Predicting exercise patterns

PM

Objective:

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://groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Loading libraries.

```
library(randomForest)
```

```
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(ggplot2)
library(caret)
```

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

```
library(rpart)
library(rattle)
```

```
## Rattle: A free graphical interface for data mining with R.
## Version 4.0.5 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(reshape2)
library(gbm)
```

```
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:caret':
##
##     cluster
##
## Loading required package: splines
## Loading required package: parallel
## Loaded gbm 2.1.1
```

Setting multi-core processing.

```
library(doMC)
```

```
## Loading required package: foreach
## Loading required package: iterators
```

```
registerDoMC(8)
```

Loading data files. These data is provided by (Ugulino et al. 2012).

```
url_1 <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
url_2 <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

file <- download.file(url = url_1, destfile = "trainData.csv", method = "curl")
file <- download.file(url = url_2, destfile = "testData.csv", method = "curl")

train <- read.csv(file = "trainData.csv", na.strings= c('#DIV/0!', '#DIV/0', '', 'NA', ' '))
test <- read.csv(file = "testData.csv", na.strings= c('#DIV/0!','#DIV/0', '', 'NA', ' '))
```

Removing the Na columns and all columns having 60% of Na's.

```
train <- train[,colSums(is.na(train)) <nrow(train)*0.6]
test <- test[,colSums(is.na(test)) <nrow(test)*0.6]



train <- train[,-(1:8)]
test <- test[,-(1:8)]
```

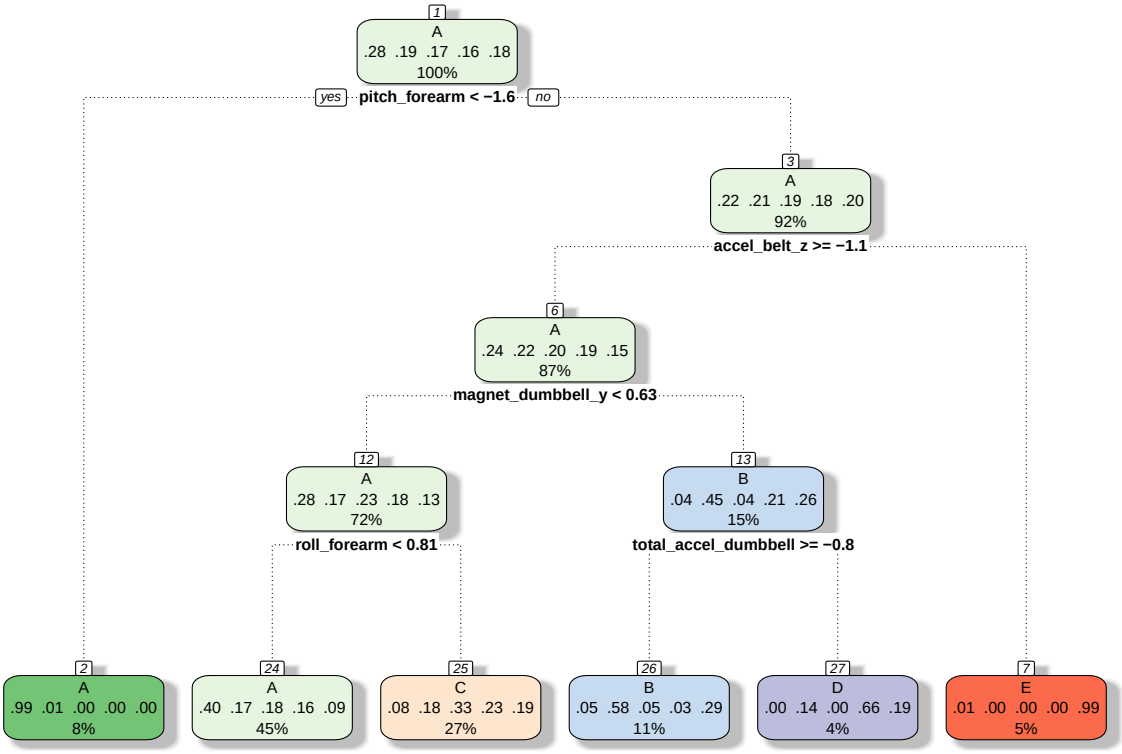
Partitioning training in two data sets for validation.

```
inTrain <- createDataPartition(train$classe, p = 3/4)[[1]]
train.training <- train[ inTrain,]
train.testing <- train[-inTrain,]
```

Training algorithms

Get the decision tree from recursive partitioning and regression trees.

```
set.seed(13579)
model <- train(classe ~ ., data = train.training, method = "rpart", preProcess=c("center", "scale"))
fancyRpartPlot(model$finalModel, sub = "")
```

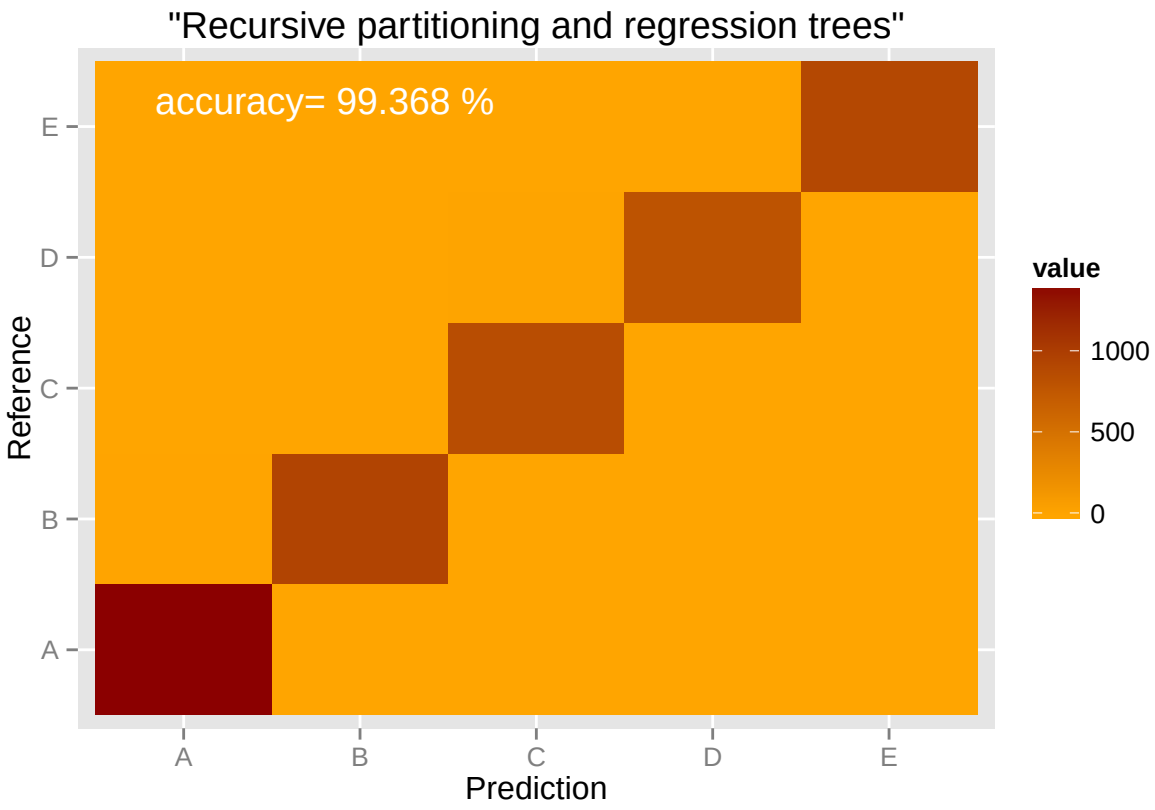


Random forest with 5 cross-validations and pre-processing treatment.

```
set.seed(13579)
model_1 <- train(classe ~ ., data = train.training, method = "rf",
                 preProcess=c("center", "scale"),
                 trControl = trainControl(method="cv", number= 5))

fit_1 <- predict(model_1, train.testing)

conf_1 <- confusionMatrix(fit_1, train.testing$classe)
```

Extracting the prediction vs reference data to create a heat-plot.

```
ttable_1 <- conf_1$table
ttable_1 <- melt(ttable_1)

ggplot(data = ttable_1, aes(x=Prediction, y=Reference,
                            title='Recursive partitioning and regression trees',
                            fill=value)) + geom_tile() + scale_fill_gradient(low="orange", high="darkre
        annotate("text",  x = 1.8, y = 5.2,
                 label = paste("accuracy=",round(conf_1$overall[[1]]*100,3), "%"),
                 colour='white')
```



Generalized Boosted Regression Modelling.

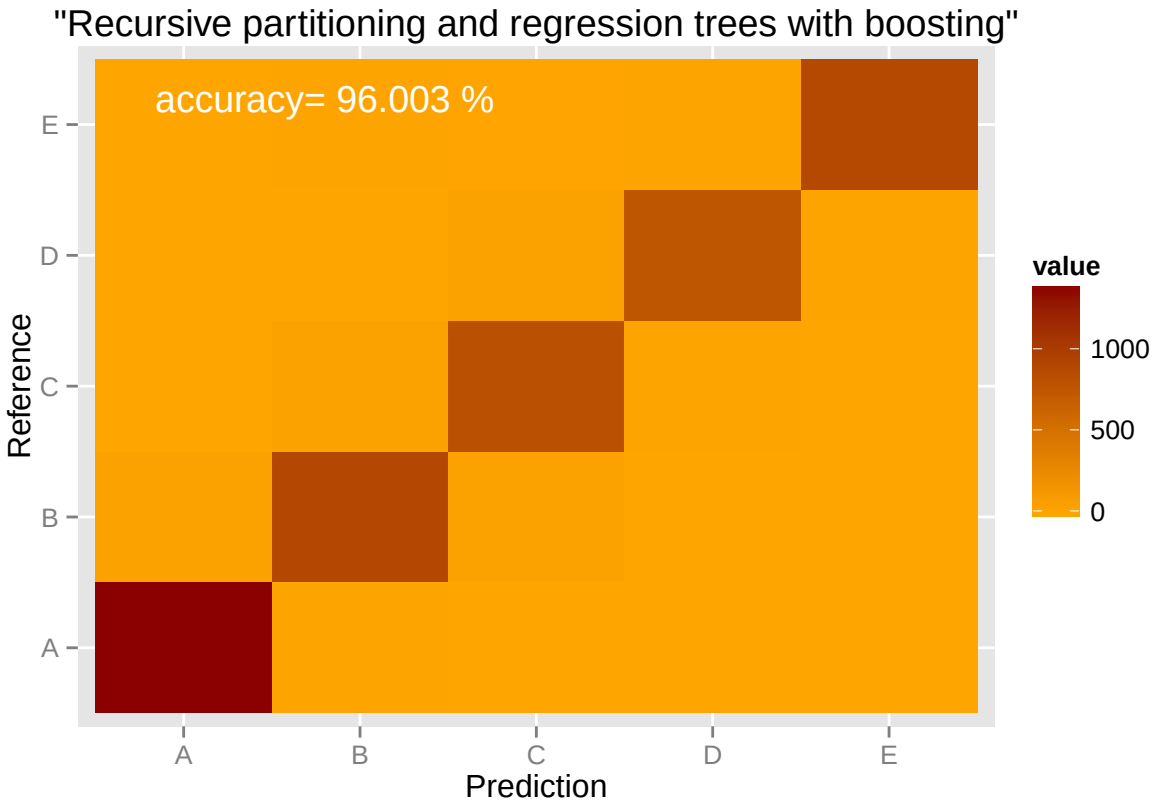
```
set.seed(13579)
model_2 <- train(classe ~ ., data = train.training, method = "gbm", verbose=FALSE, preProcess=c("center
                 trControl = trainControl(method="repeatedcv", number=5, repeats=2),  metric="Accuracy"
```

```
## Loading required package: plyr
```

```
fit_2 <- predict(model_2, train.testing)

conf_2 <- confusionMatrix(fit_2, train.testing$classe)
```

Extracting the prediction vs reference data to create a heat-plot.

```
ttable_2 <- conf_2$table
ttable_2 <- melt(ttable_2)

ggplot(data = ttable_2, aes(x=Prediction, y=Reference,
                            title='Recursive partitioning and regression trees with boosting',
                            fill=value)) + geom_tile() + scale_fill_gradient(low="orange", high="darkre
        annotate("text",  x = 1.8, y = 5.2,
                 label = paste("accuracy=",round(conf_2$overall[[1]]*100,3), "%"),
                 colour='white')
```



The random forest is giving a higher accuracy than Generalized Boosted Regression Modeling (gmb).

Predicting

```
predict(model_1, test)
```

```
##  [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
```

```
answers <- c('B', 'A', 'B', 'A', 'A', 'E', 'D', 'B', 'A', 'A', 'B',
'C', 'B', 'A', 'E', 'E', 'A', 'B', 'B', 'B')
pml_write_files = 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)
        }
}
pml_write_files(answers)
```

References:

Ugulino, Wallace, Débora Cardador, Katia Vega, Eduardo Velloso, Ruy Milidiú, and Hugo Fuks. 2012. "Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements." In *Advances*

in Artificial Intelligence - SBIA 2012, 52–61. Springer Science + Business Media. doi:10.1007/978-3-642-34459-6_6.