SAS® GLOBAL FORUM 2020 #SASGF

Paper 5043-2020

Relationship analysis and customer churn forecasting in a financial cooperative institution

Carolina Silva, Sicoob Confederação; Brunno Sousa Ramos, Brazilian Air Force

ABSTRACT

The analysis of financial institutions, facing the new era of information access technologies, opened several ways to be addressed and explored, in particular to increase understanding about clients needs and improve the relationship with customers. In view of this, the objective of the present paper is to analyze the exit of clients in a financial cooperative, in order to understand the behaviors related to churn and identify in advance its possible occurrence. The development of the analysis was done following the steps of the SEMMA methodology. SAS® Institute defines SEMMA as a data mining process of Sampling, Exploring, Modifying, and Assessing large amounts of data. The analyzed clients were initially separated in three clusters, using the K-means algorithm. In each group, a predictive modeling was done with algorithms of Random Forest, Decision Tree and Logistic Regression. Predictive modeling provided greater understanding about churn incidence reasons. One of the clear points in this regard is that the constant use of products offered decreases the probability of the client leaving the institution. In addition, business rules were created and their application was performed in conjunction with the best results achieved with the predictive modeling in order to provide a better churn classification.

INTRODUCTION

In the present Brazilian economy scene, a segment that strengthens and plays a major role for is financial institutions. Buying and selling mobile values, raising resources from surplus agents and lending to deficit agents, they contribute to the national economy development, creating jobs and moving the economy (Reis, 2018).

According to the Brazilian Central Bank (2018), a credit cooperative is a financial institution formed by an autonomous association of people united voluntarily, with their own legal form, civil in nature, non-profit, meant to deliver financial services in a simple and advantageous way to their associates.

What is common between banks and credit cooperatives is that both are financial intermediation institution. Both offer financial services and products as checking account, credit card, investments, loans, consortia, insurance, etc. The main differences are in the way they are compounds and performance purposes. Commercial Banks are private or public financial institutions that have as main goal provide resources supply needed to finance trade, industry, service providers and people. Banks have few owners, which seek return on invested capital. Clients has no decision power over their way of acting, they are just users of the financial services offered.

While exploring those main differences, it is important to notice that the customer retention analysis has increasingly gained focus in the financial market, when the subject is relationship marketing. The emergence of fintech’s, services digitalization and greater access of people to more and more financial institutions, have been leading companies in this industry to invest even more in managing customer relationship. The main goal is to understand their needs, what makes them change from an institution to another and try to guarantee their satisfaction before they gone, that is, before churn occurs.

That said, this paper aims to carry out a churn study, focusing on the creation of a predictive model that allows the early identification of costumers exit in a financial cooperative.

SEMMA

SEMMA is the data mining methodology used to get business advantage in customer retention. SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, and Assessing (SEMMA) large amounts of data, as it follows:

- Sample the data by creating in or more data tables. The samples should be large enough to contain the significant information, yet small enough to process;
- Explore the data by searching for anticipated relationships, unanticipated trend, and anomalies in order to gain understanding and ideas;
- Modify the data by creating, selection and transforming the variables to focus the model selection process;
- Model the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome;
- Assess the data by evaluating the usefulness and reliability of the findings from the data mining process.

DESCRIBE ANALYSIS

To develop this work a database was structured containing both costumer registration variables and financial movement variables. More than 2.8 million clients associated were analyzed.

An interesting association that we could notice was the relation between the members' number of products and the target churn in three different moments throughout the year (Table 1). In January of 2018, the mean of the products from associates that stayed in the company was 1,61, while the mean from those whom left was 0,97. The standard deviations were, respectively, 1,94 and 1,38. In July and December of the same year the average product of costumers who stayed suffered slight change (1,63 and 1,73, respectively), meantime the average products of costumers with yes to churn fell sharply registering 0,70 in July and 0,37 in December. That is, those records show that there is a possibility of a product deactivation behavior as the client's withdrawal data is approaching. Meanwhile, among those who stayed, the average product consumption increased.

2018 moment **Client situation**	**Jan/2018**	**Jun/2018**	**Dec/2018**
Stayed	1,61	1,63	1,73
Left	0,97	0,70	0,37

Table 1. Average amount of products in three moments along 2018, separated by churn

CLUSTER ANALYSIS

As the analysis moves on it gets clear that there are big differences between the associates. We are dealing with costumers with no financial activities and also with some extremely active. How the costumers should be prioritized through customer loyalty and retention? To answer this important question customer segmentation was performed through cluster analysis. Using SAS® Enterprise Guide® the PROC FASTCLUS procedure produces the k-means results and the source code used is as follow.

```
proc fastclus
      data= data.churndatabase
      out= data.cluster
      RANDOM = 0
      maxiter= 100
      converge= 0.0001
      maxclusters= 3
      summary
      outstat= data.OutClus;
      var  VAR_1 VAR_2 VAR_3 VAR_4 VAR_5 VAR_6
           VAR_7 VAR_8 VAR_9 VAR_10 VAR_11  VAR_12
           VAR_13 VAR_14 VAR_15 VAR_16 VAR_17;
run;
```

In this case, PROC FASTCLUS uses 17 different variables to split the database in three clusters. These variables are information about clients as age, number of products and others confidential data. The Table 2 below shows the members distribution by clusters and summary information.

Cluster	**Number of associates**	**Percentage**	**Average age (Years)**	**Average time in the company (Years)**	**Average number of products (Dec/2018)**
Bronze	972.502	33,75	61	11,3	1,16
Silver	1.329.278	46,12	35	4,9	1,79
Gold	579.869	20,13	47	8,6	2,49

Table 2. Members distribution by clusters and summary information

PREDICTIVE ANALYSIS

In order to better understand the main reasons that lead a costumer to leave the institution, and also to be able to predict it, predictive analysis models were used in each of the clusters. The chosen analysis techniques were Random Forest, Decision Tree and Logistic Regression. The data was prepared to fit in all the models, and, noting the existence of rare events in the target variable, a key step that was taken was to balance the samples using Random Undersampling.

BEST RESULTS WITH PREDICTIVE ANALYSIS MODELS

The evaluation of each model was preceded using the confusion matrix, accuracy and ROC curve. The confusion matrix is a table that shows the observed and predicted classification frequency to each class of the target variable. It allows the visualization of positives and negatives cases that were correctly predicted which also are known as true positives (TP) and true negatives (TN), respectively, they represent the model’s right classification. The positives that were predicted as negatives and the negatives that were predicted as positives, they are known as false negatives (FN) and false positives (FP), respectively, and they are the model’s wrong classification. From the matrix, some statistics can be used to better choose the best model:

- Sensibility: It’s the ability of the model to correctly predict positive values among those that are truly positives. It’s true positive proportion:

$$\text{Sensibility} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Specificity: It’s the ability of the model to predict negative values among those that are truly negative. It’s true negative proportion:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Accuracy: That is how the model got it right in the forecast, both positive and negative.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Area Under Curve: It’s a measure used to evaluate a model, which represent the area under curve ROC. The closer to 1, the better the model fit.

Table 3 shows the matrix confusion obtained using the results from the best models. The selected models to Cluster Gold and Cluster Silver were Random Forest, while the best model to Cluster Bronze was a Decision Tree.

	Predicted					
	Cluster Gold		**Cluster Silver**		**Cluster Bronze**	
Observed	**Left**	**Stayed**	**Left**	**Stayed**	**Left**	**Stayed**
Left	75%	25%	80%	20%	77%	23%
Stayed	9%	91%	19%	81%	20%	80%

Table 3. Confusion matrix of the best models results to all clusters

The Cluster Gold accuracy, with random forest model, is 75% and true positive estimative was the higher among the three clusters (91%). Cluster Silver accuracy is 80%, True Positive and True Negative estimative were quite similar with 81% and 80%, respectively. Cluster Bronze accuracy is 77%.

The following figures shows how all three models were created using SAS® Enterprise Miner™: starting with a import of a balanced sample of each cluster; impute step to fill missing data; transform variable that creates categories based on quartiles; data partition (75% training, 25% validation); Decision Tree, Random Forest and Logistic Regression; model comparison to choose the best model; and Score the best model result in the cluster data.

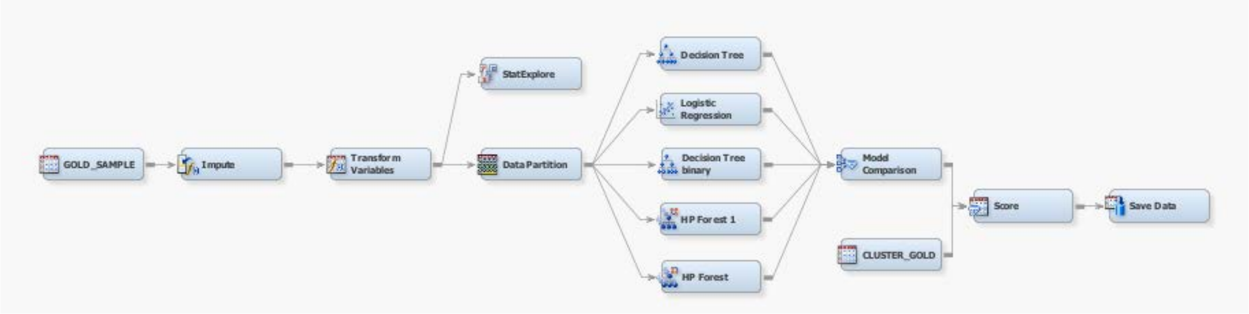

Figure 1. Gold Cluster model workflow

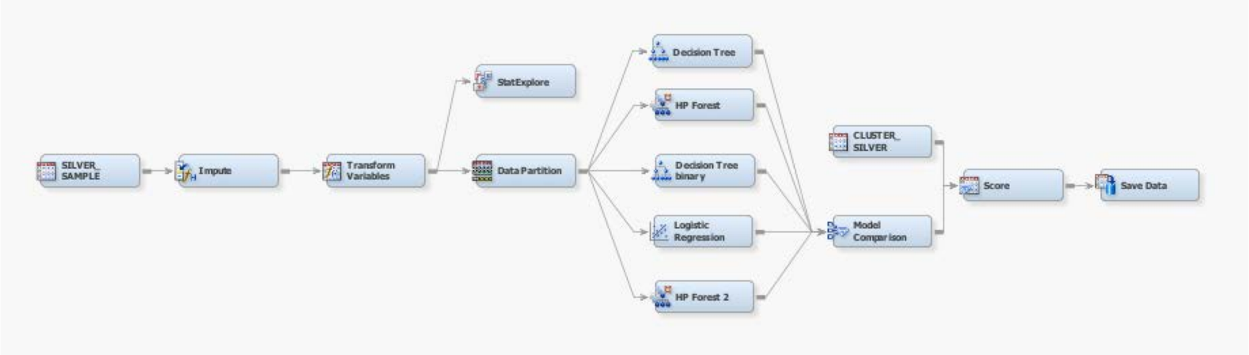

Figure 2. Silver Cluster model worklow

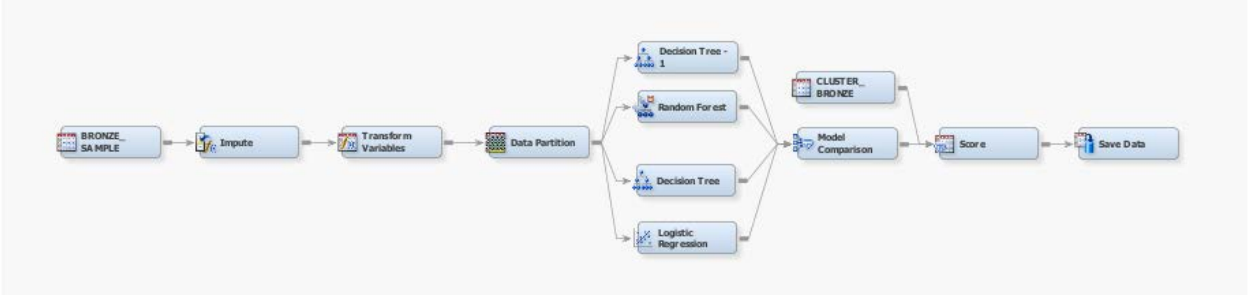

Figure 3. Bronze Cluster model workflow

To improve the results obtained business rules were created. These rules reflect the relationship models between customers and companies and also the knowledge about the business. These precepts not only have the intuition to clearly show who as the bigger probability to leave the company, but also to separate this group from the one who shows no sign to be about to left. For this reason, positive and negative rules were created in order to give positive points to those who fit in rules related to churn possibilities, and give negative points to the ones that fit the rules that show evidences to stay.

1. Rule 1 (Positive): Number of financial institutions greater than 0 and quantity of products throughout the year equal to 0.
2. Rule 2 (Positive): The sum of transactions in all service channels throughout the year is equal to 0;
3. Rules 3 (Positive): Number of transactions in December 2018 is lower than the monthly average transactions;
4. Rule 4 (Negative): Product variation between January and July is greater than 0 and the product variation between July and December is also greater than 0;
5. Rule 5 (Negative): Number of products in December is greater than the average quantity of clusters products in December; and
6. Rule 6 (Negative): active checking account and number of capital account pay-outs in the last 6 months greater than 6.

In addition to being positives and negatives, the rules also have weights. The weight was also attributed to the predicted value calculated by the model selected in each cluster, so that a final score would be created later, which is the result of a weighted sum of the predictive model score and business rules. Table 4 shows the summary of sign and weight to each business rule.

RULE	POLARITY	WEIGHT
$\mathbf{Rule_1}$	POSITIVE	5
$\mathbf{Rule_2}$	POSITIVE	7
$\mathbf{Rule_3}$	POSITIVE	3
$\mathbf{Rule_4}$	NEGATIVE	4
$\mathbf{Rule_5}$	NEGATIVE	7
$\mathbf{Rule_6}$	NEGATIVE	4

Table 4. Weight and signs definition to each business rule

Based on the information in Table 4, the Final Score is calculated using the following equation:

$$SCORE_{Final} = CHURN_{predict} * (70) + Rule_1 * (5) + Rule_2 * (7) + Rule_3 * (3) - Rule_4 * (3) - Rule_5 * (7) - Rule_6 * (4).$$

BUSINESS RULES INTO THE CLUSTERS

The Final Score was applied in all clusters, and the results were positive. It was possible to enhance the accuracy in Clusters Gold and Silver, and the results in Cluster Bronze were not so different than those obtained by the Random Forest algorithm, but it was expected since this cluster consisted mostly of low active associates.

Business rules increased the Cluster Gold accuracy to 86%, 11% better than the original random forest model. In absolute numbers it means that with the business rules we reach out more than 61.000 new correct predictions. These numbers to Cluster Silver and Cluster Bronze were, respectively:

- Cluster Silver: accuracy 87% (7% better than the original) and more than 27.000 new correct predictions; and
- Cluster Bronze: accuracy 80% (3% better than the original) and more than 93.000 new correct predictions.

CONCLUSION

This paper focused on analyzing the churn of clients of an institution cooperative in order to develop methods to identify possible customer exit before it occurs so you can work on your retention. This issue is quite relevant in the current market scenario as consumers are increasingly offered financial services and more information, which facilitates the decision to change institutions if the current one is not attending as expected. Given this, the use of information to support strategic actions and guide the relationship management with the client becomes even more important.

The use of descriptive analysis, cluster analysis, three predictive modeling techniques and business rules allowed a vast scenario of strategies to be used when searching for the model with the most assertiveness. It also made possible the interpretation of the results in different ways, which further enriched the understanding about the various behaviors that can lead to churn.

REFERENCES

Reis, T. "Instituições Financeiras: Clique e conheça as principais do Brasil". 2018. Available at https://www.sunoresearch.com.br/artigos/instituicoes-financeiras.

Banco Central do Brasil. "Perguntas frequentes:Cooperativas de crédito". 2018. Available at https://www.bcb.gov.br/acessoinformacao/perguntasfrequentes-respostas/faq_cooperativascredito

SAS Institute. “SAS® Enterprise Miner™ 14.3: Reference Help”. Available at https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en.

CONTACT INFORMATION

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

Carolina Silva
Sicoob Confederação
+55 61 9 82146605
carolina_andrade5@hotmail.com

Brunno Sousa Ramos
Brazilian Air Force
+55 61 9 86074913
bks_ramos@hotmail.com