2017 Coursera Practical Machine Learning Assigment

Analysis of the Weight Lifting Exercises Dataset

Carlos Crosetti (carlos.crosetti@gmail.com) - August 28, 2017.

Executive Summary

People understand how much of a particular physical activity they do, but they rarely quantify how well they do it. This project goal is to analyze data from accelerometers on the belt, forearm, arm, and dumbell of six participants. They were asked to perform barbell lifts correctly and incorrectly in five different ways. For more information see the “Weight Lifting Exercises Dataset” in the following location:

http://groupware.les.inf.puc-rio.br/har

Specifically, the goal of this machine learning capstone project is to predict the manner in which the participants did the exercise–that is, to predict the “classe” variable found in the training set. The prediction model will then be used to predict twenty different test cases, as provided in the testing dataset. Such test cases are ultiimately submitted back to the Coursera platform, to be used for both grading and as evidence of the completion of the produced predictive model developed by this project author.

As a secondary technical objective, is to convey this work by the means of deliberately using "Rattle" (an open-source front-end GUI to R) as the primary R package, that will load "on the fly" many other packages to accomplish the required tasks. The R code generated by Rattle will be included In Appendix C.

To learn about Rattle, please visit https://rattle.togaware.com/

Platform and Tools

To run this project an Intel Pentium dual core PC with 4mb or RAM and Windows 10 was used, loaded with 64-bit of R software version is 3.4.1 and Rattle version 5.0.19.

Exploratory data analysis

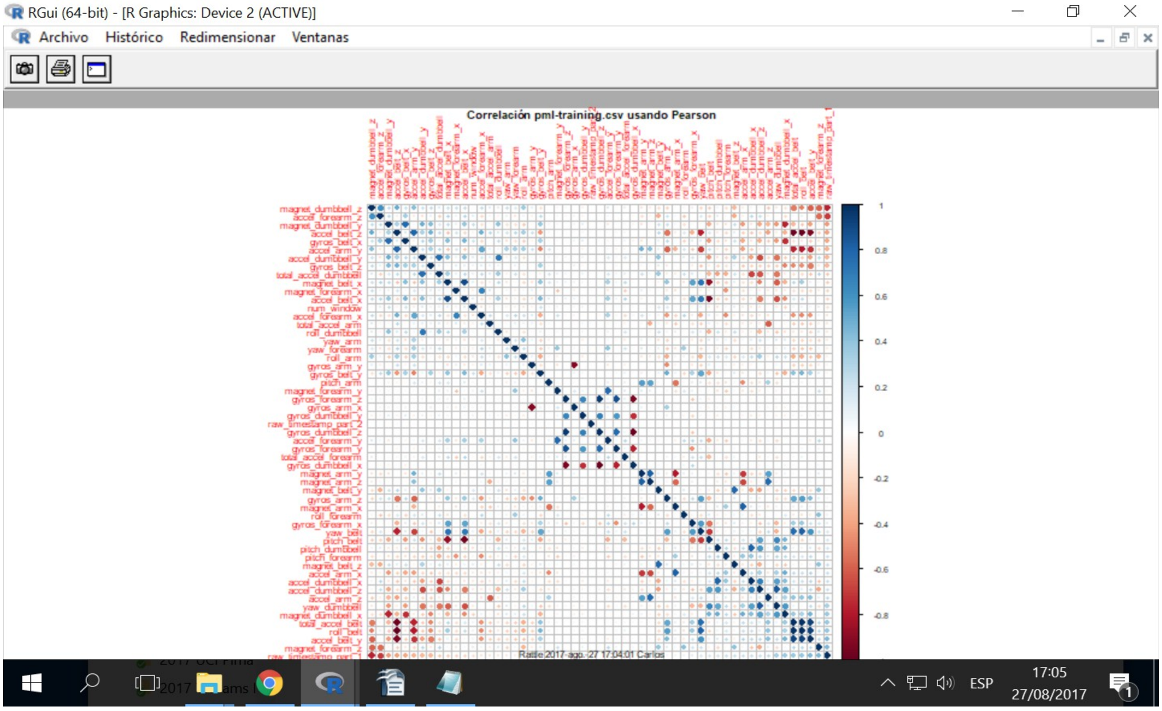
The training and testing datasets used in the analysis may be found at:

Training dataset:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv

Testing dataset:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

A quick inspection if the training dataset structure reveals 160 columns, with lots of variables populated with missing values. Variables showing more than 90% missing, were marked to be ignored by Rattle. Some variables (like "new window") showing constant values were also exluded from the model. Looking to the statistical summary (Appendix B) variables were checked on their mean, min-max and percentiles. This process helped to continue rejecting some variables that were not adeuqate to be kept as predictors to feed the models.

Second, a correlation analysis was made to graphically appreciate that a significant number of useful variables (57 total) showed loose correlation among them.


RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
Archivo
Histórico
Redimensionar
Ventanas
Correlación pml-training.csv usando Pearson
magnet_dumbbell_z
accel_forearm_z
magnet_dumbbell_y
accel_belt_z
gyros_belt_x
accel_arm_y
accel_dumbbell_y
gyros_belt_z
total_accel_dumbbell
magnet_belt_x
magnet_forearm_x
accel_belt_x
num_window
accel_forearm_x
total_accel_arm
roll_dumbbell
yaw_arm
yaw_forearm
roll_arm
gyros_arm_y
gyros_belt_y
pitch_arm
magnet_forearm_y
gyros_forearm_z
gyros_arm_x
gyros_dumbbell_y
raw_timestamp_part_2
gyros_dumbbell_z
accel_forearm_y
gyros_forearm_y
total_accel_forearm
gyros_dumbbell_x
magnet_arm_y
magnet_arm_z
magnet_belt_y
gyros_arm_z
magnet_arm_x
roll_forearm
gyros_forearm_x
yaw_belt
pitch_belt
pitch_dumbbell
pitch_forearm
magnet_belt_z
accel_arm_x
accel_dumbbell_x
accel_dumbbell_z
accel_arm_z
yaw_dumbbell
magnet_dumbbell_x
total_accel_belt
roll_belt
accel_belt_y
magnet_forearm_z
raw_timestamp_part_1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
Rattle 2017-ago.-27 17:04:01 Carlos
ESP
17:05
27/08/2017

Model building

Based on the number of predictors and the data distributions observed, two model candidates were chosen: Support Vector Machine (SVM) and Random Forest (RF) to start with.

The training set was partitioned into three subsets: 75% for training, 15% for validation and 15% for testing.

A initial run of rhe SVM showed poor results in terms of error (10% approx). Below is the confusion matrix for the SVM run on the validation set.

```
Error matrix for SVM model in pml-training.csv [validate] (count):


      Predicted
Actual   A   B   C   D   E Error
     A 848  23   4   0   0   3.1
     B  54 451  31   0   0  15.9
     C   3  32 457   5   0   8.0
     D   0   5  44 404  34  17.0
     E   0   1   1  33 513   6.4

Error matrix for the SVM model on pml-training.csv [validate] (proportions):


      Predicted
Actual    A    B    C    D    E Error
     A 28.8  0.8  0.1  0.0  0.0   3.1
     B  1.8 15.3  1.1  0.0  0.0  15.9
     C  0.1  1.1 15.5  0.2  0.0   8.0
     D  0.0  0.2  1.5 13.7  1.2  17.0
     E  0.0  0.0  0.0  1.1 17.4   6.4


Overall error: 9.3%, Averaged class error: 10.08%
```

The above SVM based model was discarded to immediately try the Random Forest algorithm (RF, set to 500 trees), that showed a significant improvement, an OOB estimate of 0.14% . Below is the RF summary:

```
Summary of the Random Forest Model
==================================
Number of observations used to build the model: 13735
Call:
 randomForest(formula = classe ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 500, mtry = 100, importance = TRUE, replace = FALSE, na.action = na.omit)
Type of random forest: classification
              Number of trees: 500
No. of variables tried at each split: 57
        OOB estimate of  error rate: 0.14%
Confusion matrix:
     A    B    C    D    E class.error
A 3898    0    0    0    0 0.000000000
B    1 2661    3    0    0 0.001500938
C    0    2 2411    1    0 0.001242751
D    0    0    5 2250    2 0.003101462
E    0    0    0    5 2496 0.001999200
```

Conclusion: predicting weight lifting with Random Forest

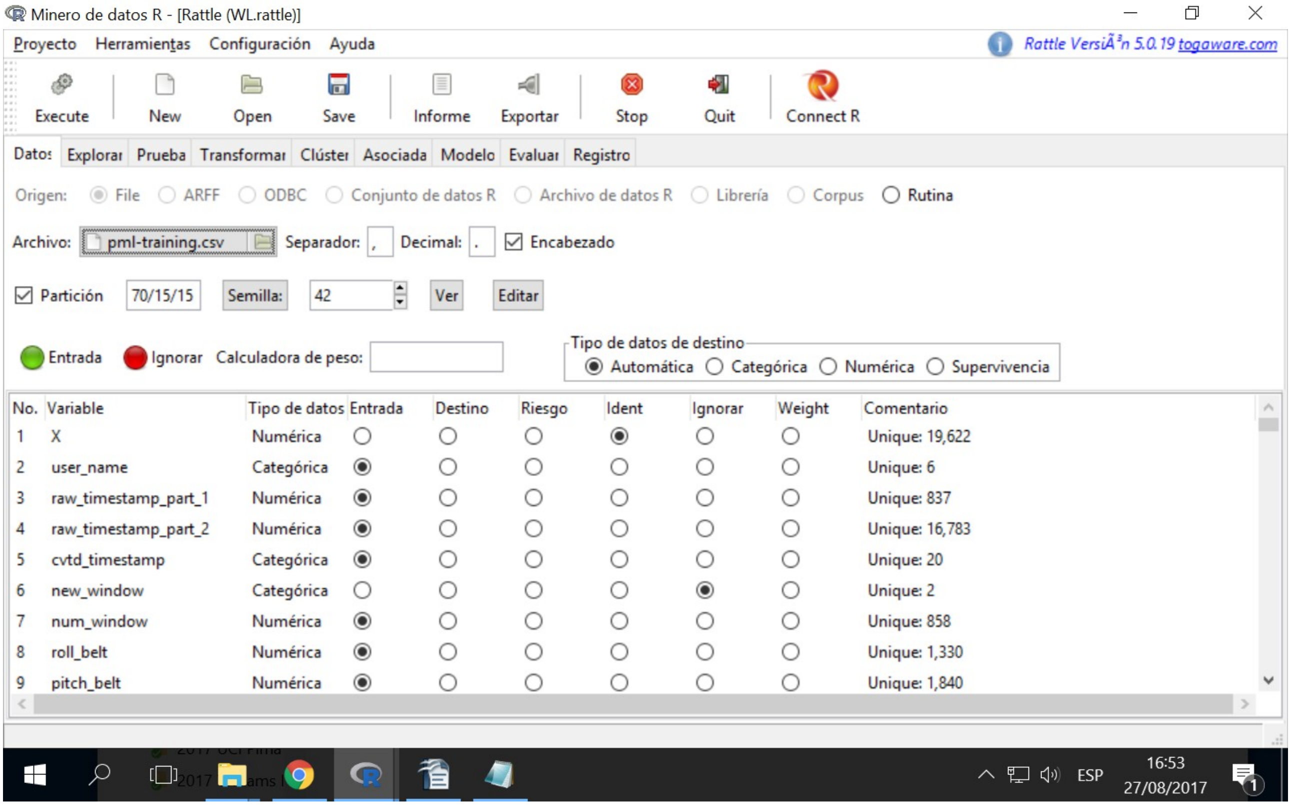
Random Forest was found as an adequate model to predict the weight lifting excercise.
The answer variable to be predicted was categorical with 5 levels, all possible values were A, B, C, D and E. OOB error rates are charted in Appendix D.

For both the SVM and RF models, the test set of 20 values (found in file "pml-testing.csv") were scored with Rattle into a CSV output file (named "pml-training_score_idents.csv"), producing the following results. The column #2 "rf" was used to provide feedback to the Coursera assignemt quiz.

```
"X"   "rf"   "ksvm"
   1 "A"    "A"
   2 "A"    "B"
   3 "B"    "B"
   4 "A"    "A"
   5 "A"    "A"
   6 "C"    "E"
   7 "C"    "D"
   8 "B"    "B"
   9 "A"    "D"
  10 "A"    "B"
  11 "B"    "C"
  12 "C"    "A"
  13 "B"    "D"
  14 "A"    "A"
  15 "C"    "E"
  16 "E"    "A"
  17 "E"    "E"
  18 "B"    "B"
  19 "A"    "A"
  20 "B"    "B"
```

Note: in order to evaluate the test set (file "pml-testing.csv") with Rattle, it was required to add the "classe" column to the file filling it with dummy values including at least one for each level (from Ä" to "E").

Appendix A. Rattle – Variable classification (type, input, ignore, answer)


Minero de datos R - [Rattle (WL.rattle)]
Proyecto Herramientas Configuración Ayuda
Rattle VersiÃ³n 5.0.19 togaware.com
Execute New Open Save Informe Exportar Stop Quit Connect R
Datos Explorar Prueba Transformar Clúster Asociada Modelo Evaluar Registro
ESP
16:53
27/08/2017

Origen: ◉ File ○ ARFF ○ ODBC ○ Conjunto de datos R ○ Archivo de datos R ○ Librería ○ Corpus ○ Rutina

Archivo: pml-training.csv Separador: , Decimal: . ☑ Encabezado

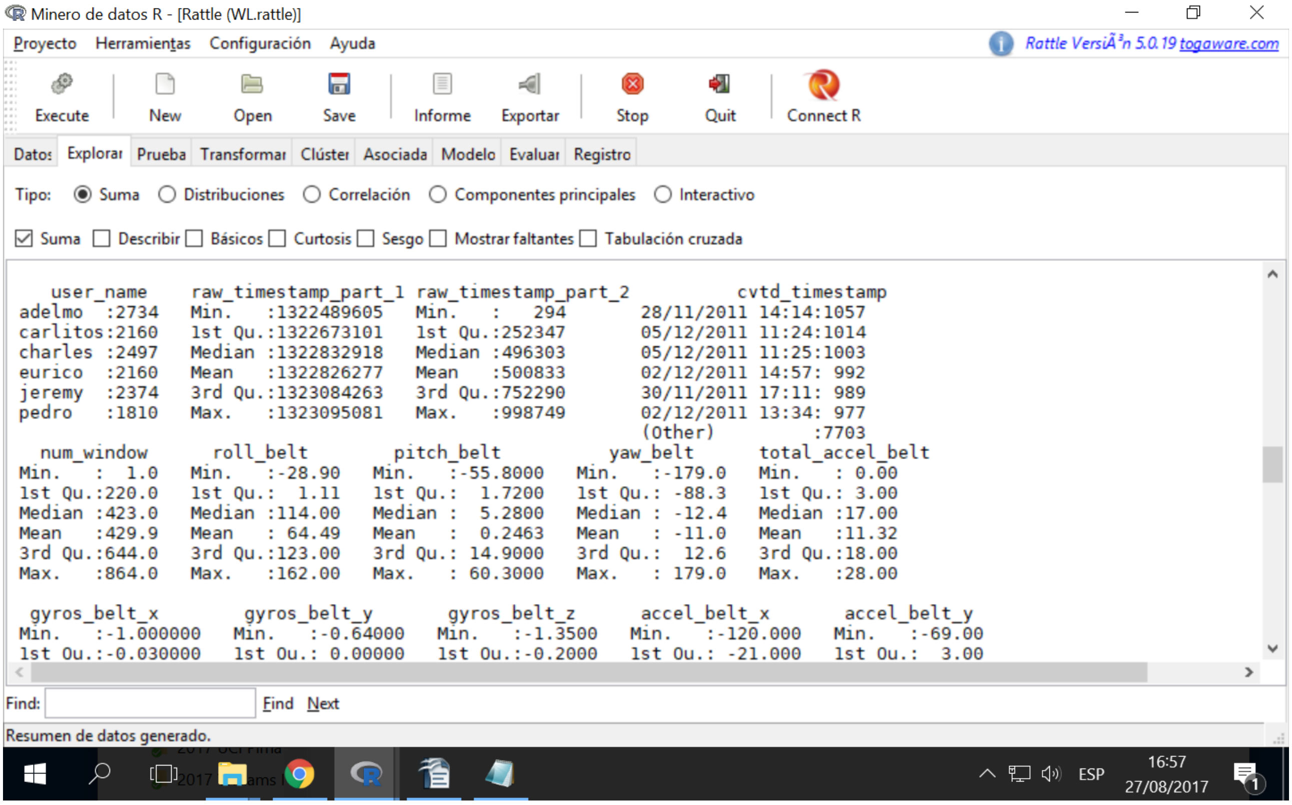
☑ Partición 70/15/15 Semilla: 42 Ver Editar

🟢 Entrada 🔴 Ignorar Calculadora de peso:

Tipo de datos de destino: ● Automática ○ Categórica ○ Numérica ○ Supervivencia

No.	Variable	Tipo de datos	Entrada	Destino	Riesgo	Ident	Ignorar	Weight	Comentario
1	X	Numérica	○	○	○	●	○	○	Unique: 19,622
2	user_name	Categórica	●	○	○	○	○	○	Unique: 6
3	raw_timestamp_part_1	Numérica	●	○	○	○	○	○	Unique: 837
4	raw_timestamp_part_2	Numérica	●	○	○	○	○	○	Unique: 16,783
5	cvtd_timestamp	Categórica	●	○	○	○	○	○	Unique: 20
6	new_window	Categórica	○	○	○	○	●	○	Unique: 2
7	num_window	Numérica	●	○	○	○	○	○	Unique: 858
8	roll_belt	Numérica	●	○	○	○	○	○	Unique: 1,330
9	pitch_belt	Numérica	●	○	○	○	○	○	Unique: 1,840

Appendix B. Rattle - Statistical summary.



Tipo: ● Suma ○ Distribuciones ○ Correlación ○ Componentes principales ○ Interactivo

☑ Suma □ Describir □ Básicos □ Curtosis □ Sesgo □ Mostrar faltantes □ Tabulación cruzada

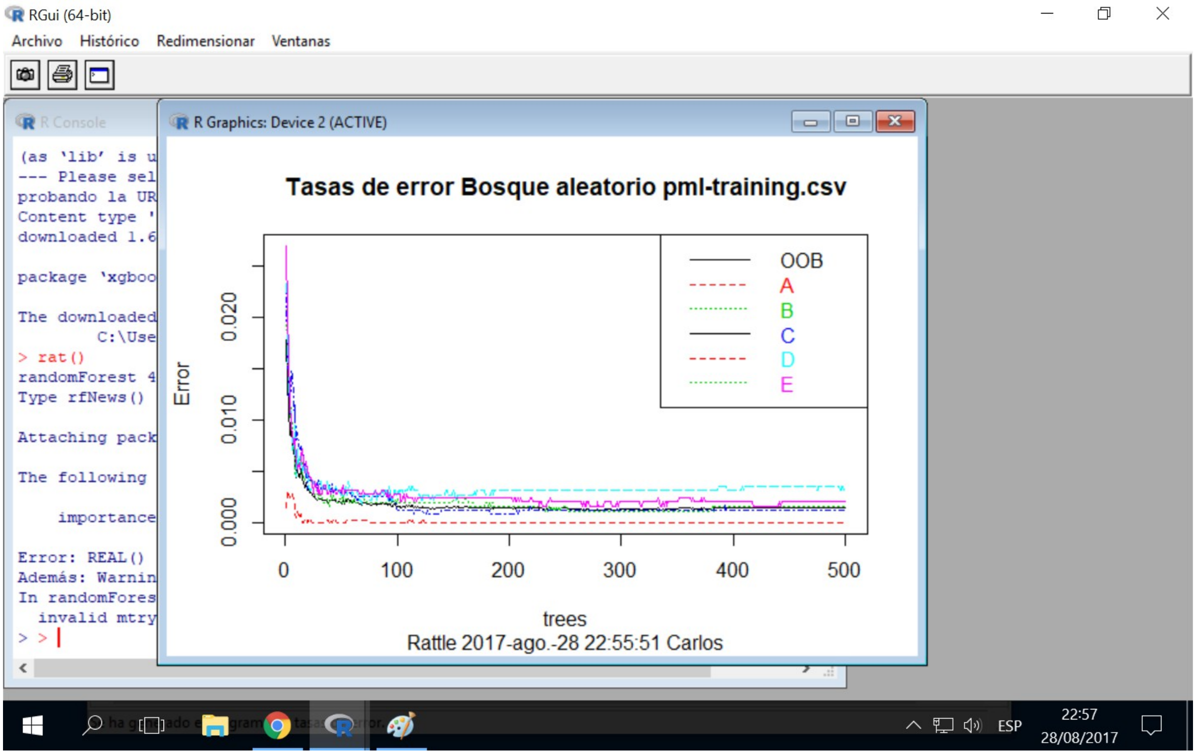
```
   user_name    raw_timestamp_part_1 raw_timestamp_part_2          cvtd_timestamp
 adelmo  :2734   Min.   :1322489605   Min.   :   294        28/11/2011 14:14:1057
 carlitos:2160   1st Qu.:1322673101   1st Qu.:252347        05/12/2011 11:24:1014
 charles :2497   Median :1322832918   Median :496303        05/12/2011 11:25:1003
 eurico  :2160   Mean   :1322826277   Mean   :500833        02/12/2011 14:57: 992
 jeremy  :2374   3rd Qu.:1323084263   3rd Qu.:752290        30/11/2011 17:11: 989
 pedro   :1810   Max.   :1323095081   Max.   :998749        02/12/2011 13:34: 977
                                                            (Other)         :7703
   num_window       roll_belt         pitch_belt           yaw_belt      total_accel_belt
 Min.   :  1.0   Min.   :-28.90   Min.   :-55.8000   Min.   :-179.0   Min.   : 0.00
 1st Qu.:220.0   1st Qu.:  1.11   1st Qu.:  1.7200   1st Qu.: -88.3   1st Qu.: 3.00
 Median :423.0   Median :114.00   Median :  5.2800   Median : -12.4   Median :17.00
 Mean   :429.9   Mean   : 64.49   Mean   :  0.2463   Mean   : -11.0   Mean   :11.32
 3rd Qu.:644.0   3rd Qu.:123.00   3rd Qu.: 14.9000   3rd Qu.:  12.6   3rd Qu.:18.00
 Max.   :864.0   Max.   :162.00   Max.   : 60.3000   Max.   : 179.0   Max.   :28.00

  gyros_belt_x        gyros_belt_y       gyros_belt_z       accel_belt_x       accel_belt_y
 Min.   :-1.000000   Min.   :-0.64000   Min.   :-1.3500   Min.   :-120.000   Min.   :-69.00
 1st Qu.:-0.030000   1st Qu.: 0.00000   1st Qu.:-0.2000   1st Qu.: -21.000   1st Qu.:  3.00
```

Resumen de datos generado.

Appendix C. Rattle generated R code.

Refer ro file "wl.R" file in the repository.

Appendix D. Rattle OOB error rates chart for training set.



```
(as ‘lib’ is u
--- Please sel
probando la UR
Content type '
downloaded 1.6

package ‘xgboo

The downloaded
        C:\Use
> rat()
randomForest 4
Type rfNews()

Attaching pack

The following

    importance

Error: REAL()
Además: Warnin
In randomFores
  invalid mtry
> >
```