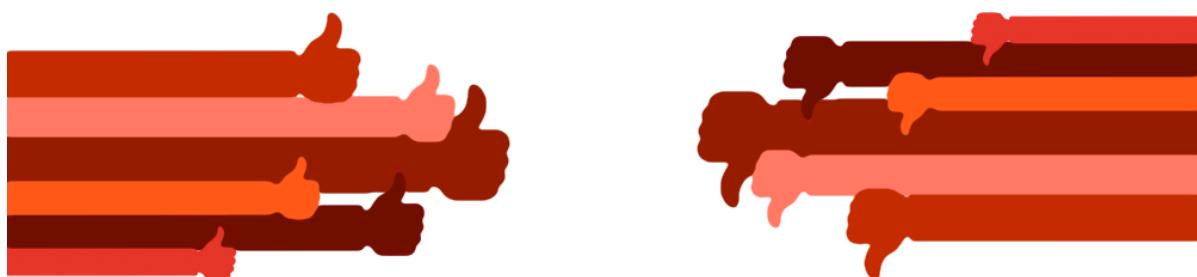

Kaggle / Santander Project



UNIVERSIDAD COMPLUTENSE
MADRID

Data analysis and customer satisfaction predictions of the Kaggle/Santander dataset using SAS.



Complutense University of Madrid

Caio Fernandes Moreno (caiofern@ucm.es)
Madrid, Spain, September, 3, 2016.

The Santander Project	3
The dataset.....	4
<i>Understanding the data</i>	4
Feature Selection	7
<i>Using R and SAS to find the important variables</i>	7
Data Preparation	22
<i>Preparing the data to make better predictions</i>	22
Logistic Regression.....	35
<i>Predicting using Logistic Regression</i>	35
Neuronal Networks	51
<i>Using neuronal networks to predict</i>	51
Trees	64
<i>Using Bagging, Random Forest and Gradient Boosting to predict</i>	64
Ensemble Models	80
<i>Using neuronal networks to predict</i>	80
Conclusions	86
References.....	87
Extras.....	88

The Santander Project

The Santander Project is an academic research project developed by Caio Moreno student in the Master of Science in Data Mining and Business Intelligence from the Statistics Department at “Universidad Complutense de Madrid” or Complutense University of Madrid.

This project is a class assignment requested by Prof. Dr. (Phd) Javier Portela in the class of “Neuronal Networks and Genetic Algorithms”.

For more detailed information about the class assignment goals please contact the author by e-mail.

The project is oriented to predict a binary variable using different classification algorithms using SAS, SAS Enterprise Miner and R.

Chapter Number One

The dataset

Understanding the data

The author decided to use this dataset because it is a real problem for a real company published in the Kaggle Platform.

Santander, a prominent Spanish bank with Worldwide presence and headquarters in Spain, launched a competition at Kaggle, asking data scientist from all over the world to predict customer satisfaction using the provided dataset.

The screenshot shows the Kaggle interface for the Santander Customer Satisfaction competition. At the top, there's a navigation bar with links for Host, Competitions, Datasets, Kernels, Jobs, Community, and Logout. The user 'caiomsouza' is logged in. Below the navigation is a red header with the Santander logo and the competition title 'Santander Customer Satisfaction'. It also displays the completion status: 'Completed • \$60,000 • 5,123 teams' and the duration: 'Wed 2 Mar 2016 – Mon 2 May 2016 (4 months ago)'. On the left, there's a sidebar with links for Dashboard, Home, Data, Make a submission, Information, Description, Evaluation, Rules, Prizes, Timeline, Forum, Kernels, New Script, New Notebook, Leaderboard, Public, Private, My Team, Your model, and My Submissions. The main content area has a heading 'Which customers are happy customers?' followed by a descriptive paragraph about customer satisfaction and a quote from Santander Bank. It also mentions that Kagglers will work with anonymized features to predict customer satisfaction. At the bottom, there are two decorative icons: one showing thumbs up and another showing thumbs down.

Figure 1 - Kaggle Santander Competition

Competition link:

<https://www.kaggle.com/c/santander-customer-satisfaction>

The dataset provided by Santander Bank is anonymised and contains 371 variables (all continuous variables).

A continuous variable is a variable that has an infinite number of possible values. In other words, any value is possible for the variable. A continuous variable is the opposite of a discrete variable, which can only take on a certain number of values.

The TARGET column is the variable to predict. It equals 1 (one) for unsatisfied customers and 0 for satisfied customers.

The Kaggle Competition Objective is to predict who are satisfied and unsatisfied clients.

Numbers of observations (Row number):

- Train: 76020 rows
- Test: 75818 rows

Number of 1s (train): 3008 (3.95%) (Imbalanced Dataset Problem)

Variables:

- 34 variables with one single value; (Action: Delete all of them)
- 100 variables with two unique values; (binary variables)
- 157 variables with values between 3 y 101 unique values; (categorical variables)
- 80 variables has more than 101 distinct values; (continuous variables)

Files:

- train.csv - with TARGET variable
- test.csv - without TARGET variable

CSV Files also available at: <https://github.com/caiomsouza/kaggle-competitions/tree/master/santander-customer-satisfaction/dat>

Unbalanced Dataset Problem

A dataset is defined by unbalanced when the class of interest (minority class) is much rarer than normal behaviour (majority class). The cost of missing a minority class is typically much higher than missing a majority class. Most learning systems are not

prepared to cope with unbalanced data and several techniques have been proposed to rebalance the classes.

Some links about Unbalanced Dataset Problem:

- <http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>
- <http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>
- <http://florianhartl.com/thoughts-on-machine-learning-dealing-with-skewed-classes.html>
- <https://cran.r-project.org/web/packages/unbalanced/unbalanced.pdf>
- <https://www3.nd.edu/~dial/publications/chawla2005data.pdf>

Chapter Number Two

Feature Selection

Using R and SAS to find the important variables

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

In this project I will use R, SAS and SAS Enterprise Miner to the Feature Selection Phase.

The first attempt to find the important variables was made using R and H2o. It was published at github on June, 18, 2016.

As we can see in the link below it is possible to find 6 CSV files, all of them show us the important variables according with different algorithms.

https://github.com/caiomsouza/kaggle-competitions/tree/master/santander-customer-satisfaction/outputs/variables_importance

[kaggle-competitions / santander-customer-satisfaction / outputs / variables_importance /](#)

 caiomsouza	Add Kaggle Santander Competion	Latest commit 4da8a14 on Jun 18
..		
variable_importances-h2o.glm.model.2.csv	Add Kaggle Santander Competion	3 months ago
variable_importances-h2o.glm.model1.csv	Add Kaggle Santander Competion	3 months ago
variable_importances-h2o.glm.model2.csv	Add Kaggle Santander Competion	3 months ago
variable_importances-h2o.randomForest.model1.csv	Add Kaggle Santander Competion	3 months ago
variable_importances-h2o.randomForest.model2.csv	Add Kaggle Santander Competion	3 months ago
variable_importances-h2o.randomForest.model3.csv	Add Kaggle Santander Competion	3 months ago

I used H2o GLM and Random Forest algorithms to find the important variables, all R codes to perform it are available at <https://github.com/caiomsouza/kaggle-competitions/tree/master/santander-customer-satisfaction/src>

H2o GLM (Generalized Linear Models)

In the image below we can find the results for the H2o GLM algorithm showing the first 25 variables ranked by importance.

[kaggle-competitions](#) / santander-customer-satisfaction / outputs / variables_importance / variable_importances-h2o.glm.model1.csv

names	coefficients	sign
var38	0.66704519277191	NEG
num_var42	0.510454334820693	NEG
ind_var13	0.502951833851412	NEG
num_meses_var5_ult3	0.490782579151765	NEG
var15	0.489397707798164	POS
ind_var30	0.428169375246911	NEG
num_var5	0.419621180386955	POS
saldo_medio_var8_hace2	0.414560751936344	NEG
saldo_medio_var8_ult1	0.356956311455839	NEG
saldo_var5	0.355705422181016	NEG
saldo_medio_var8_ult3	0.293700838139008	POS
ind_var17	0.267243897839386	POS
num_meses_var13_largo_ult3	0.265601479633978	NEG
num_meses_var17_ult3	0.260761783613886	NEG
num_var30	0.219071391185197	POS
saldo_medio_var8_hace3	0.208042273405786	NEG
num_aport_var13_hace3	0.185985659629869	NEG
ind_var30_0	0.178127322892657	POS
num_var40_0	0.158647511504273	POS
num_meses_var39_vig_ult3	0.151681924644899	POS
num_var1_0	0.14493257533079	POS
num_var22_ult3	0.137618097039769	POS
saldo_medio_var13_corto_hace2	0.13369133576451	POS
num_reemb_var17_ult1	0.130156191286175	POS

The variable var38 has a coefficient of 0.6670 in the H2o GLM algorithm. It means that for this model this variable is the most important variable to predict the target variable in this case customer success. After this variable we can see others like num_var42, ind_var13, num_meses_var5_ult3, var15, ind_var30, etc.

Please check all variables list in model 1 in the link below:

https://github.com/caiomsouza/kaggle-competitions/blob/master/santander-customer-satisfaction/outputs/variables_importance/variable_importances-h2o.glm.model1.csv

H2o Random Forest

The next algorithm used to check variable importance was H2o Random Forest.

In the image below we can find the results for the H2o Random Forest algorithm showing the first 25 variables ranked by importance.

Using H2o Random Forest the variable list as we can see is different.

We need to ignore the ID variable, it needs to be removed. The others variables we can consider as important variables.

[kaggle-competitions](#) / [santander-customer-satisfaction](#) / [outputs](#) / [variables_importance](#) / [variable_importances-h2o.randomForest.model1.csv](#)

caiomsouza Add Kaggle Santander Competition 1354372 on Jun 18
1 contributor

338 lines (337 sloc) | 22.4 KB Raw Blame History

Q Search this file...

variable	relative_importance	scaled_importance	percentage
var15	10909.0517578125	1	0.148798832359088
ID	2659.28442382812	0.243768613704091	0.0362724850849623
num_meses_var5_ult3	2206.24829101562	0.20224015248948	0.0300930985465585
num_var4	1848.6728515625	0.169462286237534	0.0252157903210467
num_var35	1719.08117675781	0.157583006747282	0.0234481674036298
num_var22_ult3	1535.74816894531	0.14077742071812	0.0209475158253803
num_var30	1507.79919433594	0.138215422184254	0.0205662934350355
var38	1505.36340332031	0.137992140539827	0.0205330693870575
num_var42	1481.63098144531	0.135816660727111	0.0202093605311045
ind_var30	1343.49267578125	0.123153937263072	0.0183251620651696
num_var45_ult3	1282.47436523438	0.117560572055764	0.0174928758533641
num_var45_hace3	1276.70922851562	0.11703209929326	0.01741423972337
num_var45_hace2	1242.11157226562	0.113860636088383	0.0169423297016145
num_meses_var39_vig_ult3	1196.62353515625	0.109690884388672	0.0163218755174701
num_var22_hace3	1194.80212402344	0.109523921102288	0.0162970315754094
num_var22_ult1	1191.45703125	0.109217286497586	0.0162514047042688
num_var22_hace2	1164.24267578125	0.106722628293287	0.0158802024763341
num_med_var45_ult3	1088.37194824219	0.099767786642203	0.0148453301594105
num_var5	1064.62060546875	0.0975905724075719	0.0145213632235017
num_var45_ult1	1023.42626953125	0.093814411394494	0.0139594748739559
var36	997.096069335938	0.0914008010477967	0.0136003324725975
num_med_var22_ult3	984.830993652344	0.0902764984084945	0.0134330375526511
ind_var5	883.565124511719	0.0809937604227567	0.0120517769792779
num_op_var39_efect_ult3	681.407287597656	0.0624625588662798	0.00929435582546325

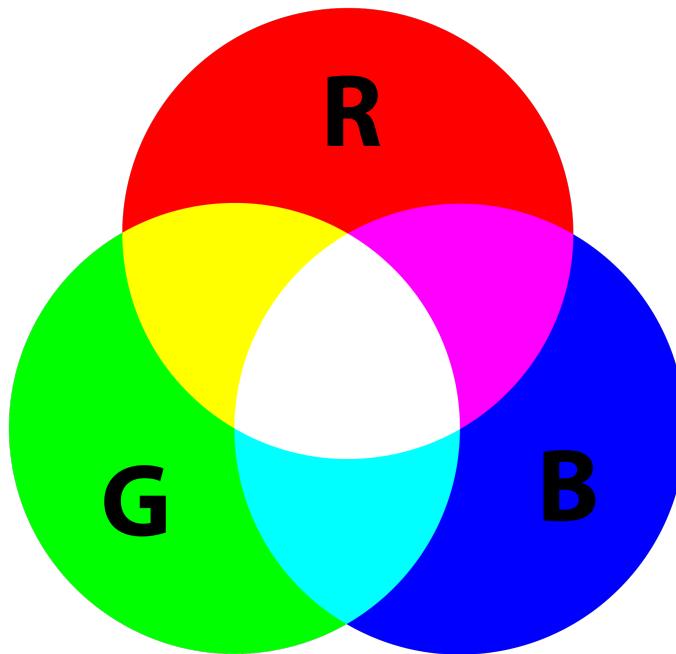
Please check all variables list in Random Forest model in the link below:

https://github.com/caiomsouza/kaggle-competitions/blob/master/santander-customer-satisfaction/outputs/variables_importance/variable_importances-h2o.randomForest.model1.csv

Mixing variable importance results of different algorithms to find the best features for our model.

The image below give us the idea of mixing R, G and B. Using the same principle, we can use this same technique to find the best variables for our model.

Let's imagine that R, G and B are different algorithm. We could consider only variables that are important in all models.



Using SAS to Feature Selection

In the image below we can see a snapshot of some variables and data we have in the dataset.

The SAS System											
Obs	ID	VAR2	VAR3	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_comer_ult3	imp_op_var40_efc	imp_op_var40_efe	
1	1	2	23		0		0		0		
2	3	2	34		0		0		0		
3	4	2	23		0		0		0		
4	8	2	37		0	195	195		0		
5	10	2	39		0		0		0		
6	13	2	23		0		0		0		
7	14	2	27		0		0		0		
8	18	2	26		0		0		0		
9	20	2	45		0		0		0		
10	23	2	25		0		0		0		
11	25	2	42		0		0		0		
12	26	2	26		0		0		0		
13	29	2	51		0		0		0		
14	31	2	43		0		0		0		
15	32	2	33	600		1086.48	1952.91		0		
16	34	2	30		0		0		0		
17	36	2	44		0		0		0		
18	39	2	36		0	55.2	70.95		0		

We will use the dataset “train”. Using SAS we can see that the dataset has 76020 rows and 371 variables.

The SAS System			
The CONTENTS Procedure			
Data Set Name	UCM.SANTANDER	Observations	76020
Member Type	DATA	Variables	371
Engine	V9	Indexes	0
Created	07/09/2016 20:30:43	Observation Length	2968
Last Modified	07/09/2016 20:30:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

A snapshot of some variables.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	ID	Num	8	BEST12.	BEST32.
371	TARGET	Num	8	BEST12.	BEST32.
2	VAR2	Num	8	BEST12.	BEST32.
3	VAR3	Num	8	BEST12.	BEST32.
196	delta_imp_amort_var18_1y3	Num	8	BEST12.	BEST32.
197	delta_imp_amort_var34_1y3	Num	8	BEST12.	BEST32.
198	delta_imp_aport_var13_1y3	Num	8	BEST12.	BEST32.
199	delta_imp_aport_var17_1y3	Num	8	BEST12.	BEST32.
200	delta_imp_aport_var33_1y3	Num	8	BEST12.	BEST32.
201	delta_imp_compra_var44_1y3	Num	8	BEST12.	BEST32.
202	delta_imp_reemb_var13_1y3	Num	8	BEST12.	BEST32.
203	delta_imp_reemb_var17_1y3	Num	8	BEST12.	BEST32.
204	delta_imp_reemb_var33_1y3	Num	8	BEST12.	BEST32.

The metrics about the variables.

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ID	76020	75964	43782	5774787136	1.00000	151838
VAR2	76020	-1523	39033	-115793609	-999999	238.00000
VAR3	76020	33.21287	12.95649	2524842	5.00000	105.00000
imp_ent_var16_ult1	76020	86.20827	1615	6553552	0	210000
imp_op_var39_comer_ult1	76020	72.36307	339.31583	5501040	0	12888
imp_op_var39_comer_ult3	76020	119.52963	546.26629	9086643	0	21025
imp_op_var40_comer_ult1	76020	3.55913	93.15575	270565	0	8238
imp_op_var40_comer_ult3	76020	6.47270	153.73707	492054	0	11074
imp_op_var40_efect_ult1	76020	0.41295	30.60486	31392	0	6600
imp_op_var40_efect_ult3	76020	0.56735	36.51351	43130	0	6600
imp_op_var40_ult1	76020	3.16072	95.26820	240278	0	8238
imp_op_var41_comer_ult1	76020	68.80394	319.60552	5230475	0	12888
imp_op_var41_comer_ult3	76020	113.05693	512.15482	8594588	0	16567
imp_op_var41_efect_ult1	76020	68.20514	531.89792	5184955	0	45990

The frequency:

The SAS System	
The FREQ Procedure	
Number of Variable Levels	
Variable	Levels
ID	76020
VAR2	208
VAR3	100
imp_ent_var16_ult1	596
imp_op_var39_comer_ult1	7551
imp_op_var39_comer_ult3	9099
imp_op_var40_comer_ult1	293
imp_op_var40_comer_ult3	346
imp_op_var40_efect_ult1	23
imp_op_var40_efect_ult3	29
imp_op_var40_ult1	224
imp_op_var41_comer_ult1	7421
imp_op_var41_comer_ult3	8961
imp_op_var41_efect_ult1	331
imp_op_var41_efect_ult3	454

Analysing the number of missing values:

The SAS System		
The MEANS Procedure		
Variable	N	N Miss
ID	76020	0
VAR2	76020	0
VAR3	76020	0
imp_ent_var16_ult1	76020	0
imp_op_var39_comer_ult1	76020	0
imp_op_var39_comer_ult3	76020	0
imp_op_var40_comer_ult1	76020	0
imp_op_var40_comer_ult3	76020	0
imp_op_var40_efect_ult1	76020	0
imp_op_var40_efect_ult3	76020	0
imp_op_var40_ult1	76020	0
imp_op_var41_comer_ult1	76020	0
imp_op_var41_comer_ult3	76020	0
imp_op_var41_efect_ult1	76020	0
imp_op_var41_efect_ult3	76020	0
imp_op_var41_ult1	76020	0
imp_op_var39_efect_ult1	76020	0
imp_op_var39_efect_ult3	76020	0
imp_op_var39_ult1	76020	0
imp_sal_var16_ult1	76020	0

There is no missing value in this dataset.

SAS Code:

```
PROC IMPORT OUT= UCM.santander_
    DATAFILE= "\\\\vmware-host\\Shared Folders\\git\\Bitbucket\\santander-kaggle\\dataset\\train.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

/* Show 20 rows */
proc print data=ucm.santander (obs=20);
run;

/* Describe the dataset variables */
proc contents data=ucm.santander out=sal;
data; set sal; put name @@; run;

/*
```

```

proc corr data=ucm.santander;
run;
 */

/* Variables
ID TARGET VAR2 VAR3 delta_imp_amort_var18_1y3 delta_imp_amort_var34_1y3
delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3
delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_reemb_var13_1y3 delta_num_reemb_var17_1y3
delta_num_reemb_var33_1y3
delta_num_trasp_var17_in_1y3 delta_num_trasp_var17_out_1y3
delta_num_trasp_var33_in_1y3
delta_num_trasp_var33_out_1y3 delta_num_venta_var44_1y3
imp_amort_var18_hace3 imp_amort_var18_ult1
imp_amort_var34_hace3 imp_amort_var34_ult1 imp_aport_var13_hace3
imp_aport_var13_ult1
imp_aport_var17_hace3 imp_aport_var17_ult1 imp_aport_var33_hace3
imp_aport_var33_ult1
imp_compra_var44_hace3 imp_compra_var44_ult1 imp_ent_var16_ult1
imp_op_var39_comer_ult1
imp_op_var39_comer_ult3 imp_op_var39_efect_ult1 imp_op_var39_efect_ult3
imp_op_var39_ult1
imp_op_var40_comer_ult1 imp_op_var40_comer_ult3 imp_op_var40_efect_ult1
imp_op_var40_efect_ult3
imp_op_var40_ult1 imp_op_var41_comer_ult1 imp_op_var41_comer_ult3
imp_op_var41_efect_ult1
imp_op_var41_efect_ult3 imp_op_var41_ult1 imp_reemb_var13_hace3
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_hace3
imp_reemb_var33_ult1
imp_sal_var16_ult1 imp_trans_var37_ult1 imp_trasp_var17_in_hace3
imp_trasp_var17_in_ult1
imp_trasp_var17_out_hace3 imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1
imp_trasp_var33_out_hace3 imp_trasp_var33_out_ult1 imp_var43_emit_ult1
imp_var7_emit_ult1
imp_var7_recib_ult1 imp_venta_var44_hace3 imp_venta_var44_ult1 ind_var1
ind_var2 ind_var5 ind_var6
ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var18 ind_var19
ind_var20 ind_var24 ind_var25
ind_var26 ind_var27 ind_var28 ind_var29 ind_var30 ind_var31 ind_var32
ind_var33 ind_var34 ind_var37
ind_var39 ind_var40 ind_var41 ind_var44 ind_var46 ind_var10_ult1
ind_var10cte_ult1 ind_var12_0
ind_var13_0 ind_var13_corto ind_var13_corto_0 ind_var13_largo
ind_var13_largo_0 ind_var13_medio
ind_var13_medio_0 ind_var14_0 ind_var17_0 ind_var18_0 ind_var1_0
ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0 ind_var26_cte ind_var27_0 ind_var28_0
ind_var29_0 ind_var2_0 ind_var30_0
ind_var31_0 ind_var32_0 ind_var32_cte ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0
ind_var40_0 ind_var41_0 ind_var43_emit_ult1 ind_var43_recib_ult1
ind_var44_0 ind_var46_0 ind_var5_0
```

ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1 ind_var9_ult1
num_aport_var13_hace3 num_aport_var13_ult1 num_aport_var17_hace3
num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3
num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_hace3
num_reemb_var13_ult1 num_reemb_var17_hace3
num_reemb_var17_ult1 num_reemb_var33_hace3 num_reemb_var33_ult1
num_sal_var16_ult1
num_trasp_var11_ult1 num_trasp_var17_in_hace3 num_trasp_var17_in_ult1
num_trasp_var17_out_hace3
num_trasp_var17_out_ult1 num_trasp_var33_in_hace3 num_trasp_var33_in_ult1
num_trasp_var33_out_hace3
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var18 num_var20 num_var24 num_var25 num_var26 num_var27
num_var28 num_var29 num_var30
num_var31 num_var32 num_var33 num_var34 num_var35 num_var37 num_var39
num_var40 num_var41 num_var42
num_var44 num_var46 num_var12_0 num_var13_0 num_var13_corto
num_var13_corto_0 num_var13_largo
num_var13_largo_0 num_var13_medio num_var13_medio_0 num_var14_0
num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2 num_var22_hace3 num_var22_ult1 num_var22_ult3
num_var24_0 num_var25_0
num_var26_0 num_var27_0 num_var28_0 num_var29_0 num_var2_0_ult1
num_var2_ult1 num_var30_0 num_var31_0
num_var32_0 num_var33_0 num_var34_0 num_var37_0 num_var37_med_ult2
num_var39_0 num_var40_0 num_var41_0
num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1 num_var44_0
num_var45_hace2 num_var45_hace3
num_var45_ult1 num_var45_ult3 num_var46_0 num_var5_0 num_var6_0
num_var7_emit_ult1 num_var7_recib_ult1
num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2 saldo_medio_var12_hace3
saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_hace3
saldo_medio_var13_medio_ult1 saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2

```

saldo_medio_var17_hace3 saldo_medio_var17_ult1 saldo_medio_var17_ult3
saldo_medio_var29_hace2
saldo_medio_var29_hace3 saldo_medio_var29_ult1 saldo_medio_var29_ult3
saldo_medio_var33_hace2
saldo_medio_var33_hace3 saldo_medio_var33_ult1 saldo_medio_var33_ult3
saldo_medio_var44_hace2
saldo_medio_var44_hace3 saldo_medio_var44_ult1 saldo_medio_var44_ult3
saldo_medio_var5_hace2
saldo_medio_var5_hace3 saldo_medio_var5_ult1 saldo_medio_var5_ult3
saldo_medio_var8_hace2
saldo_medio_var8_hace3 saldo_medio_var8_ult1 saldo_medio_var8_ult3
saldo_var1 saldo_var5 saldo_var6
saldo_var8 saldo_var12 saldo_var13 saldo_var14 saldo_var17 saldo_var18
saldo_var20 saldo_var24
saldo_var25 saldo_var26 saldo_var27 saldo_var28 saldo_var29 saldo_var30
saldo_var31 saldo_var32
saldo_var33 saldo_var34 saldo_var37 saldo_var40 saldo_var41 saldo_var42
saldo_var44 saldo_var46
saldo_var13_corto saldo_var13_largo saldo_var13_medio saldo_var2_ult1
var21 var36 var38

*/
data ucm.santander; set ucm.santander; id=_n_; run;

/* Show the number of missing values */
ods output nlevels=niveles; proc freq data=ucm.santander nlevels; tables
_all_ / noprint; run;
proc means data=ucm.santander n nmiss; run;

/*
76020*0.7 = 53.214 rows
70% of the dataset is equal 53.214 rows
*/
data ucm.santander; set ucm.santander; Run;
data uno;set ucm.santander; u=(ranuni(12355));
proc sort data=uno; by u;
data train test;
set uno; if _n_<=53214 then output train;else output test;run;

```

Using SAS base we will select the independent variables that will explain better and use them to the classification model.

The SAS procedure will be Proc Logistic with the selection method by step (called stepwise).

```

Proc logistic data=uno;
class ;
model TARGET=VAR2 VAR3 delta_imp_amort_var18_1y3
delta_imp_amort_var34_1y3 delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3

```

delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_reemb_var13_1y3 delta_num_reemb_var17_1y3
delta_num_reemb_var33_1y3
delta_num_trasp_var17_in_1y3 delta_num_trasp_var17_out_1y3
delta_num_trasp_var33_in_1y3
delta_num_trasp_var33_out_1y3 delta_num_venta_var44_1y3
imp_amort_var18_hace3 imp_amort_var18_ult1
imp_amort_var34_hace3 imp_amort_var34_ult1 imp_aport_var13_hace3
imp_aport_var13_ult1
imp_aport_var17_hace3 imp_aport_var17_ult1 imp_aport_var33_hace3
imp_aport_var33_ult1
imp_compra_var44_hace3 imp_compra_var44_ult1 imp_ent_var16_ult1
imp_op_var39_comer_ult1
imp_op_var39_comer_ult3 imp_op_var39_efect_ult1 imp_op_var39_efect_ult3
imp_op_var39_ult1
imp_op_var40_comer_ult1 imp_op_var40_comer_ult3 imp_op_var40_efect_ult1
imp_op_var40_efect_ult3
imp_op_var40_ult1 imp_op_var41_comer_ult1 imp_op_var41_comer_ult3
imp_op_var41_efect_ult1
imp_op_var41_efect_ult3 imp_op_var41_ult1 imp_reemb_var13_hace3
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_hace3
imp_reemb_var33_ult1
imp_sal_var16_ult1 imp_trans_var37_ult1 imp_trasp_var17_in_hace3
imp_trasp_var17_in_ult1
imp_trasp_var17_out_hace3 imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1
imp_trasp_var33_out_hace3 imp_trasp_var33_out_ult1 imp_var43_emit_ult1
imp_var7_emit_ult1
imp_var7_recib_ult1 imp_venta_var44_hace3 imp_venta_var44_ult1 ind_var1
ind_var2 ind_var5 ind_var6
ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var18 ind_var19
ind_var20 ind_var24 ind_var25
ind_var26 ind_var27 ind_var28 ind_var29 ind_var30 ind_var31 ind_var32
ind_var33 ind_var34 ind_var37
ind_var39 ind_var40 ind_var41 ind_var44 ind_var46 ind_var10_ult1
ind_var10cte_ult1 ind_var12_0
ind_var13_0 ind_var13_corto ind_var13_corto_0 ind_var13_largo
ind_var13_largo_0 ind_var13_medio
ind_var13_medio_0 ind_var14_0 ind_var17_0 ind_var18_0 ind_var1_0
ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0 ind_var26_cte ind_var27_0 ind_var28_0
ind_var29_0 ind_var2_0 ind_var30_0
ind_var31_0 ind_var32_0 ind_var32_cte ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0
ind_var40_0 ind_var41_0 ind_var43_emit_ult1 ind_var43_recib_ult1
ind_var44_0 ind_var46_0 ind_var5_0
ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1 ind_var9_ult1
num_aport_var13_hace3 num_aport_var13_ult1 num_aport_var17_hace3
num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3

num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_hace3
num_reemb_var13_ult1 num_reemb_var17_hace3
num_reemb_var17_ult1 num_reemb_var33_hace3 num_reemb_var33_ult1
num_sal_var16_ult1
num_trasp_var11_ult1 num_trasp_var17_in_hace3 num_trasp_var17_in_ult1
num_trasp_var17_out_hace3
num_trasp_var17_out_ult1 num_trasp_var33_in_hace3 num_trasp_var33_in_ult1
num_trasp_var33_out_hace3
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var18 num_var20 num_var24 num_var25 num_var26 num_var27
num_var28 num_var29 num_var30
num_var31 num_var32 num_var33 num_var34 num_var35 num_var37 num_var39
num_var40 num_var41 num_var42
num_var44 num_var46 num_var12_0 num_var13_0 num_var13_corto
num_var13_corto_0 num_var13_largo
num_var13_largo_0 num_var13_medio num_var13_medio_0 num_var14_0
num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2 num_var22_hace3 num_var22_ult1 num_var22_ult3
num_var24_0 num_var25_0
num_var26_0 num_var27_0 num_var28_0 num_var29_0 num_var2_0_ult1
num_var2_ult1 num_var30_0 num_var31_0
num_var32_0 num_var33_0 num_var34_0 num_var37_0 num_var37_med_ult2
num_var39_0 num_var40_0 num_var41_0
num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1 num_var44_0
num_var45_hace2 num_var45_hace3
num_var45_ult1 num_var45_ult3 num_var46_0 num_var5_0 num_var6_0
num_var7_emit_ult1 num_var7_recib_ult1
num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2 saldo_medio_var12_hace3
saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_hace3
saldo_medio_var13_medio_ult1 saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2
saldo_medio_var17_hace3 saldo_medio_var17_ult1 saldo_medio_var17_ult3
saldo_medio_var29_hace2
saldo_medio_var29_hace3 saldo_medio_var29_ult1 saldo_medio_var29_ult3
saldo_medio_var33_hace2
saldo_medio_var33_hace3 saldo_medio_var33_ult1 saldo_medio_var33_ult3
saldo_medio_var44_hace2
saldo_medio_var44_hace3 saldo_medio_var44_ult1 saldo_medio_var44_ult3
saldo_medio_var5_hace2
saldo_medio_var5_hace3 saldo_medio_var5_ult1 saldo_medio_var5_ult3
saldo_medio_var8_hace2
saldo_medio_var8_hace3 saldo_medio_var8_ult1 saldo_medio_var8_ult3
saldo_var1 saldo_var5 saldo_var6
saldo_var8 saldo_var12 saldo_var13 saldo_var14 saldo_var17 saldo_var18
saldo_var20 saldo_var24

```

saldo_var25 saldo_var26 saldo_var27 saldo_var28 saldo_var29 saldo_var30
saldo_var31 saldo_var32
saldo_var33 saldo_var34 saldo_var37 saldo_var40 saldo_var41 saldo_var42
saldo_var44 saldo_var46
saldo_var13_corto saldo_var13_largo saldo_var13_medio saldo_var2_ult1
var21 var36 var38/
selection=stepwise;
Run;

```

Output:

The LOGISTIC Procedure

Step 18. Effect num_var42_0 is removed:

Model Information	
Data Set	WORK.UNO
Response Variable	TARGET
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	76020
Number of Observations Used	76020

Response Profile		
Ordered Value	TARGET	Total Frequency
1	0	73012
2	1	3008

Probability modeled is TARGET='0'.

Note: No effects for the model in Step 18 are removed.

Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			

Model Fit Statistics			
Criterion	Intercept Only	Intercept and Covariates	
AIC	25327.378	22128.038	
SC	25336.617	22285.097	
-2 Log L	25325.378	22094.038	

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3231.3398	16	<.0001
Score	3437.6077	16	<.0001
Wald	2850.9744	16	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
361.1244	217	<.0001

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
	Entered	Removed					
1	ind_var30		1	1	1706.1314		<.0001
2	VAR3		1	2	903.8789		<.0001
3	ind_var8_0		1	3	183.9923		<.0001
4	ind_var13		1	4	116.8336		<.0001
5	num_var22_ult1		1	5	105.9687		<.0001
6	ind_var24		1	6	88.9568		<.0001
7	ind_var30_0		1	7	55.3725		<.0001
8	num_op_var39_efect_u		1	8	50.6034		<.0001
9	num_meses_var5_ult3		1	9	43.1993		<.0001
10	var38		1	10	44.0842		<.0001
11	num_var22_ult3		1	11	36.6311		<.0001
12	num_meses_var8_ult3		1	12	20.3452		<.0001
13	saldo_var42		1	13	16.3769		<.0001
14	saldo_var24		1	14	28.2210		<.0001
15	num_reemb_var17_ult1		1	15	25.6016		<.0001
16	ind_var31_0		1	16	14.3583		0.0002
17	num_var42_0		1	17	13.7752		0.0002
18	num_var42_0		1	16		2.8548	0.0911

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.7	Somers' D	0.589
Percent Discordant	19.8	Gamma	0.598
Percent Tied	1.5	Tau-a	0.045
Pairs	219620096	c	0.795

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	6.1479	0.5137	143.2346	<.0001
VAR3	1	-0.0384	0.00127	908.6606	<.0001
ind_var13	1	2.3515	0.2163	118.2074	<.0001
ind_var24	1	1.7057	0.3191	28.5647	<.0001
ind_var30	1	0.6431	0.1091	34.7561	<.0001
ind_var30_0	1	-2.8985	0.5090	32.4233	<.0001
ind_var31_0	1	2.3204	0.6837	11.5183	0.0007
ind_var8_0	1	-0.7162	0.1269	31.8726	<.0001
num_meses_var5_ult3	1	0.3443	0.0390	77.7722	<.0001
num_meses_var8_ult3	1	0.3457	0.0841	16.9048	<.0001
num_op_var39_efect_u	1	-0.0196	0.00299	43.0189	<.0001
num_reemb_var17_ult1	1	-0.7087	0.1230	33.1747	<.0001
num_var22_ult1	1	-0.0267	0.00943	7.9952	0.0047
num_var22_ult3	1	-0.0285	0.00381	55.9869	<.0001
saldo_var24	1	-0.00006	0.000010	29.7604	<.0001
saldo_var42	1	0.000061	0.000010	35.2379	<.0001
var38	1	3.807E-6	3.406E-7	124.9722	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
VAR3	0.962	0.960	0.965
ind_var13	10.501	6.873	16.045
ind_var24	5.505	2.945	10.290
ind_var30	1.902	1.536	2.356
ind_var30_0	0.055	0.020	0.149
ind_var31_0	10.180	2.665	38.878
ind_var8_0	0.489	0.381	0.627
num_meses_var5_ult3	1.411	1.307	1.523
num_meses_var8_ult3	1.413	1.198	1.666
num_op_var39_efect_u	0.981	0.975	0.986
num_reemb_var17_ult1	0.492	0.387	0.627
num_var22_ult1	0.974	0.956	0.992
num_var22_ult3	0.972	0.965	0.979
saldo_var24	1.000	1.000	1.000
saldo_var42	1.000	1.000	1.000
var38	1.000	1.000	1.000

Above we can see the list of variable selected by the model a total of 18 variables. As we can see the matrix of Analysis of Maximum Likelihood Estimates they are very significative with p value < 0.001.

In order to facilitate the understand of the codes I will not put all the variables here, please see the source code to see original code and reproduce the project.

Let's run the macro randomselect to select variables using different seed and selection criteria to use to compare with cross-validation.

```
%randomselect(data=uno,
listclass=,
vardepen=TARGET,
modelo=VAR2 VAR3 (All variables - I din not put here because they are so many and will
use a lot of space, but in the sas original code you can see them),
criterio=SBC,
sinicio=1457,
sfinal=1487,
fracciontrain=0.8,
directorio=Z:\git\Bitbucket\santander-kaggle\logs);
```

The selected model, based on SBC, is the model at Step 21.

Effects:	Intercept VAR2 VAR3 imp_op_var40_efect_u ind_var24 ind_var30 ind_var12_0 ind_var13_0 ind_var30_0 ind_var31_0 ind_var43_recib_ult1 ind_var8_0 num_med_var22_ult3 num_meses_var5_ult3 num_op_var39_efect_u num_reemb_var17_ult1 num_var8 num_var14_0 num_var22_ult1 saldo_var5 saldo_var30 var38
Analysis of Variance	
Source	DF
Model	21
Error	60794
Corrected Total	60815
Sum of Squares	
Model	109.59810
Error	2235.26574
Mean Square	
Model	5.21896
Error	0.03677
F Value	
Model	141.94

Root MSE	0.19175
Dependent Mean	0.04017
R-Square	0.0467
Adj R-Sq	0.0464
AIC	-140043
AICC	-140043
BIC	-200859
C(p)	59.50692
PRESS	2239.86679
SBC	-200663
ASE	0.03675

This model suggest us to use only this variables to predict.

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-0.082064	0.012149	-6.75
VAR2	1	6.2551209E-8	1.8770753E-8	3.33
VAR3	1	0.002015	0.000068186	29.55
imp_op_var40_efect_u	1	0.000088459	0.000020337	4.35
ind_var24	1	-0.027140	0.007474	-3.63
ind_var30	1	-0.045008	0.004371	-10.30
ind_var12_0	1	-0.020175	0.006005	-3.36
ind_var13_0	1	-0.049489	0.004575	-10.82
ind_var30_0	1	0.109673	0.011852	9.25
ind_var31_0	1	-0.048153	0.012133	-3.97
ind_var43_recib_ult1	1	-0.010995	0.002574	-4.27
ind_var8_0	1	0.097859	0.012167	8.04
num_med_var22_ult3	1	0.001996	0.000562	3.55
num_meses_var5_ult3	1	-0.007644	0.001463	-5.22
num_op_var39_efect_u	1	0.000958	0.000164	5.86
num_reemb_var17_ult1	1	0.030264	0.007293	4.15
num_var8	1	-0.029138	0.004448	-6.55
num_var14_0	1	0.007135	0.001513	4.71
num_var22_ult1	1	0.002351	0.000451	5.21
saldo_var5	1	-0.000000341	8.3584164E-8	-4.08
saldo_var30	1	-6.172577E-8	1.6740174E-8	-3.69
var38	1	-2.4359E-8	4.1773994E-9	-5.83

Chapter Number Three

Data Preparation

Preparing the data to make better predictions

In order to improve the dataset we will perform some data transformation to clean the dataset and improve the results.

Variables:

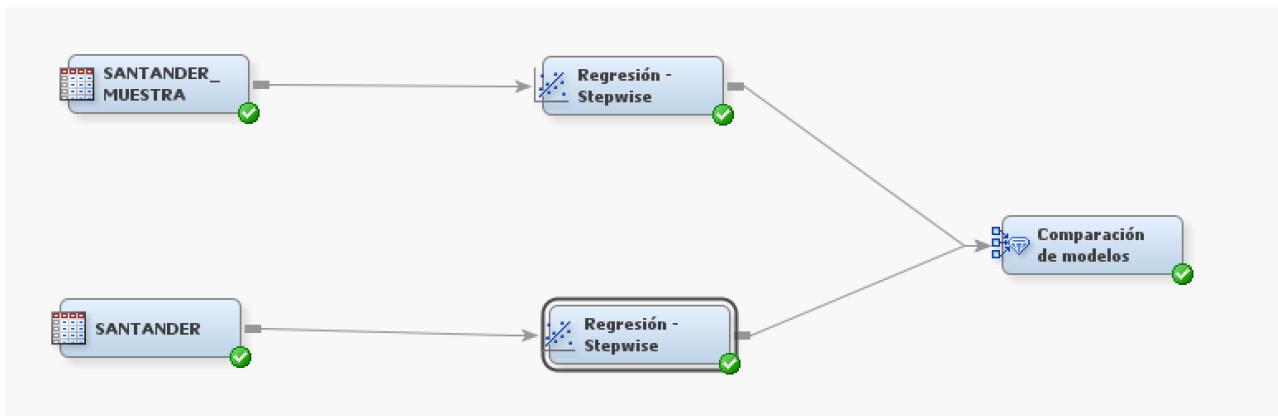
34 variables with one single value; (Action: Delete all of them)

100 variables with two unique values; (binary variables)

157 variables with values between 3 y 101 unique values; (categorical variables)

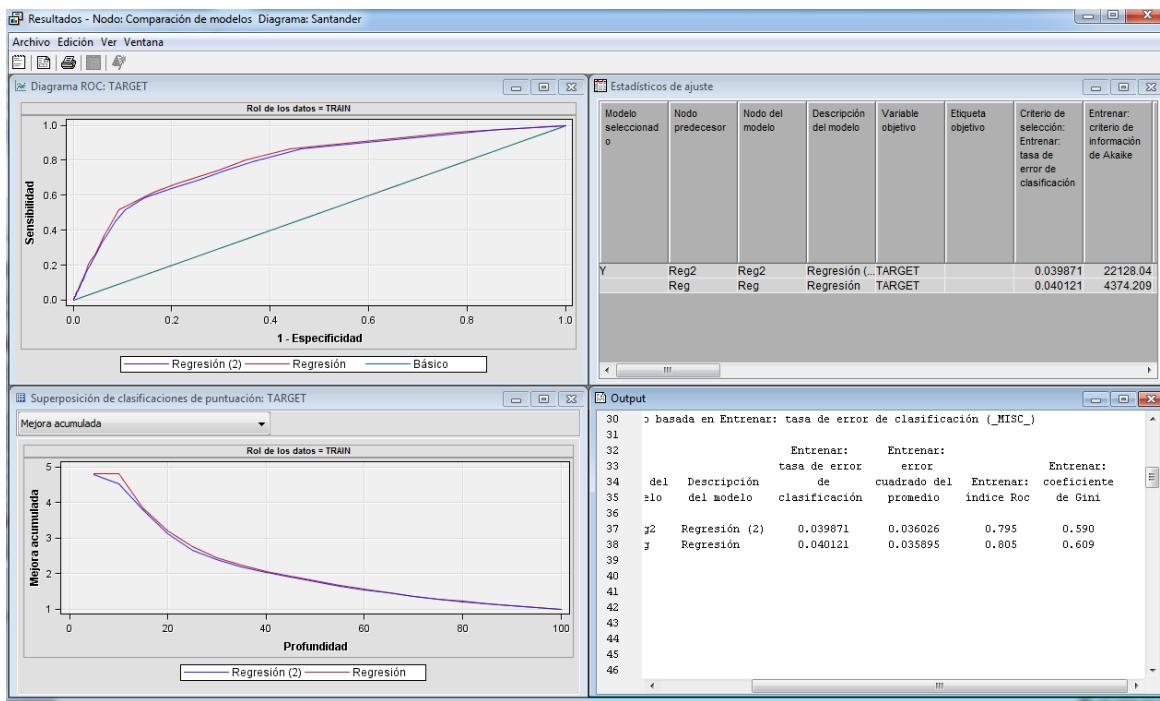
80 variables has more than 101 distinct values; (continuous variables)

Using SAS Enterprise Miner I created two datasets, one with all data and another with 20% of the train dataset that was called SANTANDER_MUESTRA.



I used Logistic Regression to predict 0 and 1 for customer satisfaction. SANTANDER_MUESTRA has 20% of SANTANDER dataset. The selection method is Stepwise for both of them.

In the image below we can see the results of both models.



Model 1 (Reg) has ROC equal of 0.805 using 20% of all rows (sample data) are better than model 2 (Reg2) with all data (all rows) with ROC equal 0.795.

So, it means that using Logistic Regression and 20% of all rows we will have better results than using the whole dataset.

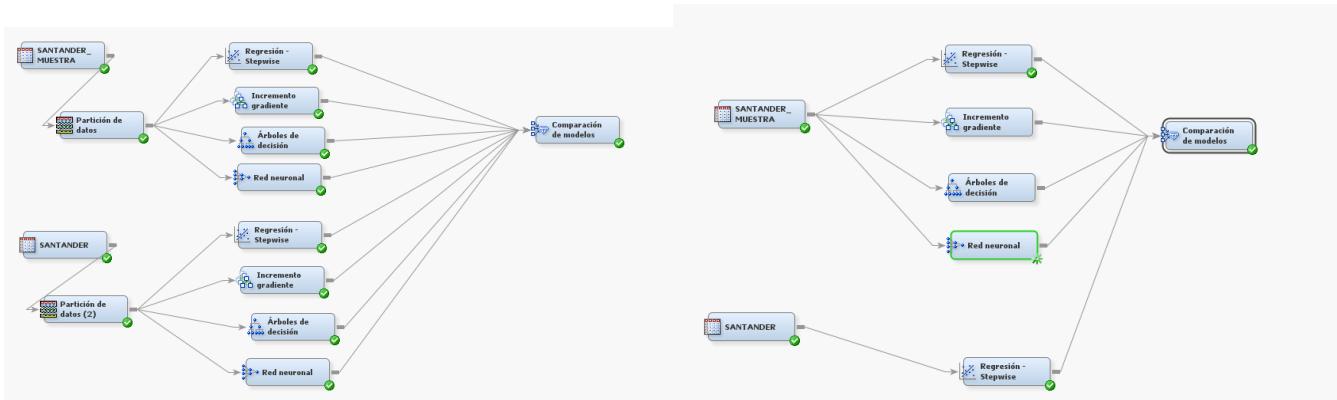
So, we can test all models using less data and then test the final model with more data using a better machine.

The problem is that this dataset is very big and we will need ways to improve performance to predict faster and I am willing make sacrifices in order to improve time to process the models.

Actual hardware:



Testing with Logistic Regression, Gradient Boosting, Decision Tree and Neuronal Network.

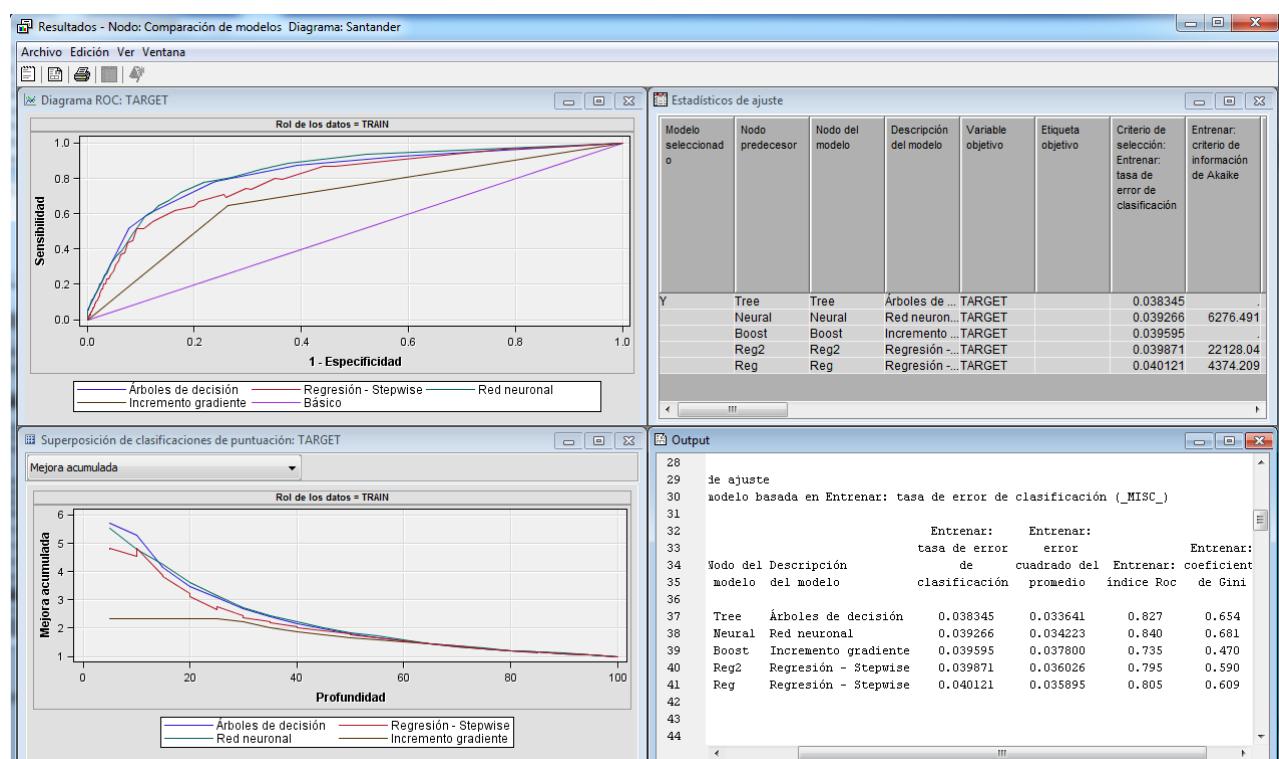


The next test is to use sample data (20% of all train dataset) and test with different algorithms.

The best algorithm in this test is Neuronal Network with ROC equals 0.840.

The next test is the same algorithms and default parameters with Train and Test with all data and 20% data sample.

The best model is Neuronal Network ROC equal 0.856 with small dataset (20% of train dataset).

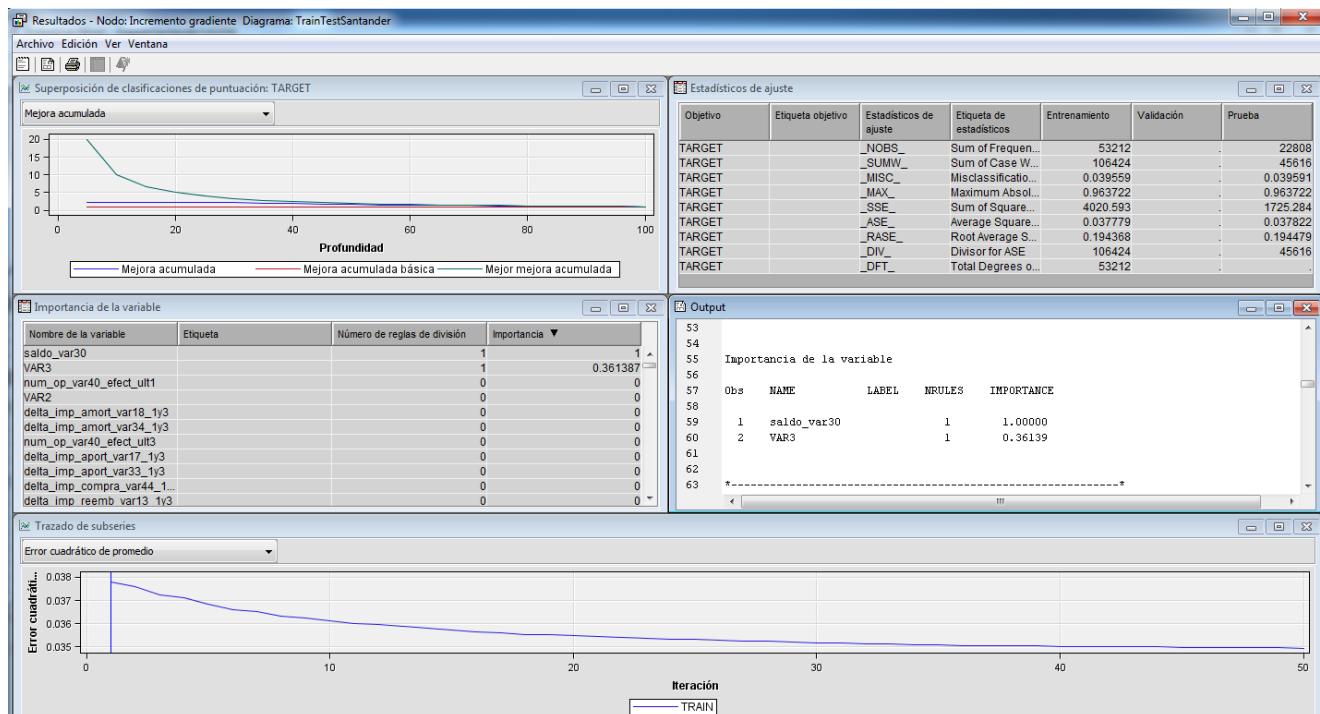


Feature Selection using SAS Enterprise Miner

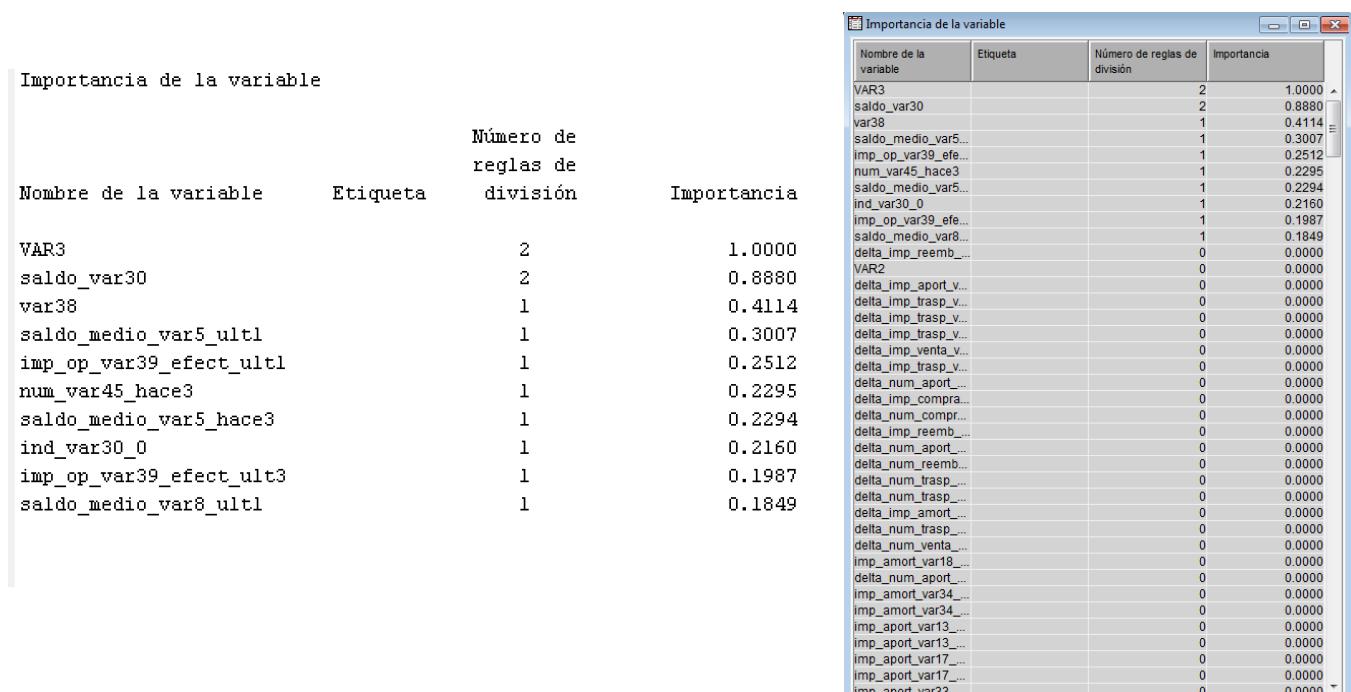
The variables selected in step 16 using logistic regression was:

var3, ind_var13, ind_var24, ind_var30, ind_var30_0, ind_var8_0, num_meses_var4_ult3, num_op_var39_efect_ult3, num_var22_ult3, var38.

Using Grading Boosting it was suggested only saldo_var30 and var3.



The variables selected by Decision Tree.

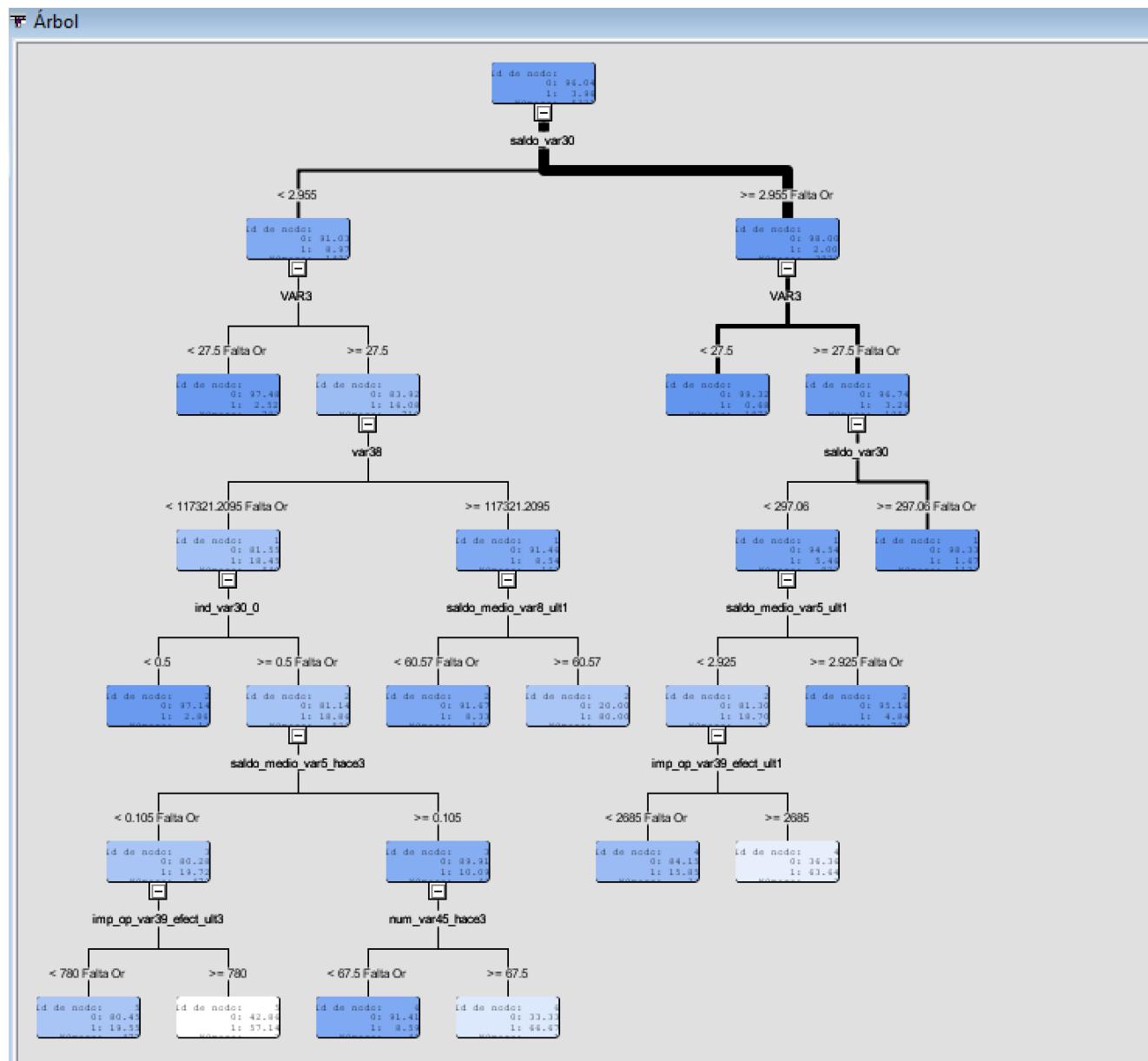


One of the advantage of using Decision Tree to predict is the power to explain the model.

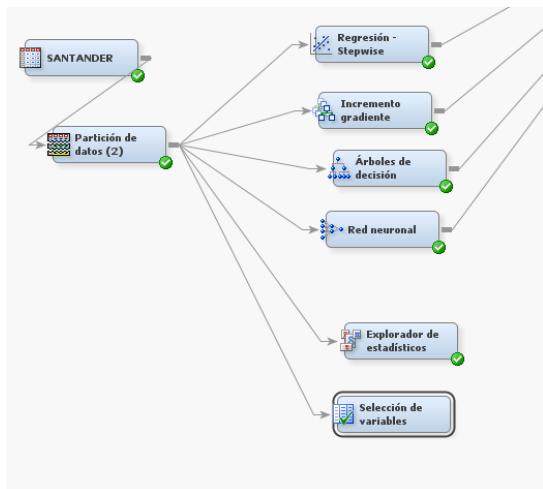
In the image below we can see in a very easy way the variables that explain the model.

As we can see in the Tree saldo_var30, var3, var38 and the others variables explain the dataset. Algorithms like Neural Network have more predictive power but they are hard to explain, they are algorithms called black box.

In the case of Decision tree they are called white box, because they are very easy to understand.



Using the node “Explorador Estadistico” and “Selección de variables” in SAS Enterprise Miner.

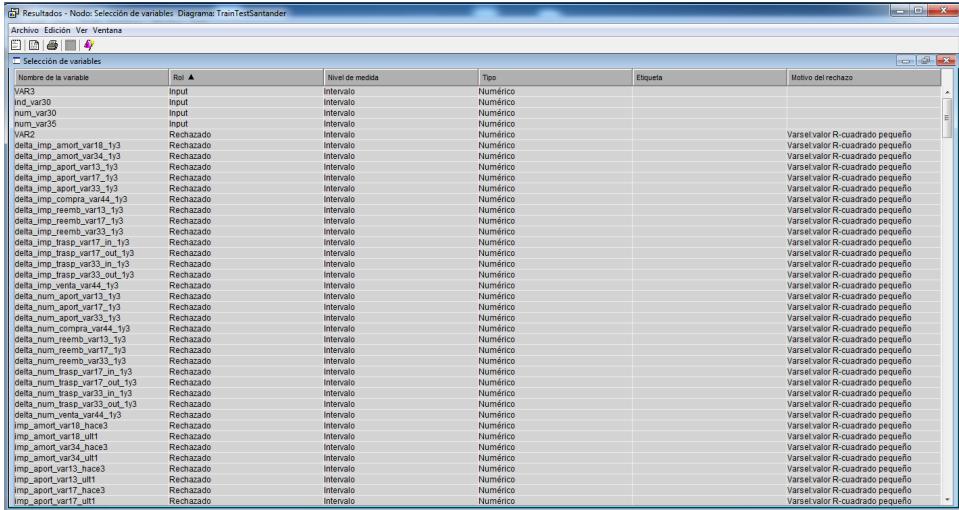


As said before, no missing values and unbalanced dataset.

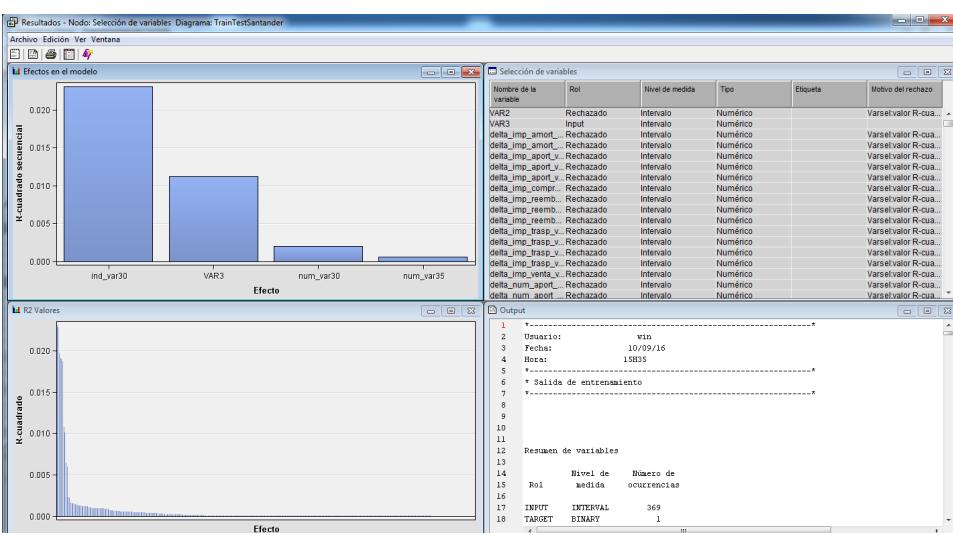
Rol de los datos=TRAIN

Rol de los datos	Nombre de la variable	Rol	Nivel	Número de ocurrencias	Porcentaje
TRAIN	TARGET	TARGET	0	51107	96.0441
TRAIN	TARGET	TARGET	1	2105	3.9559

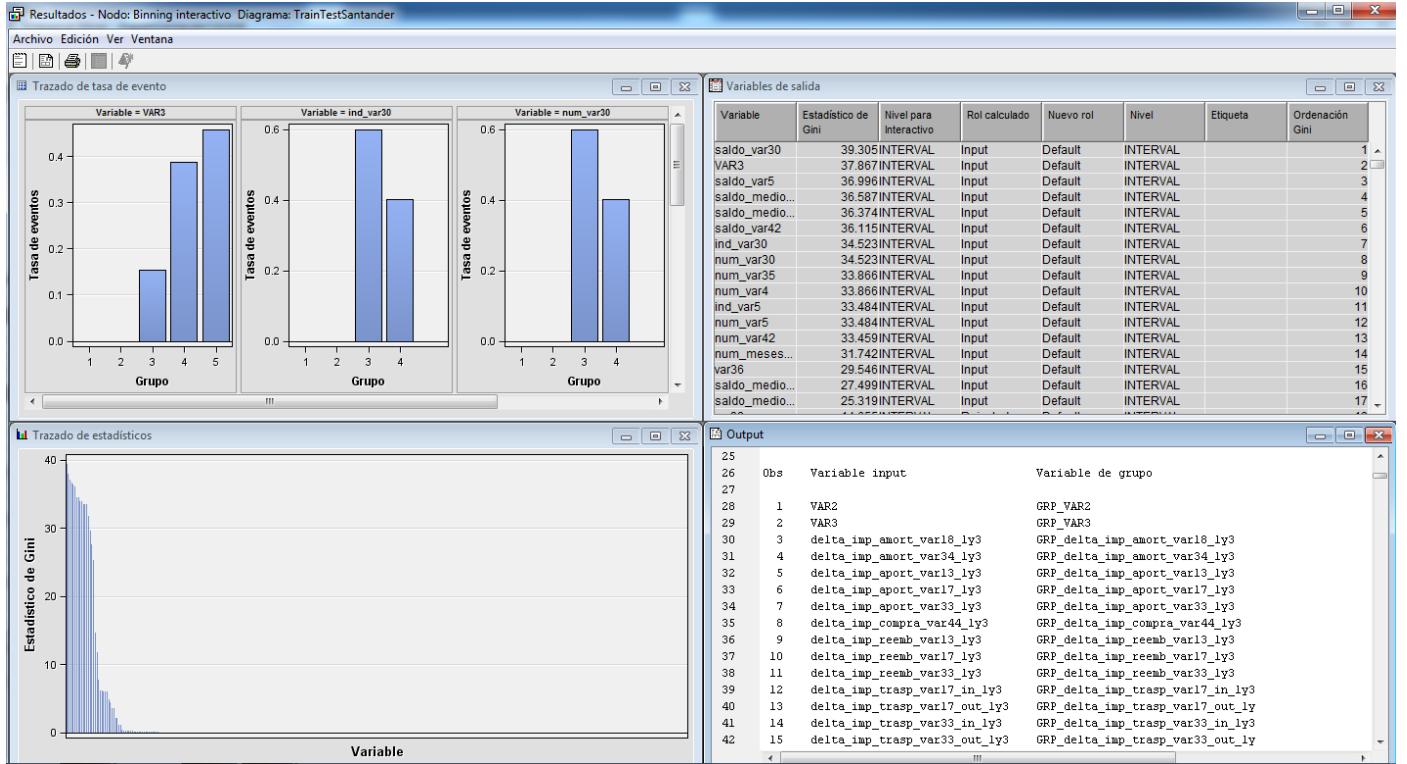
According with “Explorador Estadistico” node only VAR3, ind_var30, num_var30 and num_var35 should be used, all other variable should be excluded from the dataset.



Rol de los datos=TRAIN										
Variable	Rol	Media	Desviación estándara	No ausente	Ausente	Mínimo	Mediana	Máximo	Asimetría	Curtosis
VAR2	INPUT	-1669.85	40863.12	53212	0	-999999	2	238	-24.391	592.9451
VAR3	INPUT	33.23538	12.97851	53212	0	5	105	1.583701	2.559502	
delta_imp_aport_var13_ly3	INPUT	45666391	6.7423E8	53212	0	-1	0	1E10	14.69681	214.0042
delta_imp_compra_var44_ly3	INPUT	9020522	3.0021E8	53212	0	-1	0	1E10	33.25126	1103.688
delta_imp_compra_var44_ly3	INPUT	6013681	2.4516E8	53212	0	-0.50577	0	1E10	40.74272	1658.032
delta_num_aport_var13_ly3	INPUT	45666391	6.7423E8	53212	0	-1	0	1E10	14.69681	214.0042
delta_num_aport_var33_out_1y3	INPUT	9020522	3.0021E8	53212	0	-1	0	1E10	33.25126	1103.688
delta_num_compra_var44_ly3	INPUT	6013681	2.4516E8	53212	0	-0.5	0	1E10	40.74272	1658.032
delta_num_reemb_var13_ly3	INPUT	2816.19	25308.37	53212	0	0	0	840000	12.71536	206.7862
imp_aport_var13_hace3	INPUT	568.6917	10691.35	53212	0	0	0	450000	24.60775	722.7347
imp_compra_var44_hace3	INPUT	17.52389	132.263	53212	0	0	0	210001.4	111.3497	14569.58
imp_compra_var44_out_ult1	INPUT	84.35343	679.015	53212	0	0	0	1266766	145.6154	25059.36
imp_ent_var16_ult1	INPUT	85.71001	1435.937	53212	0	0	0	129300	54.64575	4059.737
imp_op_var39_comer_ult1	INPUT	71.20088	336.483	53212	0	0	0	12880.03	9.838977	164.1155
imp_op_var39_comer_ult3	INPUT	117.0703	540.7214	53212	0	0	0	21024.81	10.0924	170.2294
imp_op_var39_efect_ult1	INPUT	69.22688	579.760	53212	0	0	0	45990	38.56381	2348.133
imp_op_var39_efect_ult3	INPUT	115.7014	1047.935	53212	0	0	0	131100	63.28617	6388.096
imp_op_var39_ult1	INPUT	140.5715	746.3723	53212	0	0	0	47598.09	23.02025	1023.762
imp_op_var40_comer_ult1	INPUT	3.426281	88.37093	53212	0	0	0	8237.82	43.63677	2634.532
imp_op_var40_comer_ult3	INPUT	6.080654	148.695	53212	0	0	0	11073.57	40.6295	2019.349
imp_op_var40_efect_ult1	INPUT	0.39097	21.35489	53212	0	0	0	1800	67.96595	9107.238
imp_op_var40_efect_ult3	INPUT	0.577144	31.08838	53212	0	0	0	3810	76.72519	7160.516
imp_op_var40_ult1	INPUT	3.023488	90.61033	53212	0	0	0	8237.82	52.94085	3572.116
imp_op_var41_comer_ult1	INPUT	67.7746	317.2289	53212	0	0	0	12888.03	9.620393	163.2509
imp_op_var41_comer_ult3	INPUT	110.9897	506.8254	53212	0	0	0	16566.81	9.712469	154.565
imp_op_var41_efect_ult1	INPUT	68.83591	577.4515	53212	0	0	0	45990	38.87242	2382.336
imp_op_var41_efect_ult3	INPUT	115.1243	1044.856	53212	0	0	0	131100	63.76748	6462.753
imp_op_var41_ult1	INPUT	137.548	732.3614	53212	0	0	0	47598.09	23.56235	1080.47
imp_sal_var16_ult1	INPUT	4.84601	309.214	53212	0	0	0	58048.05	144.0404	24939.85
imp_trans_var37_ult1	INPUT	1818.019	23615.78	53212	0	0	0	2183564	35.40878	2151.061
imp_var43_exit_ult1	INPUT	843.6437	13694.17	53212	0	0	0	901169.1	33.74893	1557.671
imp_var7_recib_ult1	INPUT	139.0958	7090.197	53212	0	0	0	1039260	90.10152	10625.42
imp_venta_var44_ult1	INPUT	56.82892	6204.478	53212	0	0	0	1296894	181.1149	36468.22

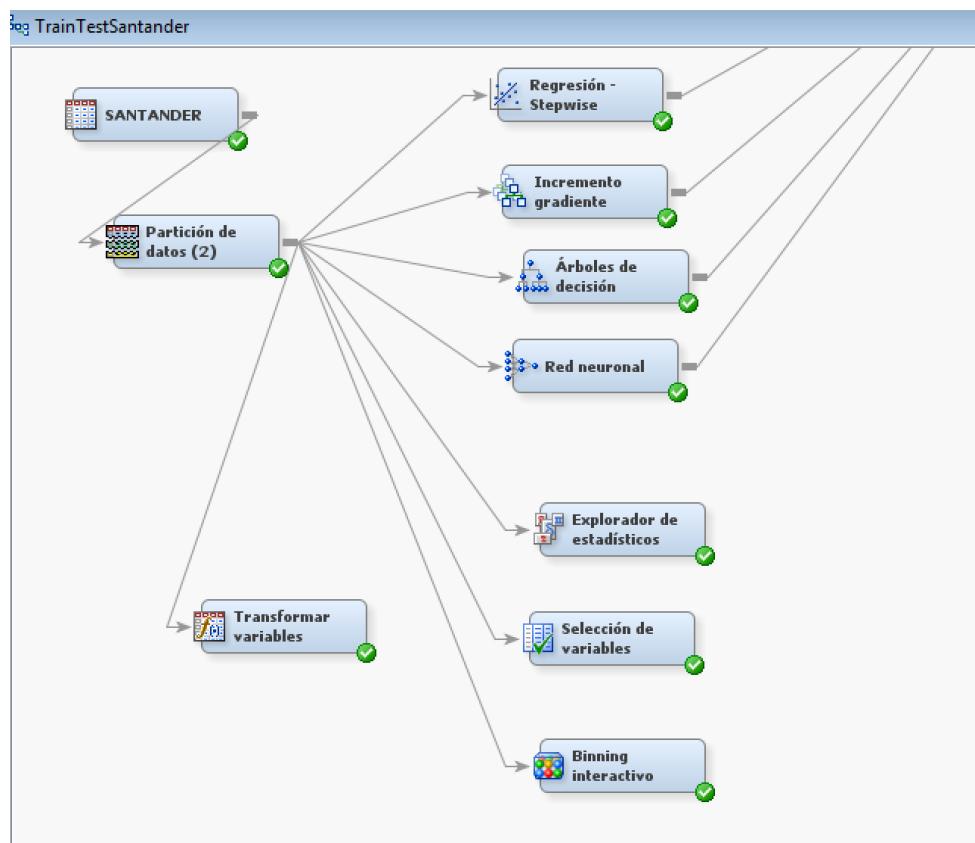


The “Binning Interactivo” node was used to understand the Gini Statistic.

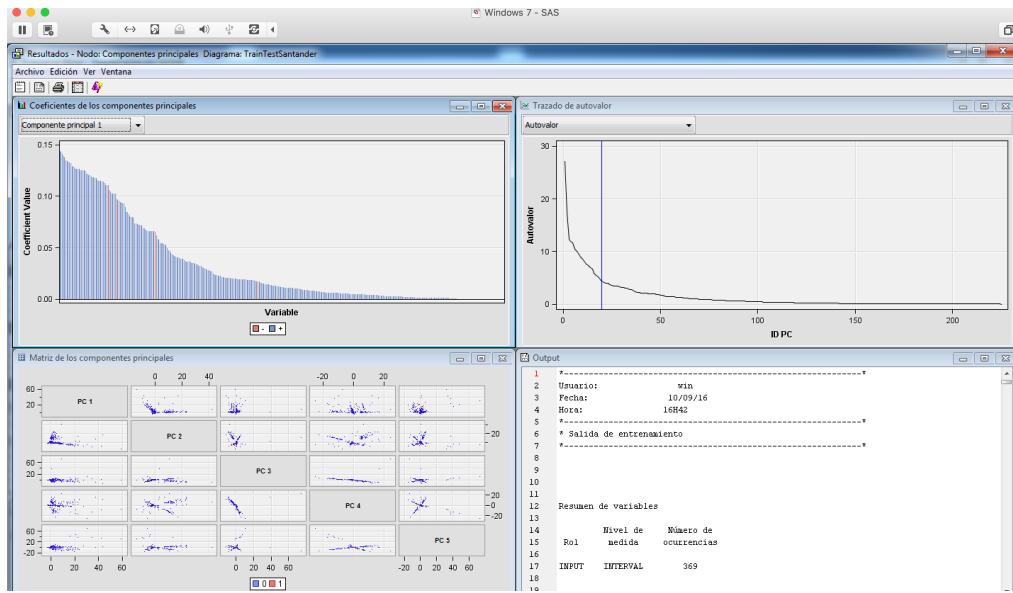


Variable	Estatístico de Gini	Nivel para Interactivo	Rol calculado
saldo_var30	39.305INTERVAL		Input
VAR3	37.867INTERVAL		Input
saldo_var5	36.996INTERVAL		Input
saldo_medio_var5_ult1	36.587INTERVAL		Input
saldo_medio_var5_hace2	36.374INTERVAL		Input
saldo_var42	36.115INTERVAL		Input
ind_var30	34.523INTERVAL		Input
num_var30	34.523INTERVAL		Input
num_var35	33.866INTERVAL		Input
num_var4	33.866INTERVAL		Input
ind_var5	33.484INTERVAL		Input
num_var5	33.484INTERVAL		Input
num_var42	33.459INTERVAL		Input
num_meses_var5_ult3	31.742INTERVAL		Input
var36	29.546INTERVAL		Input
saldo_medio_var5_hace3	27.499INTERVAL		Input
saldo_medio_var5_ult3	25.319INTERVAL		Input
var38	14.655INTERVAL		Rejected
num_meses_var39_vig_ult3	11.773INTERVAL		Rejected
num_var45_hace2	7.723INTERVAL		Rejected
ind_var39_0	6.112INTERVAL		Rejected
num_var39_0	6.112INTERVAL		Rejected

All nodes used to select understand the variables.



PCA



List of removed features and reasons

According with some algorithms more variables should be removed, but because I want to keep in this moment as much variable as I can, so I will remove only the list below and the reason is explained.

Reason to remove: removing constant features.

- ind_var2_o
- ind_var2
- ind_var27_o
- ind_var28_o
- ind_var28
- ind_var27
- ind_var41
- ind_var46_o
- ind_var46
- num_var27_o
- num_var28_o
- num_var28
- num_var27
- num_var41
- num_var46_o
- num_var46
- saldo_var28
- saldo_var27
- saldo_var41
- saldo_var46
- imp_amort_var18_hace3
- imp_amort_var34_hace3

- imp_reemb_var13_hace3
- imp_reemb_var33_hace3
- imp_trasp_var17_out_hace3
- imp_trasp_var33_out_hace3
- num_var2_o_ult1
- num_var2_ult1
- num_reemb_var13_hace3
- num_reemb_var33_hace3
- num_trasp_var17_out_hace3
- num_trasp_var33_out_hace3
- saldo_var2_ult1
- saldo_medio_var13_medio_hace3

Reason to remove: removing identical features.

- ind_var29_o
- ind_var29
- ind_var13_medio
- ind_var18
- ind_var26
- ind_var25
- ind_var32
- ind_var34
- ind_var37
- ind_var39
- num_var29_o
- num_var29
- num_var13_medio
- num_var18
- num_var26
- num_var25
- num_var32
- num_var34
- num_var37
- num_var39
- saldo_var29
- saldo_medio_var13_medio_ult1
- delta_num_reemb_var13_1y3
- delta_num_reemb_var17_1y3
- delta_num_reemb_var33_1y3
- delta_num_trasp_var17_in_1y3
- delta_num_trasp_var17_out_1y3
- delta_num_trasp_var33_in_1y3
- delta_num_trasp_var33_out_1y3

The CONTENTS Procedure

Data Set Name	UCM.SANTANDER	Observations	76020
Member Type	DATA	Variables	308
Engine	V9	Indexes	0
Created	10/09/2016 17:14:08	Observation Length	2464
Last Modified	10/09/2016 17:14:08	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

SAS code to drop this variables

```
data ucm.santander;
set ucm.santander (drop = ind_var29_0
ind_var29
ind_var13_medio
ind_var18
ind_var26
ind_var25
ind_var32
ind_var34
ind_var37
ind_var39
num_var29_0
num_var29
num_var13_medio
num_var18
num_var26
num_var25
num_var32
num_var34
num_var37
num_var39
saldo_var29
saldo_medio_var13_medio_ult1
delta_num_reemb_var13_1y3
delta_num_reemb_var17_1y3
delta_num_reemb_var33_1y3
delta_num_trasp_var17_in_1y3
delta_num_trasp_var17_out_1y3
delta_num_trasp_var33_in_1y3
delta_num_trasp_var33_out_1y3
ind_var2_0
ind_var2
ind_var27_0
ind_var28_0
ind_var28
ind_var27
ind_var41
ind_var46_0
ind_var46
num_var27_0
num_var28_0
num_var28
num_var27
num_var41
num_var46_0
num_var46
saldo_var28
saldo_var27
saldo_var41
saldo_var46
imp_amort_var18_hace3
imp_amort_var34_hace3
imp_reemb_var13_hace3
imp_reemb_var33_hace3
imp_trasp_var17_out_hace3
imp_trasp_var33_out_hace3
num_var2_0_ult1
num_var2_ult1
num_reemb_var13_hace3
num_reemb_var33_hace3
```

```
num_trasp_var17_out_hace3  
num_trasp_var33_out_hace3  
saldo_var2_ult1  
saldo_medio_var13_medio_hace3);  
run;
```

Chapter Number Four

Logistic Regression

Predicting using Logistic Regression

Creating train and testing.

Training (70%) = $76020 * 0.7 = 53.214$ rows

70% of the dataset is equal 53.214 rows

Test (30%).

SAS code:

```
data ucm.santander; set ucm.santander; Run;
data uno;set ucm.santander; u=(ranuni(12355));
proc sort data=uno; by u;
data train test;
set uno; if _n_<=53214 then output train;
else output test;
run;
```

Executing Proc Logistic with new dataset.

SAS code:

```
Proc logistic data=uno;
class ;
model TARGET=VAR2 VAR3 delta_imp_amort_var18_1y3
delta_imp_amort_var34_1y3 delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3
delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_venta_var44_1y3 imp_amort_var18_ult1 imp_amort_var34_ult1
imp_aport_var13_hace3
imp_aport_var13_ult1 imp_aport_var17_hace3 imp_aport_var17_ult1
imp_aport_var33_hace3
imp_aport_var33_ult1 imp_compra_var44_hace3 imp_compra_var44_ult1
imp_ent_var16_ult1
imp_op_var39_comer_ult1 imp_op_var39_comer_ult3 imp_op_var39_efect_ult1
imp_op_var39_efect_ult3
imp_op_var39_ult1 imp_op_var40_comer_ult1 imp_op_var40_comer_ult3
imp_op_var40_efect_ult1
```

imp_op_var40_efect_ult3 imp_op_var40_ult1 imp_op_var41_comer_ult1
imp_op_var41_comer_ult3
imp_op_var41_efect_ult1 imp_op_var41_efect_ult3 imp_op_var41_ult1
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_ult1
imp_sal_var16_ult1
imp_trans_var37_ult1 imp_trasp_var17_in_hace3 imp_trasp_var17_in_ult1
imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1 imp_trasp_var33_out_ult1
imp_var43_emit_ult1
imp_var7_emit_ult1 imp_var7_recib_ult1 imp_venta_var44_hace3
imp_venta_var44_ult1 ind_var1 ind_var5
ind_var6 ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var19
ind_var20 ind_var24 ind_var30
ind_var31 ind_var33 ind_var40 ind_var44 ind_var10_ult1 ind_var10cte_ult1
ind_var12_0 ind_var13_0
ind_var13_corto ind_var13_corto_0 ind_var13_largo ind_var13_largo_0
ind_var13_medio_0 ind_var14_0
ind_var17_0 ind_var18_0 ind_var1_0 ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0
ind_var26_cte ind_var30_0 ind_var31_0 ind_var32_0 ind_var32_cte
ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0 ind_var40_0 ind_var41_0 ind_var43_emit_ult1
ind_var43_recib_ult1 ind_var44_0
ind_var5_0 ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1
ind_var9_ult1 num_aport_var13_hace3 num_aport_var13_ult1
num_aport_var17_hace3 num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3
num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_ult1
num_reemb_var17_hace3 num_reemb_var17_ult1
num_reemb_var33_ult1 num_sal_var16_ult1 num_trasp_var11_ult1
num_trasp_var17_in_hace3
num_trasp_var17_in_ult1 num_trasp_var17_out_ult1 num_trasp_var33_in_hace3
num_trasp_var33_in_ult1
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var20 num_var24 num_var30 num_var31 num_var33 num_var35
num_var40 num_var42 num_var44
num_var12_0 num_var13_0 num_var13_corto num_var13_corto_0 num_var13_largo
num_var13_largo_0
num_var13_medio_0 num_var14_0 num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2
num_var22_hace3 num_var22_ult1 num_var22_ult3 num_var24_0 num_var25_0
num_var26_0 num_var30_0

```

num_var31_0 num_var32_0 num_var33_0 num_var34_0 num_var37_0
num_var37_med_ult2 num_var39_0 num_var40_0
num_var41_0 num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1
num_var44_0 num_var45_hace2
num_var45_hace3 num_var45_ult1 num_var45_ult3 num_var5_0 num_var6_0
num_var7_emit_ult1
num_var7_recib_ult1 num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2
saldo_medio_var12_hace3 saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2 saldo_medio_var17_hace3 saldo_medio_var17_ult1
saldo_medio_var17_ult3
saldo_medio_var29_hace2 saldo_medio_var29_hace3 saldo_medio_var29_ult1
saldo_medio_var29_ult3
saldo_medio_var33_hace2 saldo_medio_var33_hace3 saldo_medio_var33_ult1
saldo_medio_var33_ult3
saldo_medio_var44_hace2 saldo_medio_var44_hace3 saldo_medio_var44_ult1
saldo_medio_var44_ult3
saldo_medio_var5_hace2 saldo_medio_var5_hace3 saldo_medio_var5_ult1
saldo_medio_var5_ult3
saldo_medio_var8_hace2 saldo_medio_var8_hace3 saldo_medio_var8_ult1
saldo_medio_var8_ult3 saldo_var1
saldo_var5 saldo_var6 saldo_var8 saldo_var12 saldo_var13 saldo_var14
saldo_var17 saldo_var18
saldo_var20 saldo_var24 saldo_var25 saldo_var26 saldo_var30 saldo_var31
saldo_var32 saldo_var33
saldo_var34 saldo_var37 saldo_var40 saldo_var42 saldo_var44
saldo_var13_corto saldo_var13_largo
saldo_var13_medio var21 var36 var38 /
selection=stepwise;
Run;

```

The result with 308 variables were 78.7% (Percent Concordant) and 19.8% (Percent Discordant). The same result with 371 variables as we can see in the image below.

The LOGISTIC Procedure

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
VAR3	0.962	0.960	0.965
ind_var13	10.501	6.873	16.045
ind_var24	5.505	2.945	10.290
ind_var30	1.902	1.536	2.356
ind_var30_0	0.055	0.020	0.149
ind_var31_0	10.180	2.665	38.878
ind_var8_0	0.489	0.381	0.627
num_meses_var5_ult3	1.411	1.307	1.523
num_meses_var8_ult3	1.413	1.198	1.666
num_op_var39_efect_u	0.981	0.975	0.986
num_reemb_var17_ult1	0.492	0.387	0.627
num_var22_ult1	0.974	0.956	0.992
num_var22_ult3	0.972	0.965	0.979
saldo_var24	1.000	1.000	1.000
saldo_var42	1.000	1.000	1.000
var38	1.000	1.000	1.000

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.7	Somers' D	0.589
Percent Discordant	19.8	Gamma	0.598
Percent Tied	1.5	Tau-a	0.045
Pairs	219620096	c	0.795

Association of Predicted Probabilities and Observed Responses

Percent Concordant	78.7	Somers' D	0.589
Percent Discordant	19.8	Gamma	0.598
Percent Tied	1.5	Tau-a	0.045
Pairs	219620096	c	0.795

It is important to see, that removing 63 variables we obtained the same result that with all variable, with less variables it is possible to improve the time we spend to train the model.

Now let's execute Random Select with 308 variables:

```
%randomselect(data=uno,
listclass=,
vardepen=TARGET,
modelo=VAR2 VAR3 delta_imp_amort_var18_1y3 delta_imp_amort_var34_1y3
delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3
delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_venta_var44_1y3 imp_amort_var18_ult1 imp_amort_var34_ult1
imp_aport_var13_hace3
imp_aport_var13_ult1 imp_aport_var17_hace3 imp_aport_var17_ult1
imp_aport_var33_hace3
imp_aport_var33_ult1 imp_compra_var44_hace3 imp_compra_var44_ult1
imp_ent_var16_ult1
imp_op_var39_comer_ult1 imp_op_var39_comer_ult3 imp_op_var39_efect_ult1
imp_op_var39_efect_ult3
imp_op_var39_ult1 imp_op_var40_comer_ult1 imp_op_var40_comer_ult3
imp_op_var40_efect_ult1
imp_op_var40_efect_ult3 imp_op_var40_ult1 imp_op_var41_comer_ult1
imp_op_var41_comer_ult3
imp_op_var41_efect_ult1 imp_op_var41_efect_ult3 imp_op_var41_ult1
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_ult1
imp_sal_var16_ult1
imp_trans_var37_ult1 imp_trasp_var17_in_hace3 imp_trasp_var17_in_ult1
imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1 imp_trasp_var33_out_ult1
imp_var43_emit_ult1
imp_var7_emit_ult1 imp_var7_recib_ult1 imp_venta_var44_hace3
imp_venta_var44_ult1 ind_var1 ind_var5
ind_var6 ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var19
ind_var20 ind_var24 ind_var30
ind_var31 ind_var33 ind_var40 ind_var44 ind_var10_ult1 ind_var10cte_ult1
ind_var12_0 ind_var13_0
ind_var13_corto ind_var13_corto_0 ind_var13_largo ind_var13_largo_0
ind_var13_medio_0 ind_var14_0
ind_var17_0 ind_var18_0 ind_var1_0 ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0
ind_var26_cte ind_var30_0 ind_var31_0 ind_var32_0 ind_var32_cte
ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0 ind_var40_0 ind_var41_0 ind_var43_emit_ult1
ind_var43_recib_ult1 ind_var44_0
ind_var5_0 ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1
ind_var9_ult1 num_aport_var13_hace3 num_aport_var13_ult1
num_aport_var17_hace3 num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
```

num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3
num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_ult1
num_reemb_var17_hace3 num_reemb_var17_ult1
num_reemb_var33_ult1 num_sal_var16_ult1 num_trasp_var11_ult1
num_trasp_var17_in_hace3
num_trasp_var17_in_ult1 num_trasp_var17_out_ult1 num_trasp_var33_in_hace3
num_trasp_var33_in_ult1
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var20 num_var24 num_var30 num_var31 num_var33 num_var35
num_var40 num_var42 num_var44
num_var12_0 num_var13_0 num_var13_corto num_var13_corto_0 num_var13_largo
num_var13_largo_0
num_var13_medio_0 num_var14_0 num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2
num_var22_hace3 num_var22_ult1 num_var22_ult3 num_var24_0 num_var25_0
num_var26_0 num_var30_0
num_var31_0 num_var32_0 num_var33_0 num_var34_0 num_var37_0
num_var37_med_ult2 num_var39_0 num_var40_0
num_var41_0 num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1
num_var44_0 num_var45_hace2
num_var45_hace3 num_var45_ult1 num_var45_ult3 num_var5_0 num_var6_0
num_var7_emit_ult1
num_var7_recib_ult1 num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2
saldo_medio_var12_hace3 saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2 saldo_medio_var17_hace3 saldo_medio_var17_ult1
saldo_medio_var17_ult3
saldo_medio_var29_hace2 saldo_medio_var29_hace3 saldo_medio_var29_ult1
saldo_medio_var29_ult3
saldo_medio_var33_hace2 saldo_medio_var33_hace3 saldo_medio_var33_ult1
saldo_medio_var33_ult3
saldo_medio_var44_hace2 saldo_medio_var44_hace3 saldo_medio_var44_ult1
saldo_medio_var44_ult3
saldo_medio_var5_hace2 saldo_medio_var5_hace3 saldo_medio_var5_ult1
saldo_medio_var5_ult3
saldo_medio_var8_hace2 saldo_medio_var8_hace3 saldo_medio_var8_ult1
saldo_medio_var8_ult3 saldo_var1
saldo_var5 saldo_var6 saldo_var8 saldo_var12 saldo_var13 saldo_var14
saldo_var17 saldo_var18
saldo_var20 saldo_var24 saldo_var25 saldo_var26 saldo_var30 saldo_var31
saldo_var32 saldo_var33

```

saldo_var34 saldo_var37 saldo_var40 saldo_var42 saldo_var44
saldo_var13_corto saldo_var13_largo
saldo_var13_medio var21 var36 var38,
criterio=SBC,
sinicio=1457,
sfinal=1487,
fracciontrain=0.8,directorio=Z:\git\Bitbucket\santander-kaggle\logs) ;

```

Results:

The SAS System				
The FREQ Procedure				
efecto	Frequency	Percent	Cumulative Frequency	Cumulative Percent
VAR2	10	1.69	10	1.69
VAR3	31	5.25	41	6.95
delta_imp_reemb_var1	13	2.20	54	9.15
imp_op_var39_efect_u	21	3.56	75	12.71
imp_op_var40_efect_u	14	2.37	89	15.08
imp_op_var41_ult1	2	0.34	91	15.42
ind_var12_0	7	1.19	98	16.61
ind_var13	22	3.73	120	20.34
ind_var13_0	4	0.68	124	21.02
ind_var14_0	1	0.17	125	21.19
ind_var24	25	4.24	150	25.42
ind_var24_0	1	0.17	151	25.59
ind_var30	31	5.25	182	30.85
ind_var30_0	31	5.25	213	36.10
ind_var31_0	10	1.69	223	37.80
ind_var32_0	2	0.34	225	38.14
ind_var41_0	2	0.34	227	38.47
ind_var43_emit_ult1	1	0.17	228	38.64
ind_var43_recib_ult1	30	5.08	258	43.73
ind_var8	11	1.86	269	45.59
ind_var8_0	30	5.08	299	50.68
num_ent_var16_ult1	2	0.34	301	51.02
num_med_var22_ult3	5	0.85	306	51.86
num_meses_var5_ult3	31	5.25	337	57.12
num_meses_var6_ult3	1	0.17	338	57.29
num_op_var39_efect_u	10	1.69	348	58.98
num_op_var40_efect_u	8	1.36	356	60.34
num_op_var41_efect_u	1	0.17	357	60.51
num_reemb_var17_ult1	23	3.90	380	64.41
num_var14_0	12	2.03	392	66.44
num_var20_0	1	0.17	393	66.61
num_var22_hace2	9	1.53	402	68.14
num_var22_ult1	31	5.25	433	73.39
num_var22_ult3	3	0.51	436	73.90
num_var30	4	0.68	440	74.58
num_var30_0	1	0.17	441	74.75
num_var39_0	1	0.17	442	74.92
num_var42_0	11	1.86	453	76.78

The SAS System			
Obs	efecto	COUNT	PERCENT
1	VAR3	31	5.25424
2	ind_var30	31	5.25424
3	ind_var30_0	31	5.25424
4	num_meses_var5_ult3	31	5.25424
5	num_var22_ult1	31	5.25424
6	var38	31	5.25424
7	ind_var43_recib_ult1	30	5.08475
8	ind_var8_0	30	5.08475
9	saldo_var5	28	4.74576
10	ind_var24	25	4.23729
11	num_reemb_var17_ult1	23	3.89831
12	ind_var13	22	3.72881
13	imp_op_var39_efect_u	21	3.55932
14	saldo_var30	20	3.38983
15	num_var8	18	3.05085
16	imp_op_var40_efect_u	14	2.37288
17	delta_imp_reemb_var1	13	2.20339
18	num_var14_0	12	2.03390
19	saldo_var8	12	2.03390
20	ind_var8	11	1.86441
21	num_var42_0	11	1.86441
22	VAR2	10	1.63492
23	ind_var31_0	10	1.63492
24	num_op_var39_efect_u	10	1.63492
25	num_var22_hace2	9	1.52542
26	num_op_var40_efect_u	8	1.35593
27	saldo_var26	8	1.35593
28	ind_var12_0	7	1.18644
29	saldo_medio_var8_ult	7	1.18644
30	num_med_var22_ult3	5	0.84746
31	ind_var13_0	4	0.67797
32	num_var30	4	0.67797
33	num_var5	4	0.67797
34	num_var22_ult3	3	0.50847
35	num_var45_ult3	3	0.50847
36	saldo_medio_var5_ult	3	0.50847
37	imp_op_var41_ult1	2	0.33898
38	ind_var32_0	2	0.33898
39	ind_var41_0	2	0.33898
40	num_ent_var16_ult1	2	0.33898
41	ind_var14_0	1	0.16949

Output - (Untitled)		
The SAS System		
Obs	efecto	
1	Intercept	VAR2
2	Intercept	VAR2
3	Intercept	VAR2
4	Intercept	VAR2
5	Intercept	VAR2
6	Intercept	VAR2
7	Intercept	VAR2
8	Intercept	VAR2
9	Intercept	VAR2
10	Intercept	VAR3
11	Intercept	VAR3
12	Intercept	VAR3
13	Intercept	VAR3
14	Intercept	VAR3
15	Intercept	VAR3
16	Intercept	VAR3
17	Intercept	VAR3
18	Intercept	VAR3
19	Intercept	VAR3
20	Intercept	VAR3
21	Intercept	VAR3
22	Intercept	VAR3
23	Intercept	VAR3
24	Intercept	VAR3
1	1	3.22581
2	1	3.22581
3	1	3.22581
4	1	3.22581
5	1	3.22581
6	1	3.22581
7	1	3.22581
8	1	3.22581
9	1	3.22581
10	1	3.22581
11	1	3.22581
12	1	3.22581
13	1	3.22581
14	1	3.22581

Changing the seeds the results change.

```
%randomselect(data=uno,
listclass=,
vardepen=TARGET,
modelo=VAR2 VAR3 delta_imp_amort_var18_1y3 delta_imp_amort_var34_1y3
delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3
delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_venta_var44_1y3 imp_amort_var18_ult1 imp_amort_var34_ult1
imp_aport_var13_hace3
imp_aport_var13_ult1 imp_aport_var17_hace3 imp_aport_var17_ult1
imp_aport_var33_hace3
imp_aport_var33_ult1 imp_compra_var44_hace3 imp_compra_var44_ult1
imp_ent_var16_ult1
imp_op_var39_comer_ult1 imp_op_var39_comer_ult3 imp_op_var39_efect_ult1
imp_op_var39_efect_ult3
imp_op_var39_ult1 imp_op_var40_comer_ult1 imp_op_var40_comer_ult3
imp_op_var40_efect_ult1
imp_op_var40_efect_ult3 imp_op_var40_ult1 imp_op_var41_comer_ult1
imp_op_var41_comer_ult3
imp_op_var41_efect_ult1 imp_op_var41_efect_ult3 imp_op_var41_ult1
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_ult1
imp_sal_var16_ult1
imp_trans_var37_ult1 imp_trasp_var17_in_hace3 imp_trasp_var17_in_ult1
imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1 imp_trasp_var33_out_ult1
imp_var43_emit_ult1
imp_var7_emit_ult1 imp_var7_recib_ult1 imp_venta_var44_hace3
imp_venta_var44_ult1 ind_var1 ind_var5
ind_var6 ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var19
ind_var20 ind_var24 ind_var30
ind_var31 ind_var33 ind_var40 ind_var44 ind_var10_ult1 ind_var10cte_ult1
ind_var12_0 ind_var13_0
ind_var13_corto ind_var13_corto_0 ind_var13_largo ind_var13_largo_0
ind_var13_medio_0 ind_var14_0
ind_var17_0 ind_var18_0 ind_var1_0 ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0
ind_var26_cte ind_var30_0 ind_var31_0 ind_var32_0 ind_var32_cte
ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0 ind_var40_0 ind_var41_0 ind_var43_emit_ult1
ind_var43_recib_ult1 ind_var44_0
ind_var5_0 ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1
ind_var9_ult1 num_aport_var13_hace3 num_aport_var13_ult1
num_aport_var17_hace3 num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
```

```

num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3
num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_ult1
num_reemb_var17_hace3 num_reemb_var17_ult1
num_reemb_var33_ult1 num_sal_var16_ult1 num_trasp_var11_ult1
num_trasp_var17_in_hace3
num_trasp_var17_in_ult1 num_trasp_var17_out_ult1 num_trasp_var33_in_hace3
num_trasp_var33_in_ult1
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var20 num_var24 num_var30 num_var31 num_var33 num_var35
num_var40 num_var42 num_var44
num_var12_0 num_var13_0 num_var13_corto num_var13_corto_0 num_var13_largo
num_var13_largo_0
num_var13_medio_0 num_var14_0 num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2
num_var22_hace3 num_var22_ult1 num_var22_ult3 num_var24_0 num_var25_0
num_var26_0 num_var30_0
num_var31_0 num_var32_0 num_var33_0 num_var34_0 num_var37_0
num_var37_med_ult2 num_var39_0 num_var40_0
num_var41_0 num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1
num_var44_0 num_var45_hace2
num_var45_hace3 num_var45_ult1 num_var45_ult3 num_var5_0 num_var6_0
num_var7_emit_ult1
num_var7_recib_ult1 num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2
saldo_medio_var12_hace3 saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2 saldo_medio_var17_hace3 saldo_medio_var17_ult1
saldo_medio_var17_ult3
saldo_medio_var29_hace2 saldo_medio_var29_hace3 saldo_medio_var29_ult1
saldo_medio_var29_ult3
saldo_medio_var33_hace2 saldo_medio_var33_hace3 saldo_medio_var33_ult1
saldo_medio_var33_ult3
saldo_medio_var44_hace2 saldo_medio_var44_hace3 saldo_medio_var44_ult1
saldo_medio_var44_ult3
saldo_medio_var5_hace2 saldo_medio_var5_hace3 saldo_medio_var5_ult1
saldo_medio_var5_ult3
saldo_medio_var8_hace2 saldo_medio_var8_hace3 saldo_medio_var8_ult1
saldo_medio_var8_ult3 saldo_var1
saldo_var5 saldo_var6 saldo_var8 saldo_var12 saldo_var13 saldo_var14
saldo_var17 saldo_var18
saldo_var20 saldo_var24 saldo_var25 saldo_var26 saldo_var30 saldo_var31
saldo_var32 saldo_var33
saldo_var34 saldo_var37 saldo_var40 saldo_var42 saldo_var44
saldo_var13_corto saldo_var13_largo
saldo_var13_medio var21 var36 var38,
criterio=SBC,
```

```

sinicio=450,
sfinal=490,
fracciontrain=0.8,directorio=Z:\git\Bitbucket\santander-kaggle\logs);

```

Results:

The SAS System The FREQ Procedure					The SAS System			
efecto	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Obs	efecto	COUNT	PERCENT
VAR2	16	2.06	16	2.06	1	VAR3	41	5.29032
VAR3	41	5.29	57	7.35	2	ind_var30	41	5.29032
delta_imp_reemb_var1	17	2.19	74	9.55	3	ind_var30_0	41	5.29032
imp_op_var39_efect_u	27	3.48	101	13.03	4	num_var22_ult1	41	5.29032
imp_op_var39_ult1	1	0.13	102	13.16	5	var38	41	5.29032
imp_op_var40_efect_u	12	1.55	114	14.71	6	num_meses_var5_ult3	40	5.16129
imp_op_var41_efect_u	1	0.13	115	14.84	7	saldo_var5	40	5.16129
imp_op_var41_ult1	6	0.77	121	15.61	8	ind_var43_recib_ult1	39	5.03226
ind_var12_0	4	0.52	125	16.13	9	ind_var8_0	37	4.77419
ind_var13	28	3.61	153	19.74	10	ind_var24	33	4.25806
ind_var13_0	9	1.16	162	20.90	11	saldo_var30	32	4.12903
ind_var20	1	0.13	163	21.03	12	ind_var13	28	3.61290
ind_var24	33	4.26	196	25.29	13	imp_op_var39_efect_u	27	3.48387
ind_var24_0	3	0.39	199	25.68	14	num_var14_0	27	3.48387
ind_var30	41	5.29	240	30.97	15	num_reemb_var17_ult1	24	3.09677
ind_var30_0	41	5.29	281	36.26	16	num_var8	19	2.45161
ind_var31_0	13	1.68	294	37.94	17	delta_imp_reemb_var1	17	2.19355
ind_var41_0	1	0.13	295	38.06	18	ind_var8	17	2.19355
ind_var43_recib_ult1	39	5.03	334	43.10	19	VAR2	16	2.06452
ind_var8	17	2.19	351	45.29	20	saldo_medio_var8_ult	16	2.06452
ind_var8_0	37	4.77	388	50.06	21	saldo_var8	15	1.93548
num_ent_var16_ult1	1	0.13	389	50.19	22	ind_var31_0	13	1.67742
num_med_var22_ult3	8	1.03	397	51.23	23	imp_op_var40_efect_u	12	1.54839
num_meses_var5_ult3	40	5.16	437	56.39	24	num_op_var39_efect_u	11	1.41935
num_op_var39_efect_u	11	1.42	448	57.81	25	num_op_var40_efect_u	11	1.41935
num_op_var40_efect_u	11	1.42	459	59.23	26	num_var22_ult3	11	1.41935
num_op_var41_efect_u	2	0.26	461	59.48	27	ind_var13_0	9	1.16129
num_reemb_var17_ult1	24	3.10	485	62.58	28	num_var42_0	9	1.16129
num_var12	1	0.13	486	62.71	29	num_med_var22_ult3	8	1.03226
num_var14_0	27	3.48	513	66.19	30	saldo_var26	8	1.03226
num_var22_hace2	4	0.52	517	66.71	31	imp_op_var41_ult1	6	0.77419
num_var22_ult1	41	5.29	558	72.00	32	ind_var12_0	4	0.51613
num_var22_ult3	11	1.42	569	73.42	33	num_var22_hace2	4	0.51613
num_var30	3	0.39	572	73.81	34	num_var5	4	0.51613
num_var30_0	1	0.13	573	73.94	35	num_var8_0	4	0.51613
num_var32_0	1	0.13	574	74.06	36	ind_var24_0	3	0.38710
num_var39_0	3	0.39	577	74.45	37	num_var30	3	0.38710
num_var41_0	1	0.13	579	74.58	38	num_var39_0	3	0.38710
num_var42	1	0.13	579	74.71	39	num_var45_ult3	3	0.38710
num_var42_0	9	1.16	588	75.87	40	num_op_var41_efect_u	2	0.25806
num_var43_recib_ult1	2	0.26	590	76.13	41	num_var43_recib_ult1	2	0.25806
					42	imp_op_var39_ult1	1	0.12903
					43	imp_op_var41_efect_u	1	0.12903
					44	ind_var20	1	0.12903

Selection Logistic Regression with cross validation

Next step is to run the macro cruzadalogistica for all variable combination and compare all models to see the best model.

In my case I've got 24 combinations for test with first seeds and then changing seeds I got more 41 different combinations.

So, the correct would be test 65 models with different variables combinations to find the best one.

To run every cruzadalogistica it takes a lot of time because of my hardware, so I will drop variables and delete some rows and create a smaller dataset in order to continue this project.

In order to speed up I will keep only 1000 rows and create a train and test dataset.

```

data uno_small;
set ucm.santander;
if _n_>=10000 then delete;
run;

data uno_small; set uno_small; u=(ranuni(12355));
proc sort data=uno_small; by u;
data train test;
set uno_small; if _n_<=6999 then output train;
else output test;
run;

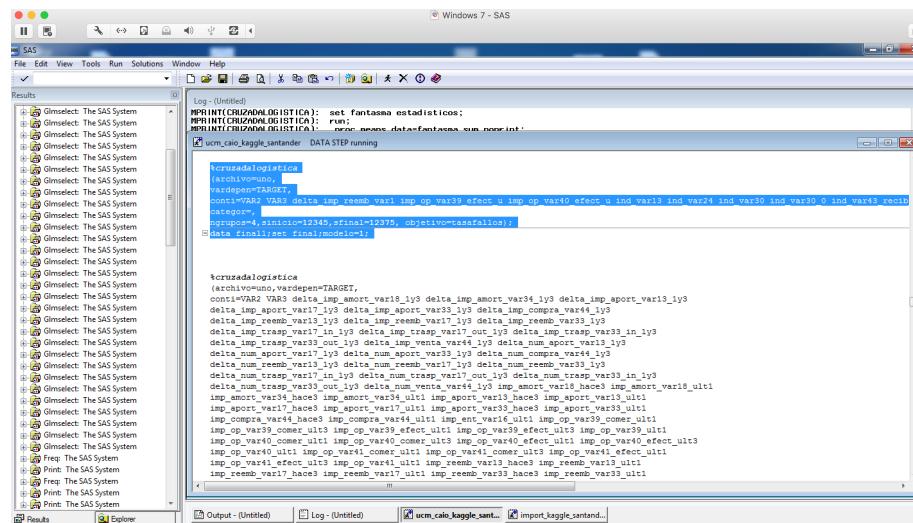
```

Feature Selection Table

Variable	Logistic Regression	Decision Tree	Gradient Boosting	Binning Interactivo	# of Present Models
VAR3	YES	YES	YES	YES	4
ind_var13	YES	NO	NO	NO	1
ind_var24	YES	NO	NO	NO	1
ind_var30	YES	NO	NO	YES	2
ind_var30_0	YES	YES	NO	NO	2
ind_var31_0	YES	NO	NO	NO	1
ind_var_8_0	YES	NO	NO	NO	1
num_meses_var4_ult3	YES	NO	NO	NO	1
num_meses_var5_ult3	YES	NO	NO	YES	2
num_meses_var8_ult3	YES	NO	NO	NO	1
num_op_var_39_efect_ult3	YES	NO	NO	NO	1
num_reemb_var17_ult1	YES	NO	NO	NO	1
num_var22_ult1	YES	NO	NO	NO	1
num_var22_ult3	YES	NO	NO	NO	1
saldo_var24	YES	NO	NO	NO	1
saldo_var42	YES	NO	NO	YES	2
var38	YES	YES	NO	NO	2
saldo_var30	NO	YES	YES	YES	3

Variable	Logistic Regression	Decision Tree	Gradient Boosting	Binning Interactivo	# of Present Models
saldo_var5	NO	NO	NO	YES	1
saldo_medio_var5_ult1	NO	YES	NO	YES	2
saldo_medio_var5_hace2	NO	NO	NO	YES	1
num_var30	NO	NO	NO	YES	1
num_var35	NO	NO	NO	YES	1
num_var4	NO	NO	NO	YES	1
ind_var5	NO	NO	NO	YES	1
num_var5	NO	NO	NO	YES	1
num_var42	NO	NO	NO	YES	1
var36	NO	NO	NO	YES	1
saldo_medio_var5_hace3	NO	YES	NO	YES	2
imp_op_var39_efect_ult1	NO	YES	NO	NO	1
num_var45_hace3	NO	YES	NO	NO	1
imp_op_var39_efect_ult3	NO	YES	NO	NO	1
saldo_medio_var8_ult1	NO	YES	NO	NO	1

In the image below we can see that I have no hardware enough to work with this dataset.



The screenshot shows the SAS software interface on a Windows 7 system. The main window title is "Windows 7 - SAS". In the center, there are two main windows: "Log - (Untitled)" and "Output - (Untitled)".

The "Log - (Untitled)" window contains the following SAS code:

```

SAS
File Edit View Tools Run Solutions Window Help
SAS
Results
Log - (Untitled)
MPRINT(CRUDPOLOGISTICA): set fantasma estadisticos;
MPRINT(CRUDPOLOGISTICA): exec, ncvs, datafantasma.sasv, nowint;
PROC DATASETS LIB=WORK MEMOPTIONS=(NOCOMPACT) MEMBER=(UCM_CAO_KAGGLE_SANTANDER);
CONTAIN=VAR3 VAR3 delta imp_reemb_var18 imp_op_var39_efect_u imp_op_var40_efect_u ind_var13 ind_var24 ind_var30 ind_var30_0 ind_var43_reemb
casemode=;
CPUTIME=4, SUBINIO=12345, SUBSTIME=12345, OBJETIVO=tarafellos;;
DATA final1/SET final1(nobr=1);
run;

```

The "Output - (Untitled)" window contains the output of the SAS code, showing various system messages and error logs related to memory issues and disk space problems. Key messages include:

- "not enough memory to complete operation"
- "not enough disk space to complete operation"
- "out of memory"
- "maximum file size exceeded"

Logistic Regression with Cross Validation using the dataset with the 33 variables

SAS Code

```
data uno_small; set uno_small; u=(ranuni(12355));  
proc sort data=uno_small; by u;
```

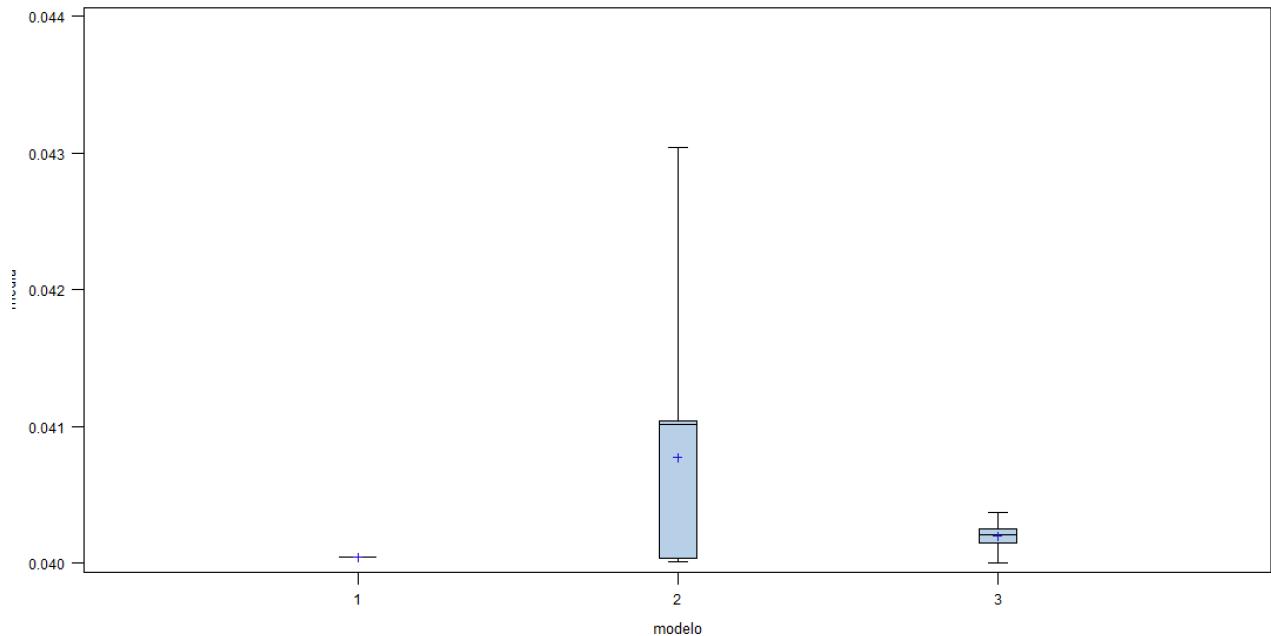
Obs	suma	media	semilla
1	0.16138	0.040345	12345
2	0.16064	0.040160	12346
3	0.16063	0.040058	12347
4	0.16064	0.040160	12348
5	0.16090	0.040200	12349
6	0.16095	0.040213	12350
7	0.16117	0.040292	12351
8	0.16027	0.040068	12352
9	0.16059	0.040147	12353
10	0.16112	0.040279	12354
11	0.16101	0.040253	12355
12	0.16064	0.040160	12356
13	0.16059	0.040147	12357
14	0.16085	0.040213	12358
15	0.16038	0.040095	12359
16	0.16096	0.040239	12360
17	0.16064	0.040160	12361
18	0.16101	0.040253	12362
19	0.16065	0.040160	12363
20	0.16101	0.040253	12364
21	0.16091	0.040226	12365
22	0.16091	0.040226	12366
23	0.16048	0.040121	12367
24	0.16085	0.040213	12368
25	0.16001	0.040003	12369
26	0.16096	0.040239	12370
27	0.16017	0.040042	12371
28	0.16059	0.040147	12372
29	0.16148	0.040371	12373
30	0.16096	0.040239	12374
31	0.16138	0.040345	12375

We will use cross validation with 3 models.

Model 1 - All 33 variables using 7000 rows for train and 3000 for test.

Model 2 - Only 2 variables (Gradient Boosting) using 7000 rows for train and 3000 for test.

Model 3 - All 371 variables and all data (Original Dataset);



As we can see mode 1 is the best model.

SAS Code:

```
%cruzadalogistica  
(archivo=uno_small,  
vardepen=TARGET,  
conti=VAR3  
ind_var13  
ind_var24  
ind_var30
```

```

ind_var30_0
ind_var31_0
ind_var_8_0
num_meses_var4_ult3
num_meses_var5_ult3
num_meses_var8_ult3
num_op_var_39_efect_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
categor=,
ngrupos=4,sinicial=12345,sfinal=12375, objetivo=tasafallos);
data final1;set final;modelo=1;

%cruzadalogistica
(archivo=uno_small,vardepen=TARGET,
conti=saldo_var30 var3,
categor=,
ngrupos=4,sinicial=12345,sfinal=12375, objetivo=tasafallos);
data final2;set final;modelo=2;

%cruzadalogistica
(archivo=uno,vardepen=TARGET,
conti=VAR2 VAR3 delta_imp_amort_var18_1y3 delta_imp_amort_var34_1y3
delta_imp_aport_var13_1y3
delta_imp_aport_var17_1y3 delta_imp_aport_var33_1y3
delta_imp_compra_var44_1y3
delta_imp_reemb_var13_1y3 delta_imp_reemb_var17_1y3
delta_imp_reemb_var33_1y3
delta_imp_trasp_var17_in_1y3 delta_imp_trasp_var17_out_1y3
delta_imp_trasp_var33_in_1y3
delta_imp_trasp_var33_out_1y3 delta_imp_venta_var44_1y3
delta_num_aport_var13_1y3
delta_num_aport_var17_1y3 delta_num_aport_var33_1y3
delta_num_compra_var44_1y3
delta_num_venta_var44_1y3 imp_amort_var18_ult1 imp_amort_var34_ult1
imp_aport_var13_hace3
imp_aport_var13_ult1 imp_aport_var17_hace3 imp_aport_var17_ult1
imp_aport_var33_hace3
imp_aport_var33_ult1 imp_compra_var44_hace3 imp_compra_var44_ult1
imp_ent_var16_ult1
imp_op_var39_comer_ult1 imp_op_var39_comer_ult3 imp_op_var39_efect_ult1
imp_op_var39_efect_ult3

```

imp_op_var39_ult1 imp_op_var40_comer_ult1 imp_op_var40_comer_ult3
imp_op_var40_efect_ult1
imp_op_var40_efect_ult3 imp_op_var40_ult1 imp_op_var41_comer_ult1
imp_op_var41_comer_ult3
imp_op_var41_efect_ult1 imp_op_var41_efect_ult3 imp_op_var41_ult1
imp_reemb_var13_ult1
imp_reemb_var17_hace3 imp_reemb_var17_ult1 imp_reemb_var33_ult1
imp_sal_var16_ult1
imp_trans_var37_ult1 imp_trasp_var17_in_hace3 imp_trasp_var17_in_ult1
imp_trasp_var17_out_ult1
imp_trasp_var33_in_hace3 imp_trasp_var33_in_ult1 imp_trasp_var33_out_ult1
imp_var43_emit_ult1
imp_var7_emit_ult1 imp_var7_recib_ult1 imp_venta_var44_hace3
imp_venta_var44_ult1 ind_var1 ind_var5
ind_var6 ind_var8 ind_var12 ind_var13 ind_var14 ind_var17 ind_var19
ind_var20 ind_var24 ind_var30
ind_var31 ind_var33 ind_var40 ind_var44 ind_var10_ult1 ind_var10cte_ult1
ind_var12_0 ind_var13_0
ind_var13_corto ind_var13_corto_0 ind_var13_largo ind_var13_largo_0
ind_var13_medio_0 ind_var14_0
ind_var17_0 ind_var18_0 ind_var1_0 ind_var20_0 ind_var24_0 ind_var25_0
ind_var25_cte ind_var26_0
ind_var26_cte ind_var30_0 ind_var31_0 ind_var32_0 ind_var32_cte
ind_var33_0 ind_var34_0 ind_var37_0
ind_var37_cte ind_var39_0 ind_var40_0 ind_var41_0 ind_var43_emit_ult1
ind_var43_recib_ult1 ind_var44_0
ind_var5_0 ind_var6_0 ind_var7_emit_ult1 ind_var7_recib_ult1 ind_var8_0
ind_var9_cte_ult1
ind_var9_ult1 num_aport_var13_hace3 num_aport_var13_ult1
num_aport_var17_hace3 num_aport_var17_ult1
num_aport_var33_hace3 num_aport_var33_ult1 num_compra_var44_hace3
num_compra_var44_ult1
num_ent_var16_ult1 num_med_var22_ult3 num_med_var45_ult3
num_meses_var12_ult3
num_meses_var13_corto_ult3 num_meses_var13_largo_ult3
num_meses_var13_medio_ult3 num_meses_var17_ult3
num_meses_var29_ult3 num_meses_var33_ult3 num_meses_var39_vig_ult3
num_meses_var44_ult3
num_meses_var5_ult3 num_meses_var8_ult3 num_op_var39_comer_ult1
num_op_var39_comer_ult3
num_op_var39_efect_ult1 num_op_var39_efect_ult3 num_op_var39_hace2
num_op_var39_hace3
num_op_var39_ult1 num_op_var39_ult3 num_op_var40_comer_ult1
num_op_var40_comer_ult3
num_op_var40_efect_ult1 num_op_var40_efect_ult3 num_op_var40_hace2
num_op_var40_hace3
num_op_var40_ult1 num_op_var40_ult3 num_op_var41_comer_ult1
num_op_var41_comer_ult3
num_op_var41_efect_ult1 num_op_var41_efect_ult3 num_op_var41_hace2
num_op_var41_hace3
num_op_var41_ult1 num_op_var41_ult3 num_reemb_var13_ult1
num_reemb_var17_hace3 num_reemb_var17_ult1
num_reemb_var33_ult1 num_sal_var16_ult1 num_trasp_var11_ult1
num_trasp_var17_in_hace3
num_trasp_var17_in_ult1 num_trasp_var17_out_ult1 num_trasp_var33_in_hace3
num_trasp_var33_in_ult1
num_trasp_var33_out_ult1 num_var1 num_var4 num_var5 num_var6 num_var8
num_var12 num_var13 num_var14
num_var17 num_var20 num_var24 num_var30 num_var31 num_var33 num_var35
num_var40 num_var42 num_var44
num_var12_0 num_var13_0 num_var13_corto num_var13_corto_0 num_var13_largo
num_var13_largo_0
num_var13_medio_0 num_var14_0 num_var17_0 num_var18_0 num_var1_0
num_var20_0 num_var22_hace2

```

num_var22_hace3 num_var22_ult1 num_var22_ult3 num_var24_0 num_var25_0
num_var26_0 num_var30_0
num_var31_0 num_var32_0 num_var33_0 num_var34_0 num_var37_0
num_var37_med_ult2 num_var39_0 num_var40_0
num_var41_0 num_var42_0 num_var43_emit_ult1 num_var43_recib_ult1
num_var44_0 num_var45_hace2
num_var45_hace3 num_var45_ult1 num_var45_ult3 num_var5_0 num_var6_0
num_var7_emit_ult1
num_var7_recib_ult1 num_var8_0 num_venta_var44_hace3 num_venta_var44_ult1
saldo_medio_var12_hace2
saldo_medio_var12_hace3 saldo_medio_var12_ult1 saldo_medio_var12_ult3
saldo_medio_var13_corto_hace2
saldo_medio_var13_corto_hace3 saldo_medio_var13_corto_ult1
saldo_medio_var13_corto_ult3
saldo_medio_var13_largo_hace2 saldo_medio_var13_largo_hace3
saldo_medio_var13_largo_ult1
saldo_medio_var13_largo_ult3 saldo_medio_var13_medio_hace2
saldo_medio_var13_medio_ult3
saldo_medio_var17_hace2 saldo_medio_var17_hace3 saldo_medio_var17_ult1
saldo_medio_var17_ult3
saldo_medio_var29_hace2 saldo_medio_var29_hace3 saldo_medio_var29_ult1
saldo_medio_var29_ult3
saldo_medio_var33_hace2 saldo_medio_var33_hace3 saldo_medio_var33_ult1
saldo_medio_var33_ult3
saldo_medio_var44_hace2 saldo_medio_var44_hace3 saldo_medio_var44_ult1
saldo_medio_var44_ult3
saldo_medio_var5_hace2 saldo_medio_var5_hace3 saldo_medio_var5_ult1
saldo_medio_var5_ult3
saldo_medio_var8_hace2 saldo_medio_var8_hace3 saldo_medio_var8_ult1
saldo_medio_var8_ult3 saldo_var1
saldo_var5 saldo_var6 saldo_var8 saldo_var12 saldo_var13 saldo_var14
saldo_var17 saldo_var18
saldo_var20 saldo_var24 saldo_var25 saldo_var26 saldo_var30 saldo_var31
saldo_var32 saldo_var33
saldo_var34 saldo_var37 saldo_var40 saldo_var42 saldo_var44
saldo_var13_corto saldo_var13_largo
saldo_var13_medio var21 var36 var38,
categor=,
ngrupos=4,sinicio=12345,sfinal=12375, objetivo=tasafallos);
data final3;set final; modelo=3;

```

```
Proc print data=final; run;
```

```

data union;set final1 final2 final3;
ods graphics off;
proc boxplot data=union;plot media*modelo;run;

```

```

%include '\\vmware-host\Shared Folders\git\Bitbucket\santander-kaggle
\macros-and-code-samples\neural binarias basicas.sas';

%binarialogistic(archivo=uno_small,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
ind_var_8_0
num_meses_var4_ult3
num_meses_var5_ult3

```

```

num_meses_var8_ult3
num_op_var_39_efect_ult3
num_reemb_var17_ulti
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
      listclass=,
vardep=TARGET,
corte=50,semilla=12345,porcen=0.80) ;

```

The SAS System													
Obs	vp	vn	fp	fn	suma	porcen_VN	porcenFN	porcen_vp	porcen_fp	sensi	especif	tasafallos	tasaciertos
1	0	18221	0	784	19005	0.95875	0.041252	0	0	0	1	0.041252	0.95875

Output - (Untitled)

The SAS System

The SURVEYSELECT Procedure

Selection Method Simple Random Sampling

Input Data Set ARCHIVOBASE

Random Number Seed 12345

Sample Size 799

Selection Probability 0.7998

Sampling Weight 0

Output Data Set MUESTRA

Table of pdepen by TARGET

pdepen	TARGET		Total
Frequency	0	1	
Percent			
Row Pct			
Col Pct			
0	18221	784	19005
	95.87	4.13	100.00
	95.87	4.13	
	100.00	100.00	
Total	18221	784	19005
	95.87	4.13	100.00

The “tasafallos” is 0.041% and 0.958% of “tasaciertos”, because we are not using all variables and all observation. If we use all variables and observation the results probably will be different, but using this small dataset model 1 is better.

Chapter Number Five

Neuronal Networks

Using neuronal networks to predict

Now we will try to improve results using Neuronal Network without cross validation.

We will try with different optimisation methods, seeds, etc.

We will first the number of nodes with the best results (tasadefallos).

Number of nodes (Test with 1 to 30 nodes)

I will try from 0 to 30 nodes to find the best node with the best results (tasadefallos).

```
/*Número de nodos - 1 to 30 node */

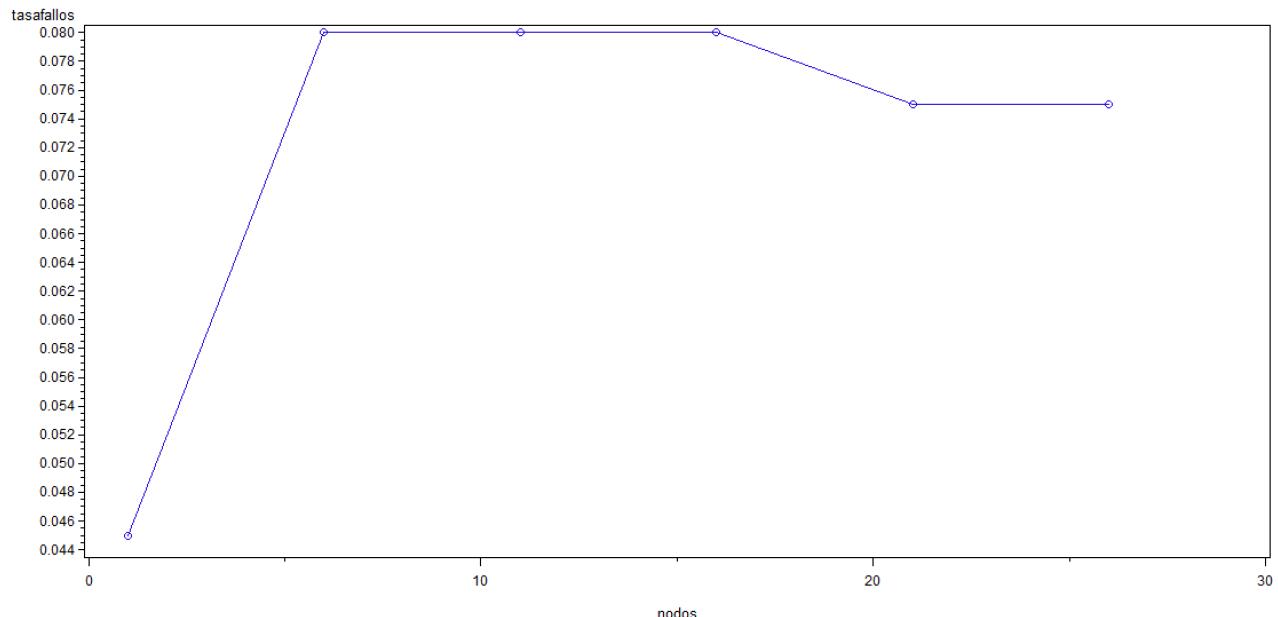
%macro numeronodos(inicionodos=,finalnodos=,increnodos=);
data union;run;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
  %neuralbinariabasica(archivo=uno_small,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
listclass=,
```

```

vardep=TARGET,
nodos=&nodos,corte=50,semilla=12345,porcen=0.80);
  data estadisticos;set estadisticos;nodos=&nodos;run;
  data union;set union estadisticos;run;
%end;
data union;set union ;if _n_=1 then delete;run;
symbol v=circle i=join;
proc gplot data=union;plot (porcenVN porcenFN porcenVP porcenFP sensi
especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%numeronodos(inicionodos=1,finalnodos=30,increnodos=5);

```



Number of nodes (Test with 3 to 20 nodes)

Now with 3 to 20 nodes.

```

/*Número de nodos*/

%macro numeronodos(inicionodos=,finalnodos=,increnodos=);
data union;run;
%do nodos=&inicionodos %to &finalnodos %by &increnodos;
  %neuralbinariabasica(archivo=uno_small,
listconti=VAR3
ind_var13
ind_var24

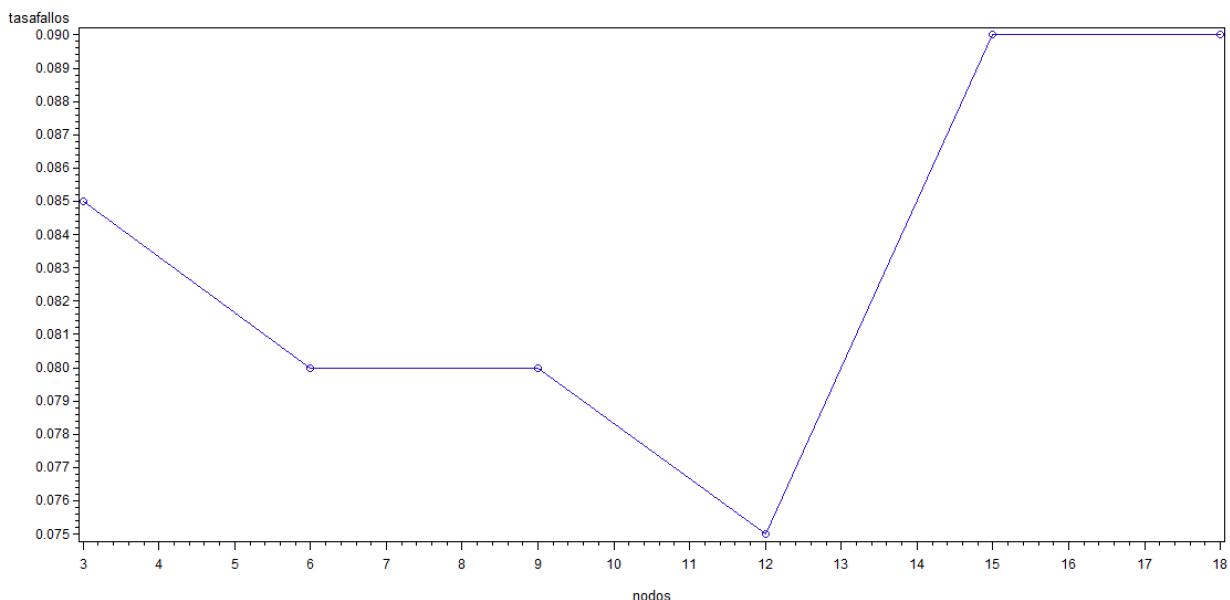
```

```

ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
      listclass=,
vardep=TARGET,
nodos=&nodos,corte=50,semilla=12345,porcen=0.80);
      data estadisticos;set estadisticos;nodos=&nodos;run;
      data union;set union estadisticos;run;
%end;
data union;set union ;if _n_=1 then delete;run;
symbol v=circle i=join;
proc gplot data=union;plot (porcenVN porcenFN porcenVP porcenFP sensi
especif tasafallos tasaciertos precision F_M)*nodos;run;
%mend;

%numeronodos(inicionodos=3,finalnodos=20,increnodos=3);

```



We will now use 12 nodes, because they have the best results (tasafallos).

Next step is to test with different optimisation methods.

Method = Levmar

```
/* Levmar 12 nodos semilla 12345 prelim=5 porcen=0.8
*/
%neuralbinariabasica(archivo=uno_small,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
listclass=,
vardep=TARGET,
nodos=12,corte=50,semilla=12345,porcen=0.80);
```

The FREQ Procedure				
Table of predi1 by TARGET				
		TARGET		
Frequency		0	1	Total
Percent	0	1889	80	1969
	1	94.45	4.00	98.45
	0	95.94	4.06	
	1	98.54	96.39	
Row Pct	0			
	1			
	0			
	1			
Col Pct	0			
	1			
	0			
	1			
Total		1917	83	2000
		95.85	4.15	100.00

Obs	vp	vn	fp	fn	suma	porcen VN	porcen FN	porcen VP	porcen FP	sensi	especif	tasafallos	tasaciertos
1	3	1889	28	80	2000	0.9445	0.04	.0015	0.014	0.036145	0.98539	0.054	0.946

For the other models I will only put the code line different, you can see the original code in the src folder.

`nodos=12, corte=50, algoritmo=levmar, semilla=456, porcen=0.80);`

Table of predit by TARGET				
predit		TARGET		
Frequency				
Percent				
Row Pct				
Col Pct		0	1	Total
	0	1894 94.70 96.09 98.65	77 3.85 3.91 96.25	1971 98.55
	1	26 1.30 89.66 1.35	3 0.15 10.34 3.75	29 1.45
Total		1920 96.00	80 4.00	2000 100.00

Obs	vp	vn	fp	fn	suma	porcen VN	porcen FN	porcen VP	porcen FP	sensi	especif	tasafallos	tasaciertos	precision
1	3	1894	26	77	2000	0.947	0.0385	.0015	0.013	0.0375	0.98646	0.0515	0.9485	0.10345

Method = Back Propagation

```
nodos=12,corte=50,semilla=12345,algoritmo=bprop mom=0.8  
learn=0.2,porcen=0.80);
```

The FREQ Procedure			
Table of predit by TARGET			
predit		TARGET	
Frequency		0	1
Percent			Total
Row Pct			
Col Pct			
0	1889	80	1969
	94.45	4.00	98.45
	95.94	4.06	
	98.54	96.39	
1	28	3	31
	1.40	0.15	1.55
	90.32	9.68	
	1.46	3.61	
Total	1917	83	2000
	95.85	4.15	100.00

Obs	vp	vn	fp	fn	suma	porcen VN	porcen FN	porcen VP	porcen FP	sensi	especif	tasafallos	tasaciertos	precision
1	3	1894	26	77	2000	0.947	0.0385	.0015	0.013	0.0375	0.98646	0.0515	0.9485	0.10345

nodos=12,corte=50,semilla=456,algoritmo=BPROP mom=0.8
learn=0.2,porcen=0.80);

The FREQ Procedure				
Table of pred1 by TARGET				
pred1		TARGET		
Frequency				
Percent				
Row Pct				
Col Pct				
		0	1	Total
		1894	77	1971
		94.70	3.85	98.55
		96.09	3.91	
		98.65	96.25	
		1		
		26	3	29
		1.30	0.15	1.45
		89.66	10.34	
		1.35	3.75	
	Total	1920	80	2000
		96.00	4.00	100.00

Obs	vp	vn	fp	fn	suma	porcen VN	porcen FN	porcen VP	porcen FP	sensi	especif	tasafallos	tasaciertos	precisi
1	3	1889	28	80	2000	0.9445	0.04	.0015	0.014	0.036145	0.98539	0.054	0.946	0.0967

Method = Quasi Newton

nodos=12,corte=50,semilla=12345,algoritmo=quanew,porcen=0.80) ;

The SAS System

The FREQ Procedure

Table of pred1 by TARGET

		TARGET		Total
		0	1	
Frequency	0	1889	80	
	1	28	3	
	Percent	94.45	4.00	
Row Pct	95.94	4.06		
Col Pct	98.54	96.39		
0	1917	83	2000	
Total	95.85	4.15	100.00	

Obs	vp	vn	fp	fn	suma	porcen_VN	porcen_FN	porcen_VP	porcen_FP	sensi	especif	tasafallos	tasaciertos	precisi
1	3	1889	28	80	2000	0.9445	0.04	.0015	0.014	0.036145	0.98539	0.054	0.946	0.0967

nodos=12,corte=50,semilla=456,algoritmo=quanew,porcen=0.80) ;

The SAS System

The FREQ Procedure

Table of pred1 by TARGET

		TARGET		Total
		0	1	
Frequency	0	1894	77	
	1	26	3	
	Percent	94.70	3.85	
Row Pct	96.09	3.91		
Col Pct	98.65	96.25		
0	1920	80	2000	
Total	96.00	4.00	100.00	

Obs	vp	vn	fp	fn	suma	porcen_VN	porcen_FN	porcen_VP	porcen_FP	sensi	especif	tasafallos	tasaciertos	precisi
1	3	1889	28	80	2000	0.9445	0.04	.0015	0.014	0.036145	0.98539	0.054	0.946	0.0967

Method = Trust Region

nodos=12,corte=50,semilla=12345,algoritmo=trureg,porcen=0.80);

Obs	vp	vn	fp	fn	suma	porcen_VN	porcen_FN	porcen_VP	porcen_FP	sensi	especif	tasafallos	tasaciertos	precision
1	3	1894	26	77	2000	0.947	0.0385	.0015	0.013	0.0375	0.98646	0.0515	0.9485	0.10345

The SAS System					
The FREQ Procedure					
Table of predil by TARGET					
predil		TARGET			
Frequency	Percent	0	1	Total	
Row Pct	Col Pct				
0		1889 94.45 95.94 98.54	80 4.00 4.06 96.39	1969 98.45	
1		28 1.40 90.32 1.46	3 0.15 9.68 3.61	31 1.55	
Total		1917 95.85	83 4.15	2000 100.00	

nodos=12,corte=50,semilla=456,algoritmo=trureg,porcen=0.80);

The FREQ Procedure					
Table of predil by TARGET					
predil TARGET					
Frequency	Percent	0	1	Total	
Row Pct	Col Pct				
0		1894 94.70 96.09 98.65	77 3.85 3.91 96.25	1971 98.55	
1		26 1.30 89.66 1.35	3 0.15 10.34 3.75	29 1.45	
Total		1920 96.00	80 4.00	2000 100.00	

The SAS System														
Obs	vp	vn	fp	fn	suma	porcen VN	porcen FN	porcen VP	porcen FP	sensi	especif	tasafallos	tasaciertos	precision
1	3	1894	26	77	2000	0.947	0.0385	.0015	0.013	0.0375	0.98646	0.0515	0.9485	0.10345

Method	Tasa Fallos	Tasa Aciertos	Seed
Levmar	5,4 %	94,6 %	12345
Levmar	5,15 %	94,85 %	456
bprob	5,4 %	94,6 %	12345
bprob	5,15 %	94,85 %	456
quasi newton	5,15 %	94,85 %	12345
quasi newton	5,15 %	94,85 %	456
trust region	5,4 %	94,50 %	12345
trust region	5,15 %	94,85 %	456

Conclusion:

I will select quasi newton method because both with different seeds the results were the same and better results.

All models were executed with train dataset with 7999 rows.

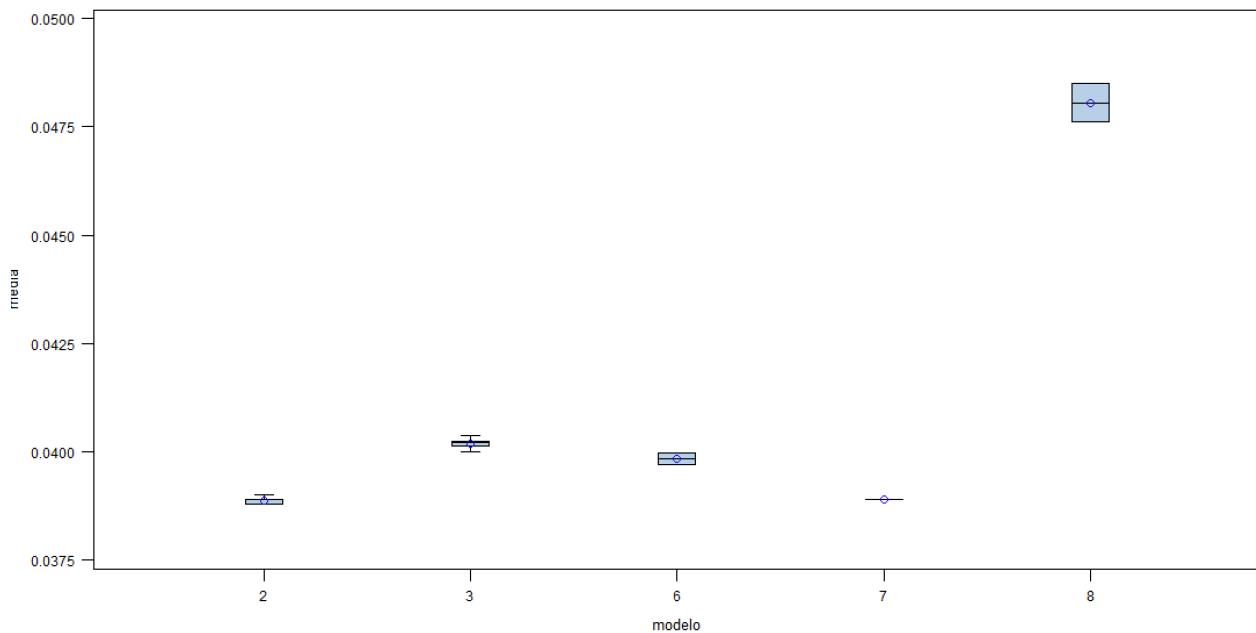
I also executed with the same SAS code but with more rows and less rows, but the results will not be described here in this project. Changing the size of the dataset also change the results (tasadefallos y tasasaciertos).

Neuronal Network using cross validation

When I used Cross Validation the time to process was very slow and even some models I had to leave the computer for more than 8 days without shutting down.

Even with this challenge I was able to generate some results and below we can see the models.

As we can see models 2 and 7 are better.



Model #	# of Variable	# of rows	Algorithm
2	2	7000 train / 3000 test	Logistic Regression
3	371	All dataset	Logistic Regression
6	30	7000 train / 3000 test	NN / nodos=12,meto=quanew
7	30	7000 train / 3000 test	NN / nodos=12,meto=bprop mom=0.8 learn=0.2
8	30	7000 train / 3000 test	NN / nodos=12,meto=levmar

SAS Code:

Model 6

```
%cruzadabinarianeural(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
categor=,
ngrupos=4,sinicio=123,sfinal=12456,nodos=12,meto=quanew,objetivo=tasafall
os);
data final6;set final;modelo=6;
```

Model 7

```
%cruzadabinarianeural(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
```

```

saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
categor=,
ngrupos=4,sinicio=123,sfinal=124,nodos=12,meto=bprop mom=0.8
learn=0.2,objetivo=tasafallos);
data final7;set final;modelo=7;

```

Model 8

```

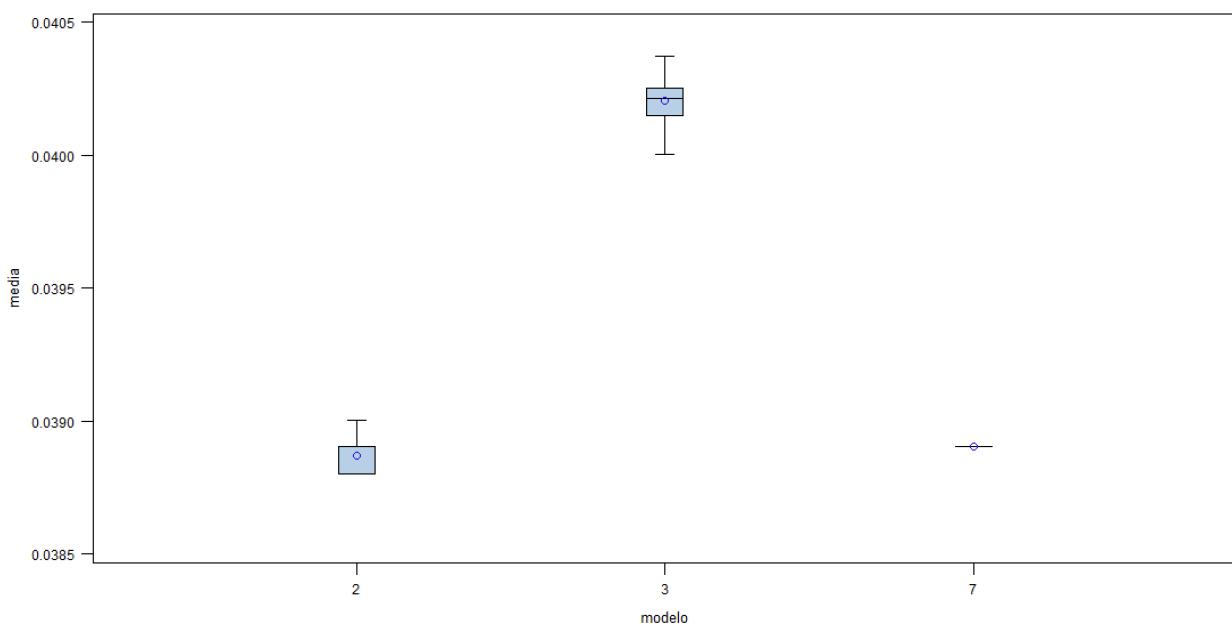
%cruzadabinarianeural(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
categor=,
ngrupos=4,sinicio=123,sfinal=124,nodos=12,meto=levmar,objetivo=tasafallos
);
data final8;set final;modelo=8;

data union;set final2 final3 final6 final7 final8 ;
ods graphics off;

```

```
proc boxplot data=union;plot media*modelo;run;  
  
data union;set final2 final3 final7;  
ods graphics off;  
proc boxplot data=union;plot media*modelo;run;
```

Models 2 and 7 are better.



Chapter Number Six

Trees

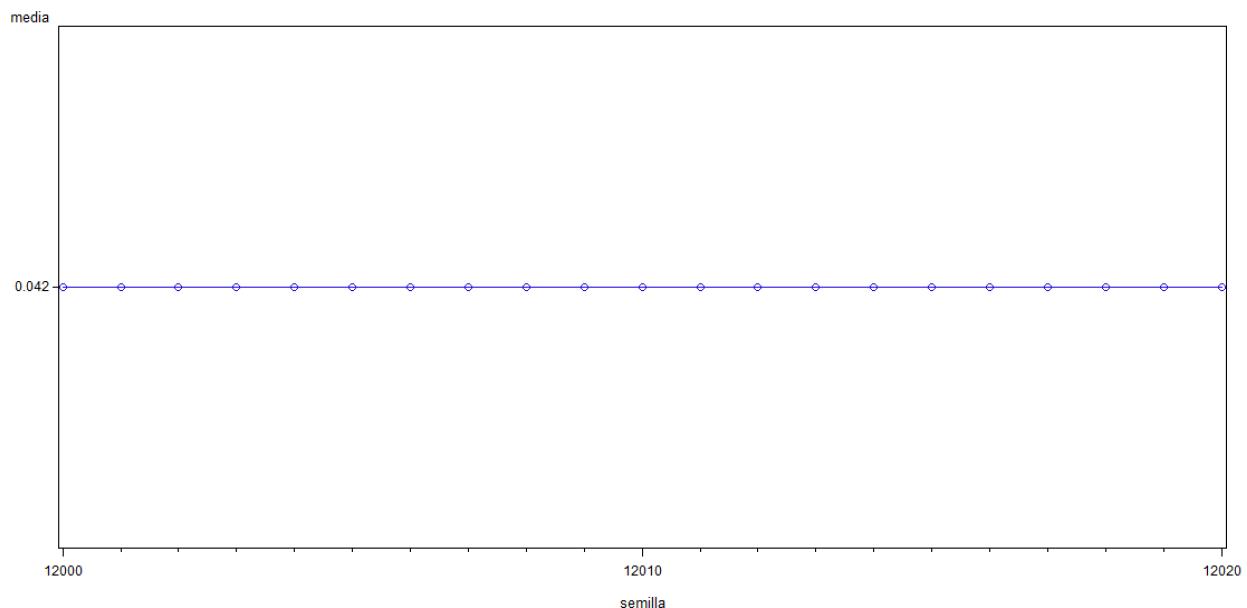
Using Bagging, Random Forest and Gradient Boosting to predict

As we can see the code below I tried others algorithms like Bagging, Randon Forest and Gradient Boosting.

SAS Code:

```
%baggingbin(archivo=uno_small,
vardep=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=
semilla1=12345,porcen1=0.80,
sinicial=12000,sfinal=12020,
porcenbag=1,maxbranch=5,
nleaves=20,tamhoja=10,
reemplazo=1,listatalog=1,compara=1);
```

BAGGING iteraciones=21 porcenbag=1 maxbranch=5 nleaves=20 tamhoja=10



BAGGING 12000 - 12020 . Iteraciones=21

Selection Indicator=0

The MEANS Procedure

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
1999	0.0420210	0.2006873	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0380000	0.1912082	0	1.0000000

```
%randomforestbin(archivo=uno_small,
vardep=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
semilla1=12345,porcen1=0.8,
maxtrees=50,variables=3,porcenbag=0.5,maxbranch=50,tamhoja=30,maxdepth=15
,pvalor=0.1,compara=1);
```

RANDOM FOREST Iteraciones=50

Selection Indicator=0

The MEANS Procedure

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
1999	0.0420210	0.2006873	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0380000	0.1912082	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0386250	0.1927116	0	1.0000000

```
%boostingbin(archivo=uno_small,
vardep=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
semillal=12345,porcenl=0.80,
iterations=1000,shrink=0.05,maxbranch=10,tamhoja=10,maxdepth=4,compara=1)
;
```

RESULTADOS LOGISTICA

Selection Indicator=0

The MEANS Procedure

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
1999	0.0420210	0.2006873	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0386250	0.1927116	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0386250	0.1927116	0	1.0000000

RESULTADOS LOGISTICA

Selection Indicator=0

The MEANS Procedure

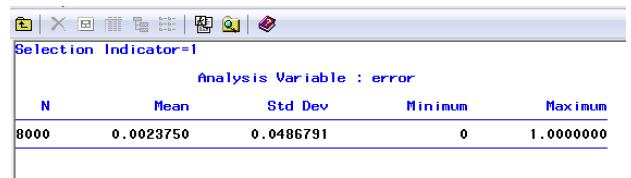
Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
1999	0.0530265	0.2241424	0	1.0000000

Selection Indicator=1

Analysis Variable : error

N	Mean	Std Dev	Minimum	Maximum
8000	0.0023750	0.0486791	0	1.0000000



```
%cruzadabaggingbin(archivo=uno_small,
vardepen=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
```

```

var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
ngrupos=4,sinicio=12345,sfinal=12355,
siniciobag=12345,sfinalbag=12355,
porcenbag=0.8,maxbranch=2,
nleaves=30,tamhoja=10,
reemplazo=1,objetivo=tasafallos);
data final9;set final;modelo=9;
proc print data=final9;run;

```

```

%cruzadabaggingbin(archivo=uno_small,
vardepen=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
ngrupos=4,sinicio=12345,sfinal=12355,
siniciobag=12345,sfinalbag=12355,
porcenbag=0.8,maxbranch=3,
nleaves=10,tamhoja=6,
reemplazo=1,objetivo=tasafallos);
data final10;set final;modelo=10;
proc print data=final10;run;

```

```

%cruzadabaggingbin(archivo=uno_small,
vardepen=TARGET,
listconti=VAR3
ind_var13
ind_var24

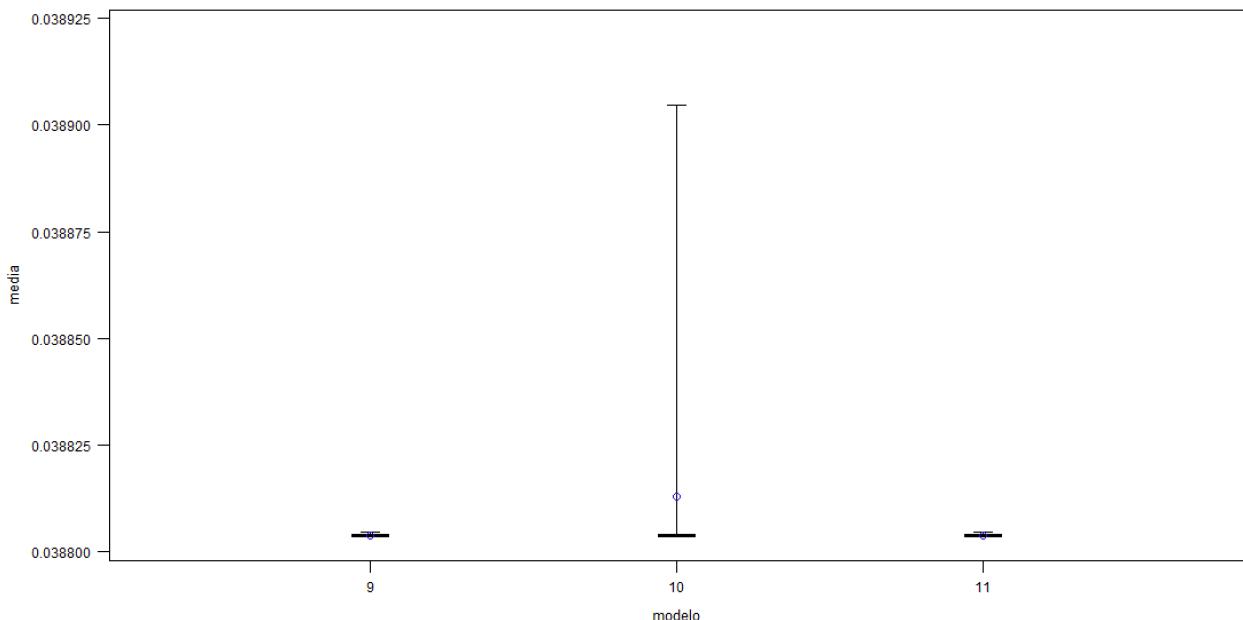
```

```

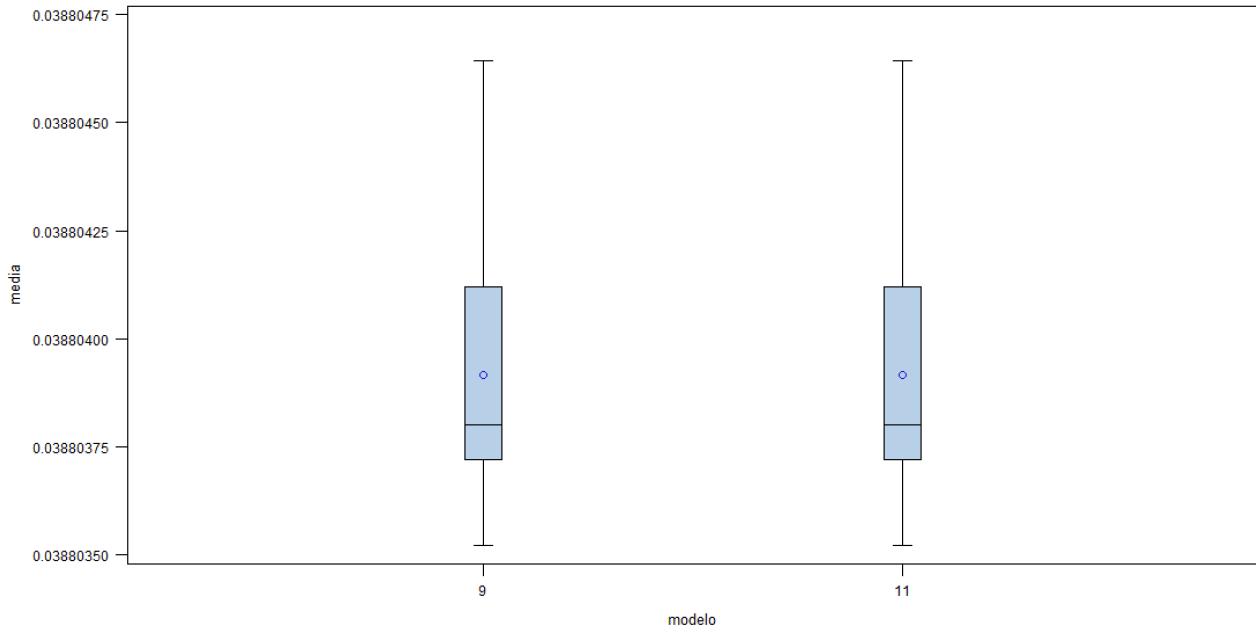
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
ngrupos=4,sinicio=12345,sfinal=12355,
siniciobag=12345,sfinalbag=12355,
porcenbag=0.8,maxbranch=5,
nleaves=20,tamhoja=10,
reemplazo=1,objetivo=tasafallos);
data final11;set final;modelo=11;
proc print data=final11;run;

data union;set final9 final10 final11;
proc boxplot data=union;plot media*modelo;run;

```



```
data union;set final9 final11;
proc boxplot data=union;plot media*modelo;run;
```



Now I tried Gradient Boosting.

```
/*Gradient boosting*/
%cruzadatreeboostbin(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
```

```

var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1
,categor=,ngrupos=4,sinicio=12345,sfinal=12365,leafsize=5,
iteraciones=300,shrink=0.02,maxbranch=2,maxdepth=4,mincatsize=15,minobs=2
0,objetivo=tasafallos);
data final12;set final;modelo=12;

%cruzadatreeboostbin(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1 ,
categor=,ngrupos=4,sinicio=12345,sfinal=12365,leafsize=5,
iteraciones=500,shrink=0.02,maxbranch=2,maxdepth=4,mincatsize=15,minobs=2
0,objetivo=tasafallos);
data final13;set final;modelo=13;

%cruzadatreeboostbin(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5

```

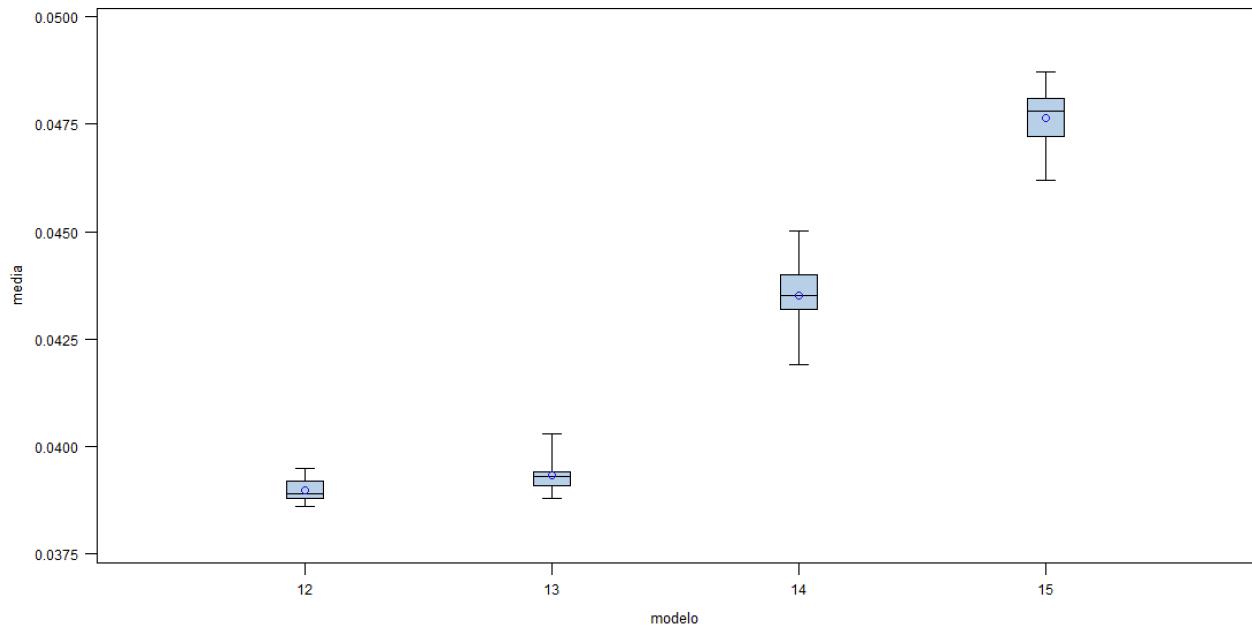
```

saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1
,categor=,ngrupos=4,sinicio=12345,sfinal=12365,leafsize=10,
iteraciones=1000,shrink=0.01,maxbranch=10,maxdepth=4,mincatsize=15,minobs
=20,objetivo=tasafallos);
data final14;set final;modelo=14;

%cruzadatreeboostbin(archivo=uno_small,vardepen=TARGET,
conti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1
,categor=,ngrupos=4,sinicio=12345,sfinal=12365,leafsize=10,
iteraciones=1000,shrink=0.05,maxbranch=10,maxdepth=4,mincatsize=15,minobs
=20,objetivo=tasafallos);
data final15;set final;modelo=15;

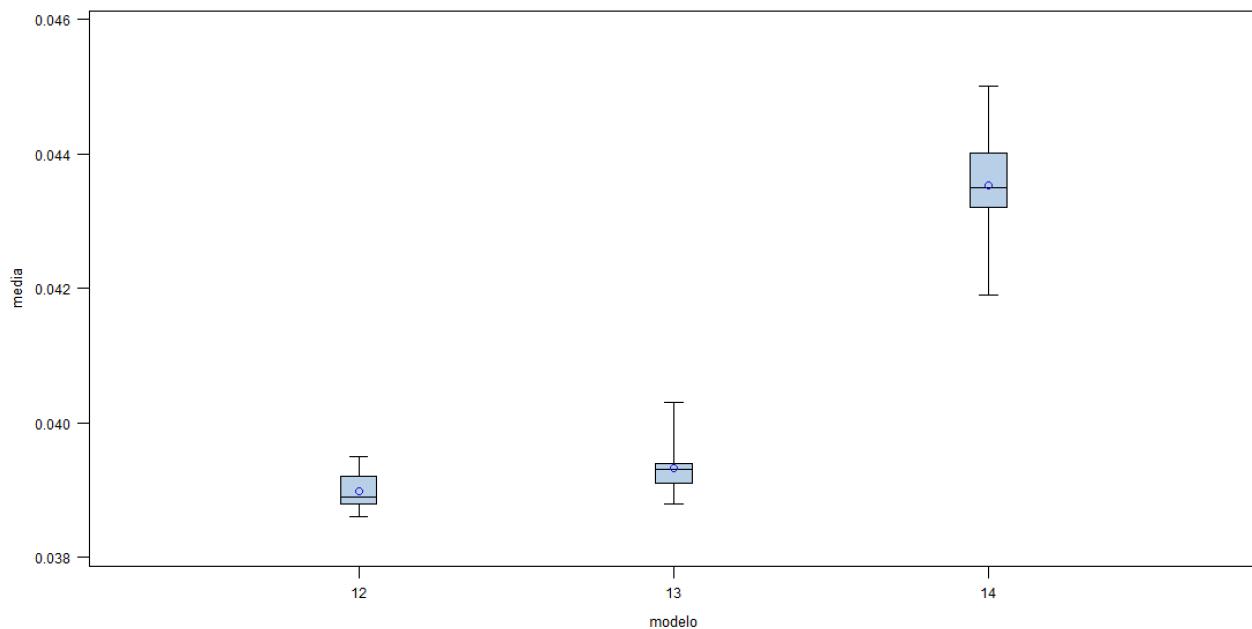
```

```
data union;set final12 final13 final14 final15;
proc boxplot data=union;plot media*modelo;run;
```



Model 12 and 13 are better.

```
data union;set final12 final13 final14;
proc boxplot data=union;plot media*modelo;run;
```



Now I tried Random Forest.

```
/*Random Forest*/  
  
%randomforestbin(archivo=uno_small,  
vardep=TARGET,  
listconti=VAR3  
ind_var13  
ind_var24  
ind_var30  
ind_var30_0  
ind_var31_0  
num_meses_var5_ult3  
num_meses_var8_ult3  
num_reemb_var17_ult1  
num_var22_ult1  
num_var22_ult3  
saldo_var24  
saldo_var42  
var38  
saldo_var30  
saldo_var5  
saldo_medio_var5_ult1  
saldo_medio_var5_hace2  
num_var30  
num_var35  
num_var4  
ind_var5  
num_var5  
num_var42  
var36  
saldo_medio_var5_hace3  
imp_op_var39_efect_ult1  
num_var45_hace3  
imp_op_var39_efect_ult3  
saldo_medio_var8_ult1,  
listcategor=,  
semilla1=12345,porcen1=0.8,  
maxtrees=50,variables=3,porcenbag=0.5,maxbranch=50,tamhoja=30,maxdepth=15  
,pvalor=0.1,compara=1);  
data final16;set final;modelo=16;  
  
%randomforestbin(archivo=uno_small,  
vardep=TARGET,  
listconti=VAR3  
ind_var13  
ind_var24  
ind_var30  
ind_var30_0  
ind_var31_0  
num_meses_var5_ult3  
num_meses_var8_ult3  
num_reemb_var17_ult1  
num_var22_ult1  
num_var22_ult3  
saldo_var24  
saldo_var42  
var38  
saldo_var30
```

```

saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
semilla1=12345,porcen1=0.8,
maxtrees=30,variables=3,porcenbag=0.5,maxbranch=10,tamhoja=10,maxdepth=15
,pvalor=0.1,compara=1);
data final17;set final;modelo=17;

%randomforestbin(archivo=uno_small,
vardep=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
semilla1=12345,porcen1=0.8,
maxtrees=50,variables=3,porcenbag=0.3,maxbranch=2,tamhoja=5,maxdepth=15,p
valor=0.1,compara=1);
data final18;set final;modelo=18;

%randomforestbin(archivo=uno_small,
vardep=TARGET,
listconti=VAR3
ind_var13

```

```

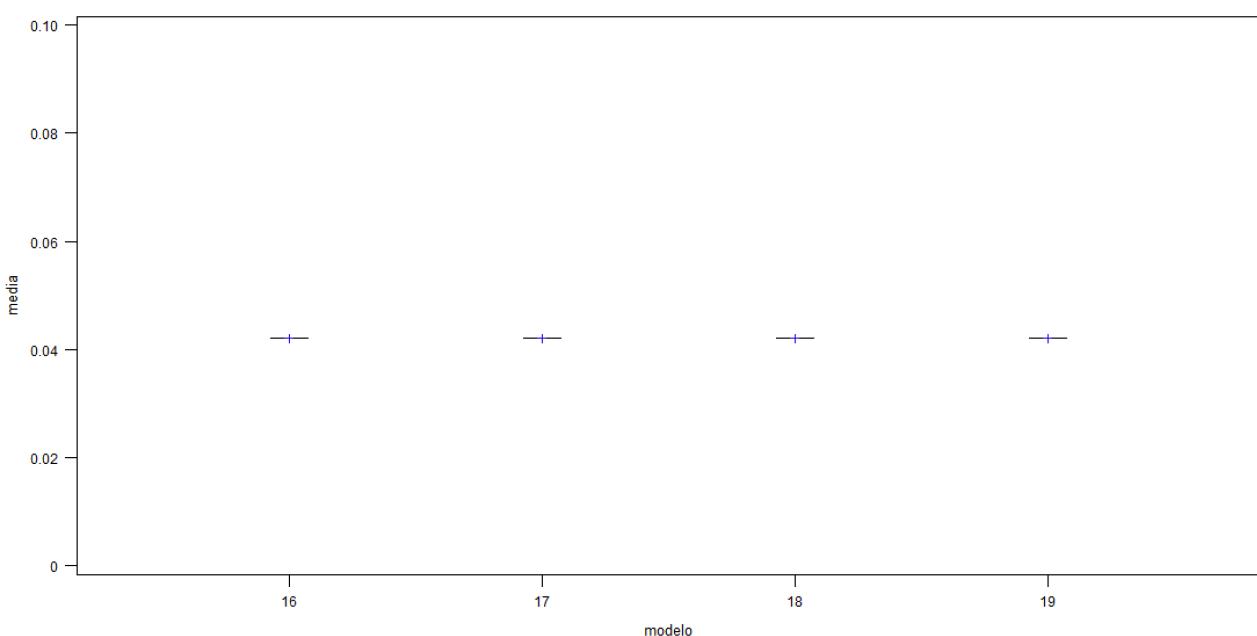
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
listcategor=,
semilla1=12345,porcen1=0.8,
maxtrees=50,variables=4,porcenbag=0.5,maxbranch=2,tamhoja=5,maxdepth=15,p
valor=0.1,compara=1);
data final19;set final;modelo=19;

data union;set final16 final17 final18 final19;
proc boxplot data=union;plot media*modelo;run;

proc print data=union;
run;

```

All models 16, 17, 18 and 19 have similar results.



```

data union;set final3 final6 final9 final13 final16;
proc boxplot data=union;plot media*modelo;run;
ods graphics off;
data union;set final3 final6 final9 final13;
proc boxplot data=union;plot media*modelo;run;

```

Now I tried KNN.

```

/*KNN*/
%cruzadakNNbin(archivo=uno_small,vardepen=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
ngrupos=4,seminicio=12345,semifinal=12385,k=7);
data final23;set final;modelo='kNN';

```

```

%cruzadakNNbin(archivo=uno_small,vardepen=TARGET,
listconti=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1

```

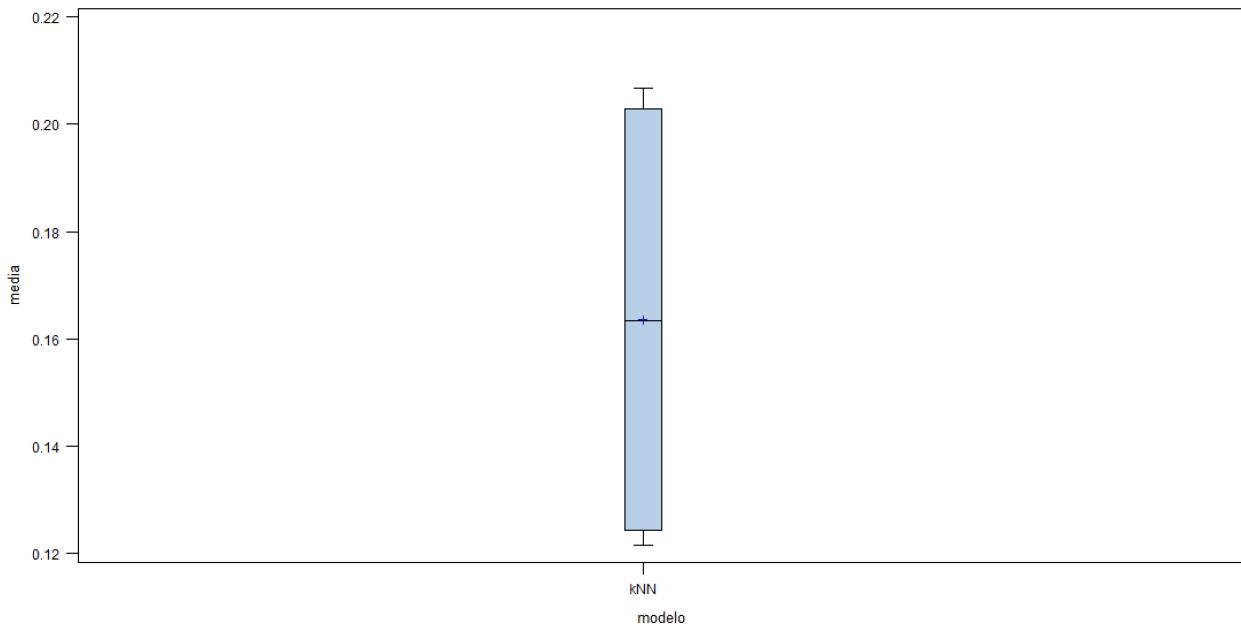
```

num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1,
ngrupos=4,seminicio=12345,semifinal=12385,k=3);
data final24;set final;modelo='kNN2';

data union;set final23 final24;
ods graphics off;

proc boxplot data=union;plot media*modelo;run;

```



Chapter Number Seven

Ensemble Models

Using neuronal networks to predict

The last try was to create an ensemble model.

```
/* Load the dataset to ucm library */
libname ucm '\\vmware-host\Shared Folders\git\Bitbucket\santander-kaggle
\dataset\' ;
run;

PROC IMPORT OUT= UCM.ensamblado
    DATAFILE= "\\vmware-host\Shared Folders\git\Bitbucket\santan
der-kaggle\dataset\train.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

data UCM.ensamblado; set UCM.ensamblado; id=_n_; run;

/* Show the number of missing values */
ods output nlevels=niveles; proc freq data=ucm.ensamblado nlevels; tables
_all_ / noprint; run;
proc means data=ucm.ensamblado n nmiss; run;

/*
76020*0.7 = 53.214 rows
70% of the dataset is equal 53.214 rows

data ucm.santander; set ucm.santander; Run;
data uno;set ucm.santander; u=(ranuni(12355));
proc sort data=uno; by u;
data train test;
set uno; if _n_<=53214 then output train;else output test;run;

*/
/* para pruebas

999*0.7 = 699 rows
70% of the dataset is equal 699 rows

*/


data ucm.ensamblado;
set ucm.ensamblado (drop = ID);
run;

data ucm.ensamblado;
set ucm.ensamblado (drop = ind_var29_
ind_var29
```

```

ind_var13_medio
ind_var18
ind_var26
ind_var25
ind_var32
ind_var34
ind_var37
ind_var39
num_var29_0
num_var29
num_var13_medio
num_var18
num_var26
num_var25
num_var32
num_var34
num_var37
num_var39
saldo_var29
saldo_medio_var13_medio_ult1
delta_num_reemb_var13_1y3
delta_num_reemb_var17_1y3
delta_num_reemb_var33_1y3
delta_num_trasp_var17_in_1y3
delta_num_trasp_var17_out_1y3
delta_num_trasp_var33_in_1y3
delta_num_trasp_var33_out_1y3
ind_var2_0
ind_var2
ind_var27_0
ind_var28_0
ind_var28
ind_var27
ind_var41
ind_var46_0
ind_var46
num_var27_0
num_var28_0
num_var28
num_var27
num_var41
num_var46_0
num_var46
saldo_var28
saldo_var27
saldo_var41
saldo_var46
imp_amort_var18_hace3
imp_amort_var34_hace3
imp_reemb_var13_hace3
imp_reemb_var33_hace3
imp_trasp_var17_out_hace3
imp_trasp_var33_out_hace3
num_var2_0_ult1
num_var2_ult1
num_reemb_var13_hace3
num_reemb_var33_hace3
num_trasp_var17_out_hace3
num_trasp_var33_out_hace3
saldo_var2_ult1
saldo_medio_var13_medio_hace3);
run;

```

```

data ucm.ensamblado;
set ucm.ensamblado;
if _n_>=10000 then delete;
run;

/* Describe the dataset variables */
proc contents data=ucm.santander out=sal;
data; set sal; put name @@; run;

data ucm.ensamblado; set ucm.ensamblado; Run;
data uno;set ucm.ensamblado; u=(ranuni(12355));
proc sort data=uno; by u;
data train test;
set uno; if _n_<=7000 then output train;
else output test;
run;

proc surveyselect data=uno out=muestra method=srs n=3000 outall
seed=12345;
data train valida;set muestra; if selected=1 then output train;else
output valida;run;

/*Ensambllo red y logistica */

proc printto print='Z:\git\Bitbucket\santander-kaggle\logs\logs.txt';run;
PROC DMDB DATA=train dmdbcatt=catares;
target TARGET;
var VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1;
class TARGET;
run;

proc neural data=train dmdbcatt=catares random=9999;

```

```

input VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5
num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1;
input;
target TARGET /level=nominal;
hidden 9 /act=arc;
netoptions randist=normal ranscale=0.15 random=15459;
prelim 15 preiter=10 pretech=levmar;
train maxiter=100 technique=levmar;
score data=valida role=test out=sall;
run;

```

```

proc logistic data=train ;
class ;
model TARGET=VAR3
ind_var13
ind_var24
ind_var30
ind_var30_0
ind_var31_0
num_meses_var5_ult3
num_meses_var8_ult3
num_reemb_var17_ult1
num_var22_ult1
num_var22_ult3
saldo_var24
saldo_var42
var38
saldo_var30
saldo_var5
saldo_medio_var5_ult1
saldo_medio_var5_hace2
num_var30
num_var35
num_var4
ind_var5

```

```

num_var5
num_var42
var36
saldo_medio_var5_hace3
imp_op_var39_efect_ult1
num_var45_hace3
imp_op_var39_efect_ult3
saldo_medio_var8_ult1;
score data=valida out=sal2;
run;

data union;merge sal1 sal2;pensemble=(p_TARGET1+p_1)/2;run;

data salfin;set union;
if p_TARGET1>0.5 then prered=1; else prered=0;
if p_1>0.5 then prelog=1; else prelog=0;
if pensemble>0.5 then prensemble=1; else prensemble=0;
run;

proc freq data=salfin;tables prered*TARGET/out=s1;run;
proc freq data=salfin;tables prelog*TARGET/out=s2;run;
proc freq data=salfin;tables prensemble*TARGET/out=s3;run;

data estadisticos1(drop=count percent prered );
    retain vp vn fp fn suma 0;
    set s1 nobs=nume;
    suma=suma+count;
    if prered=0 and TARGET=0 then vn=count;
    if prered=0 and TARGET=1 then fn=count;
    if prered=1 and TARGET=0 then fp=count;
    if prered=1 and TARGET=1 then vp=count;
    if _n_=nume then do;
        tasafallos=1-(vp+vn)/suma;
        modelo='RED';
        output;
    end;
run;
data estadisticos2(drop=count percent prered );
    retain vp vn fp fn suma 0;
    set s2 nobs=nume;
    suma=suma+count;
    if prelog=0 and TARGET=0 then vn=count;
    if prelog=0 and TARGET=1 then fn=count;
    if prelog=1 and TARGET=0 then fp=count;
    if prelog=1 and TARGET=1 then vp=count;
    if _n_=nume then do;
        tasafallos=1-(vp+vn)/suma;
        modelo='LOG';
        output;
    end;
run;
data estadisticos3(drop=count percent prered );
    retain vp vn fp fn suma 0;
    set s3 nobs=nume;
    suma=suma+count;
    if prensemble=0 and TARGET=0 then vn=count;
    if prensemble=0 and TARGET=1 then fn=count;
    if prensemble=1 and TARGET=0 then fp=count;
    if prensemble=1 and TARGET=1 then vp=count;
    if _n_=nume then do;
        tasafallos=1-(vp+vn)/suma;
        modelo='ENSEMBLE';
        output;
    end;

```

```

run;

data u;set estadisticos1 estadisticos2 estadisticos3;run;
title 'ENSEMBLE RED-LOG';
ods graphics off;

proc print data=u;run;

```

As we can see in the results below the “tasafallos” was 0.05% (Neuronal Networks), 0.03% (Regression Logistic) and 0.04% (Ensemble Model).

ENSEMBLE RED-LOG 16:45 Wednesday, September 14, 2016 92										
Obs	vp	vn	fp	fn	suma	TARGET	tasafallos	modelo	prelog	prensemble
1	12	6631	115	241	6999	1	0.050864	RED	.	.
2	2	6737	9	251	6999	1	0.037148	LOG	i	.
3	5	6706	40	248	6999	1	0.041149	ENS	.	i

Chapter Number Eighth

Conclusions

It was a very difficult task to predict if a client is satisfied or unsatisfied. This binary classification problem had many features and the dataset was big enough to slow down SAS, some models stayed 8 days running and did not finished, only when I stopped it because I could not wait more time.

When I was using SAS Enterprise Miner without Cross Validation I was able to execute different types of algorithms and compare results.

But, when I started using the technique cross validation the time to process became very slow and hard to process, specially with Neuronal Network with Quanew.

I can concluded that the best models were created using R and the algorithm called xgboost (eXtreme Gradient Boosting), this model can be visualised in the first project available in the link: <https://github.com/caiomsouza/kaggle-competitions/tree/master/santander-customer-satisfaction>. This was the model submitted to Kaggle.

Using SAS, I developed a feature selection study in order to define the variables that I used to try different models.

I decided to stay only with 30 variables for some reasons. First, it was fast to process but the results with less variables were worst than with more variables. Second, I did a feature study to find the most important variables and tried hard to keep only the variables selected by the algorithms, I was trying to predict using different algorithms only with the features with prediction power.

I can conclude that the number of variable will improve the results, so models with more variables were better to predict than models with less variables. Some variables were not necessary and could be discard without any impact.

But because of computer power I decided to stay with only the variables that were more important according to the feature selection study developed in the chapter number two.

This dataset was a big challenge for me and I had to reduce the number of observations and features.

I really recommend the use of H2O to predict, it is very fast and in my opinion is fast than SAS and easy to use. But unfortunately it does not have all algorithms.

Chapter Number Nine

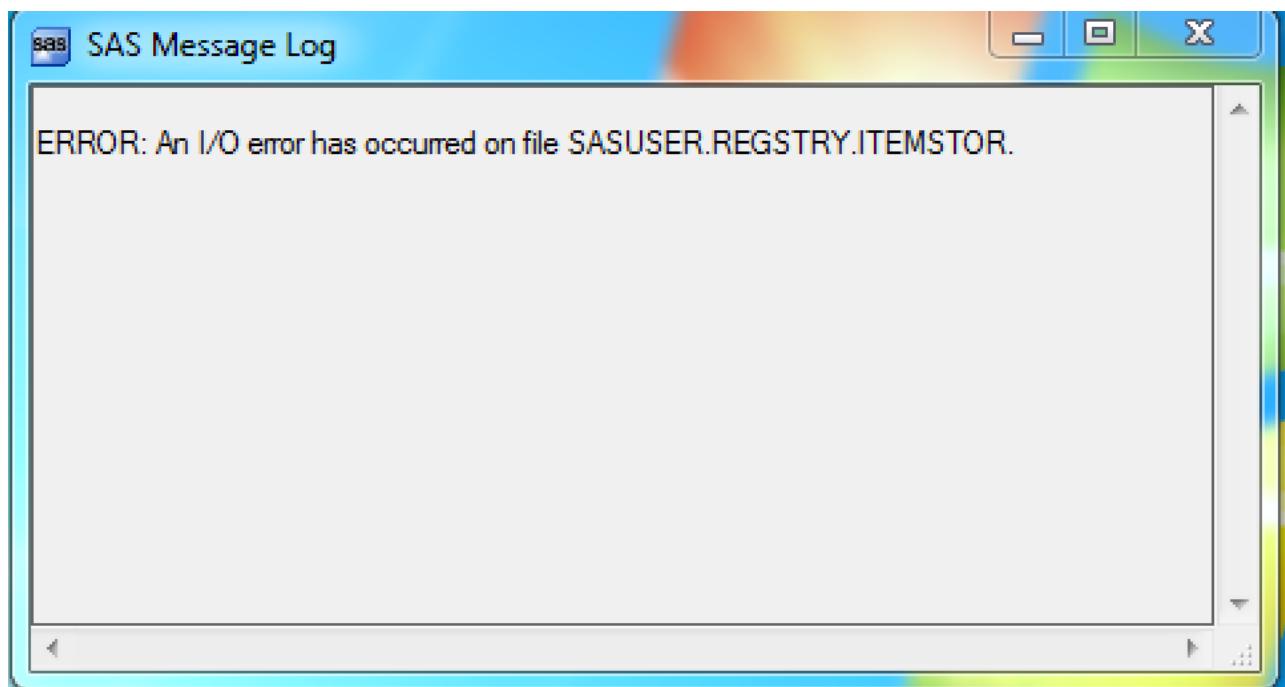
References

1. Slides - UCM. Prof. Dr. (Phd) Javier Portella, 2016.
2. Friedman, J., Hastie, T. and Tibshirani, R., 2001. The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.
3. H2o.ai - Gradient Boosting Machine. Available at: <http://www.h2o.ai/verticals/algos/gbm/>
4. Dal Pozzolo, Andrea, et al. "Racing for unbalanced methods selection." Intelligent Data Engineering and Automated Learning - IDEAL 2013. Springer Berlin Heidelberg, 2013. 24-31.

Chapter Number Ten

Extras

Below we can see some error that I had during the project. I had others errors but I did not put here. Some of this errors I was able to find the solutions.



ERROR: Undetermined I/O failure

<http://support.sas.com/kb/36/644.html>

```
7500 records read
7500 records read
7500 records read
ERROR: Floating Point Overflow.
ERROR: Termination due to Floating Point Exception
NOTE: The SAS System stopped processing this step because of errors.
WARNING: The data set WORK.SAL6 may be incomplete. When this step was stopped there were 0 observations and 316 variables.
WARNING: Data set WORK.SAL6 was not replaced because this step was stopped.
NOTE: PROCEDURE SVM used (Total process time):
      real time          0.28 seconds
      cpu time           0.28 seconds

7348  data final20;set final;modelo='SVM';
```