

# PML Project1

## Executive Summary

## 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. The goal of this project is to use data from accelerometers on the belt, forearm, arm, and dumbbell of 6 participants as they perform barbell lifts correctly and incorrectly 5 different ways.

Six young healthy participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: \* Class A - exactly according to the specification \* Class B - throwing the elbows to the front \* Class C - lifting the dumbbell only halfway \* Class D - lowering the dumbbell only halfway \* Class E - throwing the hips to the front

More information is available from the website here: [linked phrase](#) (see the section on the Weight Lifting Exercise Dataset).

## Data

The training data for this project are available here: [linked phrase](#)

The test data are available here: [\[linked phrase\]](#) (<https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv>)

```
library(RCurl)
```

```
## Loading required package: bitops
```

```
train_url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
test_url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
train_data <- read.csv(text=getURL(train_url), na.strings=c("", "NA"))
test_data <- read.csv(text=getURL(test_url), na.strings=c("", "NA"))
```

The first column of the data is just index. We remove it from training data frame.

```
train_data$X <- NULL
```

Similarly the user and time information should not have any effect on whether barbell lifts are performed correctly or not.

```
cols_to_remove <- c("user_name", "raw_timestamp_part_1",
                    "raw_timestamp_part_2", "cvtd_timestamp")
for (col in cols_to_remove) {
  train_data[, col] <- NULL
}
```

Many columns in the dataset have mostly missing values. We remove features from the training and testing data that have too many missing values, where imputing is not an option.

```
NAs <- apply(train_data, 2, function(x) {sum(is.na(x))})
train_data <- train_data[,which(NAs == 0)]
```

We also remove features that don't have many missing values but have one unique value (i.e. zero variance predictors) or have few unique values relative to the number of samples and the ratio of frequency of the most common value to the frequency of second most common value is large.

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
nsv <- nearZeroVar(train_data)
train_data <- train_data[-nsv]
test_data <- test_data[-nsv]
```

The final set of predictors used for classification are as follows.

```
names(train_data)
```

```
## [1] "num_window"      "roll_belt"      "pitch_belt"
## [4] "yaw_belt"        "total_accel_belt" "gyros_belt_x"
## [7] "gyros_belt_y"    "gyros_belt_z"   "accel_belt_x"
## [10] "accel_belt_y"    "accel_belt_z"   "magnet_belt_x"
## [13] "magnet_belt_y"   "magnet_belt_z"  "roll_arm"
## [16] "pitch_arm"       "yaw_arm"        "total_accel_arm"
## [19] "gyros_arm_x"     "gyros_arm_y"    "gyros_arm_z"
## [22] "accel_arm_x"     "accel_arm_y"    "accel_arm_z"
## [25] "magnet_arm_x"    "magnet_arm_y"   "magnet_arm_z"
## [28] "roll_dumbbell"   "pitch_dumbbell" "yaw_dumbbell"
## [31] "total_accel_dumbbell" "gyros_dumbbell_x" "gyros_dumbbell_y"
## [34] "gyros_dumbbell_z" "accel_dumbbell_x" "accel_dumbbell_y"
## [37] "accel_dumbbell_z" "magnet_dumbbell_x" "magnet_dumbbell_y"
## [40] "magnet_dumbbell_z" "roll_forearm"    "pitch_forearm"
## [43] "yaw_forearm"      "total_accel_forearm" "gyros_forearm_x"
## [46] "gyros_forearm_y"  "gyros_forearm_z"  "accel_forearm_x"
## [49] "accel_forearm_y"  "accel_forearm_z"  "magnet_forearm_x"
## [52] "magnet_forearm_y" "magnet_forearm_z" "classe"
```

## The model

We build a random forest classifier to predict the action class. To measure the accuracy of the model, we do 10-fold cross validation with 80:20 split, on each fold, 80% of the data is used for training the random forest and remaining 20% is used for testing.

```
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:ggplot2':
##
##     margin
```

```
set.seed(1)
obs <- c()
preds <- c()
for(i in 1:10) {
  intrain = sample(1:dim(train_data)[1], size=dim(train_data)[1] * 0.8, replace=F)
  train_cross = train_data[intrain,]
  test_cross = train_data[-intrain,]
  rf <- randomForest(classe ~ ., data=train_cross)
  obs <- c(obs, test_cross$classe)
  preds <- c(preds, predict(rf, test_cross))
}
```

The confusion matrix for predictions on cross validation folds is given below.

```
conf_mat <- confusionMatrix(table(preds, obs))
conf_mat$table
```

```
##      obs
## preds  1    2    3    4    5
##  1 11365   11    0    0    0
##  2    3 7542   15    0    0
##  3    0    4 6810   35    0
##  4    0    0    0 6307   17
##  5    1    0    0    2 7138
```

The proposed model seems classifying well enough. The accuracy is 99.8293% and it misclassifies only few instances. Finally, we train the random forest with whole dataset so that the classifier can be used to predict the class of an action, given the set of activity measurements.

```
model <- randomForest(classe ~ ., data=train_data)
```

## References

Velloso, Eduardo, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. 2013. "Qualitative Activity Recognition of Weight Lifting Exercises." In Proceedings of the 4th Augmented Human International Conference, 116–123. AH '13. New York, NY, USA: ACM. doi:10.1145/2459236.2459256. <http://doi.acm.org/10.1145/2459236.2459256>.