Course Project - Practical Machine Learning

Seher Can Akay

12/4/2019

Background and Overview

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, my 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.

Loading Libraries and Data

```r
library(caret)
library(rpart)
library(randomForest)
library(rattle)

train <- read.csv("pml-training.csv") # will be used for prediction on this project
quiz_testing <- read.csv("pml-testing.csv") # will be used for the quiz of the project
```

Cleaning Data for Prediction

I am removing the columns with NA values:

```r
train <- train[, colSums(is.na(train)) == 0]
```

I am cleaning the columns which has nearly zero variance:

```r
#Remove Near-Zero Values from Data
nearZero <- nearZeroVar(train)
train <- train[, -nearZero]
```

I am removing first 5 variables, since they dont seem to have an impact on the variable "classe"

```r
train <- train[, -c(1:5)]
dim(train)
```

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

Now, I will split the data to training and testing sets:

```r
inTrain <- createDataPartition(train$classe, p = 0.75, list = FALSE)
training <- train[inTrain,]
testing <- train[-inTrain,]

dim(training)
```

```
## [1] 14718    54
```

```
dim(testing)
```

```
## [1] 4904   54
```

Modelling

I will apply 3 different models on the data and try to find the best model to predict the "classe".

Random Forest Model

```
modFit1 <- randomForest(classe ~., method = "class", data = training)
pred1 <- predict(modFit1, newdata = testing, type = "class")

confMatrix1 <- confusionMatrix(pred1, testing$classe)
confMatrix1$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 1395    2    0    0    0
##          B    0  946    0    0    0
##          C    0    1  855    5    0
##          D    0    0    0  799    4
##          E    0    0    0    0  897
```

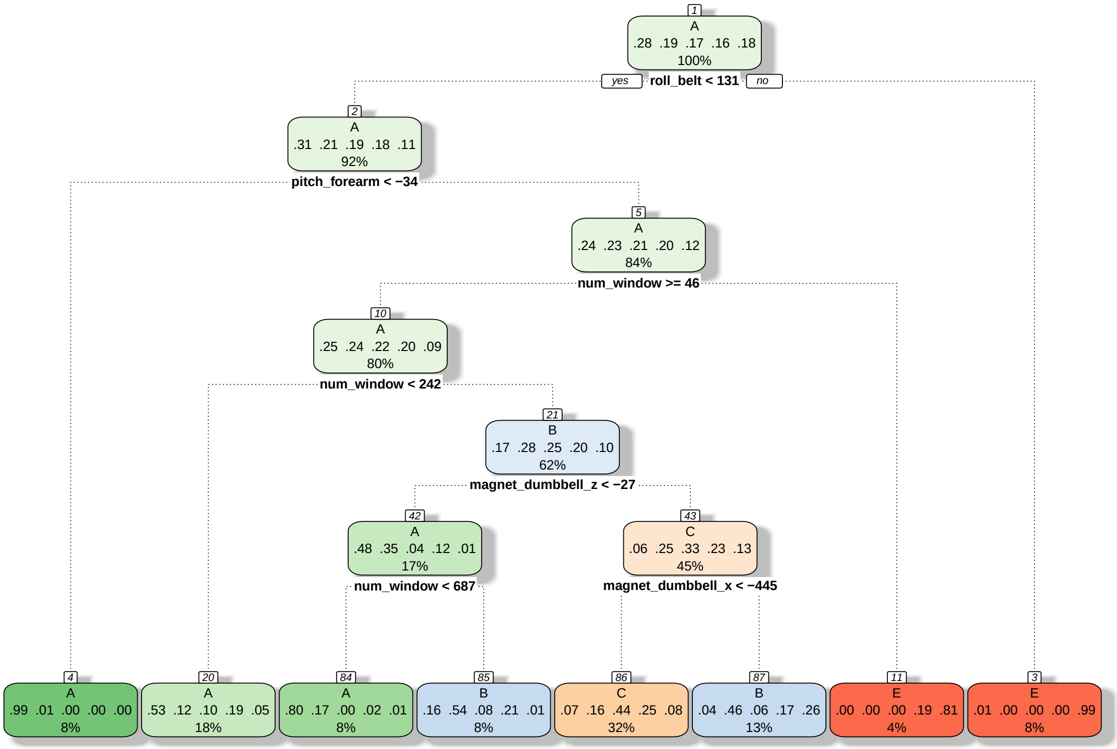
```
confMatrix1$overall[1]
```

```
## Accuracy
## 0.997553
```

0.997553 is preety good accuracy. But let's try other models as well.

Decision Tree

```
modFit2 <- train(classe ~., method = "rpart", data = training)
fancyRpartPlot(modFit2$finalModel)
```



Rattle 2019−Dec−05 14:13:54 trcanseh

Let's see confuison matrix and accuracy:

```
pred2 <- predict(modFit2, newdata = testing)

confMatrix2 <- confusionMatrix(pred2, testing$classe)
confMatrix2$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 1164  167  106  182   56
##          B  101  533   84  208  161
##          C  126  249  665  369  141
##          D    0    0    0    0    0
##          E    4    0    0   45  543
```

```
confMatrix2$overall[1]
```

```
##  Accuracy
## 0.5923736
```

As seen above accuracy of 0.5923736 is not so clear to select this model.

Conclusion

2 different model has been tested. Random Forest seems the best model to fit the data. So Random Forest Model will be used during quiz.

```
predQuiz <- predict(modFit1, newdata = quiz_testing, type = "class")
predQuiz
```

```
##  1  2  3  4  5  6  7  8  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
```