

# Practical Machine Learning Course Project

*Lee Davies*

*October 27, 2016*

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.3.1
```

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

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

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.3.1
```

```
## randomForest 4.6-12
```

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

```
##
```

```
## Attaching package: 'randomForest'
```

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

```
##
```

```
##      margin
```

## Download the data files and load the data

```
trainUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
```

```
testUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
```

```
trainFile <- "./data/pml-training.csv"
```

```
testFile <- "./data/pml-testing.csv"
```

```
#create the directory if it does not exists
```

```
if (!file.exists("./data")) {
```

```
  dir.create("./data")
```

```
}
```

```
#download the training file
```

```
if (!file.exists(trainFile)) {
```

```
  download.file(trainUrl, destfile=trainFile)
```

```
}
```

```
#download the test file
```

```

if (!file.exists(testFile)) {
  download.file(testUrl, destfile=testFile)
}

trainSet <- read.csv("./data/pml-training.csv",header=TRUE, as.is = TRUE, stringsAsFactors = FALSE, sep=";", as.is=TRUE)
finaltestSet <- read.csv("./data/pml-testing.csv", header=TRUE, as.is = TRUE, stringsAsFactors = FALSE, sep=";", as.is=TRUE)

str(trainSet)

```

```

## 'data.frame': 19622 obs. of 160 variables:
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...
## $ user_name : chr "carlitos" "carlitos" "carlitos" "carlitos" ...
## $ raw_timestamp_part_1 : int 1323084231 1323084231 1323084231 1323084232 1323084232 1323084232 ...
## $ raw_timestamp_part_2 : int 788290 808298 820366 120339 196328 304277 368296 440390 484323 484323 ...
## $ cvtd_timestamp : chr "05/12/2011 11:23" "05/12/2011 11:23" "05/12/2011 11:23" "05/12/2011 11:23" ...
## $ new_window : chr "no" "no" "no" "no" ...
## $ num_window : int 11 11 11 12 12 12 12 12 12 12 ...
## $ roll_belt : num 1.41 1.41 1.42 1.48 1.48 1.45 1.42 1.42 1.43 1.45 ...
## $ pitch_belt : num 8.07 8.07 8.07 8.05 8.07 8.06 8.09 8.13 8.16 8.17 ...
## $ yaw_belt : num -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 ...
## $ total_accel_belt : int 3 3 3 3 3 3 3 3 3 3 ...
## $ kurtosis_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_pitch_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_yaw_belt : logi NA NA NA NA NA NA ...
## $ skewness_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_roll_belt.1 : num NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_yaw_belt : logi NA NA NA NA NA NA ...
## $ max_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ max_pitch_belt : int NA NA NA NA NA NA NA NA NA NA ...
## $ max_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ min_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ min_pitch_belt : int NA NA NA NA NA NA NA NA NA NA ...
## $ min_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_pitch_belt : int NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ var_total_accel_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ avg_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ var_roll_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ avg_pitch_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_pitch_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ var_pitch_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ avg_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ var_yaw_belt : num NA NA NA NA NA NA NA NA NA NA ...
## $ gyros_belt_x : num 0 0.02 0 0.02 0.02 0.02 0.02 0.02 0.02 0.03 ...
## $ gyros_belt_y : num 0 0 0 0 0.02 0 0 0 0 0 ...
## $ gyros_belt_z : num -0.02 -0.02 -0.02 -0.03 -0.02 -0.02 -0.02 -0.02 -0.02 0 ...
## $ accel_belt_x : int -21 -22 -20 -22 -21 -21 -22 -22 -20 -21 ...
## $ accel_belt_y : int 4 4 5 3 2 4 3 4 2 4 ...
## $ accel_belt_z : int 22 22 23 21 24 21 21 21 24 22 ...
## $ magnet_belt_x : int -3 -7 -2 -6 -6 0 -4 -2 1 -3 ...

```

```

## $ magnet_belt_y      : int  599 608 600 604 600 603 599 603 602 609 ...
## $ magnet_belt_z      : int -313 -311 -305 -310 -302 -312 -311 -313 -312 -308 ...
## $ roll_arm           : num -128 -128 -128 -128 -128 -128 -128 -128 -128 -128 ...
## $ pitch_arm          : num  22.5 22.5 22.5 22.1 22.1 22 21.9 21.8 21.7 21.6 ...
## $ yaw_arm            : num -161 -161 -161 -161 -161 -161 -161 -161 -161 -161 ...
## $ total_accel_arm     : int   34 34 34 34 34 34 34 34 34 34 ...
## $ var_accel_arm       : num  NA NA NA NA NA NA NA NA NA NA ...
## $ avg_roll_arm        : num  NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_roll_arm     : num  NA NA NA NA NA NA NA NA NA NA ...
## $ var_roll_arm        : num  NA NA NA NA NA NA NA NA NA NA ...
## $ avg_pitch_arm       : num  NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_pitch_arm    : num  NA NA NA NA NA NA NA NA NA NA ...
## $ var_pitch_arm       : num  NA NA NA NA NA NA NA NA NA NA ...
## $ avg_yaw_arm         : num  NA NA NA NA NA NA NA NA NA NA ...
## $ stddev_yaw_arm      : num  NA NA NA NA NA NA NA NA NA NA ...
## $ var_yaw_arm         : num  NA NA NA NA NA NA NA NA NA NA ...
## $ gyros_arm_x         : num   0 0.02 0.02 0.02 0 0.02 0 0.02 0.02 0.02 ...
## $ gyros_arm_y         : num   0 -0.02 -0.02 -0.03 -0.03 -0.03 -0.03 -0.02 -0.03 -0.03 ...
## $ gyros_arm_z         : num  -0.02 -0.02 -0.02 0.02 0 0 0 0 -0.02 -0.02 ...
## $ accel_arm_x         : int -288 -290 -289 -289 -289 -289 -289 -289 -288 -288 ...
## $ accel_arm_y         : int  109 110 110 111 111 111 111 111 109 110 ...
## $ accel_arm_z         : int -123 -125 -126 -123 -123 -122 -125 -124 -122 -124 ...
## $ magnet_arm_x        : int -368 -369 -368 -372 -374 -369 -373 -372 -369 -376 ...
## $ magnet_arm_y        : int  337 337 344 344 337 342 336 338 341 334 ...
## $ magnet_arm_z        : int  516 513 513 512 506 513 509 510 518 516 ...
## $ kurtosis_roll_arm   : num  NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_pitch_arm  : num  NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_yaw_arm    : num  NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_roll_arm   : num  NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_pitch_arm  : num  NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_yaw_arm    : num  NA NA NA NA NA NA NA NA NA NA ...
## $ max_roll_arm        : num  NA NA NA NA NA NA NA NA NA NA ...
## $ max_pitch_arm       : num  NA NA NA NA NA NA NA NA NA NA ...
## $ max_yaw_arm         : int  NA NA NA NA NA NA NA NA NA NA ...
## $ min_roll_arm        : num  NA NA NA NA NA NA NA NA NA NA ...
## $ min_pitch_arm       : num  NA NA NA NA NA NA NA NA NA NA ...
## $ min_yaw_arm         : int  NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_roll_arm  : num  NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_pitch_arm : num  NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_yaw_arm   : int  NA NA NA NA NA NA NA NA NA NA ...
## $ roll_dumbbell       : num  13.1 13.1 12.9 13.4 13.4 ...
## $ pitch_dumbbell      : num -70.5 -70.6 -70.3 -70.4 -70.4 ...
## $ yaw_dumbbell        : num -84.9 -84.7 -85.1 -84.9 -84.9 ...
## $ kurtosis_roll_dumbbell : num  NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_pitch_dumbbell : num  NA NA NA NA NA NA NA NA NA NA ...
## $ kurtosis_yaw_dumbbell : logi  NA NA NA NA NA NA ...
## $ skewness_roll_dumbbell : num  NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_pitch_dumbbell : num  NA NA NA NA NA NA NA NA NA NA ...
## $ skewness_yaw_dumbbell : logi  NA NA NA NA NA NA ...
## $ max_roll_dumbbell   : num  NA NA NA NA NA NA NA NA NA NA ...
## $ max_pitch_dumbbell  : num  NA NA NA NA NA NA NA NA NA NA ...
## $ max_yaw_dumbbell    : num  NA NA NA NA NA NA NA NA NA NA ...
## $ min_roll_dumbbell   : num  NA NA NA NA NA NA NA NA NA NA ...
## $ min_pitch_dumbbell  : num  NA NA NA NA NA NA NA NA NA NA ...

```

```
## $ min_yaw_dumbbell      : num  NA NA NA NA NA NA NA NA NA NA NA ...
## $ amplitude_roll_dumbbell : num  NA NA NA NA NA NA NA NA NA NA NA ...
## [list output truncated]
```

The training set has the 19622 rows and 160 columns.

```
#str(train)
```

## Splitting the training set

Going to split the training set into two, so that I can train and test the model before I run the test set through the model. For reproducible results I am going to first set the seed. As I have 19622 results I am going to use 60% of the training set to generate the model and then the remaining 40% to test the model.

```
set.seed(8872)
```

As there are a large number of features in the data set, I am going to reduce this number by removing the number of variables.

```
#train the model
#randForest <- train(classe~., data=trainSet, method="rf", ntree=500, proxy=TRUE)

#randomForest
#randomForest <- train(classe ~., data = trainSet, method="rf", ntree=500, allowParallel = TRUE)

#randomForset
```

```
# validate the model on the
#predictTrainingSet <- predict(randomForset, validationSet)
#confusionMatrix(predictTrainingSet, validationSet$classe)
```

```
# predict on the test set
#predict(randomForest, testSet)
```