A mathematical model for fraud prediction and control planning of electrical company clients

UNIVERSIDAD COMPLUTENSE
MADRID

-Camilla Bruni (Italy). Università degli Studi di Firenze

-Herberth Espinoza Bernardo(Peru). Universidad Complutense de Madrid

-Antoine Levitt (France). University of Oxford

-María Loreto Luque (Spain). Universidad Complutense de Madrid

-Carlos Parra (Spain). Universidad Complutense de Madrid

-Aranzazu Pérez (Spain). Universidad Complutense de Madrid

-Elisa Pérez (Spain) Universidad Complutense de Madrid

Coordinators:

-Dr. Benjamin Ivorra (Universidad Complutense de Madrid)

-Dr. Juan Tejada (Universidad Complutense de Madrid)

-D. Jorge Juan Suerias (Neo Metrics)

-D. Fernando Fernández (Neo Metrics)

-Pilar (Neo Metrics)

neometrics

UNIVERSIDAD COMPLUTENSE
MADRID

OUTLINE

- **PROBLEM DESCRIPTION**
- **DATA ANALYSIS**
    - CD_EMPALME CODIFICATION
    - DATA DESCRIPTION
    - CATEGORIZE NON-LINEAR CONTINUOUS VARIABLES
    - PRINCIPAL COMPONENT ANALYSIS (PCA)
    - STUDY OF SIGNIFICANCE
    - PEARSON CORRELATION COEFICIENT
    - STUDY OF MULTICOLINEARITY
    - BINARY TREE
- **MATHEMATICAL MODELING**
    - REGRESSION
    - TRAINING
    - ROC CURVE
    - LIFT CHART
- **NUMERICAL VALIDATION**
    - OPTIMAL CONTROL CAMPAIGN
- **SUMMARY**
- **FUTURE WORK**

PROBLEM DESCRIPTION

In some countries it is a common practice to manipulate electrical meters to reduce the electrical invoice in a fraudulent way. An electrical company in Chile keeps a crew of inspectors to check whether customers are manipulating their electrical meters. They have a 100% safe method to identify and cancel a fraud where it exists. Each inspection has a cost and the company wishes to identify the customers with higher risk of fraud in order to reduce de cost of the investigation.

The current company policy is to check randomly on customers achieving a 6.6% of successful checks (success being "finding a fraudulent customer"). Using the data set provided by the company, with variables associated to each customer and the target variable (fraud or no fraud), we have:

1. Applied data mining techniques using SAS.
2. Fit a logistic regression model using SAS to calculate the lift chart.
3. Optimized number of inspections using Matlab.
4. Performed validation analysis.

DATA ANALYSIS

We received from *Neometrics* one *.csv* file with 79,459 records and 49 variables.
This file was composed of seven groups of variables. Those are:

- Geographic variables
- Connexion identifiers
- Customers characteristics
- Calculated variables of debt
- Calculated variables of payment
- Calculated variables of consume
- Informative variables that should not be use to construct the model, only for splitting samples.

According to the advices of *Neometrics'* workers we had to divide the dataset by *falso_target* in two samples as follows:

- One train sample composed of 0 and 1.
- Other valid sample composed of missing values.

Because of the great amount of records, we could not use a spreadsheet, such as Excel, for managing it, so we used Access.

In Access, we imported the *.csv* file and through a simple query similar to

> *Select falso_target from dataset where falso_target<>""*
> *Select falso_target from dataset where falso_target=""*

we split the former dataset.

Afterwards, we used SAS software to study our problem.

CD_EMPALME CODIFICATION

Within the connexion identifiers group there is a variable called *Cd_empalme*. This variable has its own code. The first figure means Overhead (A) or Underground (S) lines, the second figure correspond to the maximum electric current.

E.g., A6 is an overhead connection of six amperes of maximum current.

We considered more comfortable to divide those values in two new variables, which will be within this group. The two new ones are: *Codigo* and *Amperios*.

DATA DESCRIPTION

We have a total of 49 variables plus the variable target. These 49 variables are divided into several groups, as shown below.

grupo	variable
Cross variables	nis
	fe_corte
target	Resultado
Geographic variables	Cd_Sector
	Cd_area
	Nm_comuna
	Nm_zona
	Nr_consumidor
Conexion identifiers	SSEE
	Alimentador
	SED
	SED_numero
	Cd_empalme
Customers caracteristic	fc_ini_vigencia
	fc_ultima_lectura
	Cd_Estado_Cyr
	Nm_Tarifa
	Nm_tipo_suministro
	num_cortes
Calculated variables of debt	deuda_max_min
	deuda_ult_mean
	dif_max_min_deuda
	max_deuda
	max_dif_deuda
	dif_ult_mean_deuda
	mean_deuda
	min_deuda
	min_dif_deuda
	ultima_deuda
	ult_dif_deuda

grupo	variable
Calculated variables of payments	dif_max_min_pago
	dif_pago_alto
	dif_ult_mean_pago
	max_dif_pago
	max_pago
	mean_dif_pago
	mean_pago
	min_dif_pago
	min_pago
	pago_max_min
	pago_ult_mean
	ult_dif_pago
	ult_pago
Calculated variables of consume	tasa_estimados
	tasa_leidos
	tasa_resto
Informative variables that should not be use	fc_recepcion
	mes_recepcion
	St_actual
	falso_target

In this group only we stay with those that refer to geographic information, connection type, Customers characteristic, calculate variables of Debt and Calculated variables of payment. The other groups are not going to provide information to choose the variables of our model.

We select and simplify the more representing variables. To do so:

- We categorize non-linear continuous variables according to fraud proportion in population.
- We analyze the groups that have both categorical and quantitative variables in discrimination techniques.

– Groups containing only quantitative variables apply principal components (PCA) to see these correlations.

The variables we have to work with are divided into 9 groups.

- The first group is composed by the variables that are in common with the table ‘Consumos’, that contains the data about the consume of energy of every customers.
- The group with the geographic variables gives some information about the place where the clients live.
- The variables of the ‘connection identifiers’ kind tell us something about the supplier and the transformer used to provide the energy to the clients.
- Then there is a group of variables about the contract and the state of every client.
- There are two groups that contain significant values of debt and payments.
- The last to groups of variables are not used in building the model.
- The model must express the target variable ‘resultado’ in function of the other ones; the most relevant variables have to be found in order not to build a too-complex model.

CATEGORIZE NON-LINEAR CONTINUOUS VARIABLES

The dependence of ‘resultado’ on each variable has to be checked to decide on which variable the model will be built.

The variable ‘resultado’ has a linear dependence on some variables (‘num_cortes’ for instance), so these variables are the ones that influence the value of ‘resultado’ more than others.

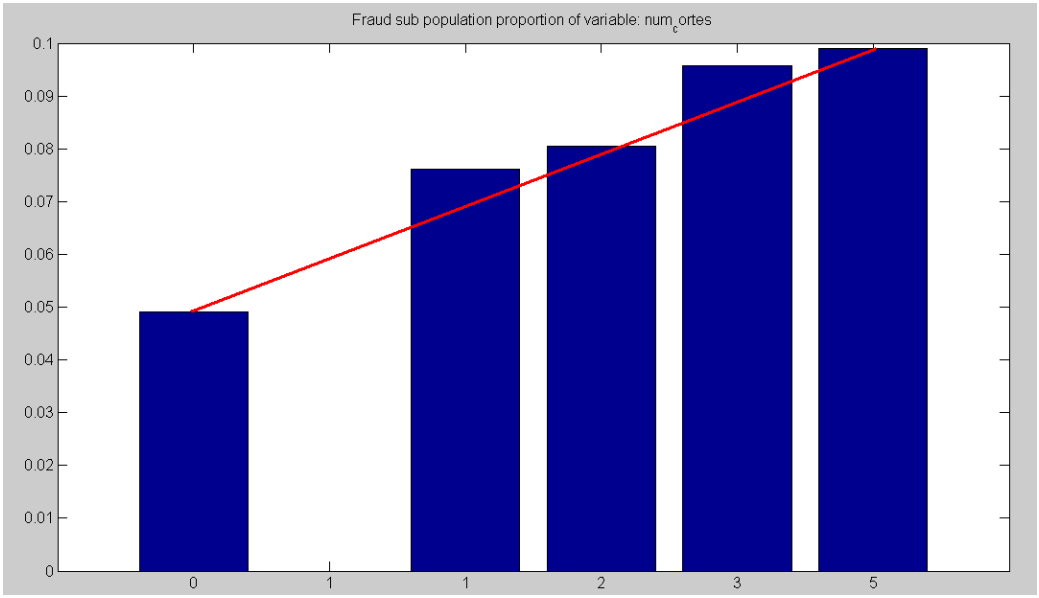


Figure: on the x axis there are the values of num_cortes, on the y axis there is the fraud proportion.

The variable 'mean_deuda' influence directly 'resultado': if a customer has a big debt with the electrical company, he could be more motivated to fraud the company itself.

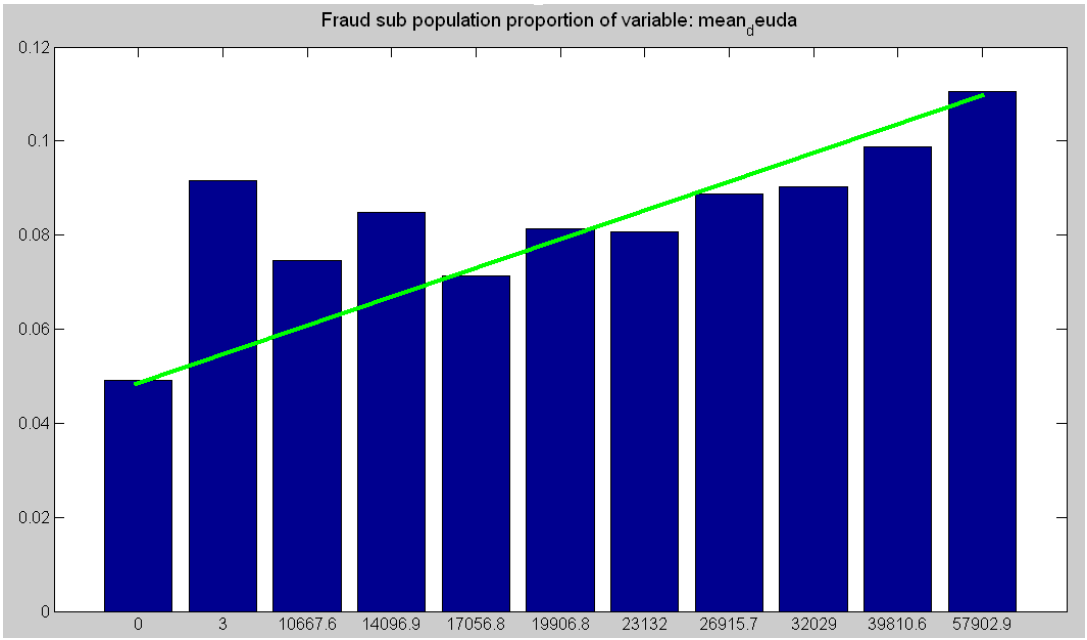


Figure: on the x axis there are the values of mean_deuda, on the y axis there is the fraud proportion.

Before using this variable in the model, the values have to be divided in the null ones and in the positive ones because of the particular code used.

Other variables that belong to the same group and have the same trend, like ‘max_deuda’, can be omitted because their contribute is very similar to the ‘mean_deuda’ one.

The dependence of ‘resultado’ on other variables is not linear, so they have to be categorized in groups to make their analysis easier.
For instance, the variable ‘pago_ult_mean’, that represents the difference between last and mean payment, has a parabolic trend. It has been divided in 4 groups: 0, values less than -5888, between -5888 and -582 and greater than -582.

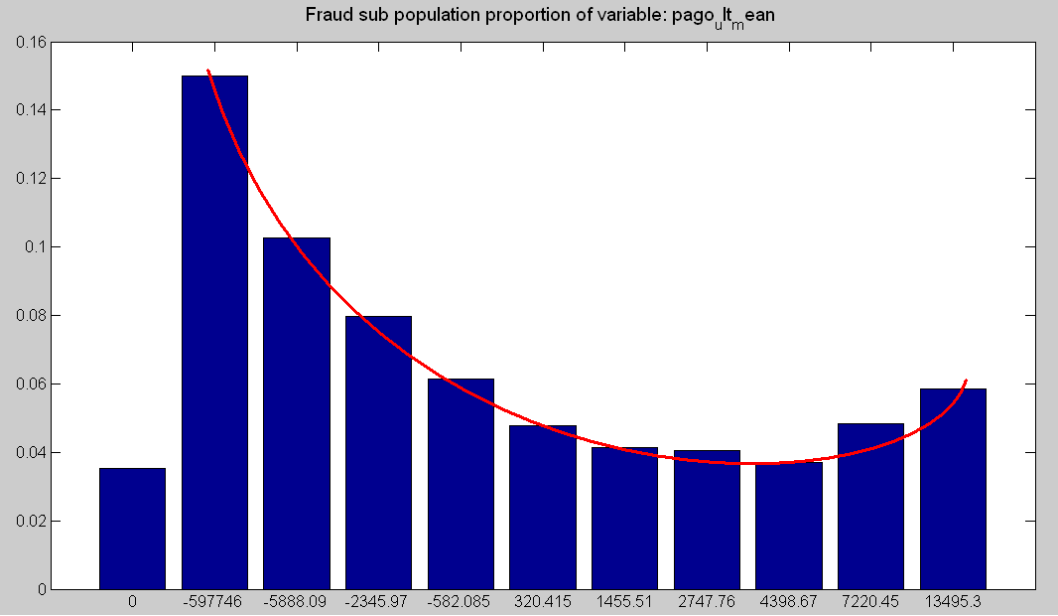


Figure: on the x axis there are the values of pago_ult_mean, on the y axis there is the fraud proportion.

Some variables have a particular trend: after an initial increasing, they decrease and increase again. To be used in the model, they have to be categorized in at least 3 groups that include the different types of trend.

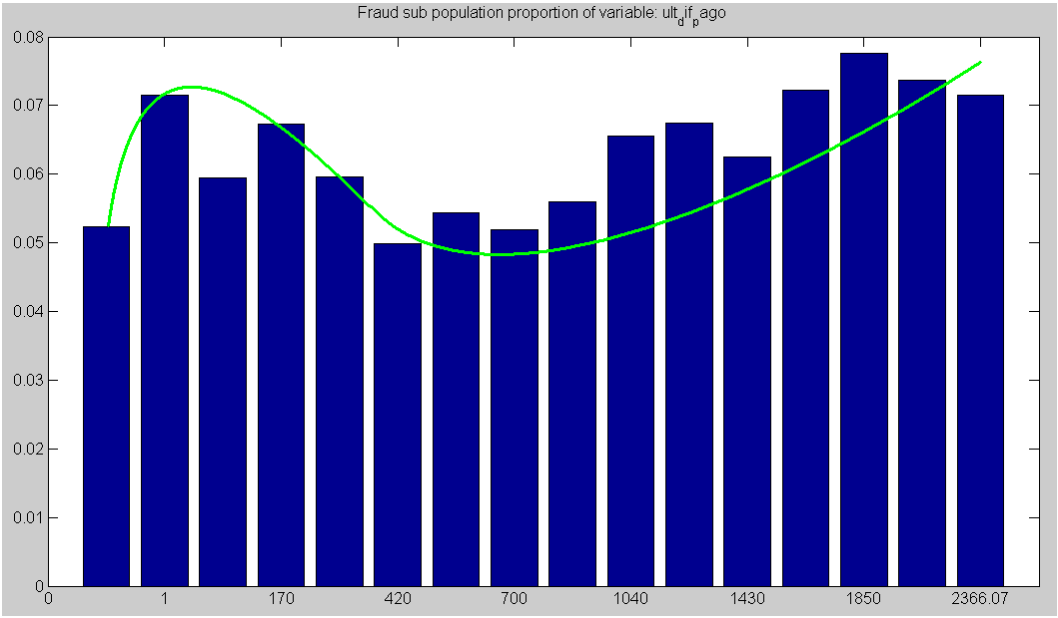


Figure: on the x axis there are the values of ult_dif_pago, on the y axis there is the fraud proportion.

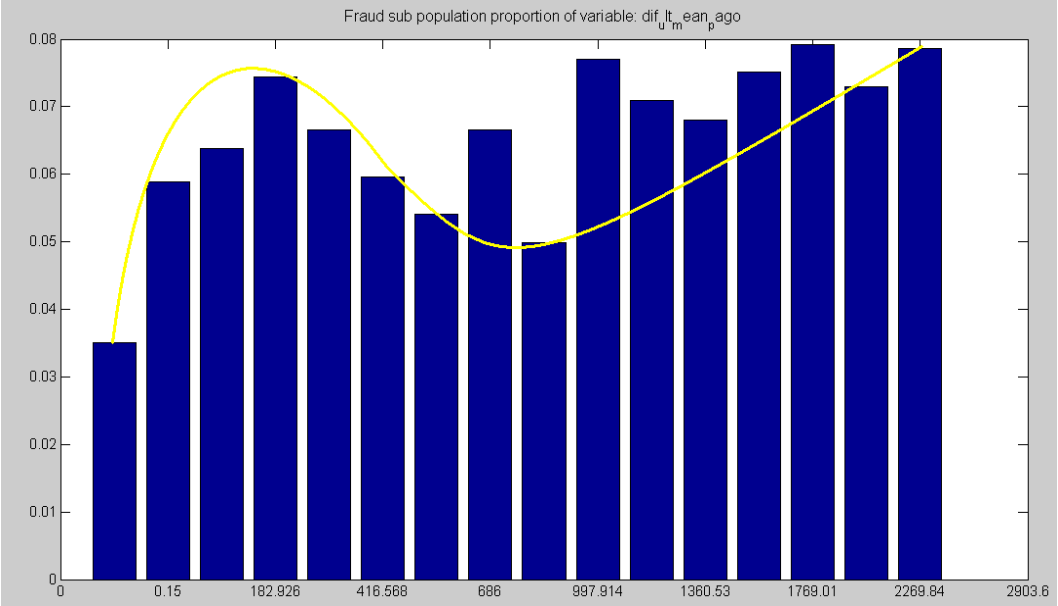


Figure: on the x axis there are the values of dif_ult_mean_pago, on the y axis there is the fraud proportion.

Also the categorical variables give useful information: for instant from the following histogram it can be found that some zones are more fraudulent (San Antonio) and others are more honest (Los Andes). So this information can be used by the model too.

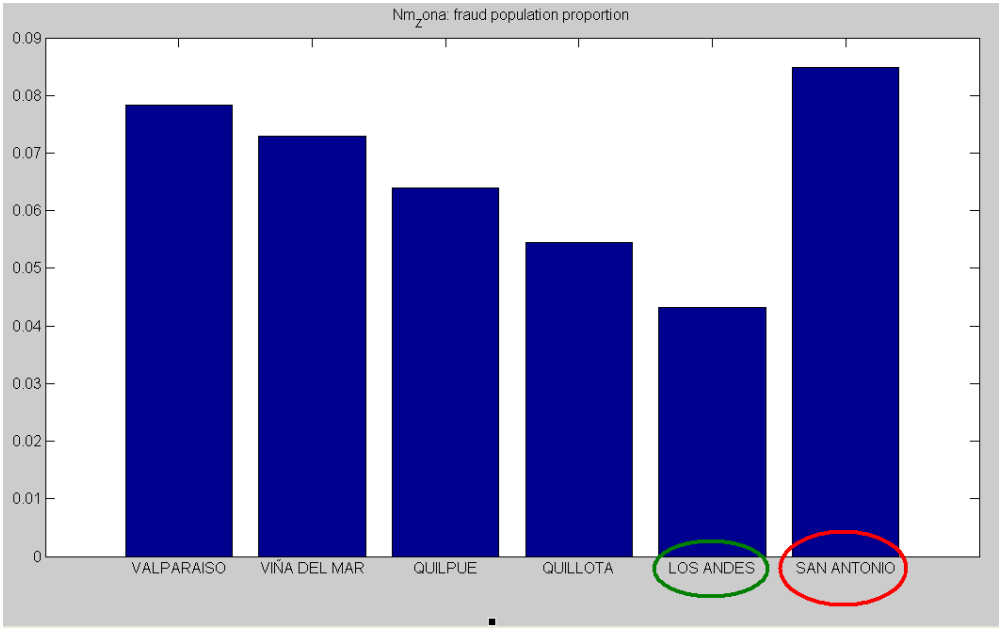


Figure: on the x axis there are the values of Nm_zona, on the y axis there is the fraud proportion.

You can obtain interesting information also comparing different types of histogram.

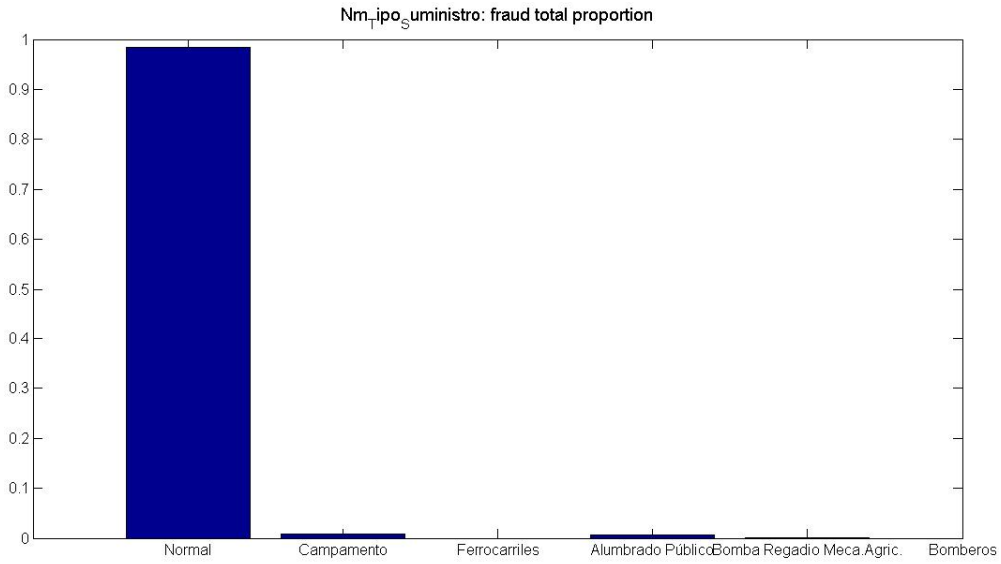


Figure: on the x axis there are the values of Nm_Tipo_Suministro, on the y axis there is the total fraud proportion.

This histogram seems to suggest that people with a normal type of supply are more fraudulent than the ones with other types of supply, but it is not true. It must be noted that the majority of the people has a normal supply, so it is obvious that the majority of

the fraudulent people is in this class; looking at the following histogram it can be noted that the greatest proportion of fraud is in the ‘camping’ class.

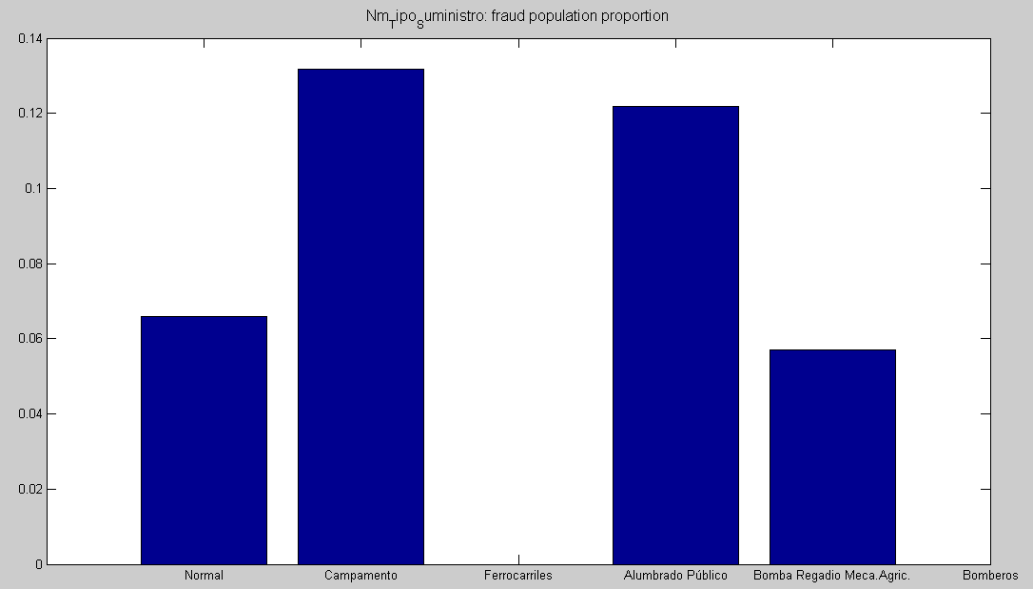
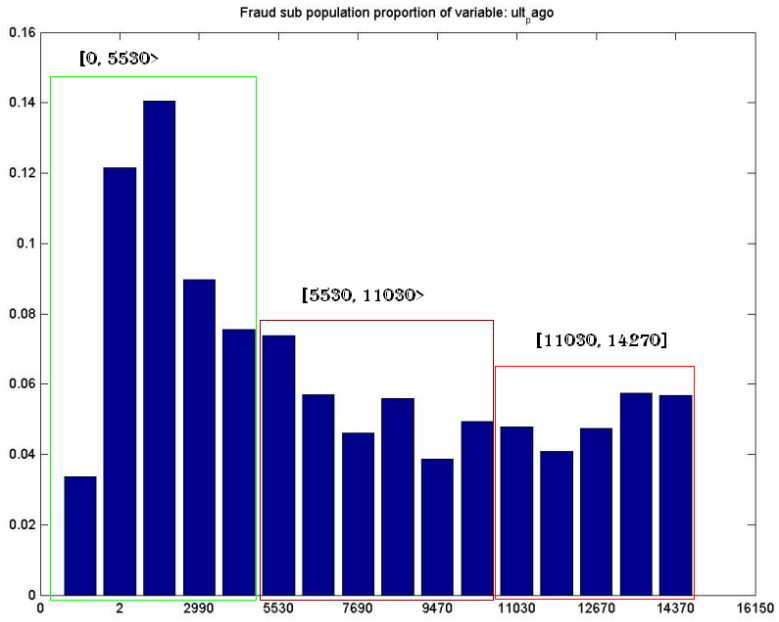


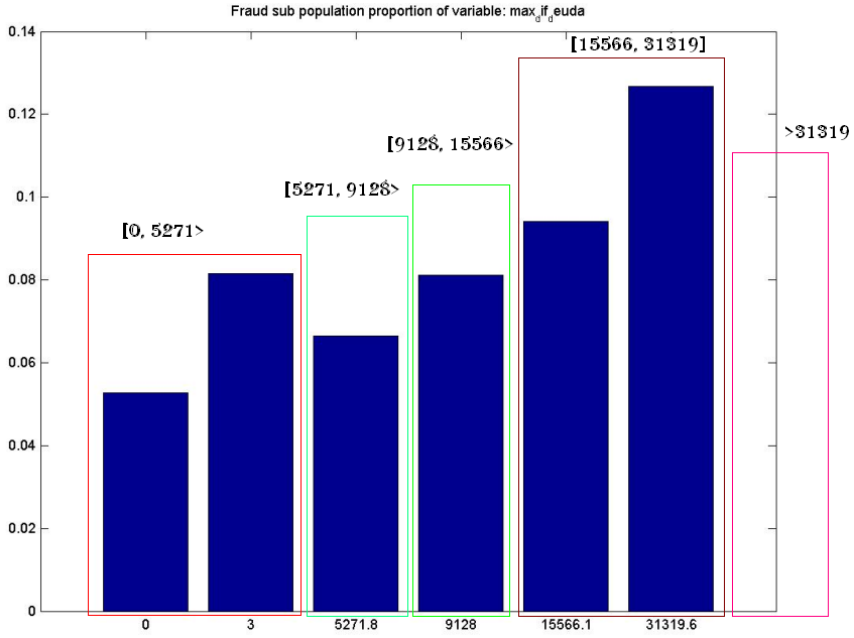
Figure: on the x axis there are the values of Nm_Tipo_Suministro, on the y axis there is the fraud proportion in each type of supply.

The categorization will be as follows:

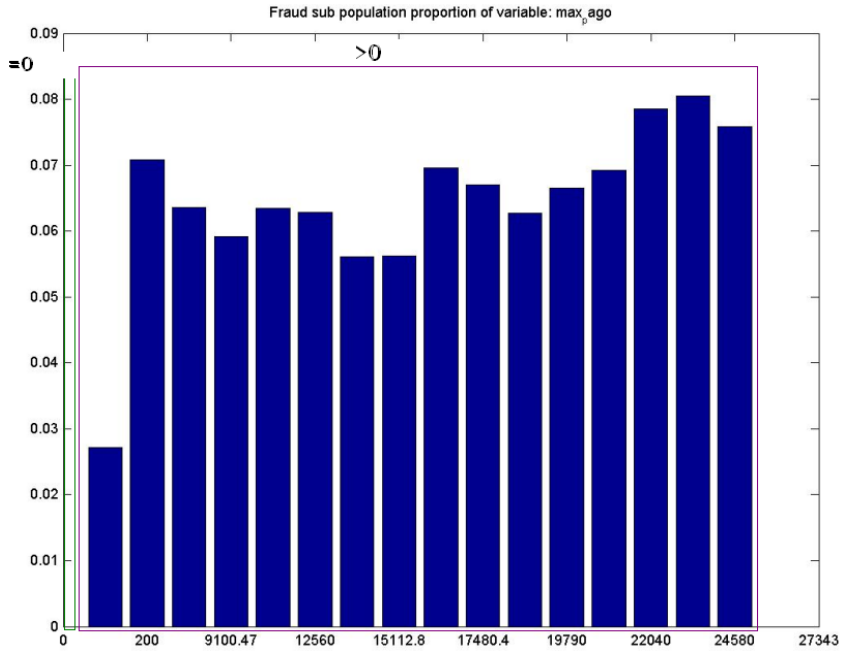
- For *ult_pago* assign variable "A", "B" and "C" to the elements belonging to [0.5530>, [5530.11030> and [11030.14270] respectively.



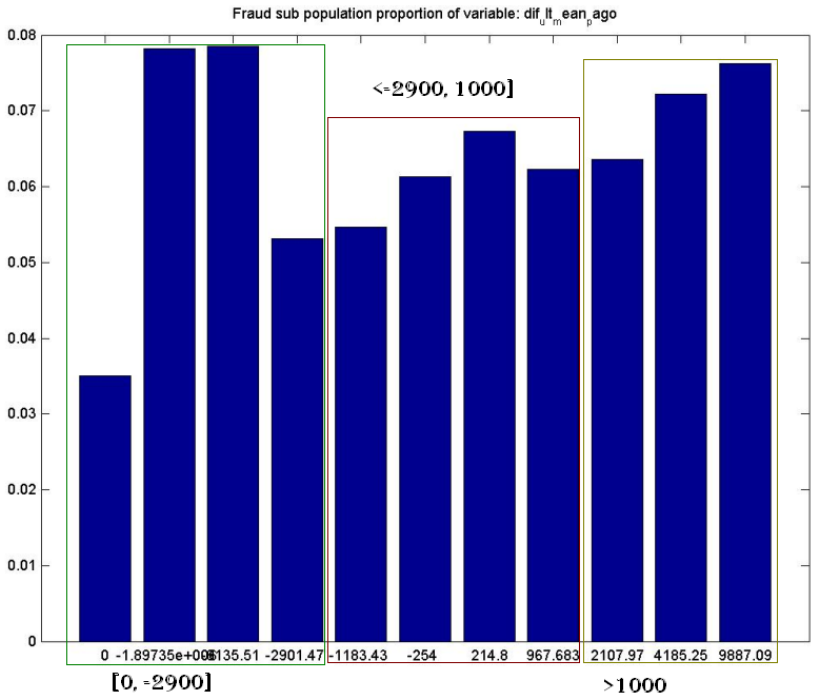
For *max_dif_deuda* assign variable "A", "B", "C", "D" and "E" for items that belong to [00.5271>, [5271.9128>, [9128, 15 566> [15566 31319> and> 31 319 respectively.



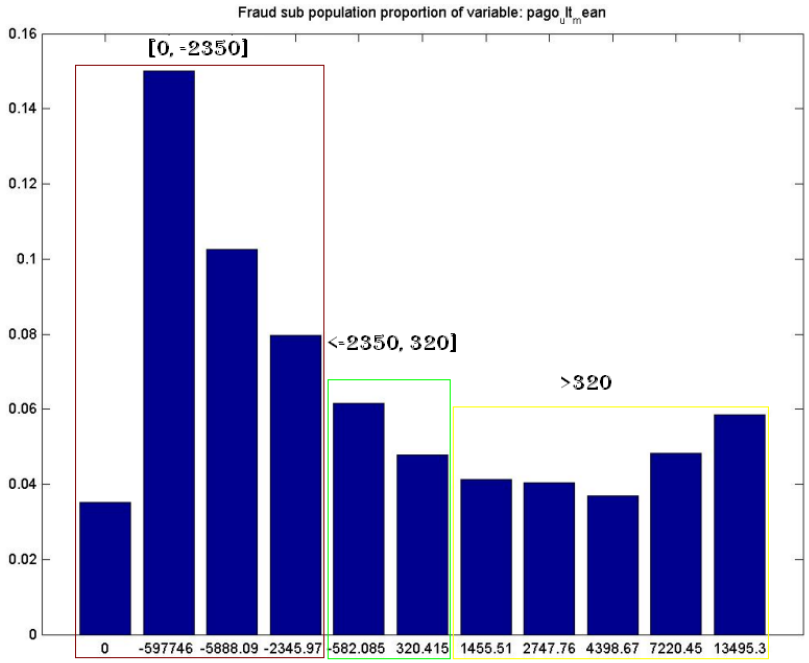
For *max_pago* assign variable "A" and "B" to the elements belonging to (0) and which are greater than zero respectively.



For *dif_ult_mean_pago* assign variable "A", "B" and "C" to the elements belonging to [0.-2900], <-2900, 1000] and over 1000 respectively.



For *pago_ult_mean* assign variable “A”, “B” y “C” to the elements belonging to a [0.-2350], <-2350, 320] and over 320 respectively.



PRINCIPAL COMPONENT ANALYSIS (PCA)

Due to the large number of observations, we need to do a preliminary analysis of the variables to see which of them are more correlated, i.e. we have to determine which variables are redundant.

We have different types of variables, as some are categorical and other quantitative. We are going to separate this analysis into two parts. On the one hand, we analyze the groups that have both categorical and quantitative variables in discrimination techniques, in this case Binary tree. While in groups containing only quantitative variables, we apply principal components to see these correlations.

We start with the principal component analysis for groups of variables "Calculated variables of Debt and Payment".

Remember that the goal of PCA is to reduce the size of the observed variables for each individual, keeping the greater variability. That is, we will reduce the size of the data without losing information of these.

If you look at the eigenvalues, we see that 10 of the 12 variables of this group it grouped 100% of the data. The increased variability of the latter variables is very small so we can eliminate variables.

Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	6.02482787	3.58140216	0.5021	0.5021
2	2.44342571	0.85372651	0.2036	0.7057
3	1.58969920	0.80409449	0.1325	0.8382
4	0.78560471	0.21134933	0.0655	0.9036
5	0.57425538	0.18726560	0.0479	0.9515
6	0.38698978	0.22701377	0.0322	0.9837
7	0.15997601	0.13947855	0.0133	0.9971
8	0.02049746	0.01034643	0.0017	0.9988
9	0.01015104	0.00557818	0.0008	0.9996
10	0.00457285	0.00457285	0.0004	1.0000
11	0.00000000	0.00000000	0.0000	1.0000
12	0.00000000		0.0000	1.0000

Obtain the following correlation matrix for all variables in the group of Debt Calculated variables:

Matriz de correlación

	deuda_max_min	deuda_ult_mean	dif_max_min_deuda	max_deuda	max_dif_deuda
deuda_max_min	1.0000	0.4634	0.7878	0.7400	0.6706
deuda_ult_mean	0.4634	1.0000	0.1389	0.3393	0.4509
dif_max_min_deuda	0.7878	0.1389	1.0000	0.5778	0.5676
max_deuda	0.7400	0.3393	0.5778	1.0000	0.8419
max_dif_deuda	0.6706	0.4509	0.5676	0.8419	1.0000
mean_dif_deuda	0.2923	0.4214	0.0380	0.6447	0.8376
dif_ult_mean_deuda	-.0847	0.4529	-.1376	-.0650	-.0192
mean_deuda	0.4496	0.2008	0.3242	0.9287	0.7521
min_deuda	0.0617	0.0232	0.0409	0.7169	0.5543
min_dif_deuda	-.1588	0.3220	-.5000	0.2518	0.4292
ultima_deuda	0.4410	0.4321	0.2592	0.7309	0.6461
ult_dif_deuda	0.1709	0.4270	-.0179	0.4155	0.5609

Matriz de correlación

	mean_dif_deuda	dif_ult_mean_deuda	mean_deuda	min_deuda	min_dif_deuda	ultima_deuda	ult_dif_deuda
deuda_max_min	0.2923	-.0847	0.4496	0.0617	-.1588	0.4410	0.1709
deuda_ult_mean	0.4214	0.4529	0.2008	0.0232	0.3220	0.4321	0.4270
dif_max_min_deuda	0.0380	-.1376	0.3242	0.0409	-.5000	0.2592	-.0179

	mean_dif_deuda	dif_ult_mean_deuda	mean_deuda	min_deuda	min_dif_deuda	ultima_deuda	ult_dif_deuda
max_deuda	0.6447	-.0650	0.9287	0.7169	0.2518	0.7309	0.4155
max_dif_deuda	0.8376	-.0192	0.7521	0.5543	0.4292	0.6461	0.5609
mean_dif_deuda	1.0000	0.0390	0.7096	0.6538	0.8395	0.6102	0.6894
dif_ult_mean_deuda	0.0390	1.0000	-.0604	-.0087	0.1308	0.0926	0.3394
mean_deuda	0.7096	-.0604	1.0000	0.9123	0.4355	0.7390	0.4610
min_deuda	0.6538	-.0087	0.9123	1.0000	0.5383	0.6275	0.4395
min_dif_deuda	0.8395	0.1308	0.4355	0.5383	1.0000	0.3953	0.6097
ultima_deuda	0.6102	0.0926	0.7390	0.6275	0.3953	1.0000	0.8145
ult_dif_deuda	0.6894	0.3394	0.4610	0.4395	0.6097	0.8145	1.0000

With this first correlation matrix, we can see that there are three variables that are highly correlated: *max_deuda with mean_deuda and mean_deuda with min_deuda*. Of these three variables, we are left with *max_deuda* because it accumulates the most variability.

If we repeat the process with all variables except *mean_deuda min_deuda* and we see that the new correlation matrix is:

Matriz de correlación

	deuda_max_min	deuda_ult_mean	dif_max_min_deuda	max_deuda
deuda_max_min	1.0000	0.4634	0.7878	0.7400
deuda_ult_mean	0.4634	1.0000	0.1389	0.3393
dif_max_min_deuda	0.7878	0.1389	1.0000	0.5778
max_deuda	0.7400	0.3393	0.5778	1.0000
max_dif_deuda	0.6706	0.4509	0.5676	0.8419
mean_dif_deuda	0.2923	0.4214	0.0380	0.6447
dif_ult_mean_deuda	-.0847	0.4529	-.1376	-.0650
min_dif_deuda	-.1588	0.3220	-.5000	0.2518
ultima_deuda	0.4410	0.4321	0.2592	0.7309
ult_dif_deuda	0.1709	0.4270	-.0179	0.4155

Matriz de correlación

	max_ dif_ deuda	mean_ dif_ deuda	dif_ult_ mean_ deuda	min_ dif_ deuda	ultima_ deuda	ult_ dif_ deuda
deuda_max_min	0.6706	0.2923	-.0847	-.1588	0.4410	0.1709
deuda_ult_mean	0.4509	0.4214	0.4529	0.3220	0.4321	0.4270
dif_max_min_deuda	0.5676	0.0380	-.1376	-.5000	0.2592	-.0179
max_deuda	0.8419	0.6447	-.0650	0.2518	0.7309	0.4155
max_dif_deuda	1.0000	0.8376	-.0192	0.4292	0.6461	0.5609

	max_ dif_ deuda	mean_ dif_ deuda	dif_ult_ mean_ deuda	min_ dif_ deuda	ultima_ deuda	ult_ dif_ deuda
mean_dif_deuda	0.8376	1.0000	0.0390	0.8395	0.6102	0.6894
dif_ult_mean_deuda	-.0192	0.0390	1.0000	0.1308	0.0926	0.3394
min_dif_deuda	0.4292	0.8395	0.1308	1.0000	0.3953	0.6097
ultima_deuda	0.6461	0.6102	0.0926	0.3953	1.0000	0.8145
ult_dif_deuda	0.5609	0.6894	0.3394	0.6097	0.8145	1.0000

Now, there is much correlation between *max_dif_deuda* with *mean_dif_deuda*, *mean_dif_deuada* with *min_dif_deauda* and *ult_dif_deuda* with *ultima_deuda*.

It also continues to see the value of the eigenvalues can still be excluded as variables with only 5 of them up in 95% of variability.

Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	4.84643131	2.45507702	0.4846	0.4846
2	2.39135429	1.11414416	0.2391	0.7238
3	1.27721014	0.62112958	0.1277	0.8515
4	0.65608056	0.23243991	0.0656	0.9171
5	0.42364065	0.16954842	0.0424	0.9595
6	0.25409223	0.12676994	0.0254	0.9849
7	0.12732229	0.11174634	0.0127	0.9976
8	0.01557595	0.00728339	0.0016	0.9992
9	0.00829257	0.00829257	0.0008	1.0000
10	0.00000000		0.0000	1.0000

Repeating the process we observe that the variables which we get to keep more information to simplify the model are: *deuda_ult_mean, max_deuda dif_ult_mean_deuda and mean_dif_deuda*.

Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	1.99998829	0.75408512	0.5000	0.5000
2	1.24590316	0.83845789	0.3115	0.8115
3	0.40744527	0.06078200	0.1019	0.9133
4	0.34666328		0.0867	1.0000

Then, we accumulate 90% of the model information with just three variables.

Now, we have to do the same exercise for the group of Payment Calculated variables.

Thus, we see that with 9 of the 13 variables of this group accumulate 100% of the observations information.

Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	6.73214303	3.69288001	0.5179	0.5179
2	3.03926302	1.40907616	0.2338	0.7516
3	1.63018687	0.79233520	0.1254	0.8770
4	0.83785166	0.34525795	0.0645	0.9415
5	0.49259371	0.34376109	0.0379	0.9794
6	0.14883262	0.09789228	0.0114	0.9908
7	0.05094034	0.00959460	0.0039	0.9948
8	0.04134574	0.01450274	0.0032	0.9979
9	0.02684300	0.02684300	0.0021	1.0000
10	0.00000000	0.00000000	0.0000	1.0000
11	0.00000000	0.00000000	0.0000	1.0000
12	0.00000000	0.00000000	0.0000	1.0000
13	0.00000000		0.0000	1.0000

The full correlation matrix is:

Matriz de correlación

	dif_max_min_pago	dif_pago_alto	dif_ult_mean_pago	max_dif_pago	max_pago	mean_dif_pago
dif_max_min_pago	1.0000	-.5899	-.0102	0.9694	0.9242	0.0060
dif_pago_alto	-.5899	1.0000	-.0487	-.5750	-.8063	-.1501
dif_ult_mean_pago	-.0102	-.0487	1.0000	0.0533	0.0625	0.2190
max_dif_pago	0.9694	-.5750	0.0533	1.0000	0.9081	0.1909
max_pago	0.9242	-.8063	0.0625	0.9081	1.0000	0.0821
mean_dif_pago	0.0060	-.1501	0.2190	0.1909	0.0821	1.0000
mean_pago	0.6075	-.9570	0.0476	0.5938	0.8349	0.1474
min_dif_pago	-.9671	0.5673	0.0754	-.8750	-.8811	0.1859
min_pago	0.2203	-.7725	-.0032	0.2257	0.5037	0.2301
pago_max_min	0.9638	-.6949	0.0698	0.9447	0.9761	0.0327
pago_ult_mean	0.0488	-.1597	0.6096	0.1813	0.1731	0.4475
ult_dif_pago	-.0080	-.0805	0.9728	0.0944	0.0771	0.4392
ult_pago	0.4653	-.7811	0.3883	0.5324	0.7010	0.3660

	mean_pago	min_dif_pago	min_pago	pago_max_min	pago_ult_mean	ult_dif_pago	ult_pago
dif_max_min_pago	0.6075	-.9671	0.2203	0.9638	0.0488	-.0080	0.4653
dif_pago_alto	-.9570	0.5673	-.7725	-.6949	-.1597	-.0805	-.7811
dif_ult_mean_pago	0.0476	0.0754	-.0032	0.0698	0.6096	0.9728	0.3883
max_dif_pago	0.5938	-.8750	0.2257	0.9447	0.1813	0.0944	0.5324
max_pago	0.8349	-.8811	0.5037	0.9761	0.1731	0.0771	0.7010
mean_dif_pago	0.1474	0.1859	0.2301	0.0327	0.4475	0.4392	0.3660
mean_pago	1.0000	-.5824	0.8104	0.7169	0.1738	0.0788	0.8202
min_dif_pago	-.5824	1.0000	-.2004	-.9213	0.0916	0.1136	-.3657
min_pago	0.8104	-.2004	1.0000	0.3040	0.1328	0.0517	0.6600
pago_max_min	0.7169	-.9213	0.3040	1.0000	0.1575	0.0720	0.6071
pago_ult_mean	0.1738	0.0916	0.1328	0.1575	1.0000	0.6675	0.7059
ult_dif_pago	0.0788	0.1136	0.0517	0.0720	0.6675	1.0000	0.4445
ult_pago	0.8202	-.3657	0.6600	0.6071	0.7059	0.4445	1.0000

We conclude that we repeat the previous process, the election for this group is: *dif_ult_mean_pago, mean_pago, mean_dif_pago and pago_ult_mean*.

Well, in this case with three of the variables accumulate 90% of the model information.

Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	1.91968708	0.93563411	0.4799	0.4799
2	0.98405296	0.21977021	0.2460	0.7259
3	0.76428275	0.43230554	0.1911	0.9170
4	0.33197721		0.0830	1.0000

Now we analyze the three groups where we have both categorical and quantitative variables.

STUDY OF SIGNIFICANCE

We are going to study the most relevant variables for the model. Therefore, we constructed a *logistic model* in SAS. This model will be explained in detail in following sections.

We are using a **Stepwise selection**: start with a single variable and will include the rest one by one until obtain the equation of the logistic model. Calculate the Chi-square statistic for each variable that is not in the model. If it is significant at the level set (less than 5%), that variable is added to the model.

In this procedure, the variables that have entered into the model may come out.

For example, for the first three categories: *Geographic Variables, Conexion Identifiers, Customer Caracteristic Groups.*

As we see in the p-value column, all these variables are significances.

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Cd_Estado_Cyr		1	1	199.9451		<.0001
2	num_cortes		1	2	107.7762		<.0001
3	Nm_Zona		5	3	115.2781		<.0001
4	CODIGO		1	4	66.8962		<.0001
5	AMPERIOS		1	5	9.1535		0.0025
6	Cd_Sector		1	6	6.7835		0.0092
7	Fc_Ini_Vigencia		1	7	4.9953		0.0254

This is another selection of variables to use later in our final model, as we see all are significant:

Pr > ChiSq	Variable Label
<.0001	num_cortes
<.0001	Nm_Zona
<.0001	CODIGO
0.0031	AMPERIOS
0.0143	Cd_Sector

PEARSON CORRELATION COEFICIENT

Similarly, we also conducted a study on the correlation between variables in order to decide which variables are candidates to be in our model.

The Pearson correlation coefficient is an index that measures the linear relationship between two random variables quantitative.

So we have done a procedure in SAS, **proc corr**, to obtain the Matrix of variance and Covariances. The main diagonal contains the information of the variance.

For example, between **Amperios** and **cd_sector**, there is a very low positive correlation. Between amp and cut the negative correlation that exists is also very low.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Cd_Sector	AMPERIOS	num_cortes
Cd_Sector	1.00000	0.00543	0.00338
Cd_Sector		0.2809	0.5019
	39496	39495	39496
AMPERIOS	0.00543	1.00000	-0.02528
AMPERIOS	0.2809		<.0001
	39495	39495	39495
num_cortes	0.00338	-0.02528	1.00000
num_cortes	0.5019	<.0001	
	39496	39495	39496

For this selection, do not see a significant linear relationship between numeric variables. We conducted this study for several groups of variables.

STUDY OF MULTICOLINEARITY

For example, for a linear regression model, we consider whether any independent variable is a combination of other. This phenomenon is called collinearity.

To detect complex correlations, more than two to two, we performed an analysis of multicollinearity. We add a procedure **proc reg** in SAS, with the selections **tol vif collin.**

- To conduct the study of multicollinearity, we look at the **Variance Inflation Factor (VIF).**
- The VIF represents an increase in the variance due to presence of multicollinearity
- VIF take values from a minimum of 1 when there is no degree of multicollinearity.

$$TOL = \frac{1}{VIF}$$

$$VIF = \frac{1}{1 - R_j^2}$$

- The first thing to do to eliminate variables is to see if they cause multicollinearity. For the variables of *payment and debt group*, we see that the VIF values are high. They are causing multicollinearity.

Parameter Estimates

Label	DF	Tolerance	Variance Inflation
Intercept	1	.	0
deuda_ult_mean	1	0.61949	1.61423
max_deuda	1	0.66545	1.50274
dif_ult_mean_deuda	1	0.75167	1.33038
min_dif_deuda	1	0.84112	1.18889
dif_ult_mean_pago	1	0.59742	1.67387
max_dif_pago	1	0.04780	20.91977
max_pago	B	0.02139	46.75719
mean_dif_pago	1	0.37257	2.68406
mean_pago	B	0.07299	13.70027
min_dif_pago	1	0.06362	15.71864
min_pago	B	0.20056	4.98597
pago_max_min	0	.	.
pago_ult_mean	B	0.41327	2.41973
ult_pago	0	.	.

- If we eliminate these variables now we obtain a model without multicollinearity, with VIF values close to 1.

Parameter Estimates

Label	DF	Tolerance	Variance Inflation
Intercept	1	.	0
max_deuda	1	0.80274	1.24573
dif_ult_mean_deuda	1	0.98296	1.01733
min_dif_deuda	1	0.89327	1.11948
dif_ult_mean_pago	1	0.61272	1.63207
mean_dif_pago	1	0.78983	1.26609
mean_pago	1	0.84671	1.18105
pago_ult_mean	1	0.50962	1.96225

Discriminant Analysis

Another procedure that can be used to obtain significant variables is the **Discriminant Analysis** (**proc discrim** in SAS).

Through this mechanism we have found that the variable that best discriminates the *group of payments and debts* is **Max_deuda.**

BINARY TREE

SAS has the tool **Enterprise Miner** to do Segmentation Trees.

One objective is to obtain a good predictive model that allows, from some independent variables, predict the value of a dependent one.

These models are also decision support models that can be applied in the identification of buyers of a product, fraudsters, risk analysis,...

After these identifications we can make decisions to avoid bad values and enhance good values.

- When the dependent variable is a categorical variable, we used **Classification trees.**

This is the tree that we will use as the dependent variable RESULTADO is categorical.

1 Fraud

0 Non Fraud

In our case, the category of interest is fraudulent.

- When the dependent variable is an interval variable, we use **Regression trees.**

Our Classification Tree has the following characteristics:

- Dependant variable: RESULTADO
- From the list of independent variables, we have provided possible candidates for inclusion in our model.

  For this selection we have taken into account the results obtained by other decision-making mechanisms seen before: principal component analysis, contrasts of significance, discriminants procedures...
  In the final analysis of the tree we have included these variables:

target	Resultado
Geographic variables	Cd_Sector
	Nm_zona
Conexion identifiers	CODIGO
	AMPERIOS
Customers caracteristic	num_cortes
Calculated variables of debt	max_deuda
Calculated variables of payments (categorize non-linear continuous)	cat_max_pago
	cat_pago_ult_mean
	cat_dif_ult_mean_pago
	cat_ult_pago

- As the percentage of fraudulent population is very low (approximately 6%), we use a Profit Matrix to getting a tree that gives more weight to this population.

  This can help us to identify which are the variables that best explain the fraudulent population under study.
- We are using a Training sample of 80% and Validate of 20%, with simple random.

  The training dataset is used to estimate model parameters or to obtain a decision tree.

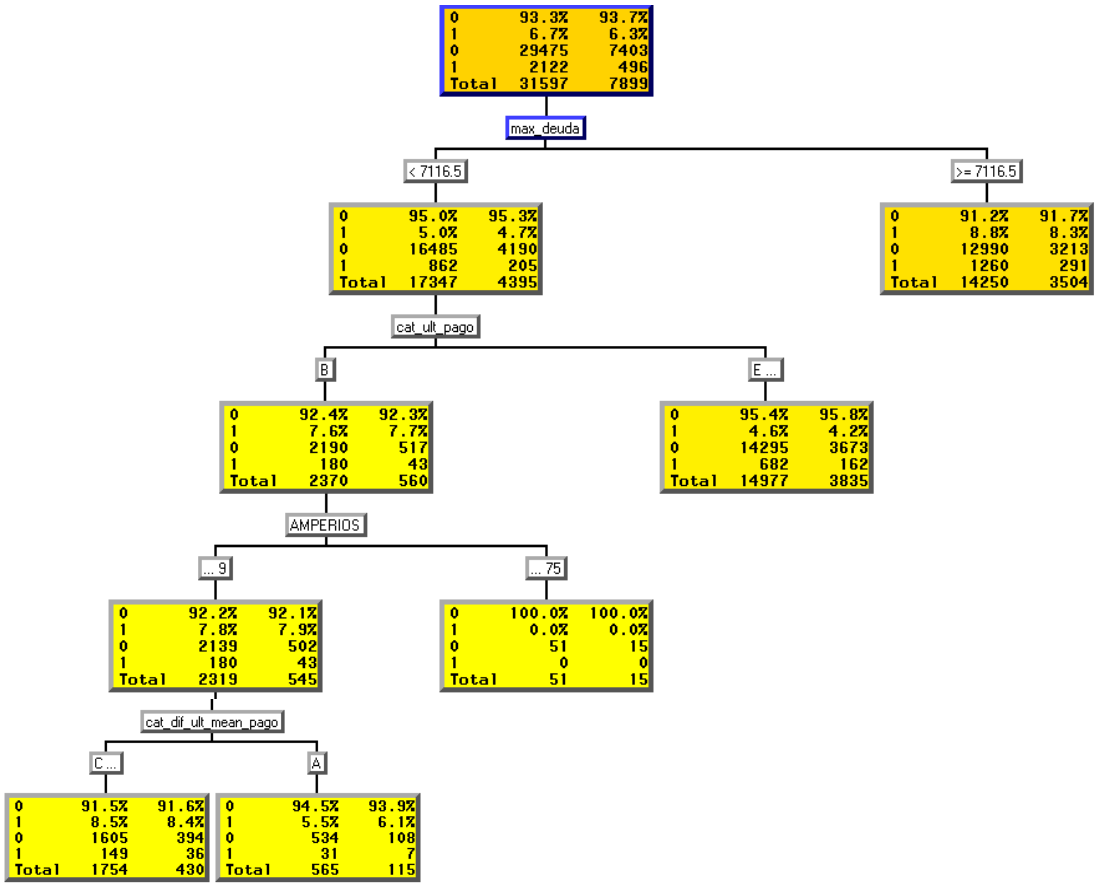
The validation dataset is used to validate the model or the decision tree and obtain a subtree of the original tree with the highest quality for independent data sets.

- **Splitting Criterion:** In our case we use the **Criterion of Entropy Reduction** based on reducing the uncertainty of the classification variable in the leaf of the tree created at each step.

- Tree Depth: 6

  The recursive process ends in a node if the depth of the node is equal to the value of this parameter, since in general, the trees too long lose their interpretability.

- SAS provides the following Classification tree:



With the characteristics given to our tree, we note that the independent variable that best classifies is:

max_deuda → cat_ult_pago →AMPERIOS → cat_ult_mean_pago

For examples, the first classification with **max_deuda,** allows us to classify individuals into two classes:

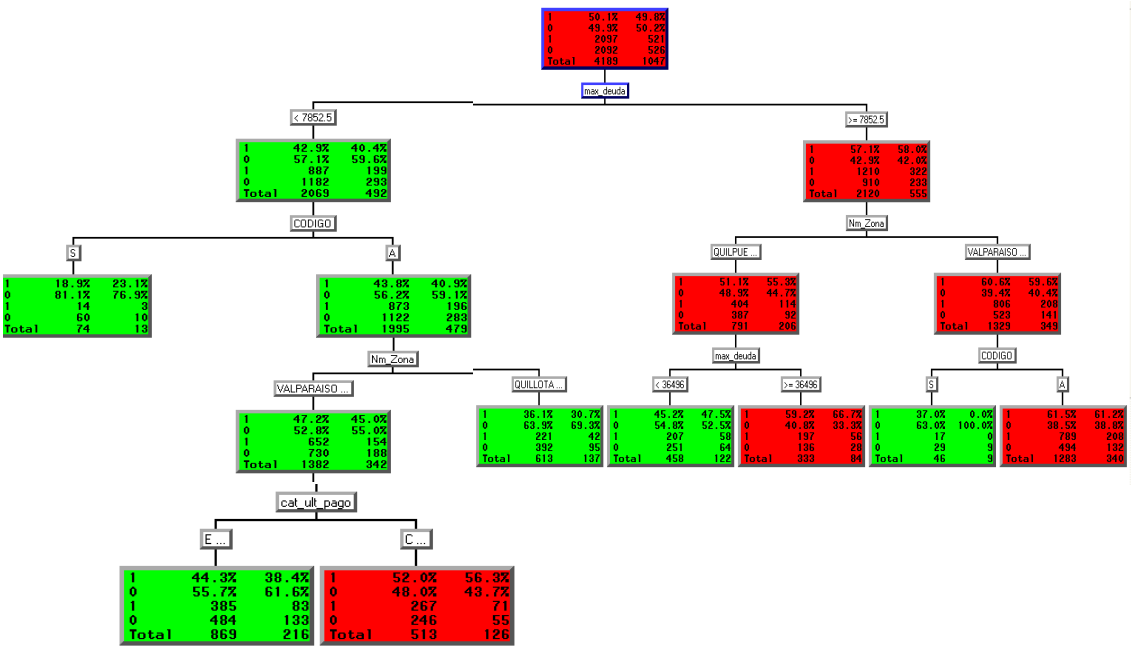
- Those with a maximum debt less than 71,165
- Those with a maximum debt greater than or equal to 71,165

" 6,7 % are fraud people in the training sample and 6,3 % in the validation sample" .

❖ Since the fraudulent population is very small, we can build a tree with a much smaller random sample. This sample contains 50% of the fraudulent population and 50% of non-fraudulent.

We can see that, from a selection of variables in each category, the one that best discriminates is **max_deuda**.

Therefore, this information can also help us to decide that this variable should be included in our final model.



Green nodes: fraud; Red Nodes: non-fraud.

MATHEMATICAL MODELING

We now turn to the problem of building a model to predict if a given consumer is fraudulent or not. To that end, we have a set of data detailing consumers characteristics, with the result of the inspection performed. The consumer data represents anything that we thought was useful. For instance, interesting variables include consumer localisation, history of payment, type of line, etc.

This type of problem fits very naturally into the framework of regression, where we want to deduce a relationship between target variables (here, fraud probability) and explanatory variables (here, consumer data). Regression has many purposes and has been used for a variety of applications. For instance, a closely related problem is spam detection: given certain characteristics of an e-mail (sender address, number of recipients, wording of the title, length of the message...), what is the probability that the message is spam, that is, unwanted mail?

In what follows, we will abstract our particular problem by calling Y the target variable (fraud probability), and X_i the explanatory variables (consumer characteristics). The X_i can be continuous or binary. For instance, the last payment of the consumer is a continuous variable, while indication that his line is underground or not is a binary variables. For categorical variables (for instance, the geographical zone of residence), we will split the possibilities by using a binary encoding of the variable, reducing a categorical variable with N possible states to N binary variables.

REGRESSION

- Linear regression:

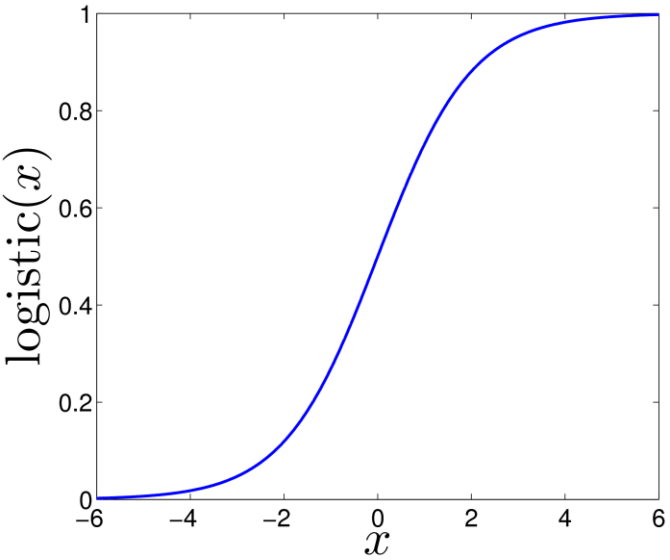
The very simplest model of regression is the linear regression. In this approach, one postulates a linear relationship between Y and X_i.
Calling $\mathbf{X}$ the vector of X_i, we get:

$$Y = a + \mathbf{b}^T\mathbf{X}$$

where a and $\mathbf{b}$ are model parameters. Given a value for the parameters a and $\mathbf{b}$, we can now predict the probability Y given the consumer data $\mathbf{X}$. We will try to obtain good values of a and $\mathbf{b}$ by training our model first on a sample data set, as we will see in the next section.

- Logistic regression

But we notice here a problem. In our application, Y represents a probability and must respect the constraint $0 \leq Y \leq 1$. However, $a + \mathbf{b}^T\mathbf{X}$ is unbounded: we cannot ensure that the constraint will be respected other than by constraining the input. But that would be non-physical, and would introduce more complexity in our model.



Rather than constraining our input, a simple solution is to replace this linear regression by an alternate model. A simple mapping is provided by the logistic curve, defined by the equation

$$\mathrm{logistic}(x) = \frac{1}{1+e^{-x}}$$

This curve provides a smooth mapping from $(-\infty, +\infty)$ to $(0, 1)$. We can now write our new model, the logistic regression:

$$Y = \mathrm{logistic}(a + \mathbf{b}^T\mathbf{X}),$$

or, equivalently,

$$\mathrm{logistic}^{-1}(Y) = a + \mathbf{b}^T\mathbf{X}.$$

This form is preferred because it is linear in the parameters, which leads to simplifications later on.

Finally, the linear form $a + \mathbf{b}^T\mathbf{X}$, while sufficient for many applications, has a number of shortcomings. In particular, it ignores any interactions between explanatory variables, who might be significant. For instance, imagining that frauds are more prevalent among rich people who live in a specific region, this correlation would not be able to be represented in our model. To account for such correlations, we can use higher-order models. For instance, including only correlations of the form $X_i X_j$ leads to

$$\text{logistic}^{-1}(Y) = a + \mathbf{b}^T\mathbf{X} + \mathbf{X}^T A\mathbf{X}$$

where A is a parameter matrix, symmetric with zero diagonal.

Other higher-order models (for instance, including terms of the form $X_i X_j X_k$, or nonlinear effects such that X_i^2, $\log(X_i)$, etc.) are possible, but the computation time increases with the complexity of the model.

TRAINING

Once we have chosen our model, our task is to find good values of the parameters. To that end, the model is fitted against a training sample, for which we know the results of the controls made by the company. This is achieved by defining an objective function, representing the error in the fit, that depends on the parameters and the training data. This function is then minimised or maximised with respect to the parameters.

In the usual approach for the linear regression problem, the objective function is defined to be the 2-norm between the measured values and the predicted values. The model that minimises this 2-norm is then said to fit the data in the least-squares sense. The minimisation can then be performed by solving the so-called normal equations, a linear system in the parameters.

Solving the system in the least squares sense is statistically justified by assuming normality of the underlying variables. In our case though, a better result is achieved by using the likelihood as our objective function, which is to be maximised. This gives

rises to the maximum likelihood parameters, which are found by optimising a nonlinear system of equations.

\section{Evaluation} Once the training is done and the parameters are determined by fitting to training data, we ask the question of the evaluation of our model. Is it actually useful, or does it just produces random results? To answer this question, we use another data set, the validation set. It is important that this validation set be distinct of the training set: if not, we could encounter the phenomenon of overfitting, whereby a model is very good for a limited set of data, but actually represents the particular data more than overall trends in it, and gives poor results on data it has not trained for. In practice, we split a dataset given to us by Neometrics into two training and validation sets of equal size.

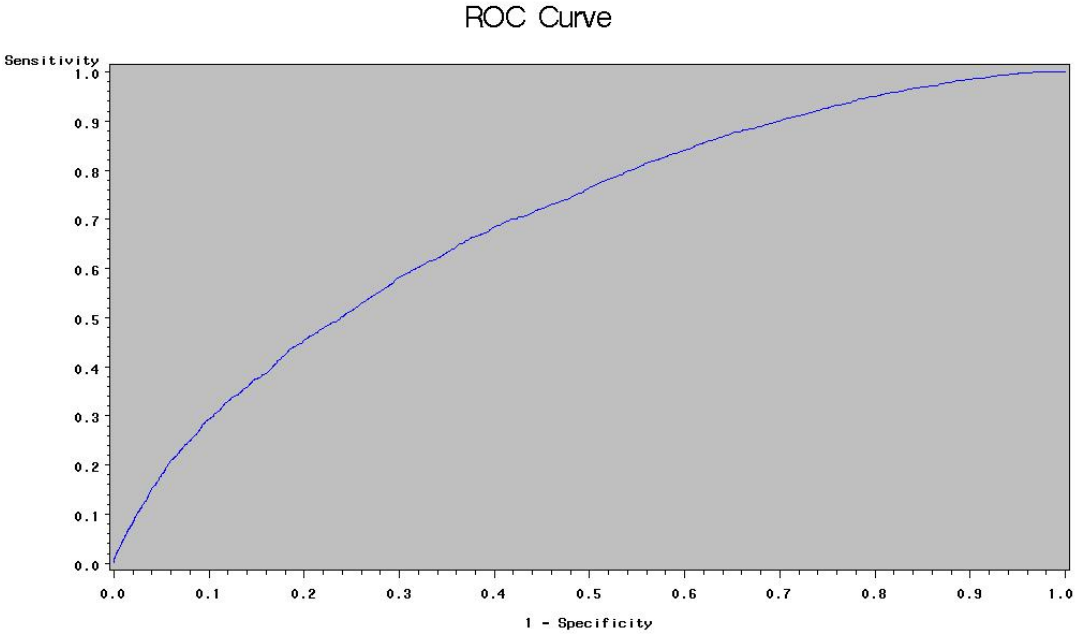
ROC CURVE

To analyse the results of our model, we will use the ROC curve. This concept is used to visualise the output of a classification algorithm. Using our model, we are able to predict the probability that each consumers commits a fraud, and then, using a threshold, we separate them into a frauding and a non-frauding group. Those groups are compared to the actual groups. Increasing the threshold will increase the false positive rate, but also increase the true positive rate. A ROC curve is a measure of this tradeoff.

We define the false positive rate to be the proportion of predicted fraudsters that were actually non-fraudsters, and the true positive rate to be the proportion of fraudsters to be predicted as such. We then plot the true positive rate against the false positive rate. Obviously, the goal of a model is to achieve a good true positive rate while keeping the false positive rate small. This means that the ROC curve should be as high as possible. A random model would give a straight line, with false positive rate equal to true positive rate.

To measure the quality of our model, a convenient quantity is the c-value, that is, the area under the ROC curve. This quantity will typically vary between 0.5 (random model) and 1 (perfect model).

For instance, here is a ROC curve generated with our final model.



LIFT CHART

Lift chart is a graph to measure a predictive model calculated as the ratio between the results obtained with and without the predictive model.

To construct this chart, customers should be order on the X-axis in descending order according to the fraud score, which the model has thrown out.

We sort the client according to their decreasing fraud probabilities.
The X-axis represents the percentage of the population according to the previous arrangement.
The Y-axis represents a rate calculated as:

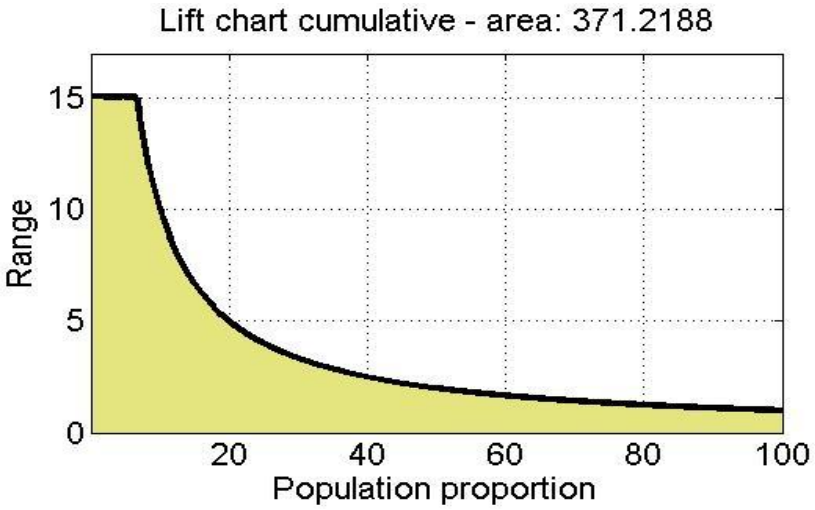
$$\text{Rate}(\alpha) = \frac{\text{Number of fraud clients in the modelint he first } \alpha\% \text{ of the population}}{(\text{Number of fraudulent clients/ population size}) * \alpha\% \text{ of the population}}$$

Recall that RESULTADO = 1 is a fraudulent person while a person classified as a RESULTADO = 0 is a fraudulent person.

If we represent, for example the graph % Response the ordinates represent the percentage of individuals of the kind RESULTADO = 1 on the subset of individuals with the percentage of probability of prediction for this senior class.

For example, with 20% of the sample must be really in a 5% of these individuals are Class RESULTADO = 1.

The usefulness of these charts, you can be to establish a cut-off point in predicting the likelihood of a sample independent predictor of individuals about their class is unknown. That is, once estimated the probability of belonging to the class RESULT = 1, the cut-off point will indicate whether that person should be or not to perform an inspection also.

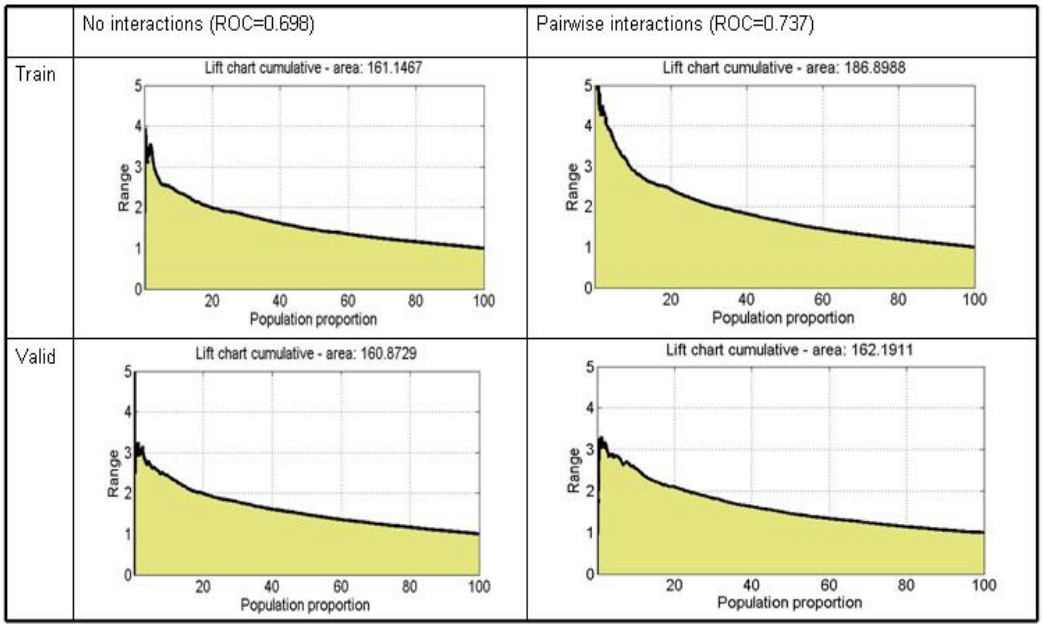


NUMERICAL VALIDATION

We present only two models the best model.

- One with interaction.
- The other without interaction.

The results are this:



The best model that we have found is using interaction. We know that because the ROC value is higher than without interactions. In the model without interactions, the ROC value is the 0.737 as long as the value with interaction is the 0.698.

The result that we have obtained in the train sample is consistent with the validate sample if you see the lift chart cumulative-area.

We observe that we lost more information in the model with interactions than without interactions.

OPTIMAL CONTROL CAMPAIGN

We want to find the optimal percent of inspections to carry out in order to maximize the return of invesment (ROI) following the score results from our model.

If fraud is detected the gross saving would be the mean income obtained by the company per customer per year (we use the variable mean_pago*12). For each inspection there will be a fixed cost of 15,000 MU. A fixed cost of 100,000,000 is required for the maintenance of the inspection crew, regardless of the number of inspections.

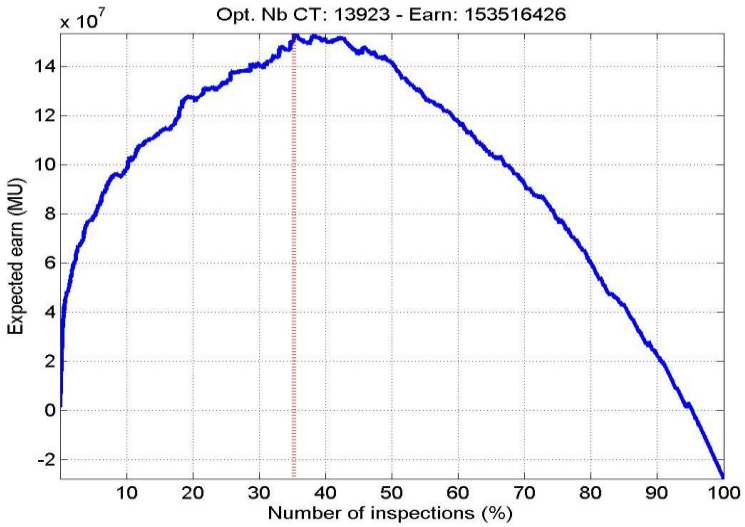
We used Matlab to calculate for each P= 0, 1, 2,...., 99, 100:

1.- The percentage of fraud (F) detected by checking a percentage of the population (P), starting with the ones with higher risk of fraud according to our logistic regression model. It can be measured by the area below the lift curve from 0 to P.

2.- Benefits and costs of checking P% of the population. N=size of the sample:
Benefits: F *N* 12 * mean_pago(fraudulentos)/100
Costs: fixed_crew_cost + fixed_inspection_cost * P*N/100

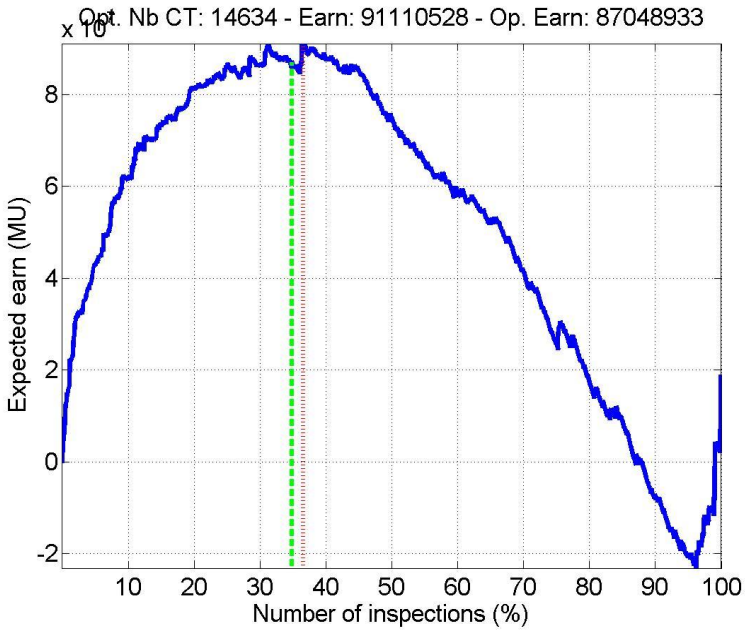
ROI (P%) = { Benefits(checking P%) - Cost(checking P%) }

The following graph shows the return of investment depending on the percentage of the population checked (in order, from higher to lower fraud risk, using the score given by the logistic regresion model with interaction) using the train data:



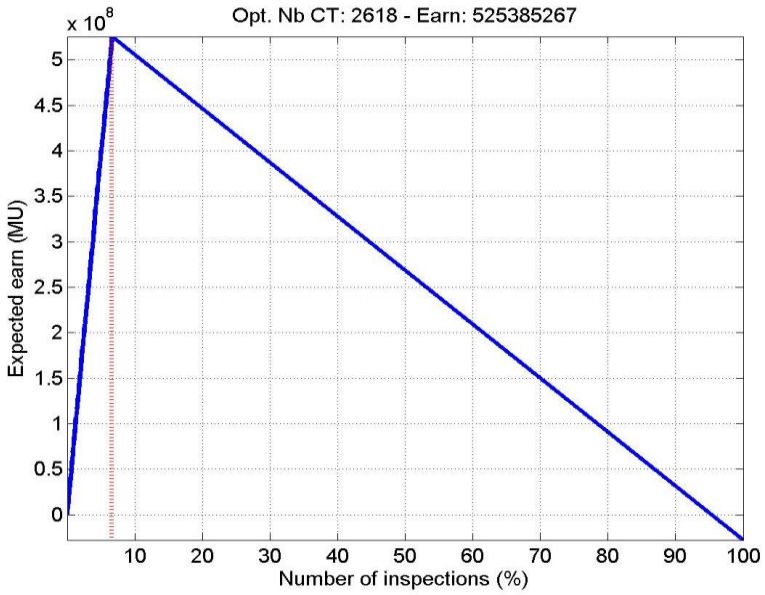
The optimal is to check 35% of the sample obtaining a ROI of 150 million MU.

The following graph shows the ROI using the validation data:

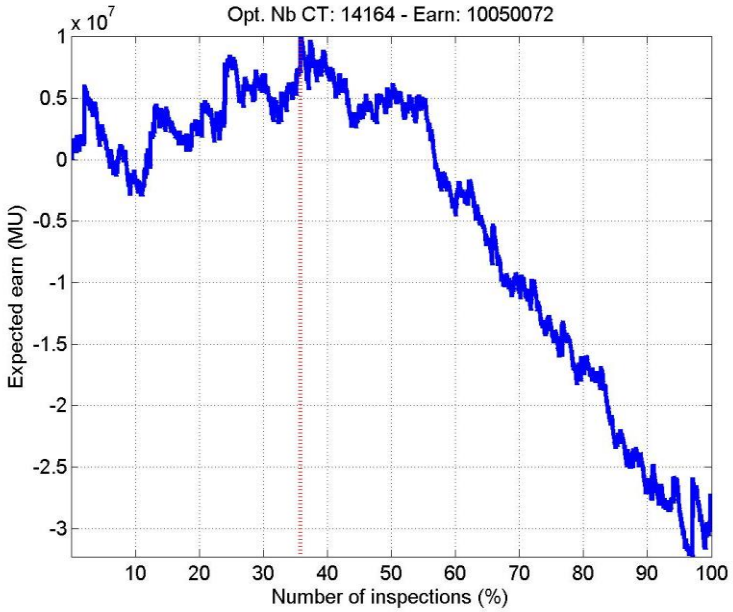


The green line shows the ROI (87 million MU) of inspecting the percent of the population given by the train data (i.e. 35%). The red one is the one that maximizes the ROI (91 million MU) of validation data.

To compare these results we will consider now the best possible model. That will be the one that captures 100% of fraudulent customers checking 6.6% of the population. That means that all fraudulent customers are perfectly defined by this (utopical) model.
The following graph shows the ROI of such a model if it existed would be 525 million MU.



The last graph we show represents the ROI with no model (i.e. checking on customer randomly). Having 35% of the population checked represents a ROI of 10 million MU.



SUMMARY

- Data analysis to isolate interesting variables
- Logistic regression to predict fraud probabilities
- Evaluation of the model (ROC, lift)
- Use in a cost-benefit analysis
- Concrete results of use to the client.

FUTURE WORK

We have considered marking some research threads to take into account:

- A deeper revising for selecting variables without loosing the greatest information. This study could be done through discriminate analysis, binary trees, principal component analysis, etc.

- Neural networks. To train some neurons basing on the training sample and validate with the valid sample.

- As a good point, we should try to add our model as a parameter with the aim of getting an optimal gain figure.