Practical Machine Learning

Tanisha Srivastava

17/09/2020

## 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. The aim of this project is to predict the manner in which participants perform a barbell lift. The data comes from <http://groupware.les.inf.puc-rio.br/har> wherein 6 participants were asked to perform the same set of exercises correctly and incorrectly with accelerometers placed on the belt, forearm, arm, and dumbell.

For the purpose of this project, the following steps would be followed:

1. Data Preprocessing
2. Exploratory Analysis
3. Prediction Model Selection
4. Predicting Test Set Output

## Data Preprocessing

First, we load the training and testing set from the online sources and then split the training set further into training and test sets.

library(caret)  
  
trainURL <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"  
testURL <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"  
training <- read.csv(url(trainURL))  
testing <- read.csv(url(testURL))  
label <- createDataPartition(training$classe, p = 0.7, list = FALSE)  
train <- training[label, ]  
test <- training[-label, ]

From among 160 variables present in the dataset, some variables have nearly zero variance whereas some contain a lot of NA terms which need to be excluded from the dataset. Moreover, other 5 variables used for identification can also be removed.

NZV <- nearZeroVar(train)  
train <- train[ ,-NZV]  
test <- test[ ,-NZV]  
label <- apply(train, 2, function(x) mean(is.na(x))) > 0.95  
train <- train[, -which(label, label == FALSE)]  
test <- test[, -which(label, label == FALSE)]  
train <- train[ , -(1:5)]  
test <- test[ , -(1:5)]

As a result of the preprocessing steps, we were able to reduce 160 variables to 54.

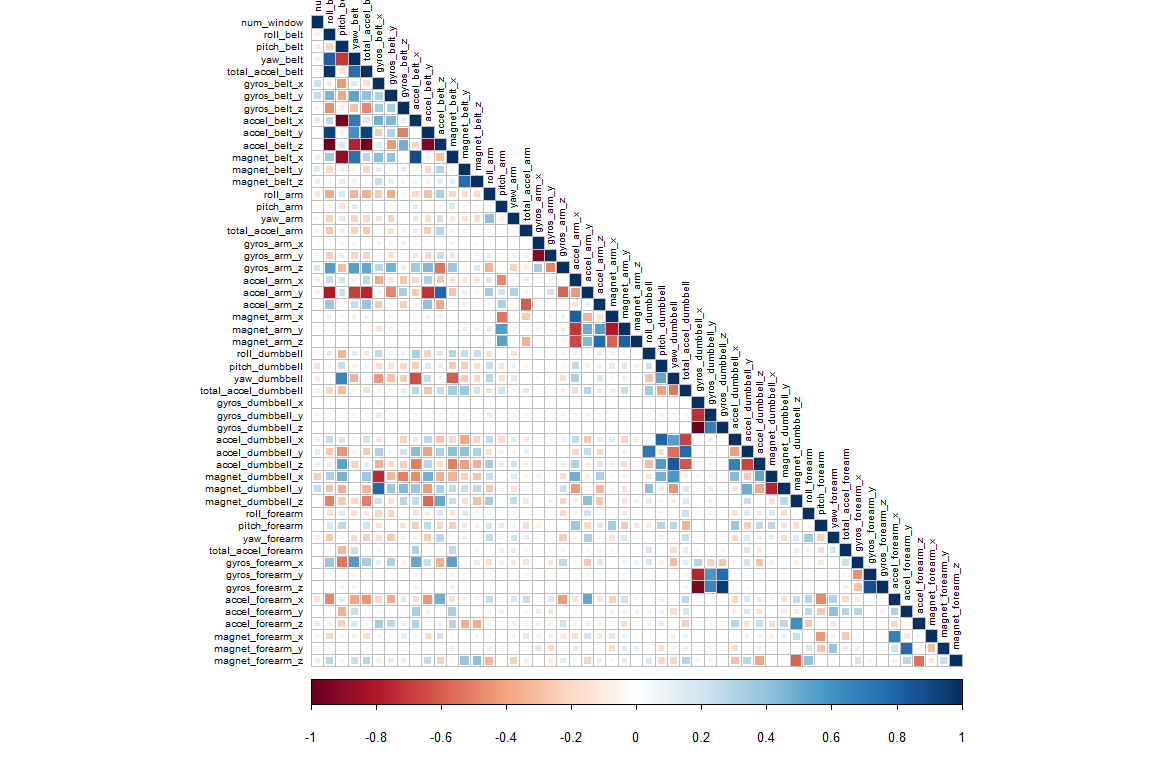
## Exploratory Analysis

Now that we have cleaned the dataset off absolutely useless varibles, we shall look at the dependence of these variables on each other through a correlation plot.

library(corrplot)

## corrplot 0.84 loaded

corrMat <- cor(train[,-54])  
corrplot(corrMat, method = "square", type = "lower", tl.cex = 0.6, tl.col = rgb(0,0,0))



In the plot above, darker gradient correspond to having high correlation. A Principal Component Analysis can be run to further reduce the correlated variables but we aren’t doing that due to the number of correlations being quite few.

## Prediction Model Selection

We will use 3 methods to model the training set and thereby choose the one having the best accuracy to predict the outcome variable in the testing set. The methods are Decision Tree, Random Forest and Generalized Boosted Model.

A confusion matrix plotted at the end of each model will help visualize the analysis better.

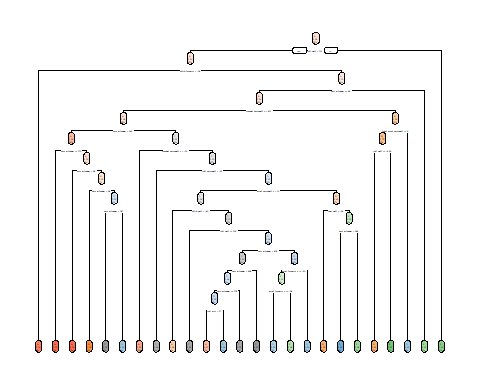
### Decision Tree

library(rpart)  
library(rpart.plot)  
library(rattle)  
set.seed(13908)

fit <- rpart(classe ~., data = train, method = "class")  
rpart.plot(fit, extra = 106)

## Warning: extra=106 but the response has 5 levels (only the 2nd level is  
## displayed)

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

 Random Forest

library(party)

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

library(randomForest)

## 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

control <- trainControl(method = "cv", number = 3, verboseIter=FALSE)  
modelRF <- train(classe ~ ., data = train, method = "rf", trControl = control)  
modelRF$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: 0.17%  
## Confusion matrix:  
## A B C D E class.error  
## A 3905 0 0 0 1 0.0002560164  
## B 5 2651 2 0 0 0.0026335591  
## C 0 5 2391 0 0 0.0020868114  
## D 0 0 6 2246 0 0.0026642984  
## E 0 0 0 5 2520 0.0019801980

predictRF <- predict(modelRF, test)

Generalized Boosted Model

library(caret)  
set.seed(13908)  
control <- trainControl(method = "repeatedcv", number = 5, repeats = 1, verboseIter = FALSE)  
modelGBM <- train(classe ~ ., data = train, trControl = control, method = "gbm", verbose = FALSE)  
modelGBM$finalModel

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