

Practical Machine Learning Assignment

Paul Martinez

2023-05-18

Executive Summary

The goal of this project is to predict the manner in which participants did exercise using devices such as Jawbone Up, Nike FuelBand, and Fitbit. This report describes in detail how the model was built, how it was cross validated, and what the expected sample error is.

First Step is to load libraries and dataset

```
library(lattice)
library(ggplot2)
library(plyr)
library(randomForest)

## randomForest 4.7-1.1

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

##
## Attaching package: 'randomForest'

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

training.raw <- read.csv('~/.pml-training.csv')
testing.raw <- read.csv('~/.pml-testing.csv')

dim(training.raw)

## [1] 19622 160
```

What we see is a lot of data with NA / empty values. Let's remove those

```
maxNAPerc = 20
maxNACount <- nrow(training.raw) / 100 * maxNAPerc
removeColumns <- which(colSums(is.na(training.raw) | training.raw=="") > maxNACount)
training.cleaned01 <- training.raw[,-removeColumns]
testing.cleaned01 <- testing.raw[,-removeColumns]
```

Also remove all time related data, since we won't use those

```
removeColumns <- grep("timestamp", names(training.cleaned01))
training.cleaned02 <- training.cleaned01[,-c(1, removeColumns )]
testing.cleaned02 <- testing.cleaned01[,-c(1, removeColumns )]
```

Then convert all factors to integers

```
classeLevels <- levels(training.cleaned02$classe)
training.cleaned03 <- data.frame(data.matrix(training.cleaned02))
training.cleaned03$classe <- factor(training.cleaned03$classe, labels=1)
testing.cleaned03 <- data.frame(data.matrix(testing.cleaned02))
```

Finally set the dataset to be explored

```
training.cleaned <- training.cleaned03
testing.cleaned <- testing.cleaned03
```

Exploratory data analyses

Since the test set provided is the the ultimate validation set, we will split the current training in a test and train set to work with.