Paper 069-2013

Predicting Women's Department Purchases in a Retail Store by using the SEMMA Methodology

Michael Soto Fuentes, Ripley Chile.

ABSTRACT

As one of the main retail companies in Chile, we are currently facing interesting challenges. Nowadays customers have so many choices than ever to select what, where, how much and how to purchase. For this reason, it is crucial to leverage what technology has to offer in terms of analytics in order to improve value of average tickets and avoid inactivity of customers. For example, it is possible to know with a high degree of certainty which products offer to customers based on buying patterns to improve the cross-selling process. It is made through propensity predictive models, a set of mathematical and statistical techniques, that allows to know what customers can buy in function of historical data.

One of our focus to improve the business is the Women Department because it is currently our most powerful department in terms of transactions originated by customers. It raises the need to implement an analytical model focused on this department for establishing what offer is the most appropriate for our customers according to buying patterns of customers, augmenting the likelihood that a customer comes back to the stores. Those patterns are calculated based on demographic, transactional data and any other interaction that our customers have had with our stores. The predictive model we used is the logistic regression and it was executed following the SEMMA methodology considered by SAS ® for projects in SAS Enterprise Miner ®.

This work is presented according to the SEMMA methodology and contains the following sections. The description of data to use or samples are expressed in the Sample section.

The second section is titled Exploration. It contains all the details of descriptive analysis carried out to formulate the predictive model and to select the input variables.

Variable transformations needed for good model implementations are described in the third section called Modification.

The Modeling section establishes some relevant aspects and preliminary results of model estimation and steps to improve that model.

Finally, the validation section shows results of the final model, measured in a validation sample. In this section a number of criteria to evaluate the model and the final solution are also detailed.

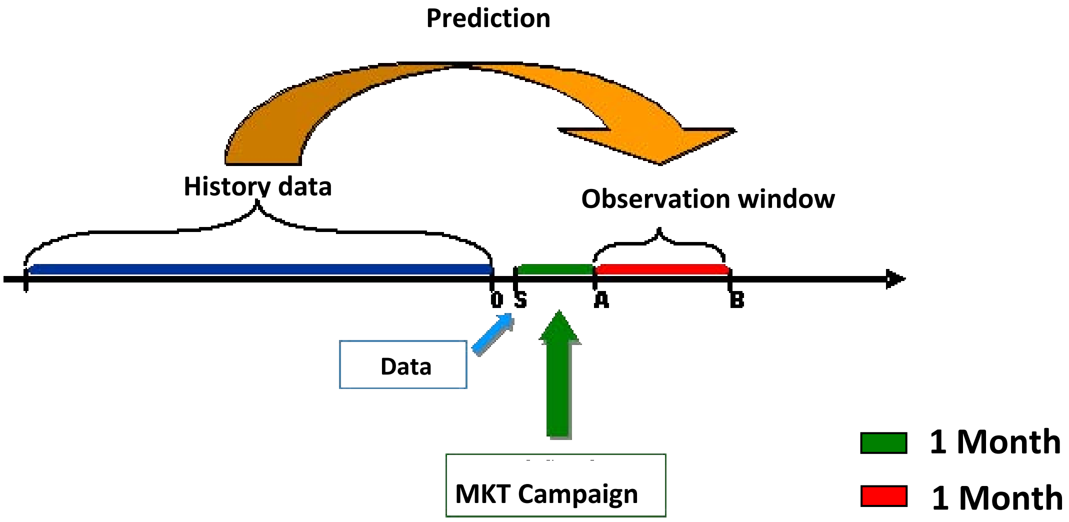
1. SAMPLE

The first step of the SEMMA methodology corresponds to the selection of the data samples to be used in building and in the subsequent propensity model validation.

The sample must contain enough data to be significant, yet an appropriate size to ensure a good processing performance.

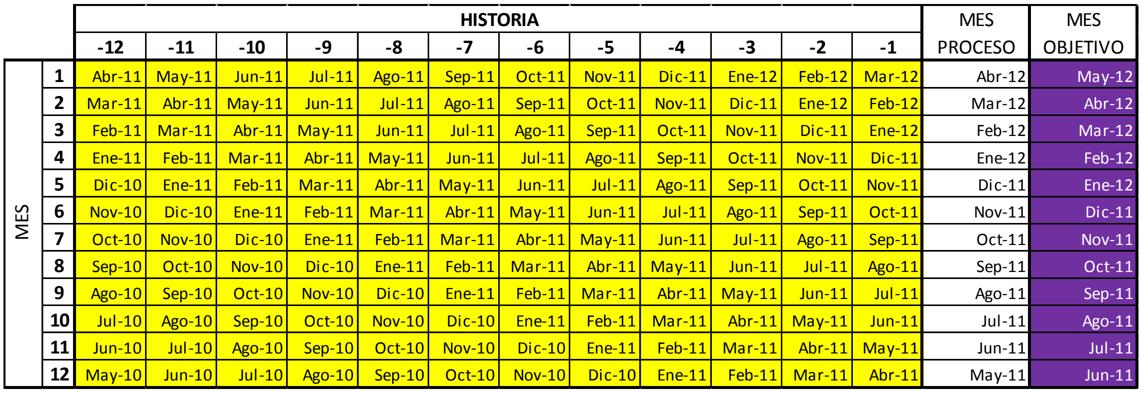
For the construction of the sample we rely on the business needs. The marketing area needs to know a month in advance what the behavior of their customers will be to manage their actions.

We selected 12 sets of data with the following structure



For each observation window we used a simple random sampling of 20,000 customers with purchases in the Women department and 20,000 customers showing no purchases in the same department.

The final data set is as follows:



		HISTORIA												MES	MES
		-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	PROCESO	OBJETIVO
MES	1	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11	Ene-12	Feb-12	Mar-12	Abr-12	May-12
	2	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11	Ene-12	Feb-12	Mar-12	Abr-12
	3	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11	Ene-12	Feb-12	Mar-12
	4	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11	Ene-12	Feb-12
	5	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11	Ene-12
	6	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11	Dic-11
	7	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11	Nov-11
	8	Sep-10	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11	Oct-11
	9	Ago-10	Sep-10	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11	Sep-11
	10	Jul-10	Ago-10	Sep-10	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11	Ago-11
	11	Jun-10	Jul-10	Ago-10	Sep-10	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11	Jul-11
	12	May-10	Jun-10	Jul-10	Ago-10	Sep-10	Oct-10	Nov-10	Dic-10	Ene-11	Feb-11	Mar-11	Abr-11	May-11	Jun-11

The final table contains a total of 480,000 observations. Each observation has a total of 200 variables that describe past behavior and customer profile.

The response variable is whether the individual purchases products within the women department in the month of observation.

To estimate the model, we define a training sample consisting of 336,000 observations (70%) obtained as a simple random sample from the initial table.

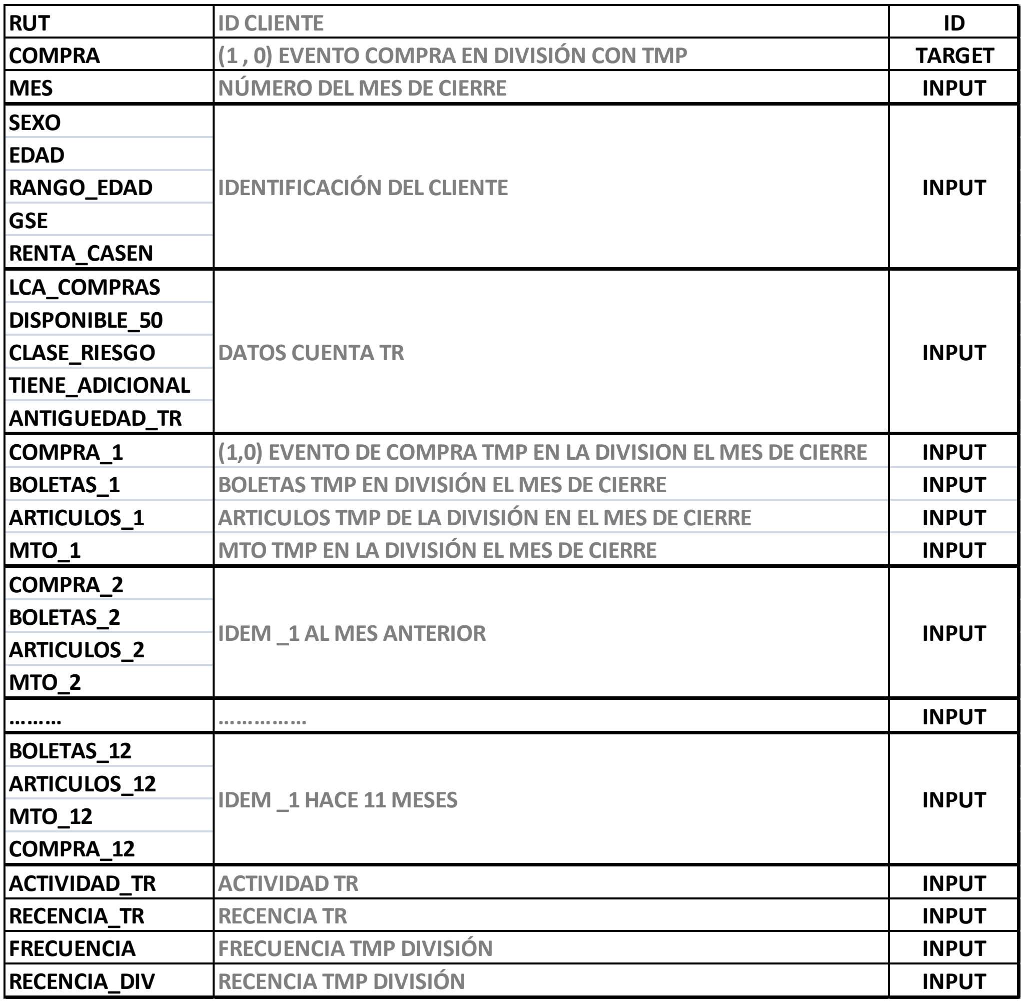
To validate the model we define a validation sample, which corresponds to the remaining 30% observations (144,000).

2. EXPLORE

Considering the 200 variables, we proceed to classify the variables according to their relevance in the event we are trying to predict.

Specifically, we consider all those variables that provide information or relate in any way to the customer's purchasing behavior in that department.

After this process we selected a total of 48 variables.



RUT	ID CLIENTE	ID
COMPRA	(1 , 0) EVENTO COMPRA EN DIVISIÓN CON TMP	TARGET
MES	NÚMERO DEL MES DE CIERRE	INPUT
SEXO	IDENTIFICACIÓN DEL CLIENTE	INPUT
EDAD		
RANGO_EDAD		
GSE		
RENTA_CASEN		
LCA_COMPRAS	DATOS CUENTA TR	INPUT
DISPONIBLE_50		
CLASE_RIESGO		
TIENE_ADICIONAL		
ANTIGUEDAD_TR		
COMPRA_1	(1,0) EVENTO DE COMPRA TMP EN LA DIVISION EL MES DE CIERRE	INPUT
BOLETAS_1	BOLETAS TMP EN DIVISIÓN EL MES DE CIERRE	INPUT
ARTICULOS_1	ARTICULOS TMP DE LA DIVISIÓN EN EL MES DE CIERRE	INPUT
MTO_1	MTO TMP EN LA DIVISIÓN EL MES DE CIERRE	INPUT
COMPRA_2	IDEM _1 AL MES ANTERIOR	INPUT
BOLETAS_2		
ARTICULOS_2		
MTO_2		
………	……………	INPUT
BOLETAS_12	IDEM _1 HACE 11 MESES	INPUT
ARTICULOS_12		
MTO_12		
COMPRA_12		
ACTIVIDAD_TR	ACTIVIDAD TR	INPUT
RECENCIA_TR	RECENCIA TR	INPUT
FRECUENCIA	FRECUENCIA TMP DIVISIÓN	INPUT
RECENCIA_DIV	RECENCIA TMP DIVISIÓN	INPUT

The next step is to carry out the exploration of the variables that have been selected according to the objectives of the model and its formulation. That is, redundant information is eliminated by creating new summary variables.

The summary variables, can be classified into two groups. The first group corresponds to those created through factor analysis. The second group, are the variables created from the originals with apparent relation to the purchasing event.

According to this exploration we decided the following actions:

- Delete outliers and replace missing values.
- Recategorize variables
- Categorize quantitative variables in order to get better results

The above activities, which define the variables to model are achieved mainly by using two statistical techniques, decision trees and factor analysis, which are listed below in the context of the problem.

2.1 Decision Trees

This technique, available in SAS Enterprise Miner is used to reclassify categorical variables or categorize quantitative variables. Specifically, the CHAID algorithm is used, which determines the optimal categories that maximize the value for the Chi-square test of association of the relationship between the variable subject to categorize and a response variable (Qualitative), maintaining an smooth distribution among the categories

2.2 Factor Analysis

In order to reduce the number of variables that enter in the final model and avoid problems of collinearity between covariates, factor analysis is used to extract principal components. This technique allows to capture in a smaller number of new variables, linear combinations of them, the information contained in many of them, without a significant loss of information for modeling porpuses.

3. MODIFY

The transformation of variables, can be classified into two main points: the creation of new ones and transformations.

3.1 Creating New Variables

As mentioned, it creates a series of variables that should beimportant factors in the purchase behavior in the woman department.

The new variables are created according to events such as: important dates in the observation month (birthdays, holidays, etc), change in purchasing behavior, current debt level, seasonal store activity, etc.

3.2 Variable Transformation

According to the criteria mentioned in the previous point, there will be transformations of the original variables and the creation of new ones.

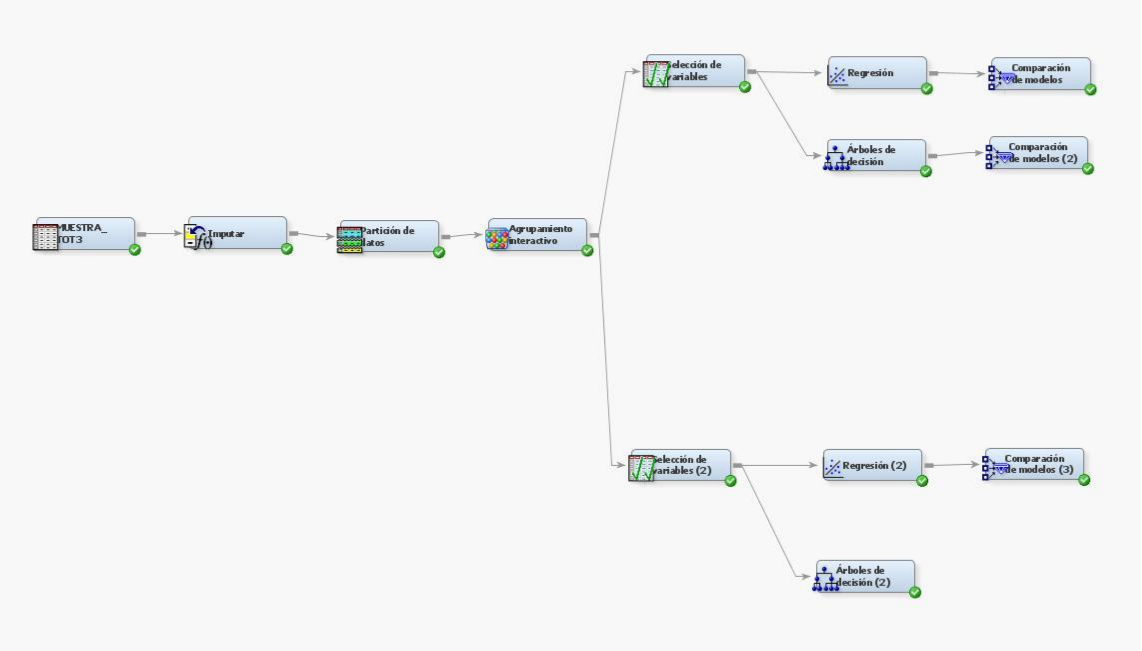
The following nodes available en SAS Enterprise Miner were used:

3.2.1 Interactive grouping

Groups variable values into classes that can be used as inputs for predictive modeling.

3.2.2 Variable Selection

Evaluates the importance of the input variables in target prediction. To select the input variables, two criteria were used: Chi-squared and R-squared. The R-squared criteria can be used to eliminate variables establishing some order of importance. Other input variables that can be eliminated at this point are the ones with high percentage of missing values and variables with a single value, in case this was not completed in prior stages of the methodology. Variables that are not related to the target appear marked with a rejected status. Although rejected variables are passed to subsequent nodes in the process these are not used in the model.



4. MODEL

With the chosen variables created and summarized in the previous chapter, we were able to test different models relating these covariates with the target variable of women's department purchase within the next month.

The SAS Entrerpise Miner nodes used to estimate the probability of purchase were:

4.1 Regressions

Regression can be used in their linear or logistic form, the target may be in continuous, ordinal or binary type and the input variables can be continuous and discrete. The node supports the stepwise , forward and backward variable selection method.

4.2 Decision Tree

This node features a variety of popular decision trees algorithms (eg CHAID, CART, C4.5, and C5.0.) Tree node supports two types of training metods: automatic and interactive. When the node is run using the automatic training mode, input variables are ranked according to the intensity of their contribution to the tree. Any automatic step can be overridden using the option to define a splitting rule and prune explicit tools or subtrees. This ranking can be used for selecting variables in subsequent models. Interactive training allows to explore and evaluate a large set of trees as they are being developed.

4.3 Neural Networks

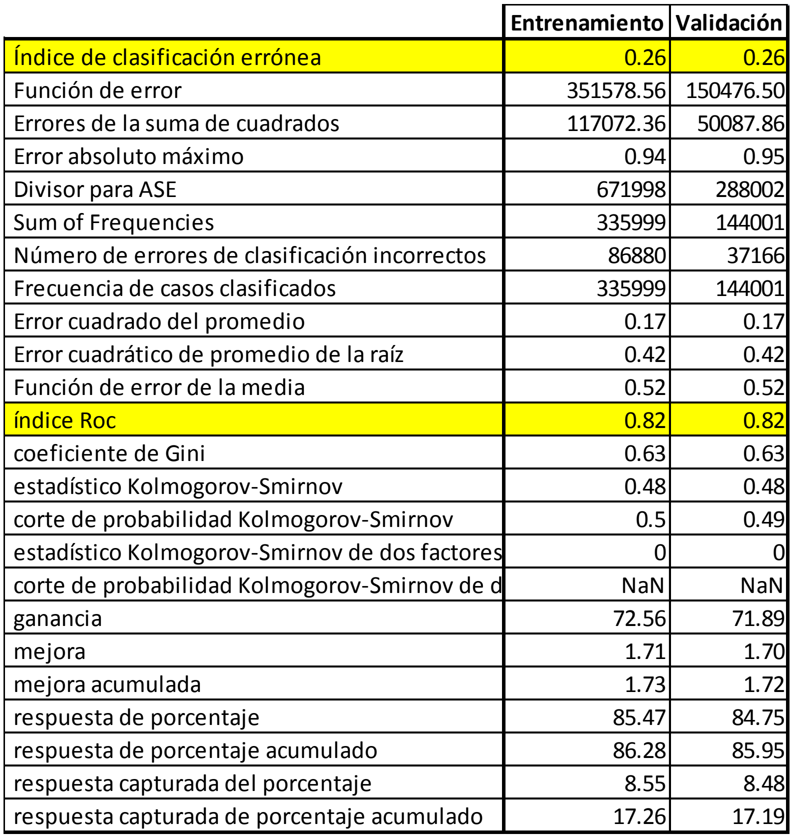
This node builds, trains and evaluates multilayer feedforward perceptron networks. The node builds networks with one hidden layer containing three neurons by default. In general, each input is completely connected to the first hidden layer, each hidden layer is connected to the next hidden layer and the last hidden layer is connected to the output.

The chosen model was a Logistic Regression, as this showed the best predictive power out of all the models that were tested. The results are presented in the following chapter.

5. ASSES

The following summarizes the main findings of the final model.

Results are presented for the training and validation sample, which allows validating the predictive ability of the model.



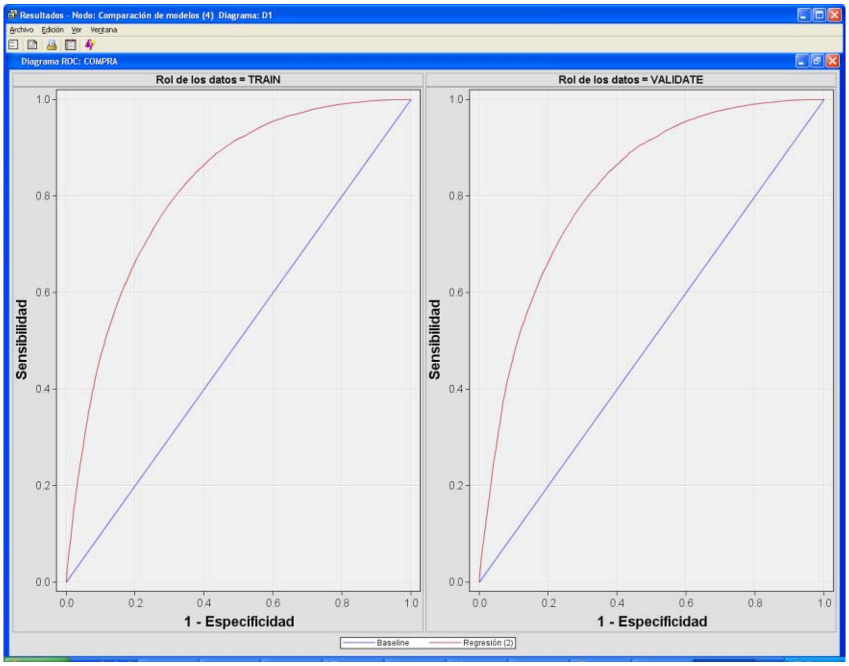
	Entrenamiento	Validación
Índice de clasificación errónea	0.26	0.26
Función de error	351578.56	150476.50
Errores de la suma de cuadrados	117072.36	50087.86
Error absoluto máximo	0.94	0.95
Divisor para ASE	671998	288002
Sum of Frequencies	335999	144001
Número de errores de clasificación incorrectos	86880	37166
Frecuencia de casos clasificados	335999	144001
Error cuadrado del promedio	0.17	0.17
Error cuadrático de promedio de la raíz	0.42	0.42
Función de error de la media	0.52	0.52
índice Roc	0.82	0.82
coeficiente de Gini	0.63	0.63
estadístico Kolmogorov-Smirnov	0.48	0.48
corte de probabilidad Kolmogorov-Smirnov	0.5	0.49
estadístico Kolmogorov-Smirnov de dos factores	0	0
corte de probabilidad Kolmogorov-Smirnov de d	NaN	NaN
ganancia	72.56	71.89
mejora	1.71	1.70
mejora acumulada	1.73	1.72
respuesta de porcentaje	85.47	84.75
respuesta de porcentaje acumulado	86.28	85.95
respuesta capturada del porcentaje	8.55	8.48
respuesta capturada de porcentaje acumulado	17.26	17.19

In the table it is first important to notice that results are similar for the training and validation samples, which is a good sign of the performance of the model.

The value of the Kolmogorov-Smirnov index and the Gini index are important signs of the model fitness quality and that using a logistic regression model is appropiate in this case.

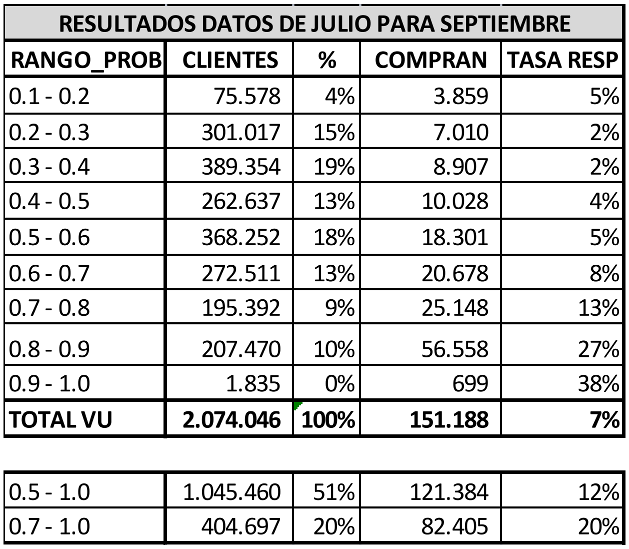
The area under the ROC curve is approximately 0.82, which is a satisfactory value to qualify the discriminatory power of the model.

As the following chart shows, the ROC curve is appropriate, which added to the previous results confirms the model as a tool for classification, these results, wich are similar in the validation sample, show the model's predictive ability.



5.1 Model results with actual data after construction

To verify the predictive ability of the model we reviewed several months of purchase information after construction of the model. As an example we can look at the propensity data obtained at the end of July which gives us the propensity of purchase in September.



RESULTADOS DATOS DE JULIO PARA SEPTIEMBRE				
RANGO_PROB	CLIENTES	%	COMPRAN	TASA RESP
0.1 - 0.2	75.578	4%	3.859	5%
0.2 - 0.3	301.017	15%	7.010	2%
0.3 - 0.4	389.354	19%	8.907	2%
0.4 - 0.5	262.637	13%	10.028	4%
0.5 - 0.6	368.252	18%	18.301	5%
0.6 - 0.7	272.511	13%	20.678	8%
0.7 - 0.8	195.392	9%	25.148	13%
0.8 - 0.9	207.470	10%	56.558	27%
0.9 - 1.0	1.835	0%	699	38%
TOTAL VU	**2.074.046**	**100%**	**151.188**	**7%**
0.5 - 1.0	1.045.460	51%	121.384	12%
0.7 - 1.0	404.697	20%	82.405	20%

In the table we can see that clients are well classified according to their real behavior. This can be obtained one month in advance.

REFERENCES

SAS Institute Inc 2009 SAS Institute Inc. Cary, NC, USA. Enterprise Miner version 6.1. SAS System Help.

SAS Institute Inc 2009 SAS Institute Inc. Cary, NC, USA. Getting Started with SAS(R) Enterprise Miner(TM) 6.1

ACKNOWLEDGMENTS

CONTACT INFORMATION

Name: Michael Soto Fuentes
Enterprise: Ripley Chile
Address: Huerfanos 1052
City, State ZIP: Santiago, Chile
Work Phone: (56 2) 26941372
Fax: (56 2) 26941442
E-mail: msotof@ripley.cl
Web: www.ripley.cl