ML\_8\_Project

Karthik Vadhri

December 2, 2017

# Practical Machine Learning - Course Project

### Karthik Vadhri

# Practical Machine Learning Project : Prediction Assignment Writeup

## 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.

## 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>

## Data Loading and Exploratory Analysis

### Dataset Overview

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 <http://groupware.les.inf.puc-rio.br/har>. Full source:

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. “Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human ’13)”. Stuttgart, Germany: ACM SIGCHI, 2013.

My special thanks to the above mentioned authors for being so generous in allowing their data to be used for this kind of assignment.

A short description of the datasets content from the authors’ website:

“Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: exactly according to the specification (Class A), throwing the elbows to the front (Class B), lifting the dumbbell only halfway (Class C), lowering the dumbbell only halfway (Class D) and throwing the hips to the front (Class E).

Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes. Participants were supervised by an experienced weight lifter to make sure the execution complied to the manner they were supposed to simulate. The exercises were performed by six male participants aged between 20-28 years, with little weight lifting experience. We made sure that all participants could easily simulate the mistakes in a safe and controlled manner by using a relatively light dumbbell (1.25kg)."

### Setting the environment

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

## Loading required package: lattice

## Loading required package: ggplot2

library(rpart)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.3.3

#library(rattle)  
library(randomForest)

## randomForest 4.6-12

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

##   
## Attaching package: 'randomForest'

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

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.3.3

### Data Loading and Cleaning

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

dim(TestSet)

## [1] 5885 160

#### Remove Variables with Zero Variance, and high NA's

NZV <- nearZeroVar(TrainSet)  
TrainSet <- TrainSet[, -NZV]  
TestSet <- TestSet[, -NZV]  
dim(TrainSet)

## [1] 13737 106

AllNA <- sapply(TrainSet, function(x) mean(is.na(x))) > 0.95  
TrainSet <- TrainSet[, AllNA==FALSE]  
TestSet <- TestSet[, AllNA==FALSE]  
dim(TrainSet)

## [1] 13737 59

dim(TestSet)

## [1] 5885 59

# remove variables that are Unique Identifiers (columns 1 to 5)  
TrainSet <- TrainSet[, -(1:5)]  
TestSet <- TestSet[, -(1:5)]  
dim(TrainSet)

## [1] 13737 54

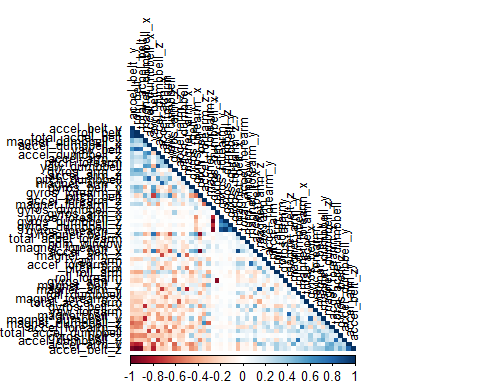
dim(TestSet)

## [1] 5885 54

### Correlation Analysis

We use the correlation plot to observe the relation among the variables.

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



### Prediction Model Building

We shall use 3 different methods, and compare the accuracy of each one of them, before deciding the final model. Confusion matrix of each model is also plotted, to give a holistic picture of the model .

#### Method 1 : Random Forest

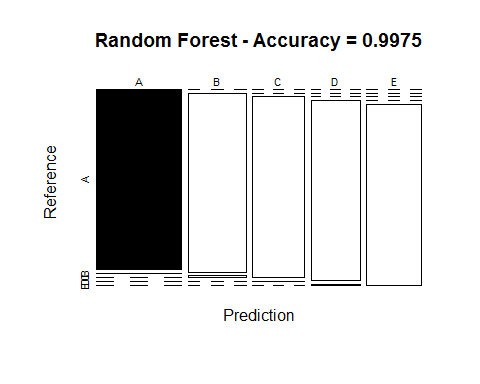
set.seed(12345)  
controlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)  
modFitRandForest <- train(classe ~ ., data=TrainSet, method="rf",  
 trControl=controlRF)  
modFitRandForest$finalModel

##   
## Call:  
## randomForest(x = x, y = y, mtry = param$mtry)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 27  
##   
## OOB estimate of error rate: 0.24%  
## Confusion matrix:  
## A B C D E class.error  
## A 3904 1 0 0 1 0.0005120328  
## B 6 2647 4 1 0 0.0041384500  
## C 0 4 2392 0 0 0.0016694491  
## D 0 0 10 2241 1 0.0048845471  
## E 0 1 0 4 2520 0.0019801980

# prediction on Test dataset  
predictRandForest <- predict(modFitRandForest, newdata=TestSet)  
confMatRandForest <- confusionMatrix(predictRandForest, TestSet$classe)  
confMatRandForest

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 1674 3 0 0 0  
## B 0 1136 9 0 0  
## C 0 0 1017 1 0  
## D 0 0 0 963 2  
## E 0 0 0 0 1080  
##   
## Overall Statistics  
##   
## Accuracy : 0.9975   
## 95% CI : (0.9958, 0.9986)  
## No Information Rate : 0.2845   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9968   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 1.0000 0.9974 0.9912 0.9990 0.9982  
## Specificity 0.9993 0.9981 0.9998 0.9996 1.0000  
## Pos Pred Value 0.9982 0.9921 0.9990 0.9979 1.0000  
## Neg Pred Value 1.0000 0.9994 0.9982 0.9998 0.9996  
## Prevalence 0.2845 0.1935 0.1743 0.1638 0.1839  
## Detection Rate 0.2845 0.1930 0.1728 0.1636 0.1835  
## Detection Prevalence 0.2850 0.1946 0.1730 0.1640 0.1835  
## Balanced Accuracy 0.9996 0.9977 0.9955 0.9993 0.9991

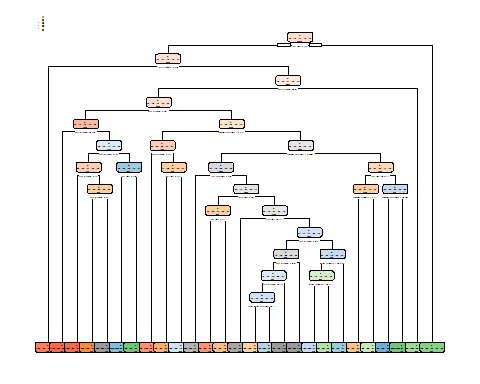
plot(confMatRandForest$table, col = confMatRandForest$byClass,   
 main = paste("Random Forest - Accuracy =",  
 round(confMatRandForest$overall['Accuracy'], 4)))



#### Method: Decision Trees

modFitDecTree <- rpart(classe ~ ., data=TrainSet, method="class")  
#plot(modFitDecTree)  
  
library(rpart.plot)  
rpart.plot(modFitDecTree)

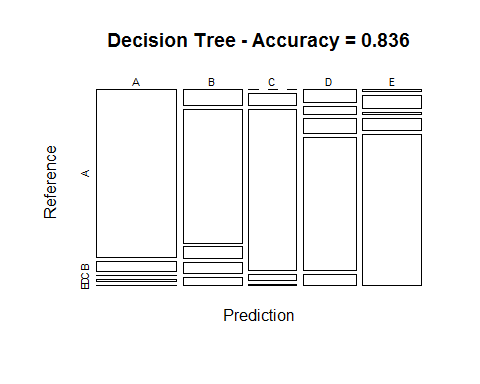
## Warning: labs do not fit even at cex 0.15, there may be some overplotting



predictDecTree <- predict(modFitDecTree, newdata=TestSet, type="class")  
confMatDecTree <- confusionMatrix(predictDecTree, TestSet$classe)  
confMatDecTree

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 1478 85 2 13 2  
## B 105 861 75 73 48  
## C 0 65 845 31 3  
## D 76 46 91 768 61  
## E 15 82 13 79 968  
##   
## Overall Statistics  
##   
## Accuracy : 0.836   
## 95% CI : (0.8263, 0.8454)  
## No Information Rate : 0.2845   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.793   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.8829 0.7559 0.8236 0.7967 0.8946  
## Specificity 0.9758 0.9366 0.9796 0.9443 0.9606  
## Pos Pred Value 0.9354 0.7410 0.8951 0.7370 0.8366  
## Neg Pred Value 0.9545 0.9411 0.9634 0.9595 0.9759  
## Prevalence 0.2845 0.1935 0.1743 0.1638 0.1839  
## Detection Rate 0.2511 0.1463 0.1436 0.1305 0.1645  
## Detection Prevalence 0.2685 0.1975 0.1604 0.1771 0.1966  
## Balanced Accuracy 0.9293 0.8463 0.9016 0.8705 0.9276

plot(confMatDecTree$table, col = confMatDecTree$byClass,   
 main = paste("Decision Tree - Accuracy =",  
 round(confMatDecTree$overall['Accuracy'], 4)))



### Method 3 : Generalized Boosted Model

controlGBM <- trainControl(method = "repeatedcv", number = 5, repeats = 1)  
modFitGBM <- train(classe ~ ., data=TrainSet, method = "gbm",  
 trControl = controlGBM, verbose = FALSE)

## Loading required package: gbm

## Loading required package: survival

## Warning: package 'survival' was built under R version 3.3.3

##   
## 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

## Loading required package: plyr

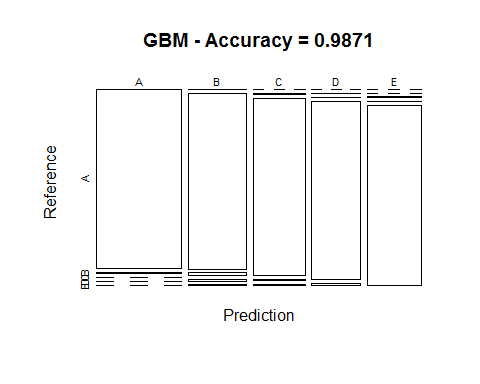
modFitGBM$finalModel

## A gradient boosted model with multinomial loss function.  
## 150 iterations were performed.  
## There were 53 predictors of which 41 had non-zero influence.

predictGBM <- predict(modFitGBM, newdata=TestSet)  
confMatGBM <- confusionMatrix(predictGBM, TestSet$classe)  
confMatGBM

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 1673 10 0 0 0  
## B 1 1116 18 8 2  
## C 0 11 1004 9 2  
## D 0 2 2 946 8  
## E 0 0 2 1 1070  
##   
## Overall Statistics  
##   
## Accuracy : 0.9871   
## 95% CI : (0.9839, 0.9898)  
## No Information Rate : 0.2845   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9837   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.9994 0.9798 0.9786 0.9813 0.9889  
## Specificity 0.9976 0.9939 0.9955 0.9976 0.9994  
## Pos Pred Value 0.9941 0.9747 0.9786 0.9875 0.9972  
## Neg Pred Value 0.9998 0.9951 0.9955 0.9963 0.9975  
## Prevalence 0.2845 0.1935 0.1743 0.1638 0.1839  
## Detection Rate 0.2843 0.1896 0.1706 0.1607 0.1818  
## Detection Prevalence 0.2860 0.1946 0.1743 0.1628 0.1823  
## Balanced Accuracy 0.9985 0.9868 0.9870 0.9894 0.9941

plot(confMatGBM$table, col = confMatGBM$byClass,   
 main = paste("GBM - Accuracy =", round(confMatGBM$overall['Accuracy'], 4)))



## Applying the Selected Model to the Test Data

The accuracy of the 3 regression modeling methods above are:

Random Forest : 0.9963 Decision Tree : 0.7368 GBM : 0.9839

It is clearly evident that the RF seems to be better classifier, so we use that for the test sets given for the exercise.