How well do you weight lift?

Authored by: Vitor Cerqueira

This report is part of the project for the Pratical Machine Learning class taught is Coursera MOOC by by Jeff Leek, PhD, Roger D. Peng, PhD, Brian Caffo, PhD from John Hopkins, Bloomberg School of Public Health.

```
FALSE randomForest 4.6-10
FALSE Type rfNews() to see new features/changes/bug fixes.
FALSE Loading required package: lattice
```

Introduction

As we enter an increasingly data driven era, there are developed many new technologies designed to improve the experience of people's lives. Examples of such technologies are the devices Jawbone Up, Nike FuelBand, and Fitbit, which allows the collection of large volumes of data about personal activities. Typical analysis on human activity recognition research rely on simply predicting which kind of activity is performed at a given time. Taking a better advantage of the devices previously mentioned, a group of researchers used the data from accelerometers on the belt, forearm, arm, and dumbells to quantify how well participants executed a weight lifting exercise. These participant were asked to execute the exercise in five different ways: according to the specification and mistakenly in other four different ways. In this report is describe one way to build a model capable of predicting the manner in which the participants did the exercise. For simplicity purposes I decided to attach the code in a section as appendix. For reproducibility issues, seed is set to "1234".

Model Building

In this section I will explain the steps of the model I built to predict the manner in which 20 of the exercises were performed. First the training and test sets are loaded into R. Url's are training (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv) and testing (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv). After that the training set is further split into training and test train sets with holdout method, in a 70/30 way. In other words, we get three datasets, a training set, a testing set, and a validation set. All three sets are preprocessed in the same ways

Preprocessing

The initial dataset contains a lot of attributes with an high percentage of missing values. For this reason, I have decided to exclude all variable that had more than seventy percent observations with missing values. Also, some variables are unnecessary for the problem context, such as temporal information or information on the names of participants. Furthermore it is run a test to check for variables with near zero variance, but there were not found any. The plot below shows the distribution of the classes in the training set. The classes are relatevely well balanced, so there is no need for balancing with stratified sampling or a similar method when running the model.