  accel_dumbbell_z  magnet_dumbbell_x
##  Min.   :-419.00   Min.   :-189.00   Min.   :-334.00   Min.   :-643.0
##  1st Qu.: -50.00   1st Qu.:  -8.00   1st Qu.:-142.00   1st Qu.:-535.0
##  Median :  -8.00   Median :  41.50   Median :  -1.00   Median :-479.0
##  Mean   : -28.62   Mean   :  52.63   Mean   : -38.32   Mean   :-328.5
##  3rd Qu.:  11.00   3rd Qu.: 111.00   3rd Qu.:  38.00   3rd Qu.:-304.0
##  Max.   : 235.00   Max.   : 315.00   Max.   : 318.00   Max.   : 592.0
##
##  magnet_dumbbell_y magnet_dumbbell_z  roll_forearm       pitch_forearm
##  Min.   :-3600     Min.   :-262.00   Min.   :-180.0000   Min.   :-72.50
##  1st Qu.:  231     1st Qu.: -45.00   1st Qu.:  -0.7375   1st Qu.:  0.00
##  Median :  311     Median :  13.00   Median :  21.7000   Median :  9.24
##  Mean   :  221     Mean   :  46.05   Mean   :  33.8265   Mean   : 10.71
##  3rd Qu.:  390     3rd Qu.:  95.00   3rd Qu.: 140.0000   3rd Qu.: 28.40
##  Max.   :  633     Max.   : 452.00   Max.   : 180.0000   Max.   : 89.80
##
##   yaw_forearm      total_accel_forearm gyros_forearm_x   gyros_forearm_y
```

```
##  Min.   :-180.00   Min.   :  0.00      Min.   :-22.000   Min.   : -7.02000
##  1st Qu.: -68.60   1st Qu.: 29.00      1st Qu.: -0.220   1st Qu.: -1.46000
##  Median :   0.00   Median : 36.00      Median :  0.050   Median :  0.03000
##  Mean   :  19.21   Mean   : 34.72      Mean   :  0.158   Mean   :  0.07517
##  3rd Qu.: 110.00   3rd Qu.: 41.00      3rd Qu.:  0.560   3rd Qu.:  1.62000
##  Max.   : 180.00   Max.   :108.00      Max.   :  3.970   Max.   :311.00000
##
##  gyros_forearm_z    accel_forearm_x   accel_forearm_y  accel_forearm_z
##  Min.   : -8.0900   Min.   :-498.00   Min.   :-632.0   Min.   :-446.00
##  1st Qu.: -0.1800   1st Qu.:-178.00   1st Qu.:  57.0   1st Qu.:-182.00
##  Median :  0.0800   Median : -57.00   Median : 201.0   Median : -39.00
##  Mean   :  0.1512   Mean   : -61.65   Mean   : 163.7   Mean   : -55.29
##  3rd Qu.:  0.4900   3rd Qu.:  76.00   3rd Qu.: 312.0   3rd Qu.:  26.00
##  Max.   :231.0000   Max.   : 477.00   Max.   : 923.0   Max.   : 291.00
##
##  magnet_forearm_x  magnet_forearm_y magnet_forearm_z classe
##  Min.   :-1280.0   Min.   :-896.0   Min.   :-973.0   A:5580
##  1st Qu.: -616.0   1st Qu.:   2.0   1st Qu.: 191.0   B:3797
##  Median : -378.0   Median : 591.0   Median : 511.0   C:3422
##  Mean   : -312.6   Mean   : 380.1   Mean   : 393.6   D:3216
##  3rd Qu.:  -73.0   3rd Qu.: 737.0   3rd Qu.: 653.0   E:3607
##  Max.   :  672.0   Max.   :1480.0   Max.   :1090.0
##
```

```
# Columns 1:7 have nothing to do with the model, so we need to subset the data.
file3<-file2[8:ncol(file2)]
```

Take a look at the clean data set

```
head(file3)
```

```
##   roll_belt pitch_belt yaw_belt total_accel_belt gyros_belt_x gyros_belt_y
## 1      1.41       8.07    -94.4               3         0.00         0.00
## 2      1.41       8.07    -94.4               3         0.02         0.00
## 3      1.42       8.07    -94.4               3         0.00         0.00
## 4      1.48       8.05    -94.4               3         0.02         0.00
## 5      1.48       8.07    -94.4               3         0.02         0.02
## 6      1.45       8.06    -94.4               3         0.02         0.00
##   gyros_belt_z accel_belt_x accel_belt_y accel_belt_z magnet_belt_x
## 1        -0.02          -21           4          22            -3
## 2        -0.02          -22           4          22            -7
## 3        -0.02          -20           5          23            -2
## 4        -0.03          -22           3          21            -6
## 5        -0.02          -21           2          24            -6
## 6        -0.02          -21           4          21             0
##   magnet_belt_y magnet_belt_z roll_arm pitch_arm yaw_arm total_accel_arm
## 1           599          -313     -128      22.5    -161              34
## 2           608          -311     -128      22.5    -161              34
## 3           600          -305     -128      22.5    -161              34
## 4           604          -310     -128      22.1    -161              34
## 5           600          -302     -128      22.1    -161              34
## 6           603          -312     -128      22.0    -161              34
##   gyros_arm_x gyros_arm_y gyros_arm_z accel_arm_x accel_arm_y accel_arm_z
## 1        0.00        0.00       -0.02        -288         109        -123
## 2        0.02       -0.02       -0.02        -290         110        -125
## 3        0.02       -0.02       -0.02        -289         110        -126
## 4        0.02       -0.03        0.02        -289         111        -123
## 5        0.00       -0.03        0.00        -289         111        -123
## 6        0.02       -0.03        0.00        -289         111        -122
##   magnet_arm_x magnet_arm_y magnet_arm_z roll_dumbbell pitch_dumbbell
## 1         -368          337          516      13.05217      -70.49400
## 2         -369          337          513      13.13074      -70.63751
## 3         -368          344          513      12.85075      -70.27812
## 4         -372          344          512      13.43120      -70.39379
## 5         -374          337          506      13.37872      -70.42856
## 6         -369          342          513      13.38246      -70.81759
##   yaw_dumbbell total_accel_dumbbell gyros_dumbbell_x gyros_dumbbell_y
## 1    -84.87394                  37                0            -0.02
## 2    -84.71065                  37                0            -0.02
## 3    -85.14078                  37                0            -0.02
## 4    -84.87363                  37                0            -0.02
## 5    -84.85306                  37                0            -0.02
## 6    -84.46500                  37                0            -0.02
##   gyros_dumbbell_z accel_dumbbell_x accel_dumbbell_y accel_dumbbell_z
## 1             0.00             -234               47             -271
## 2             0.00             -233               47             -269
## 3             0.00             -232               46             -270
## 4            -0.02             -232               48             -269
## 5             0.00             -233               48             -270
```

```
## 6             0.00             -234               48             -269
##   magnet_dumbbell_x magnet_dumbbell_y magnet_dumbbell_z roll_forearm
## 1              -559               293               -65         28.4
## 2              -555               296               -64         28.3
## 3              -561               298               -63         28.3
## 4              -552               303               -60         28.1
## 5              -554               292               -68         28.0
## 6              -558               294               -66         27.9
##   pitch_forearm yaw_forearm total_accel_forearm gyros_forearm_x gyros_forearm_y
## 1         -63.9        -153                  36            0.03            0.00
## 2         -63.9        -153                  36            0.02            0.00
## 3         -63.9        -152                  36            0.03           -0.02
## 4         -63.9        -152                  36            0.02           -0.02
## 5         -63.9        -152                  36            0.02            0.00
## 6         -63.9        -152                  36            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
```

Since the response is discrete (A/B/C/D) for various predictors, we need to

use classification models for this problem.

Model Building

```
set.seed(11)
sample<-createDataPartition(file3$classe,p=0.7,list=FALSE)
train<-file3[sample,]
ver<-file3[-sample,]
```

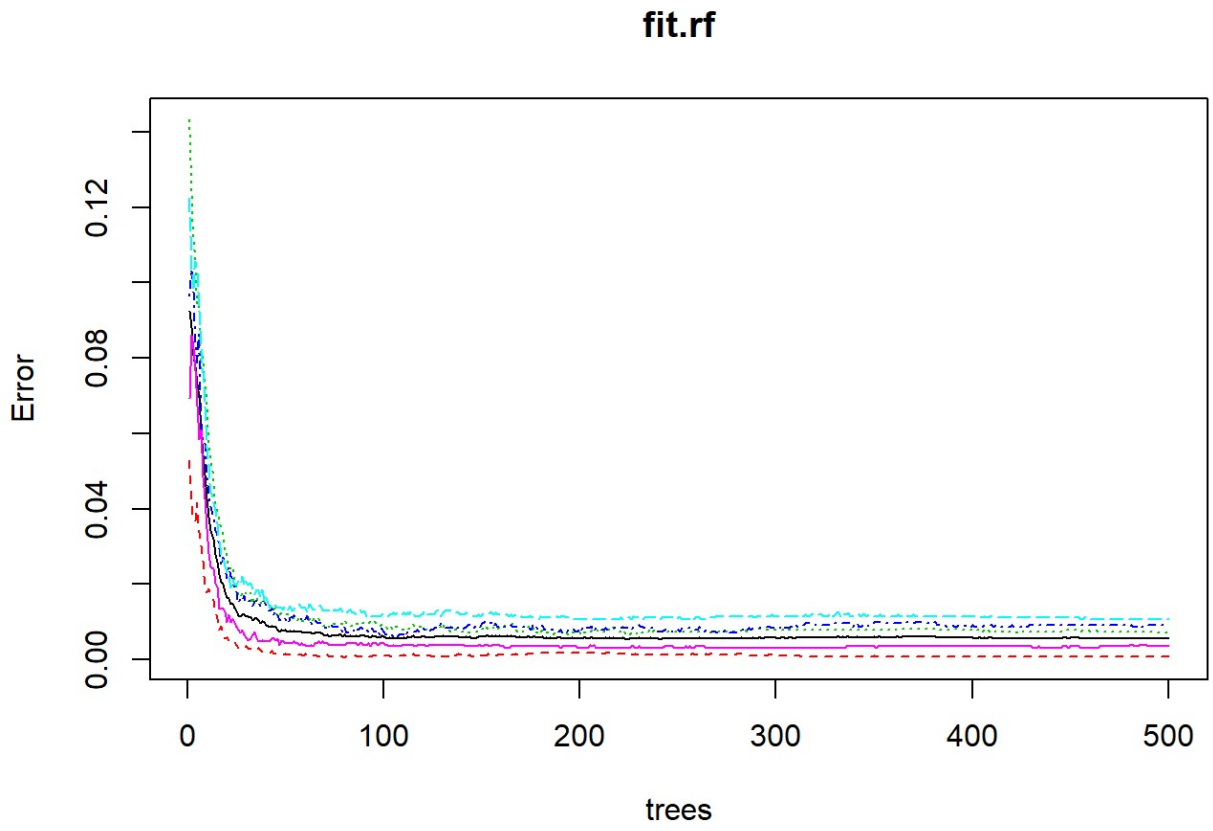
The provided data set has been divided to train data set and verification data set. The cross-validation is based on these two data sets.

1. Train the model with “randomForest” method

```
fit.rf <- randomForest(classe~., data = train, method = "class")
acc.rf<-confusionMatrix(ver$classe,predict(fit.rf,ver))
# Get the accuracy of this model
acc.rf$overall[1]
```

```
##  Accuracy
## 0.9964316
```

```
# Take a look at the plot
plot(fit.rf)
```

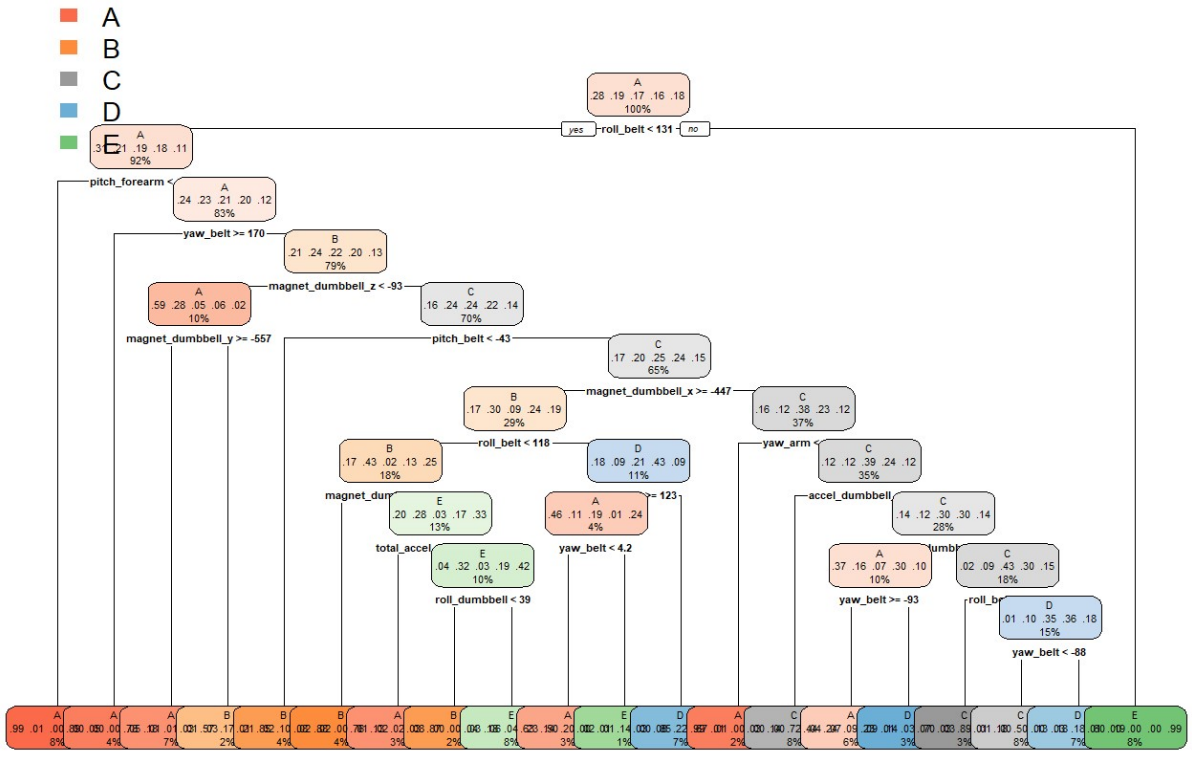


2. Train the model with decision tree method

```
fit.tree<-rpart(classe~.,data=train,method = "class")
pre.tree<-predict(fit.tree,ver,type="class")
# Get the accuracy of this model
acc.tree<-confusionMatrix(ver$classe,pre.tree)
acc.tree$overall[1]
```

```
##  Accuracy
## 0.7282923
```

```
# Take a look at the plot
rpart.plot(fit.tree,uniform=TRUE,tweak = 2.2)
```



Model Selection

Apparently, the randomForest method results in a model with a higher accuracy.and the out-of-sample error is 1-cross-validation model related accuracy = 1- 0.95=0.04, so the out-of-sample error is around 4% with *randomeForest* or *"fit.rf"* model.

Model Testing

```
test <- download.file("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testin
g.csv", destfile="test.csv")
t<-read.csv("test.csv")
pre_test<-predict(fit.rf,t)
```