Predicting "How well we do barbell lifts?" using data from Sports Devices

Sebastián Fuenzalida Garcés
August 2015

Executive Summary

Load Data

```
library(AppliedPredictiveModeling); library(caret); library(rattle); library(randomForest)

##The files are included in the github repo
##After some analysis to the data I decided to consider as NA's: "NA","#DIV/0!" and "" to avoid problems
train_data<-read.csv("pml-training.csv",header=TRUE,na.strings=c("NA","#DIV/0!",""))
test_data<-read.csv("pml-testing.csv",header=TRUE,na.strings=c("NA","#DIV/0!",""))
```

Columnas que no sirven

```
#First 7 columns aren't useful for the model
train_data1<-train_data[,8:length(train_data)]
test_data1<-test_data[,8:length(test_data)]

#NearZeroVar gives us a first approach of columns that we don't need
nsv<-nearZeroVar(train_data1,saveMetrics=TRUE)

train_data2<-train_data1[,-which(names(train_data1) %in% row.names(nsv[nsv$nzv==TRUE,]))]
test_data2<-test_data1[,-which(names(test_data1) %in% row.names(nsv[nsv$nzv==TRUE,]))]


drop_col<-c()
for(i in 1:length(train_data2))
        {
        total_no_NA<-sum(!is.na(train_data2[,i]))
        ##We look which columns have more than 70% of rows with NA's
        if(total_no_NA<nrow(train_data2)*0.3)
                {
                drop_col<-c(drop_col,names(train_data2[i]))
                }
        }

##Now we drop the columns selected
train_data3<-train_data2[,-which(names(train_data2) %in% drop_col)]
test_data3<-test_data2[,-which(names(test_data2) %in% drop_col)]
```