Paper 097-2013

Estimates of Personal Revenue from Credit and Sociodemographic information combining Decision Trees and Artificial Neural Networks.

Deybis Flórez Hormiga, Colpatria Multibanca del grupo Scotiabank

ABSTRACT

A typical problem in the information required for the different processes within a bank, is to know the customer revenue, this information is not easy to gather and update because it is highly sensitive, and the accuracy of this information in large amounts of customers is not the best. Therefore, it is very important and of high impact in the processes of the bank, to find a method to estimate the revenue of customers for validation, segmentation, profiling, business strategies, risk mitigation, regulatory compliance, or simply as information. Due to the amount of information and the high volatility of the income reported by different clients, SEMMA methodology was used with Enterprise Miner. Starting from a fine segmentation using decision trees and then using Artificial Neural Networks in each of these segments, it generates a different model for each segment. These estimated models were evaluated by their consolidated results, comparing the development within the bank, with revenue generic models of national credit bureaus, obtaining higher performance to the extent that they include credit information and customer sociodemographic variables.

INTRODUCTION

The revenue inference process involves mining and segmentation of variables related to credit history and sociodemographic information of customers, which are closely related to their monthly income.

The relations between these variables and the payment capacity of holders were evaluated in multiple scenarios, using different methodologies. This led gradually to the refinement, transformation and complementation of variables capable to reflect reliably, the payment capacity of each person.

INFORMATION CONSOLIDATION

The information used for this process includes 500,000 customers who have revenue information that fulfill the definite parameters of completeness and validity of information.

After applying minimum quality controls to outliers and eliminate others without information we obtained about 450 thousand customers. Information later confronted with credit variables that lie in the Credit Bureau to set up consistency among the data.

HOLDERS' SEGMENTATION

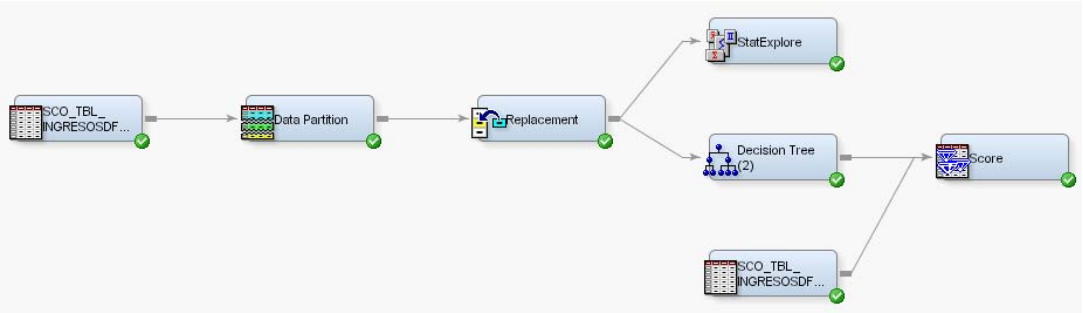
Once the base is treated, there's an intensive process that involves mining and entails the segmentation of credit profiles developed from the information bureau, these profiles are the backbone of the methodology.

These segments are broadly defined by variables such as marital status, age, time with the bank, initial values, among others. Finally, the validation of these profiles is simple: add the income information and verify that income levels are clearly different from each of these segments.

After making quality filters, we proceed to a customer segmentation using decision trees.

The methodology set nineteen different nodes which have a relatively homogeneous number of customers in each of them. The basic structure of the flow tree is shown below.

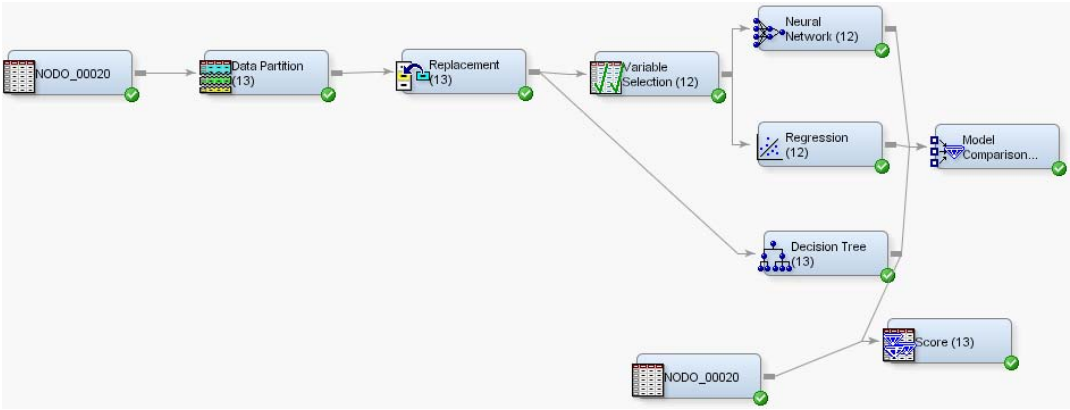
Figure 1. Structure tree flow Enterprise Miner.



MODELING IN EACH SEGMENT

The segmentation has led to different income groups, what is shown in significant differences between the average incomes for each of the 19 specified nodes. Different techniques were tested and eventually selected RNA due to its best performance.

Figure 2. Estimate node flow diagram.



GENERAL OUTCOME

The level of accuracy in the chart presented below corresponds to a count (presented in percentages).

Ratio inference to Observed Salary - Consolidated TEST sample

NODE	Ratio (Inference / Observed)			N	% (Inference / Observed)		
	< 0.6	0.6 - 1.2	> 1.2		< 0.6	0.6 - 1.2	> 1.2
1	753	2.419	352	3.524	21,4%	68,6%	10,0%
2	950	2.187	348	3.485	27,3%	62,8%	10,0%
3	828	1.432	248	2.508	33,0%	57,1%	9,9%
4	833	1.606	272	2.711	30,7%	59,2%	10,0%
5	789	1.504	256	2.549	31,0%	59,0%	10,0%
6	907	1.438	264	2.609	34,8%	55,1%	10,1%
7	949	1.297	250	2.496	38,0%	52,0%	10,0%
8	513	3.816	475	4.804	10,7%	79,4%	9,9%
9	625	1.783	267	2.675	23,4%	66,7%	10,0%
10	756	2.081	315	3.152	24,0%	66,0%	10,0%
11	922	1.679	289	2.890	31,9%	58,1%	10,0%
12	68	2.767	313	3.148	2,2%	87,9%	9,9%
13	199	2.215	270	2.684	7,4%	82,5%	10,1%
TOTAL	9.092	26.224	3.919	**39.235**	23,2%	**66,8%**	10,0%

Table 1. Results TEST Sample.

The ratio between the inference and the observed value of income is presented in order to know in each segment, how the compared prediction is distributed with the observed, giving an idea of the levels of adjustment. The chart shows that 66% of the estimated input data is in a range between 60% and 120% regarding the observed income. 23% of inferred revenue was underestimated by the methodology, while the remaining 10% was overestimated 1.2 times or more over the observed income.

CONCLUSIONS

Drawing on results in similar developments, superior performance was obtained by 14 percentage points in the ratio range of 0.6 - 1.2.

Tools like this are being used by the Bank of the Republic of Colombia as a leader in reporting country's financial stability.

In the bank using a tool like this, generates a large positive impact on their various processes.

REFERENCES

Gerencia de Riesgo Asobancaria Cifin. 2011. “Estimacion de la carga Financiera en Colombia.” Reporte de estabilidad Financiera. 37. Bogotá, Banco de la Republica de Colombia.

Matignon, Randall. 2007. “Data Mining using SAS Enterprise Miner”. Model Nodes. Wiley. New Jersey

Matignon, Randall. 2010. “Neural Network Modeling using SAS Enterprise Miner”. Neural Network Architecture. Lexington, Authorhouse.

RECOMMENDED READING

- *SAS® For Dummies®*
- *Applied Analytics using SAS®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Deybis Florez Hormiga
Enterprise: Colpatria Multibanca del grupo Scotiabank
Address: Torre Colpatria. Piso 6.
City, State ZIP: Bogota, Cundinamarca.
Work Phone: (57-1) 745 63 00 Ext. 3165
E-mail: Florezde@colpatria.com