**Practical Machine Learning Project**

**CITIZENTECH**

***13 May 21, 2019***

**Summary**

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.

Here I have used the 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).

**Data Sets**

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>

The data for this project come from this source: <http://groupware.les.inf.puc-rio.br/har>.

**Loading Data**

Download the csv sheets for both training and testing datasets, and move to the respective directories.

## Cleaning the Data

We have removed the unnecessary 7 columns in the starting of both the ‘training’ and ‘testing’ data frames. Some columns have NA and empty values “”, fot those columns a check of 95% has been made, and rest of the columns have been included.

## Further Cleaning of Data

Here we are further cleaning the data by removing the predictor variables which are showing very less value of variance. In addition to this, any 2 variables which are highly correlated will not be giving any significant advantage to our prediction model. So, we have created a correlation matrix and have removed few of the highly correlated predictor variables by putting a check of the cutoff value equal to 0.95.

# Modeling the data

We have applied the Random Forest Model and it has shown significant amount of accuracy in prediction. Now we are partitioning the ‘training’ dataset into two parts : train1A and train2A. We have created our prediction model on train1A, and then tested it on train2A.

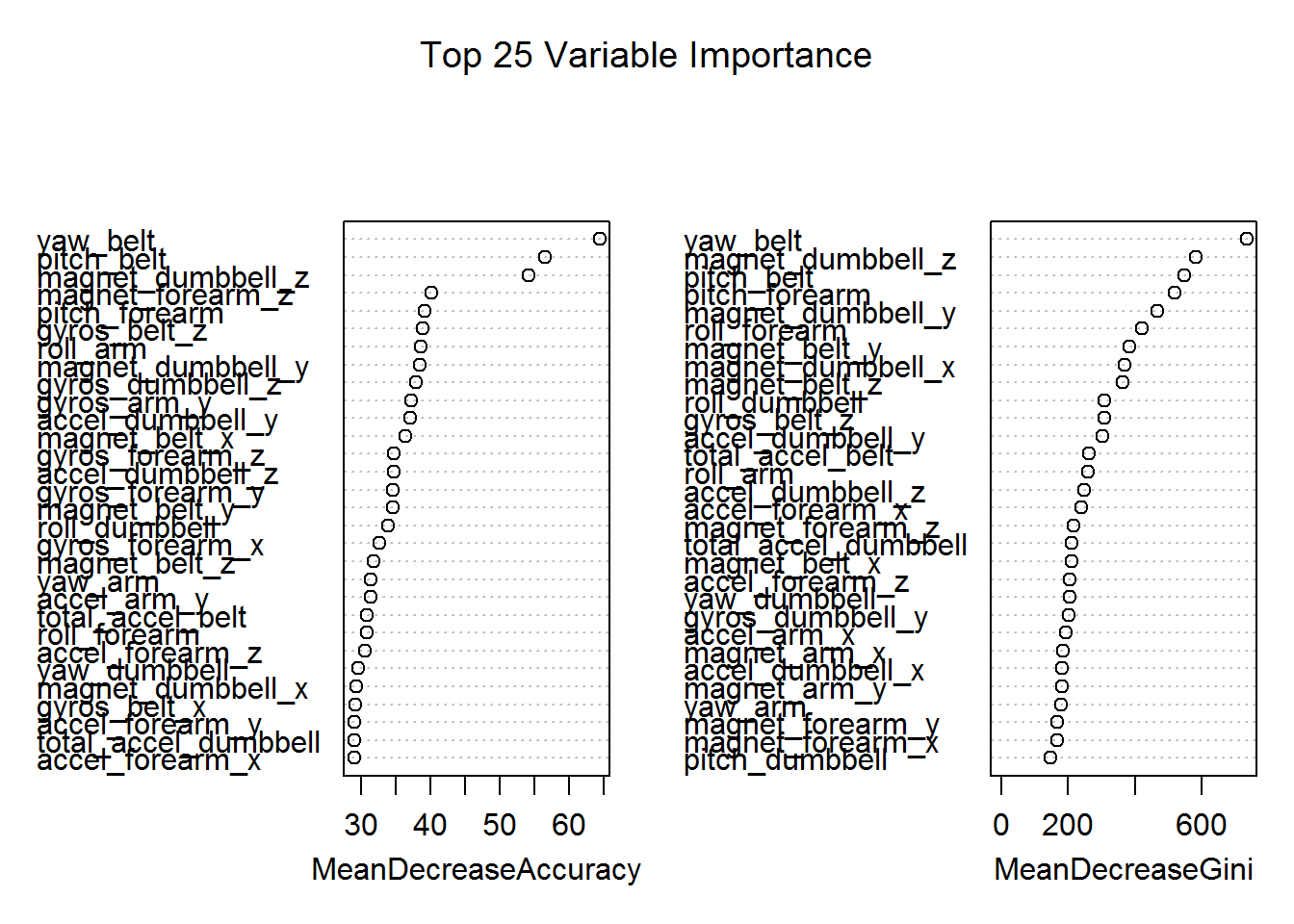
The Confusion Matrix achieved 99.59% accuracy. Here, the Out-Of-Sample Error Rate observed is 0.41%, and the OOB (Out-Of-Bag) Error Rate is 0.63%.

## Checking Out-of-Sample error

On the basis of the Confusion Matrix calculated above, showing the the out-of-sample error value:

## Variable-Importance Plot

Calculating the important predictors of the randomForest model via ‘varImp()’ function.



## Model on ‘testing’ dataset

These are the solutions for the 20 cases which have been submitted as the answers.

*# Testing the model on the 'testing' dataset*

testing\_pred <- predict(randomForMod, newdata=testing)

testing\_pred

## 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