Practical_Machine_Learning

Victor Alanis

28/7/2020

Practical Machine Learning Project : Prediction Assignment Writeup

I. Overview

This document is the final report of the Peer Assessment project from Coursera's course Practical Machine Learning, as part of the Specialization in Data Science. It was built up in RStudio, using its knitr functions, meant to be published in html format. This analysis meant to be the basis for the course quiz and a prediction assignment writeup. The main goal of the project is to predict the manner in which 6 participants performed some exercise as described below. This is the "classe" variable in the training set. The machine learning algorithm described here is applied to the 20 test cases available in the test data and the predictions are submitted in appropriate format to the Course Project Prediction Quiz for automated grading.

II.Background

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).

Read more:http://groupware.les.inf.puc-rio.br/har#ixzz3xsbS5bVX

III. Data Loading and Exploratory Analysis

Environment Preparation We first upload the R libraries that are necessary for the complete analysis.

```r
rm(list=ls())                # free up memory for the download of the data sets
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.2
```

```r
library(caret)
```

```
## Loading required package: lattice
## Warning: package 'lattice' was built under R version 4.0.2
```

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

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

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

```
## Warning: package 'rpart.plot' was built under R version 4.0.2
```

```r
library(rattle)
```

```
## Warning: package 'rattle' was built under R version 4.0.2
```

```
## Loading required package: tibble
```

```
## Loading required package: bitops
```

```
## Rattle: A free graphical interface for data science with R.
## Versión 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Escriba 'rattle()' para agitar, sacudir y  rotar sus datos.
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.2
```

```
## 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
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.2
```

```
## corrplot 0.84 loaded
```

```r
set.seed(12345)
```

Data Loading and Cleaning The next step is loading the dataset from the URL provided above. The training dataset is then partinioned in 2 to create a Training set (70% of the data) for the modeling process and a Test set (with the remaining 30%) for the validations. The testing dataset is not changed and will only be used for the quiz results generation.

```r
# set the URL for the download
UrlTrain <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
UrlTest  <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

# download the datasets
training <- read.csv(url(UrlTrain))
testing  <- read.csv(url(UrlTest))

# create a partition with the training dataset
inTrain  <- createDataPartition(training$classe, p=0.7, list=FALSE)
TrainSet <- training[inTrain, ]
TestSet  <- training[-inTrain, ]
dim(TrainSet)
```

```
## [1] 13737   160
```

```r
dim(TestSet)
```

```
## [1] 5885  160
```

Both created datasets have 160 variables. Those variables have plenty of NA, that can be removed with the cleaning procedures below. The Near Zero variance (NZV) variables are also removed and the ID variables as well.

```r
# remove variables with Nearly Zero Variance
NZV <- nearZeroVar(TrainSet)
TrainSet <- TrainSet[, -NZV]
TestSet  <- TestSet[, -NZV]
dim(TrainSet)
```

```
## [1] 13737   104
```

```r
dim(TestSet)
```

```
## [1] 5885  104
```

```r
# remove variables that are mostly NA
AllNA    <- sapply(TrainSet, function(x) mean(is.na(x))) > 0.95
TrainSet <- TrainSet[, AllNA==FALSE]
TestSet  <- TestSet[, AllNA==FALSE]
dim(TrainSet)
```

```
## [1] 13737    59
```

```r
dim(TestSet)
```

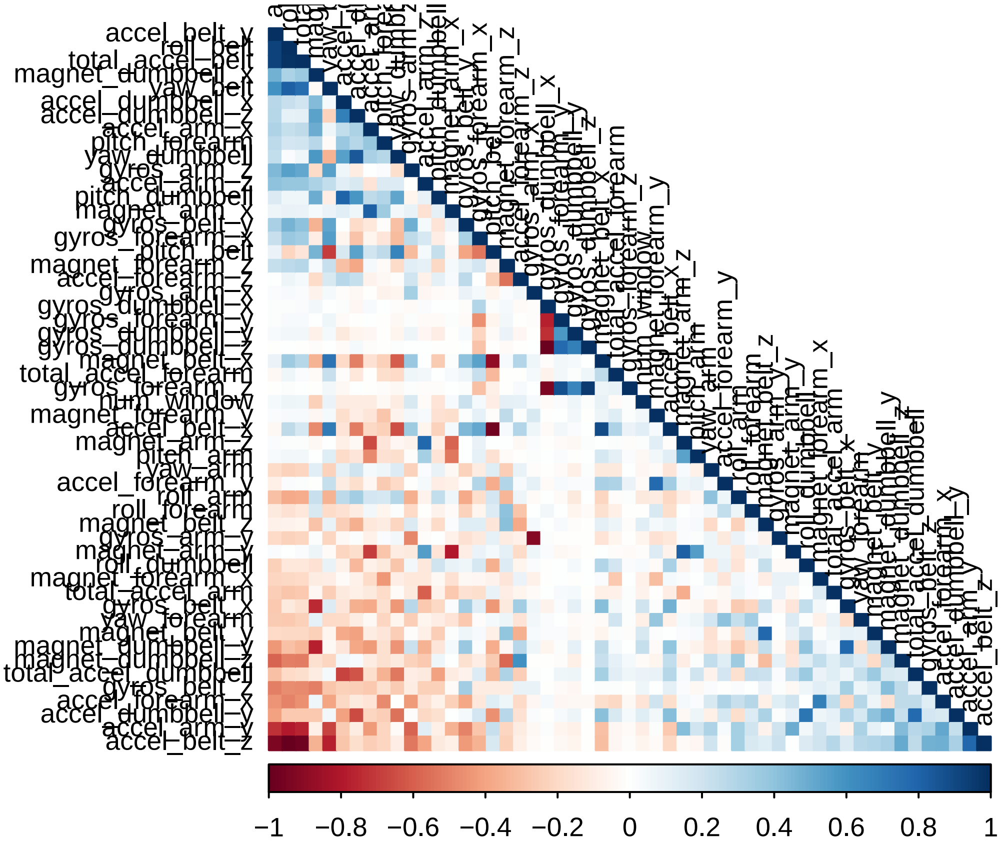
```
## [1] 5885   59
```

```r
# remove identification only variables (columns 1 to 5)
TrainSet <- TrainSet[, -(1:5)]
TestSet  <- TestSet[, -(1:5)]
dim(TrainSet)
```

```
## [1] 13737    54
```

Correlation Analysis A correlation among variables is analysed before proceeding to the modeling procedures.

```r
corMatrix <- cor(TrainSet[, -54])
corrplot(corMatrix, order = "FPC", method = "color", type = "lower",
         tl.cex = 0.8, tl.col = rgb(0, 0, 0))
```


accel_belt_y
roll_belt
total_accel_belt
magnet_dumbbell_x
yaw_belt
accel_dumbbell_x
accel_dumbbell_z
accel_arm_x
pitch_forearm
yaw_dumbbell
gyros_arm_z
accel_arm_z
pitch_dumbbell
magnet_arm_x
gyros_belt_y
gyros_forearm_x
pitch_belt
magnet_forearm_z
accel_forearm_z
gyros_arm_x
gyros_dumbbell_x
gyros_forearm_y
gyros_dumbbell_y
gyros_dumbbell_z
magnet_belt_x
total_accel_forearm
gyros_forearm_z
num_window
magnet_forearm_x
accel_belt_x
magnet_arm_z
pitch_arm
yaw_arm
accel_forearm_x
roll_arm
roll_forearm
magnet_belt_z
gyros_arm_y
magnet_arm_y
roll_dumbbell
magnet_forearm_y
total_accel_arm
gyros_belt_x
yaw_forearm
magnet_belt_y
magnet_dumbbell_y
magnet_dumbbell_z
total_accel_dumbbell
gyros_belt_z
accel_forearm_y
accel_dumbbell_y
accel_arm_y
accel_belt_z