University of California Berkeley

Information School

INFO 251

Customer attrition prediction, a real case in a Chilean Bank

Submitted by
Rivers Jenkins
Cristobal Pais
Yiyu Shi
INFO 251
December 11th
Fall 2017

Contents

Customer attrition prediction, a real case in a Chilean Bank

R.Jenkins, C.Pais, Y.Shi

December 11th 2017

Abstract

A customer attrition problem for a bank in Chile, South America is studied. Classification models are trained and tested following an underlying cost matrix structure, reflecting the real misclassification costs for the bank. The trade-off between cost and accuracy is analyzed. The Logistic regression model reaches the best overall performance when balancing cost and accuracy (86%) of the solution. Similar results are obtained when analyzing the misclassification error evolution via lift and cumulative gain charts for the positive class as more observations are included, incurring in a 9% of error when 60% of the cases are included. Using these tools, the bank is able to focus its budget and target the relevant customers via marketing campaigns and/or special offers for keeping them. Two main patterns for customers leaving the bank are identified: (1) cleaning the accounts in few & big transactions, and (2) spending the money until the account is empty. The second type is not identified by the current prediction model, leading to false negative misclassification. Similarly, false positive misclassification errors are identified as clients having a loan with the bank, information that is not available in the original dataset, leading to significant misclassification errors that should be avoided by integrating the information across the areas of the bank.

1 Introduction

In the banking world, a significant percentage of earnings are related to the size of the portfolio that the bank has, expecting more profitability with a larger number of associated clients. In this context, there are two key actions that can be performed by a bank in order to keep and increase its client base: customer attraction and customer retention.

Customer attraction is focused on incorporating new clients through a different set of strategies such as marketing campaigns, advertising, new sales points, branch offices and special offers/conditions. These strategies depend on the specific segment of the market that the bank is focused.

On the other hand, customer retention implies two processes: 1) Identification of the customer segments that are more likely to abandon the bank (client attrition), and 2) Defining a set of commercial policies and procedures/actions that will make clients desist from leaving the institution.

In general, a new client is more risky than an "old" client, thus, it is recommendable to focus resources and effort in the retention of current clients. A series of studies **([1],[2],[3], [4])** have shown that the total utility obtained from a client is proportional to the amount of years that they have been linked to the financial institution and, on the other side, the acquisition of new clients is about $5-6$ times more expensive than actually keeping an old client. New clients also increase the risk of the customer portfolio.

ABC [1], a real bank located in Chile, has a problem related to customers retention. Currently, it has an annual retention rate: $ARR = (1 - \frac{\text{leaked clients per year}}{\text{Total portfolio}}) = 90\%$, meaning that 10% of the total portfolio finish its contract with the bank annually, and they use the customers churn indicator as its main quality factor. Nowadays, the bank has implemented a basic predictive model for estimating the number of clients that will stop their relationship with the entity in a near future. It estimates that about 30% of their total portfolio may potentially finish their contract but ABC only has enough resources to contact and persuade about 5% of its total portfolio. Also, the current prediction system [2] has a very poor accuracy of about

[1] Due to special conditions imposed by the bank, we will use a fictitious name in our work.

[2] See Appendix section 7.3 for details

40%, meaning that this is the percentage of clients that were labeled as "escaping clients" and actually left the bank.

Nowadays, the bank operates with a series of simple decision rules in order to determine if a risky client should be contacted & classified as a positive case: (1) No transactions during the last 30 days, (2) low frequency of checks use during the last 3 months, and (3) devolution of products such as credit/debit cards. However, none of them seems to be very effective.

Therefore, the objective of our project is to develop an efficient and effective customer attrition prediction model for a real bank in South America, helping the bank to focus its resources for its retention actions/policies.

2 Dataset

The bank has a corporate data warehouse that records all the relevant information of the entire portfolio. However, the database presents some problems (missing values, inconsistencies, etc.) and it is also not optimized for a prediction model, thus, some variables might not be relevant or would need some transformations/engineering before using them as real and significant indicators inside our mathematical model.

We have access to a dataset with 4055 entries, including 18 attributes per client such as educational level, last transaction amount, civil state, age, etc. All observations include a binary label attribute, where 1 indicates that a client has left the bank and 0 indicates that the client is still associated to the institution.

The attributes inside the dataset are the following:

Table 1: Attributes in Data Base

Cod_cliente	Client Identification
COD_OFI	Office code
COM	Province
ED	Age in years
SX	Sex/Gender (M: male; F: female)
NIV_EDUC	Educational level
RENTA	Rent in thousand of pesos by month
E_CIVIL	Civil state
VIG	Validity (in month) — Age of the account
TRX_T	Number of transactions in month T
TRX_T-1	Number of transactions in month T-1
TRX_T-2	Number of transactions in month T-2
SALDO_T	Average balance in CL pesos in month T
SALDO_T-1	Average balance in CL pesos in month T-1
SALDO_T-2	Average balance in CL pesos month T-2
SALDO_T-3	Average balance in CL pesos month T-3
SALDO_T-4	Average balance in CL pesos month T-4
CERRO	Indicator if the person closed (1) or not (0) his/her account

a) **Bank variables**: Attributes that are related to transactional activities of the client such as the current and past balances inside their checking accounts, the number of products associated with the bank that the client has 'today', frequency of credit/debit card uses, and number of services associated with the bank that the client is frequently using. We can classify TRX and $SALDO$ as part of this set. This set of variables will be of significant relevance when developing our predictive models.

b) **Sociodemographic variables**: Relevant characteristics of the client itself (inherent to each individual) such that age, civil status, gender, socioeconomic level, and wage among others. In this set we have variables $RENTA$, NIV_EDUC, ED, COM, E_CIVIL and SX.

c) **Environment Variables**: These capture the effects associated with the relationship between the customers and the financial institution as well as with external elements from the market. In this case we have variables VIG y COD_OFI

There are 5 real variables, 8 integer ones, 2 binary, and 1 categorical variable in the data set. As indicated in the proposal, we discovered a significant discrepancy between the two datasets provided by the bank (with 1248 and 2807 observations respectively). After contacting the bank, they realized that their new IT system was not working as expected and thus, we continue our study using the correct dataset with 1248 observations [3]. From this dataset we obtained the following summary statistics.

Table 2: ABC Bank data set number 1 Summary

Attribute	**Variable**	**Type**	**Mean/Mode**	**Range and Frequencies**	**Missing**
Id	Cod_Cliente [ID]	integer	avg = 624.500 +/- 360.411	[1.000 ; 1248.000]	0.0
Label	CERRO [LABEL]	binary	mode = 0.0 (629) least = 1.0 (619)	0.0 (629)- 1.0 (619)	0.0
Regular	COD_OFI	integer	avg = 74.268 +/- 49.723	[10.000 ; 247.000]	0.0
Regular	COM	integer	avg = 122.481 +/- 91.050	[1.000 ; 516.000]	0.0
Regular	ED [years]	integer	avg = 41.319 +/- 12.013	[21.000 ; 86.000]	0.0
Regular	SX [M male,F female]	binary	mode = M (839) least = F (408)	M (839)- F (408)	1.0
Regular	NIV_EDUC	nominal	mode = UNV (525) least = BAS (5)	UNV (525), MED (298), TEC(400), BAS (5), EUN (12)	8.0
Regular	RENTA [M CLP/month]	integer	avg = 729.071 +/- 514.604	[250.000 ; 5400.000]	0.0
Regular	E_CIVIL	nominal	mode = CAS (822), least = VIU (21)	CAS (822), SOL (348), VIU (21)SEP (57)	0.0
Regular	VIG [months]	integer	avg = 55.654 +/- 53.425	[3.000 ; 366.000]	0.0
Regular	TRX_T [N°/month]	integer	avg = 38.632 +/- 36.569	[0.000 ; 417.000]	0.0
Regular	TRX_T-1 [N°/month]	integer	avg = 37.557 +/- 34.400	[0.000 ; 390.000]	0.0
Regular	TRX_T-2 [N°/month]	integer	avg = 43.248 +/- 38.089	[0.000 ; 391.000]	0.0
Regular	SALDO_T-4 [$\overline{CLP}$/month]	numeric	avg = 1174103.212 +/- 3481175.370	[-568443.750 ; 60791570.550]	0.0
Regular	SALDO_T-3 [$\overline{CLP}$/month]	numeric	avg = 1148645.272 +/- 3298359.403	[-556992.860 ; 70450696.350]	0.0
Regular	SALDO_T-2 [$\overline{CLP}$/month]	numeric	avg = 1235826.283 +/-3016768.011	[-757777.710 ; 35383283.950]	0.0
Regular	SALDO_T-1 [$\overline{CLP}$/month]	numeric	avg = 3084470.349 +/- 8149275.790	[-1359252.860 ; 98154458.100]	0.0
Regular	SALDO_T [$\overline{CLP}$/month]	numeric	avg = 4327925.344 +/-11252574.346	[-3814728.100;159122938.570]	0.0

In Table 6 we can observe that both classes (types of clients) are almost balanced (50.4% vs 49.6%) in the first dataset. Thus, classic predictive models will be useful since it is possible to train them in an efficient way with the available data, a situation that may not occur when classes are too unbalanced. In the table, we can see the mean/mode values, ranges/frequencies and the number of missing values per variable.

We can see that there are missing values associated to two variables: SX and NIV_EDUC. Depending on the relevance of these variables for the predictive model, we would select the most suitable approach for dealing with these missing values.

3 Methodology

The main objective of the project will not be focused on the (classic) average error of the predictive model, but on minimizing the *false negative* (FN) cases based on a confusion matrix with an underlying cost structure. This is due to the fact that this type of classification error will have the most significant (bad) impact for the bank profit since the idea is to identify those customers that are likely to abandon the institution and focus the bank resources (e.g via marketing campaigns) in this particular group in order to persuade/convince them to continue the relationship with the institution. On the other hand, FP cases will not be as bad as the FN since the bank will spend resources on clients that are likely to keep a longer contract with the institution and thus, they will still generate good (and better) utilities.

Therefore, we implement the following methodology in order to achieve this objective:

[3] More details in Appendix section 7.1

3.1 Pre-Processing

- Standardization of the variables: The original dataset contains several features with very different scales and thus, we standardize them in order to be able to easily compare them and obtain better algorithmic performance when applying the classification models.

  Basic standardization of the dataset is performed. The standardization is necessary because the magnitude of the $SALDO$ variables is much larger than the magnitude of most other variables (such as TRX).

- String to numerical: Variables (categorical) containing string characters such as SX or E_CIVIL are transformed into numerical ones in order to be able to implement numerical operations (and models) with them.

- Missing Values: We also drop observations that contained missing data. As this was such a small percentage of the overall dataset, it is the easiest/simplest and most straightforward approach to take.

After transforming the data, preliminary analysis is performed to look for relations between the variables we have available. We also look for simple correlation with the client attrition variable ($CERRO$) to get a sense of which variables may be more useful in subsequent analysis.

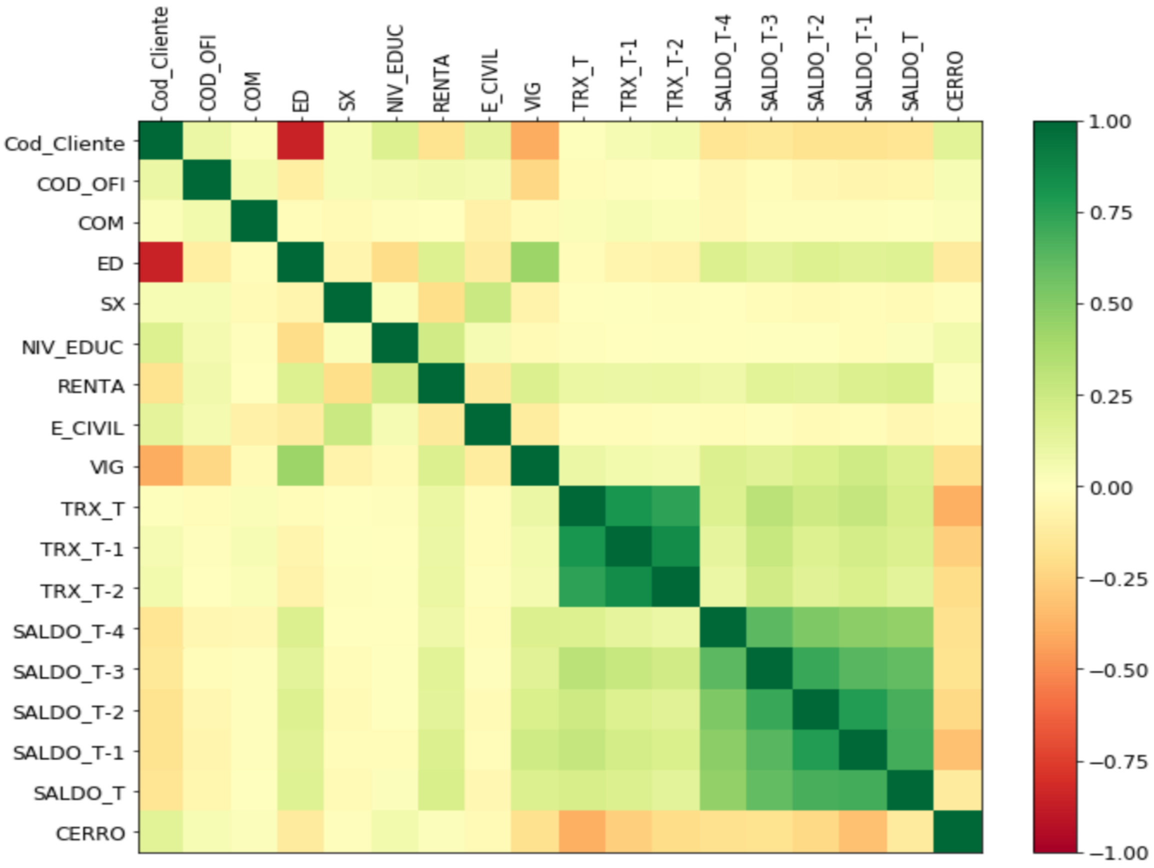

Figure 1: Correlation Matrix

As can be seen in Figure 1, none of the variables we have been given exhibit extremely strong correlations with response variable (CERRO). The strongest correlation comes from TRX_T which represents the number of transactions that the customer made in the last month reaching a correlation level of $\rho = -0.39$. This variable is followed by $SALDO_{T-1}$ with $\rho = -0.32$ and then by TRX_{T-1} with $\rho = -0.26$. The lack of strong simple correlations means that simple linear models will probably not perform especially well with the variables provided. Some variables such as gender (SX) show essentially zero correlation.[4]

[4] For extra comparisons and preliminary analysis, check Appendix section 7.4.

3.2 Feature Engineering

Because of the low correlation seen with most variables and the *CERRO* variable, we performed some feature engineering. We generated new features based off of the bank variables because they showed the most initial predictive power. The features we generated are enumerated below.

1. **1D SALDO**: Difference between the T and $T-1$ *SALDO* variables. This new feature arises as one of the main important ones in the study since - as expected - it summarizes the decreasing pattern in the average amount of money in balance for the "leaving" clients.
2. **MEAN SALDO**: Average value of the balance features.
3. **1D TRX**: Similar to the *SALDO* difference, this variable captures the evolution in the number of transactions by the end of the periods under analysis.
4. **PCA (3 components)**: In order to reflect the main three groups of variables identified, a principal component analysis is performed, selecting three components, explaining more than 80% of the variance.

3.3 Model selection: Cost Matrix approach

The typical way that machine learning models are evaluated in a classification context is with a simple accuracy score. What percent of data points were correctly classified? In most cases this is the most natural scoring mechanism. However, in the context of this study, we have an imbalanced cost of misclassification.

There are two possible misclassification contexts:

1. **False Positive** (FP): indicates that the model predicts that the customer will leave when in fact, they would have stayed. The cost in this scenario is the cost of the bank having to contact the customer (when such contact was unnecessary).
2. **False Negative** (FN): indicates that the model predicts that a customer will not leave when in fact they will leave if left untouched. This scenario is much more costly to the bank. They lose a valuable customer. The bank now has to either find a new customer (very costly compared to retaining old customers [4]) or somehow entice that customer back to the bank.

This imbalance in misclassification contexts creates an imbalance in costs which would not be reflected in a typical accuracy score. With a typical scoring mechanism, the cost of misclassification is the same no matter which type of misclassification it is. To better capture this imbalance in cost, ABC bank provided us with the following cost matrix.

Table 3: Cost matrix constructed by the ABC bank estimations.

Cost Matrix	**Predicted: Stay**	**Predicted: Leave**
Actual: Stay	TN: 0	FP: 350
Actual: Leave	FN: 1000	TP: 250

We can use Table 3 to calculate the cost associated with each model. This will allow a more accurate measure of how effective the model will be in practice. Models with lower cost should produce superior results for the bank (even if these models exhibit lower accuracies).

3.3.1 Training the models

We trained the following classification models on 80% of the dataset using a 25-fold Cross-Validation approach for optimizing their most relevant parameters based on the testing set accuracy/cost. This will allow us to analyze the trade-off between accuracy and cost of the solution.

1. Decision Tree

2. Random Forest
3. Logistic Regression
4. SVM
5. KNN
6. Neural Network

Different sets of features were tested in order to train the models, based on the performed feature engineering and their importance for each model. Hence, the previous models are trained with: (1) original dataset, (2) including $1D\ SALDO$, (3) $MEAN\ SALDO$, (4) $1D\ TRX$, (5) all the features, and (6) using the 3 component PCA dataset.

3.3.2 Class weight

The classification models (decision tree, random forest, logistic regression, svm) contain a hyper-parameter [5] which allows the user to give more or less weight to one of the classes, modeling the effect of a cost matrix. We used this parameter to tune the models to give more weight to customers who would leave the bank. This allowed us to reflect the imbalanced cost discussed earlier.

As an example, if the weight of the class 1 is selected as 4, it means that positive classifications are four times more important (translated to: FP are 4 times worse than FN) than the negative ones.

3.4 Target the relevant clients: Lift & cumulative gain charts

Based on the fact that the bank has a limited budget, we want to focus the money of the bank (e.g. via marketing campaigns or special offers) on those clients that are more likely to be classified as positive cases.

In order to do that, we construct a series of lift and cumulative gain charts for all the tested models, where we can easily see the number of misclassification for label 1 as we include more observations, sorted by their probabilities of being classified as “leaving customers”.

The lift chart is useful for determining the performance of the classification models when we need to predict the class of a client. This tool will be very useful for checking which clients have the largest probability of abandoning, such that we will be able to contact them in the first place in order to try to keep them, helping ABC to focus their resources on the most relevant groups.

Hence, we will compare the performance of all models in terms of the misclassification error evolution as we include more observations from the positive labeled cases indicating us which model is more likely to minimize the error of selecting a wrong client as we contact the predicted more risky clients.

3.5 Error Analysis

In order to conclude the study with valuable insights for the bank, a formal error analysis will be performed for the model with the best performance in order to identify the main characteristics of the relevant clients (more risky customers) as well as understanding the source of the misclassification errors, answering questions such as why some observations are classified as FN or FP.

This will help the bank to identify relevant patterns among its clients as well as guiding the re-design of their current data set in order to improve the performance of the predictive models by including crucial information/features.

[5] Implementation using Sci-kit learn package in Python. There is no support for explicit cost matrices in the current version.

4 Results & Discussion

In this section, we present the most relevant results and discussion of the study. All results are generated using Python 3.6 64 bits version with the sci-kit learn package.

4.1 Features & model selection

As expected from our preliminary analysis, many of the features (especially the categorical ones) were not useful for the predictive models. The feature importance values are shown for the Decision Tree model in Figure 2. $SALDO_{T-1}$ is clearly the most useful of the original predictor variables provided in the dataset. Any variables excluded from the graph exhibited a feature importance of 0. Similar patterns were obtained across the rest of the models.

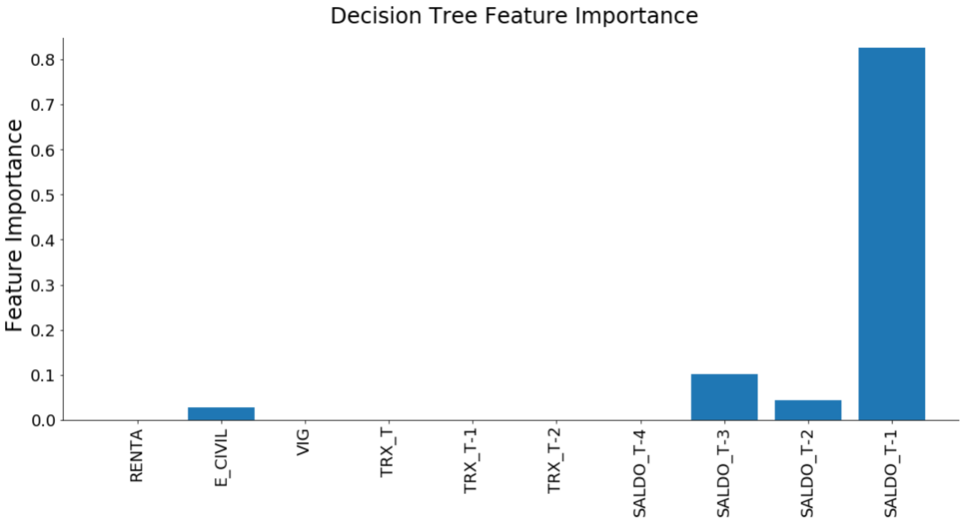

Figure 2: Feature importance for the decision tree model including the original features.

The random forest model shows similar results to the decision tree model in terms of feature importances. In Figure 3 we have included the features we engineered. When we include $1D\ SALDO$, it becomes the most important predictive feature for the model.

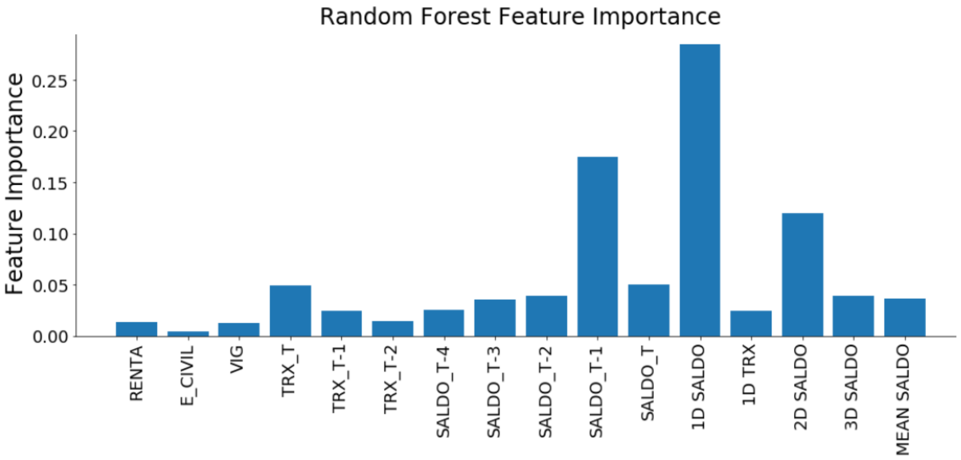

Figure 3: Feature importance for the Random Forest model including extra features.

Hence, the most relevant variables for the prediction model are related to the bank features such as the average balance in the customer account ($SALDO$ variables) as well as the number of transactions performed in the past months (TRX features). In all the tested models, $SALDO\_T - 1$ arises as the most relevant variable. This pattern is confirmed when the $1D$ $SALDO$ is added to the models, concentrating the importance along the original $SALDO\_T-1$ feature.

A clear decreasing pattern can be seen in these features for the positive cases, indicating that customers who are more likely to leave the institution are emptying their accounts before finishing the contract with the bank. Similar results can be seen with the number of transactions: tend to decrease in those clients who are more likely to leave the bank.

However, these variables are not able to capture all the customers' behaviors, leading to an unavoidable misclassification error. In addition, features that are not included in the original data set could be included in order to increase the performance of the models, situation that will be analyzed in section 4.4.

Sociodemographic and environment variables did not have a significant impact in any of the tested models and thus, were discarded due to their low predictive power.

4.2 Optimized cost models

The main results for all 6 models are summarized in Table 4. The last row indicates the GAP obtained in terms of the main cost objective function when comparing the optimized cost version of the model with the default (class weight equals to 1, maximizing accuracy) implementation. Notice that both neural network and K-NN models do not include the class weight option and thus, no trade-off analysis is performed.

Table 4: Models performance comparison

Performance/Model	Decision Tree	Random Forest	Logistic Regression	SVM	Neural Network	K-NN
Training Accuracy	84.9%	88.1%	83.4%	85.3%	80.8%	71.2%
Testing Accuracy	77.6%	83.1%	86.1%	76.5%	73.8%	61.2%
TP	109	116	114	118	87	67
TN	82	90	91	73	96	85
FP	45	37	36	27	40	60
FN	12	5	7	32	25	36
Weight	2	5	4	3	-	-
GAP Cost	6.3%	10.5%	17.4%	2.32%	-	-

From the results, it can be clearly seen that the optimal cost models do not correspond to the default implementation, obtaining better results when the positive class weight is increased due to the diminution of false negative errors.

Detailed results of the accuracy versus cost trade-off can be seen in Figures 4 and 5 for the three models (decision tree, random forest, and logistic regression) with the best overall performance.

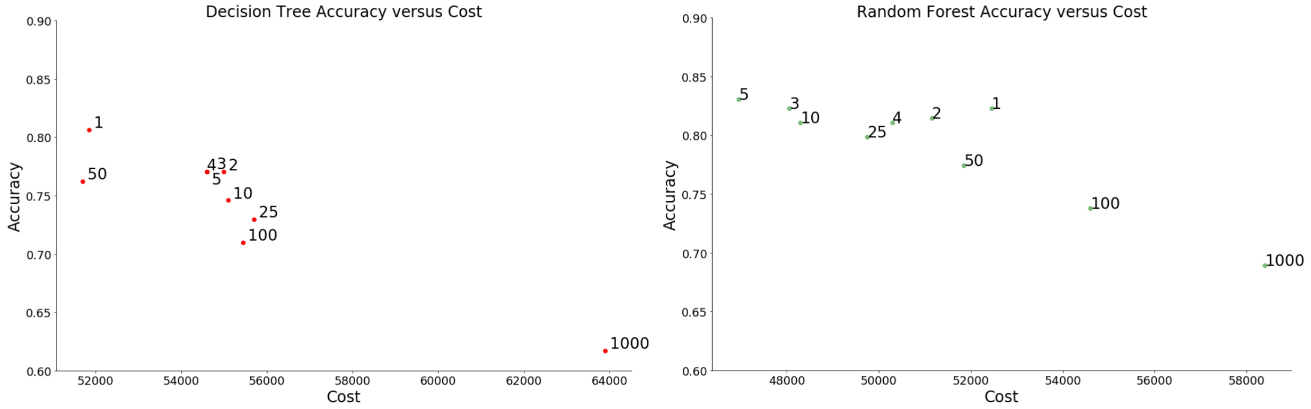

Figure 4: Comparison between accuracy and cost for the decision tree and random forest models for different weights.

Based on the previous analysis, the model that is most suitable for the bank in terms of overall performance and complexity is the logistic regression model: reaches the best testing accuracy as well as the minimum cost when using the optimal class weight value.

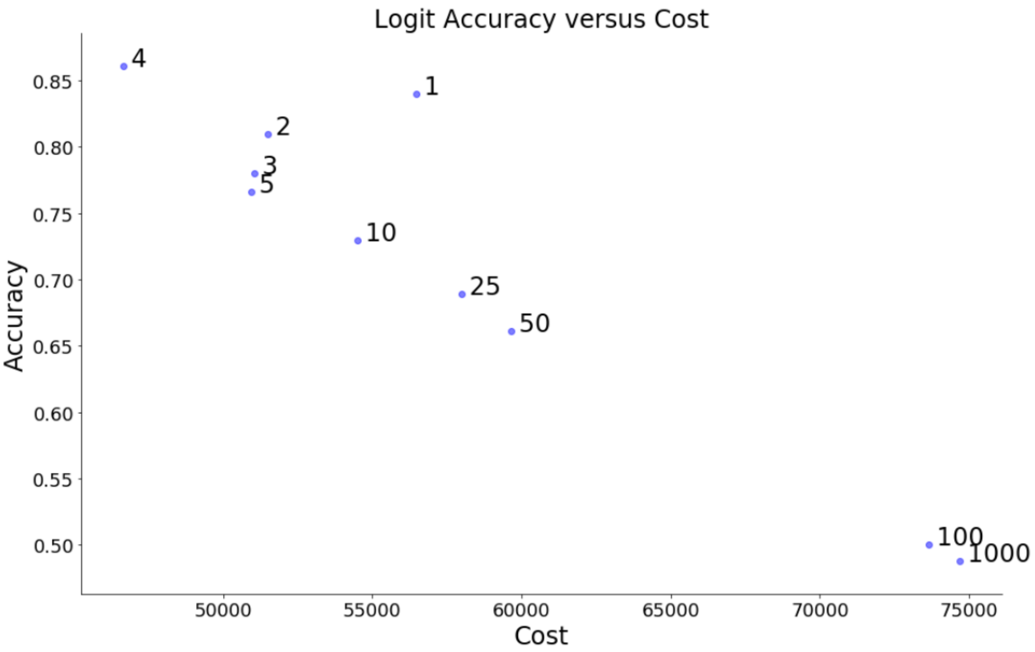

Figure 5: Comparison between accuracy and cost for the logistic regression model for different weights.

4.3 Targeting the customers

Lift and cumulative gain charts arise as one of the most important tools for the bank in order to decide which customers should be contacted for persuading them to keep the relationship with the institution obtaining a very simple but effective tool for ranking them. Clients with higher probability of closing their account should have higher priority for the marketing campaigns and/or special offers developed by the bank.

In our study (Figure 6), the model with the best performance is the Logistic regression where we have only a 9% of error when we cover 60% of the “1 class” predictions (versus an average of 19% across the rest of the models). This tool will be very helpful for the bank to focus its resources and select which clients to contact first.

In further steps, the bank should identify an optimal threshold in terms of the percentage of positive cases to contact based on the accuracy of the prediction model as well as in the effectiveness of its campaigns. More information regarding the utility and effectiveness of these actions is needed in order to determine how likely a certain client will show a positive response, meaning that he/she will continue the relationship with the bank after being contacted by the institution. Hence, a second layer of lift and cumulative gain charts analysis can be developed for those high ranked clients, ranking them by their probability of having a positive response expanding our original results.

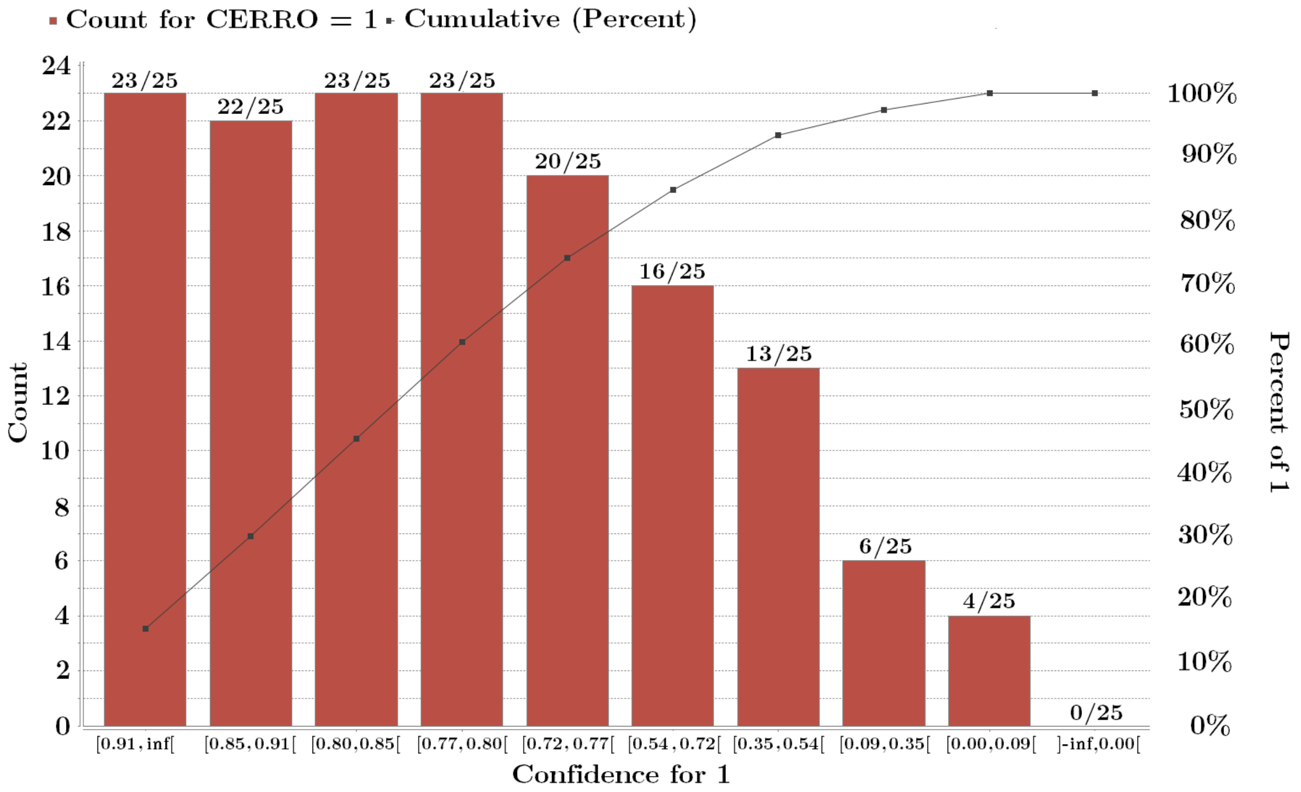

Figure 6: Lift and cumulative gain chart for the Logistic Regression Cost optimized model.

4.4 Error analysis: important insights

Finally, we wanted to understand the source of the misclassification error as well as identify the main characteristics of the more risky customers. As expected, clients who are leaving the institution tend to show a decreasing pattern in terms of their average balance as well as with the number of transactions. However, more insights were identified using the results of Table 5 and investigating the data set.

Table 5: Error analysis and misclassification interpretation (Logistic regression model).

Classification	$AVG\ SALDO_T$	$AVG\ SALDO_{T-1}$	$AVG\ TRX_T$	$AVG\ TRX_{T-1}$	$AVG\ Renta$	$AVG\ VIG$
TP	-0.072	-0.310	20.9	26.9	660.9	43.6
FN	-0.095	-0.160	61.0	42.1	1219.8	76.3
TF	0.261	0.522	62.0	54.9	743.3	73.4
FP	-0.276	-0.318	21.4	22.7	681.2	39.7

Based on the summary table, we are able to perform the following error analysis:

1. **FN**: These observations share the decreasing pattern of the TP for the average balance features ($SALDO_{T-1}$), in other words, these customers are "cleaning" their accounts in order to close them. However, they are misclassified due to the great difference in the number of transactions.

   We investigate the data set and we noticed two main customers' types:

   i) Customers who are taking out the money from their accounts in big (but few) transactions.

   ii) Customers who are spending the money from their accounts until the last day.

2. **FP**: The false positive are misclassified because of their low average balance in their accounts. Looking at the dataset, the source of the error was not clear. However, after checking these observations with the bank, we learn that these clients currently have a loan with them but there is no variable indicating this.

Therefore, we were able to identify that the FN are misclassified due to differences in the number of transactions: there are two types of customer behaviors when closing their accounts. On the other hand, the FP observations are associated with customers having a loan with the bank, information that is not included in the current dataset.

Thanks to these analysis, we obtain very important insights regarding the customers' behavior as well as a deep understanding of the predictions of the classification model. Thus, the bank should re-design their datasets in order to optimize the performance of the selected prediction models, keeping and adding the variables that have a significant predictive power.

5 Conclusions & Future Work

- The current dataset should be re-designed and optimized for applying prediction models such as the ones proposed in this study. Several variables do not have any predictive impact and can be stored in a different data set. Similarly, relevant information such like if a client currently has a loan with the bank should be included as a feature in a future implementation due to its potential significant predictive power.

- A deep understanding of the model accuracy and misclassification error has been performed, obtaining significant insights for detecting the more risky customers for the bank without "blindly" following a - limited - classification model. Two types of leaving customers are clearly identified.

- Based on the current business rules used by the bank, two implementation approaches are recommended: run the Logistic regression model on-line (automatically) versus a simple decision tree model that will be easier to understand and apply for the workers. Therefore, the trade-off between the complexity of the model and easy to follow decision rules should be analyzed by the institution in case of implementation.

- The next step of the project should be to include information regarding the effectiveness of the marketing campaigns/products offered to the clients for persuading them: using that information, we would be able to formally analyze the threshold for contacting relevant clients, taking into account the budget limitation. More information regarding the effectiveness of the offers/MKT campaigns is needed in order to determine the optimal "number of clients to contact" thresholds of the models.

- The study was successful: the group tackled a real-world problem using different approaches from the ones applied during the course using tools like lift charts and the confusion matrix optimization approach, obtaining relevant results for the bank without transforming the project into a simple "horse-race" between models' accuracies.

- Real life implementation: The next challenge of the project consists of implementing it in the ABC bank office in Chile. After talking with the ABC bank and sharing our results with them, an implementation of our model would be tested during this summer for checking its validity and usefulness to the institution.

6 References

1. Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. Journal of business research, 47(3), 191-207.

2. Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. Journal of the academy of marketing science, 26(1), 31-44.

3. Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: an examination of the differences between switchers and stayers. Journal of marketing, 64(3), 65-87.

4. Reichheld, F. F., & Sasser, J. W. (1990). Zero defections: Quality comes to services. Harvard business review, 68(5), 105-111.

5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

7 Appendix

7.1 Data set inconsistencies

In our analysis we discovered a discrepancy between the two datasets we had available to us. The second dataset showed much larger fluctuations in the amounts for each account on a month-to-month basis. For example, the average difference in the account value from month $T-4$ to $T-3$ for dataset 1 is around $650,000$ Chilean pesos with a median difference around $100,000$. For dataset 2 the mean difference for the same months is above 4 million pesos with a median around 2 million. The same pattern occurs for the differences in the other months with dataset 2 seeing differences around one order of magnitude larger than those seen in dataset 1. Because of this we may restrict our analysis to dataset 1 (which still contains around 1200 observations). Alternatively, we may perform two analyses: one with both datasets, and one focused only on dataset 1 and compare the results.

We can notice that the average values associated with the Bank variables group are significantly different from the ones detected inside the dataset 1, therefore, we suspect that this dataset is not completely consistent/correct due to the fact that the ABC bank has just implemented a new data-warehouse software and there exist the possibility that some errors are being committed in one of the steps when generating the data set. Therefore, we will perform two analysis in our project: one including only the first dataset (1248 entries), and a second one including all the observations in order to identify possible differences.

Table 6: ABC Bank data set number 2 Summary

Attribute	Variable	Type	Mean/Mode	Range and Frequencies	Missing
Id	Cod_Cliente [ID]	integer	avg = 2652.500 +/- 810.455	[1249.000 ; 4055.000]	0.0
Label	CERRO [LABEL]	binary	mode = 0.0 (1610) least = 1.0 (1196)	0.0 (1610)- 1.0 (1196)	1.0
Regular	COD_OFI	integer	avg = 59.513 +/- 41.640	[10.000 ; 247.000]	2.0
Regular	COM	integer	avg = 107.041 +/- 74.005	[1.000 ; 324.000]	6.0
Regular	ED [years]	integer	avg = 65.903 +/- 8.293	[22.000 ; 86.000]	3.0
Regular	SX [M male,F female]	binary	mode = M (2030) least = F (775)	M (2030)- F (775)	2.0
Regular	NIV_EDUC	nominal	mode = MED (1308) least = BAS (48)	UNV (950), MED (1308), TEC(499), BAS (48), EUN (0)	2.0
Regular	RENTA [M CLP/month]	integer	avg = 865.371 +/- 696.200	[110.000 ; 4451.000]	0.0
Regular	E_CIVIL	nominal	mode = CAS (2262), least = SEP (128)	CAS (2262), SOL (217), VIU (200) SEP (128)	0.0
Regular	VIG [months]	integer	avg = 115.899 +/- 91.680	[4.000 ; 366.000]	3.0
Regular	TRX_T [N°/month]	integer	avg = 31.441+/- 29.382	[0.000 ; 176.000]	8.0
Regular	TRX_T-1 [N°/month]	integer	avg = 34.354 +/- 29.705	[2.000 ; 142.000]	6.0
Regular	TRX_T-2 [N°/month]	integer	avg = 36.844 +/- 26.303	[3.000 ; 145.000]	10.0
Regular	SALDO_T-4 [$\overline{CLP}$/month]	numeric	avg = 4001844.649 +/- 7664486.178	[0.000 ; 48073799.250]	7.0
Regular	SALDO_T-3 [$\overline{CLP}$/month]	numeric	avg = 3063673.864 +/- 4829447.843	[0.000 ; 48073799.250]	6.0
Regular	SALDO_T-2 [$\overline{CLP}$/month]	numeric	avg = 3763697.893 +/- 6243180.834	[-757777.710 ; 32554700.570]	7.0
Regular	SALDO_T-1 [$\overline{CLP}$/month]	numeric	avg = 9004005.739 +/- 12129580.175	[-20422.290 ; 86020611.430]	1.0
Regular	SALDO_T [$\overline{CLP}$/month]	numeric	avg = 10899091.682 +/- 16616014.102	[-36911.910;86020611.430]	4.0

The group contacted the ABC bank and the institution realized that they are having a problem with their new IT platform for creating the datasets. Therefore, thanks to our preliminary analysis, they are now solving this very critical issue.

7.2 Performance metrics

- **Confusion matrix**

  Based on the predicted values and the real value of an observation, we can define the concepts of true positive, true negative, false positive, and false negative:

  - True Positive (TP): a positive value that is classified as positive.
  - True Negative (TN): a negative value that is classified as negative.

Table 7: Confusion matrix

Current/Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

- False Positive (FP): a negative value that is classified as positive.
- False Negative (FN): a positive value that is classified as negative.

- **Accuracy**

  Proportion of true values over the total classifications:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

  It can be misleading due to the fact that it does not take into account the classification error.

- **Precision** Probability of, given that an observation was classified as positive, it is a real positive:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

  This metric is very useful when we need to select models based on the performance of False positives.

- **Recall**

  Probability of classifying a positive observation as positive:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

  In general, we want to obtain very similar values for the Precision and Recall metrics. That's why several models try to find a threshold where these two values are close enough. Depending on the importance of each error type and their associated costs, this metric may not be the most suitable one.

- **Lift**

  Measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model.

  A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values: target response divided by average response.

- **F-Measure**

  Known also as F_1 score, it consists of one measure of the accuracy of the model. It considers the precision and recall of a model in order to compute its "score". This score can be interpreted as a weighted average of the Precision and Recall values where F_1 reaches its best value at 1 and the worst at 0.

It is defined by the following formula:

$$\frac{2 * Recall * Precision}{Recall + Precision} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

However, the F-Measure does not take into account the ratio of the TN values. Thus, it has to be complemented with other metrics in order to evaluate the overall performance of the predictive model.

7.3 ABC bank current prediction system and business rules

ABC bank has a reactive procedure for "detecting" the leaving clients of the institution. It is based on a series of alert and business rules that are activated every time a client violates a specific (or a set) condition(s). The main rules are the following:

1. **Frozen account**: If the customer has not performed any operation for the last 30 days, an alert is reported in the human resources area and one of the workers from the marketing area will analyze the case in order to check if it is needed to contact the client.
2. **Low usage of checks and checkbooks**: if the customer has not bought a checkbook for the last 3 months, an alert is triggered.
3. **Devolution of products**: Any client returning products such as credit and/or debit cards is contacted by an executive from the marketing area in order to check the reason and offer him/her a relevant offer for keeping them (adding extra benefits, points, or increasing the credit of the account).

If a client "violates" one or more of these business rules, they are immediately classified as higher risk clients (positive label). The current analysis from the bank indicates that, monthly, 30% of the total portfolio of clients is classified as risky. Because of this simple and not accurate prediction model/rule, there is a significant overload of work on the executives in charge of the retention policies of the bank knowing that in practice, there is not enough personnel for covering more than 5% of the total clients' portfolio.

Nowadays, the accuracy of this system is below 40%. Furthermore, during the last semester, the average accuracy obtained by the current implementation was around 10%, leading to an extremely poor performance with respect to the clients' retention policies.

For completeness, after conversations with the bank, the institution indicated that they used to run a simple regression model in Excel for "predicting" the customers' attrition, however, the personnel that was in charge of its implementation was relocated in a different area, leading to the end of the project.

7.4 Preliminary Data Exploration: Extra analysis

Although the response variable shows limited correlation to the predictor variables, even slightly more complex models such as decision trees or support vector machines with higher dimensionality may capture relationships that are not easily seen in the correlation heat-map.

The lack of strong simple correlations means that simple linear models will probably not perform especially well with the variables provided. Some variables such as gender (SX) show essentially zero correlation. This is further exhibited in graph below. Same pattern can be seen with the educational level and civil state features (Figure 8).

Following the previous logic, we can see in Figure 9 how the average values and the variability of the observations are significantly different depending on the $CERRO$ label: clients who are leaving the institution ($CERRO = 1$) tend to have a compact distribution around 0 Chilean pesos before closing their accounts as well as a lower amount of transactions in comparison to clients who are likely to continue the contract with the institution, with higher variability and mean values.

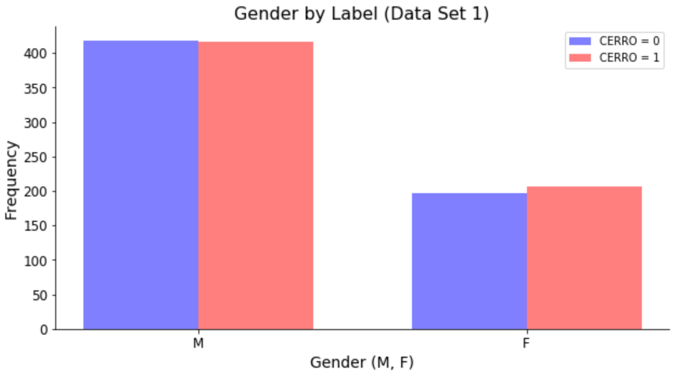

Figure 7: Gender by label

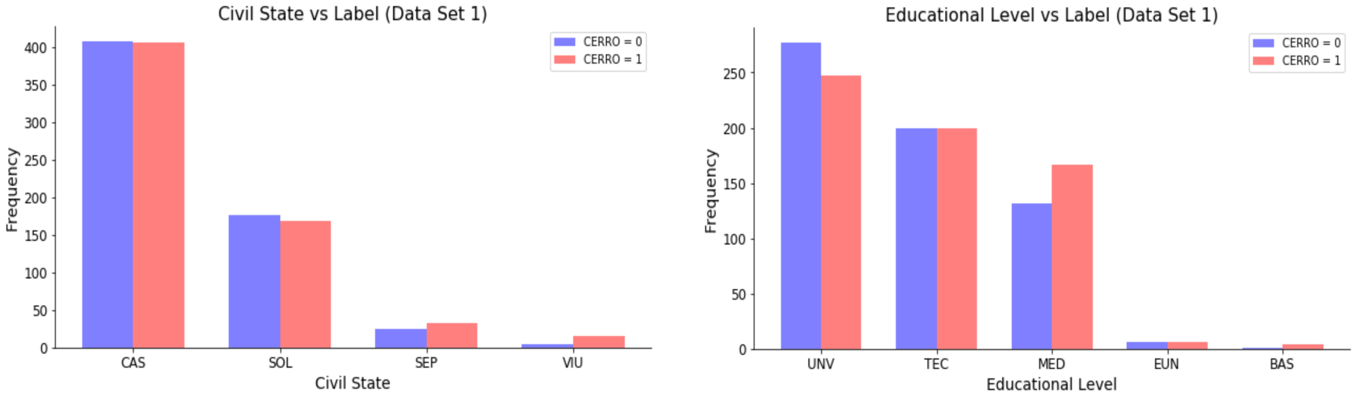

Figure 8: Civil state and Educational level by label

Therefore, we will focus our attention in these clear differences between clients in order to apply relevant feature engineering approaches such as generating new features like differences between $SALDOS$ and the variability of the distributions.

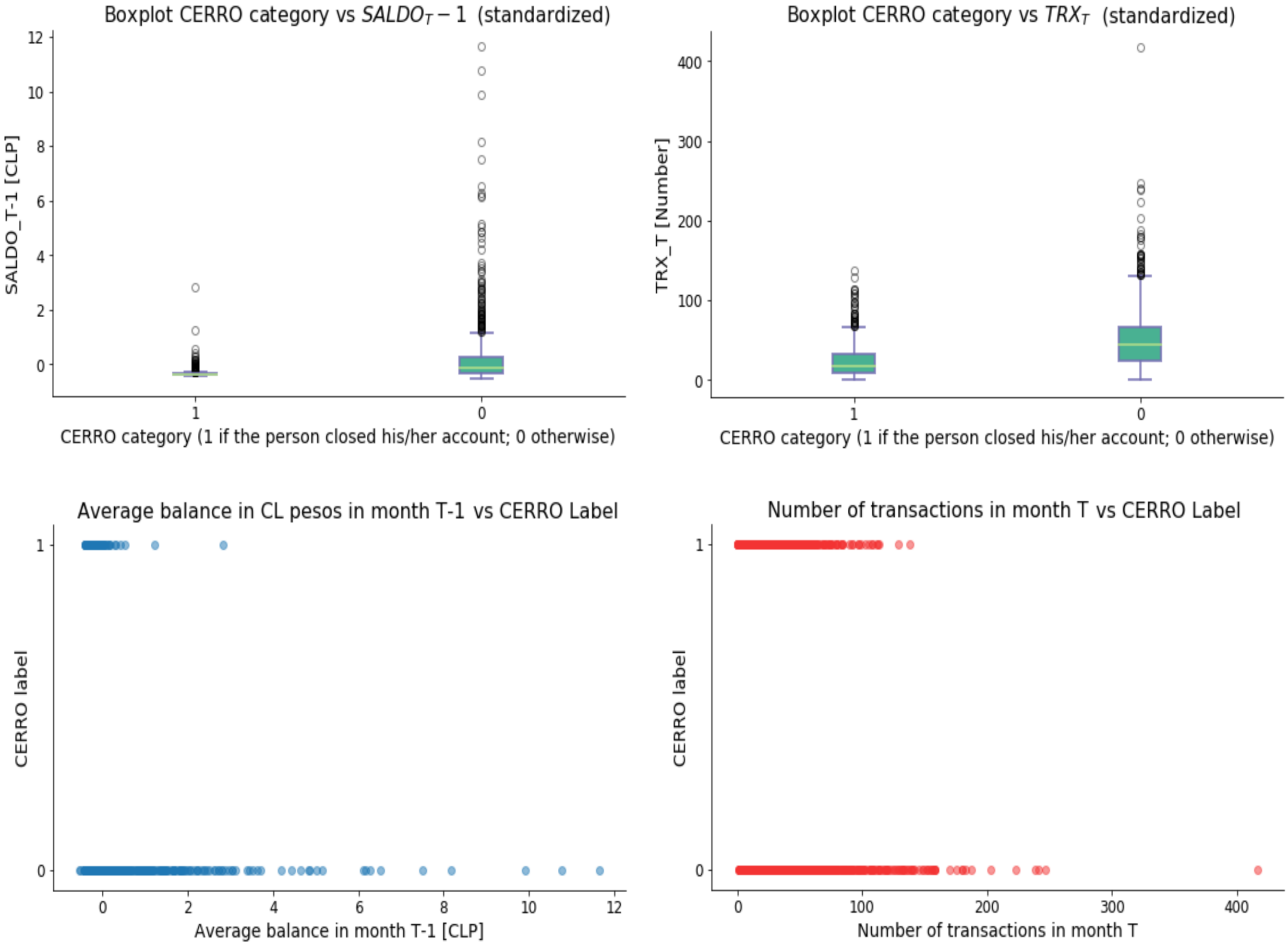

Figure 9: $Saldo_{T-1}$ and TRX_T boxplots and scatter plots comparison by label