Forecasting Credit Card Attrition using Machine Learning Models

Carlos Alvaro Rico-Poveda and Ixent Galpin

Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá, Colombia
{carlosa.ricop,ixent}@utadeo.edu.co

Abstract. The objective of this work is the implementation and evaluation of Machine Learning models to identify which customers want to cancel their credit cards. The banking industry uses this technology to obtain more reliable predictions when identifying opportunities for purchase, investment, or fraud. These models can be adapted independently, by recognizing patterns and algorithms based on mathematical calculations.
Four models (LightGBM, XGBoost, Random Forest and Logistic Regression) were implemented and evaluated to predict, using data about customers and products held pertaining to a bank in Colombia, the likelihood of customers canceling their credit cards. By analyzing the ROC curves using the AUC metric, it is concluded that, of the selected models, the model chosen for deployment would be LightGBM, since it was the one that performed best in the experiments conducted. Furthermore, the "Score Acierta" variable, a customer rating provided by the Colombian credit rating agency, was found to be the most discriminating in prediction models.

Keywords: Machine Learning · Supervised Learning · Attrition · LightGBM · XGBoost · Random Forest · Logistic Regression

1 Introduction

The materialization of the constant risk of losing customers in banking entities has been a considerable cause in the decrease of products and income for banks [1]. This is due to factors that influence on the condition of a product, benefits in interest rates and competition. In recent years, the market for new credit card customers has shrunk considerably. This means that banks are forced to increase their client base mainly by attracting users from other entities.

Their tactic has been to offer low rates through portfolio purchases, expecting the balance to remain with the bank after the interest converts to the normal rate. However, many clients, hoping to maintain an attractive interest rate, transfer their balance from one card to another before the rate returns to normal. Likewise, there is an improvement in the condition of credit such as refinancing, lower monthly payments and consolidation.

As a result, the cost of acquiring new users is increasing. This has created a major shift in banks commercialization strategies, as many banks are focusing more on customer retention because it costs less to maintain an existing customer than to acquire a new one. In this way, a long-term customer tends to consume more and is less sensitive to competition. One way to improve retention is to take appropriate action towards customers at risk of loss or leakage.

Situations like these provide the possibility of creating a model through information sources and data analysis that predict the act of reducing and ending the use of a product or service after it has been activated. Attrition models, also known as retention or churn models, predict the probability that a customer wants to cancel some or all the products with the entity through patterns or sequences of financial activity.

The retention of clients is quite important since if they do not retain their loyalty their acquisition cost is not compensated, and the expected profitability will not be achieved. To determine it an analysis is made during a specific period of time where Equation 1 can be used, which shows the retention rate [3]. This formula indicates the ability to retain customers, points to the behavior and retention strategy on which loyalty campaigns should focus on the specific target.

$$\text{Retention rate} = \left(1 - \frac{\text{Customers lost in a year}}{\text{Customer in portfolio}}\right) \cdot 100\% \tag{1}$$

Because these cancellations have such a significant impact on profitability, many companies are making these models the focus of loyalty strategies. To obtain the expected results, it is of great importance to have a standardized database where accurate and truthful information can be acquired. The study was conducted in a Colombian financial institution, which did not have this new technology to develop the model and mitigate the problem. This has been a point that has made it difficult to develop the project while organizing a complete and functional data set that can yield results and the reasons for leakage.

As mentioned above, the bank did not have the necessary data to mitigate the cancellation or surrender of financial products. The areas in charge carried out a comparison of flat files, and the result obtained was a spreadsheet with the data of the clients that were likely to abandon it. Afterwards, it was sent to the marketing area to implement loyalty campaigns. It was a purely manual process and without reliability in the information.

For this reason, the need arises to make an attrition model that helps to focus the data and understand the situation of the bank's customer churn, either due to the total and partial cancellation of the products, or inactivity for a long period of time. Having such a model will allow financial institutions to define retention and create campaigns to carry out commercial actions in a "personalized" way.

In this work, four machine learning models (LightGBM, XGBoost, Random Forest, and Logistic Regression) are implemented and evaluated to predict through customer data and their products the likelihood of them canceling their credit cards.

This paper has the following structure. Section 2 discusses existing work addressing the problem of attrition in banks. Subsequently, a subset of the phases of the CRISP-DM methodology [20] are applied. Section 3 presents the business and data understanding phases. Section 4 describes the preparation of the data and its analysis. Section 5, describes the implementation of machine learning models. Section 6 presents the results and comparison of the models, identifying the most suitable model for deployment. Section 7 presents conclusions and future work.

2 Related Work

Van *et al.* [16] is about a study carried out in England in 2003. The bank from which they obtained the data requested to remain anonymous to protect its customers. The research used the Cox proportional hazard method, also known as Cox regression [13]. The data was taken from clients with credits and insurance. The results obtained allow us to observe that there are two critical periods of attrition, the initial one in the first years after becoming a client and the secondary one after being a client for more than 20 years. As a conclusion, it was possible to predict that to reduce the rates of attrition banks must offer incentives for their customers to stay.

In [14] a multidimensional analysis was carried out of the satisfaction customers have with the financial services of their banks. The survey was conducted with consumers from five different banks in the US. Using the factor analysis technique, the authors were able to recognize four dimensions to measure customer satisfaction and thus reduce withdrawal or cancellation of their products with the banks. The dimensions were: Personal considerations (efficiency and attitude of employees, responsibility and competence), financial considerations (interest rates, handling fee, etc.), environmental considerations (location of branches, appearance of branches, etc.) and convenience of service (number of ATMs and service in branches).

It is noted that attrition prediction became the main focus of banks in China. Xie *et al.* [21] proposes a learning method called Improve Balanced Random Forests (IBRF), which is applied to a data set from a Chinese bank. As a result, it was found that it substantially improved the accuracy of the prediction of clients with high attrition likelihood compared to other machine learning algorithms such as neural networks, decision trees, support vector machines, and even other random forests algorithms.

Verbeke *et al.* [19] indicates that customer attrition models must take into account two fundamental components: the ability to predict and the ability to understand. The union of these two competencies enables the development of safe and solid strategies. It is also indicated that survival analysis and logistic regression are the most used statistical methods for this type of problems.

According to [7], for banks prediction of attrition is a subject of utmost importance. To solve this problem, the authors have proposed using classification and regression trees to obtain a better rate of categorization through the patterns

of customers who left. This knowledge was then used to assign an attrition potential rating to current customers.

In China, to increase their profits from continuing operations and improve core competitiveness, banks must avoid losing customers while acquiring new ones. In the paper [9] the prediction of customer attrition for commercial banks is analyzed using Support Vector Machines (SVMs) and uses a random sampling method to improve the results of the model, taking into account the imbalance characteristics of the customer data sets. The results show that this method can effectively improve the prediction accuracy of the selected model.

In the United Kingdom in 2014 [18] an investigation was carried out with data from a financial institution, where they showed that efficient use of information helps to predict customer attrition. Using an orthogonal polynomial approximation analysis to obtain a group of unobservable variables, which they then used as input data in a probit hazard rate model. The results obtained showed that the use of this information improves the predictive power.

Zhao *et al.* [22] proposes a new framework based on clustering and classification to help Chinese banks with the prediction of customers who will cancel their products. The proposed method is supported by the result of data exploration: it groups the characteristics of the customer and takes a decision with a classifier.

For the development of this project, data analysis was carried out using four recent algorithmic models, based on the implementation of the latest developments of each one, in order to predict a variable that identifies which customers want to cancel their credit cards. Furthermore, two techniques were implemented for the selection of definitive variables used to run the models.

3 Business and Data Understanding

The project was carried out in a Colombian bank with more than 50 years of experience in the market. This entity has more than 2.5 million clients nationwide and a wide portfolio of banking products, being one of the most important when it comes to banking Colombians. According to the economic activity of the country, the bank belongs to the tertiary sector: banking services, registering a 27% growth in loan portfolios, and increasing deposits by 29%. All these increases place it among the entities with the best growth in the bank-financial system, according to the Superintendence of Finance.

The analysis was carried out in the Customer Relationship Management (CRM) department and Analytics department, an area where all the information about the customer and their products is stored. With this data, models are developed that allow the attraction and retention of customers. The bank has its own cards with traditional franchises (generic cards and non-generic cards). The non-generic ones are associated with more than 30 brands across the country. Table 1 shows the distribution of the cards in the bank.

Table 1: Credit card distribution.

Credit Card Type	**Percentage**
Generic	25%
Non Generic	75%

The type of cards chosen to carry out the project were generic. Although these make up a lower percentage of the total distribution, more complete data about them is available, such as:

- Number of products in the bank with their respective balances and credit limits.
- Data such as the scores of the risk models to obtain the Recency, Frequency, Monetary (RFM) model [2].

As non-generic cards are from private companies, their policies prevent sharing relevant information with the bank. For this reason, this type of card is not viable for the study to be carried out. The models made from the analytics area have the client as their main focus. Therefore, the data for this study will not be taken from each credit card but from the customers to which they belong.

Table 2: Number of cards per customer.

Credit cards per customer	**Number of customers**
1	175,553
2	41,347
3	688
4 or more	39

The data obtained in Table 2 refers to the number of active cards that each customer has with the bank. In this way, we obtain the clients who are the focus of this study. An analysis was necessary to verify the cancellation percentage of credit cards at the bank. Table 3 shows how the behavior of the attrition in the bank was the quarter prior to the creation of the data set for the study. According to the analysis, it was shown that the percentage was steady but not especially high. However, this does imply a concern for the credit card area since the recurring cancellation of this product is being made on a monthly basis by customers.

4 Data Preparation and Analysis

The study uses a data set that contains a sample of approximately 220,000 records (active customers with generic credit cards as of July 2019). From this information, the financial behavior of the clients in the previous year (August

Table 3: Percentage of Attrition by Period.

Period	Percentage of Attrition
April 2019	7.46
May 2019	8.18
June 2019	7.85

2018 - July 2019) was obtained. In this data set there is detailed information that the bank obtains from different sources to analyze the use of each customer's credit cards, such as

- The balance, the invoicing, the credit card limit, product usage and its cut-off dates; provided by the Analytics and CRM area
- The scores provided by the risk area models.
- The financial information of each client at the national level provided by the DataCrédito[1] credit rating agency.

This data is complemented with information on loans, current accounts, long terms deposits, mortgage loans, revolving loans, and demographic data (age, social stratification, gender, and economic activity).

There were three main data sources: the analytical area has its information in an SQL Server 2017 database, the risk area that has its data stored in an Oracle 11g database, and finally an Excel 2017 file sent from DataCrédito every three months. As can be seen in the Fig. 1 SAS was used to integrate the sources and subsequently a .csv file was generated with the final data set.

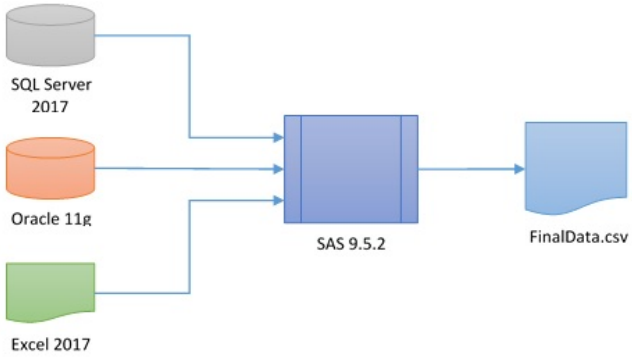


Fig. 1: Information Sources Diagram

At the same time, filters were made which allow obtaining cleaner and more relevant information for the case study. Therefore, the data set only contains credit cards that are active, that are not covered, nor provided by an SME or

[1] https://www.datacredito.com.co/

any type of company In addition to these filters, it is important to emphasize that it was not possible to differentiate between voluntary attrition and involuntary attrition since the bank did not have the relevant information to obtain this differentiation.

This results in the integrated data set, which is then analysed in order to determine the importance of the variables in terms of customer attrition, which will be used in the prediction models. For this purpose, variables are categorised thus:

- *Bank variables* comprise those that capture the financial habits of customers, with the aim of identifying changes in the transactional behavior of users. Such variables include: the amounts of products per customer, balances (average and monthly) of the liability products, the credit limit, the number and value of credit card transactions, and number of months as a customer.
- *Sociodemographic variables* are the measurable data that correspond to the general characteristics related to the personal aspects of the clients, among these are: age, economic activity, gender, occupation, and social stratification
- The *target variable* is the variable to be predicted that indicates whether a client canceled their credit card or not.

The subsequent step is to build a correlation matrix, using all the continuous variables from the data set. This matrix, shown in Fig. 2, provided an understanding of the most positively and negatively correlated variables to determine which data were best outlined to be part of the data set used in the model. According to Fig. 2 it can be observed that there is a high positive correlation between: the data of six and twelve months, the balance and usage of the credit card and the billing with the number of transactions in twelve months. Likewise, there is a negative correlation noted between the balance registered in other banks without a mortgage with the balance of all products with a mortgage.

Moreover, the analysis performed was made on categorical and continuous variables such as age, where it was observed in Fig. 3a that most credit card customers are in a range between 45 and 50 years. In [15] age is considered as a discriminatory variable. This study concludes that older people have more stable preferences, and therefore, have less tendency to switch from one financial institution to another. On the other hand, young people are more unstable in their preferences, increasing their tendency to change the financial institution.

The Fig. 3b analyzes the client's loyalty with the entity in months. In [16] It can be concluded that clients who have been a longer time at the bank have less tendency to leave.

Likewise, Fig. 3c indicates the gender distribution of the clients (female and male) against the response variable. There is a similarity observed in the percentage of canceled credit cards of both genders. However, it can be concluded that both genders behave differently when managing money but their participation in the banking market has the same proportion.

Fig. 3d shows the economic activity of each client. In this variable, the following fields were analyzed: independent workers, employees, and pensioners. It was observed that there is a scale in the cancellation of credit cards, where

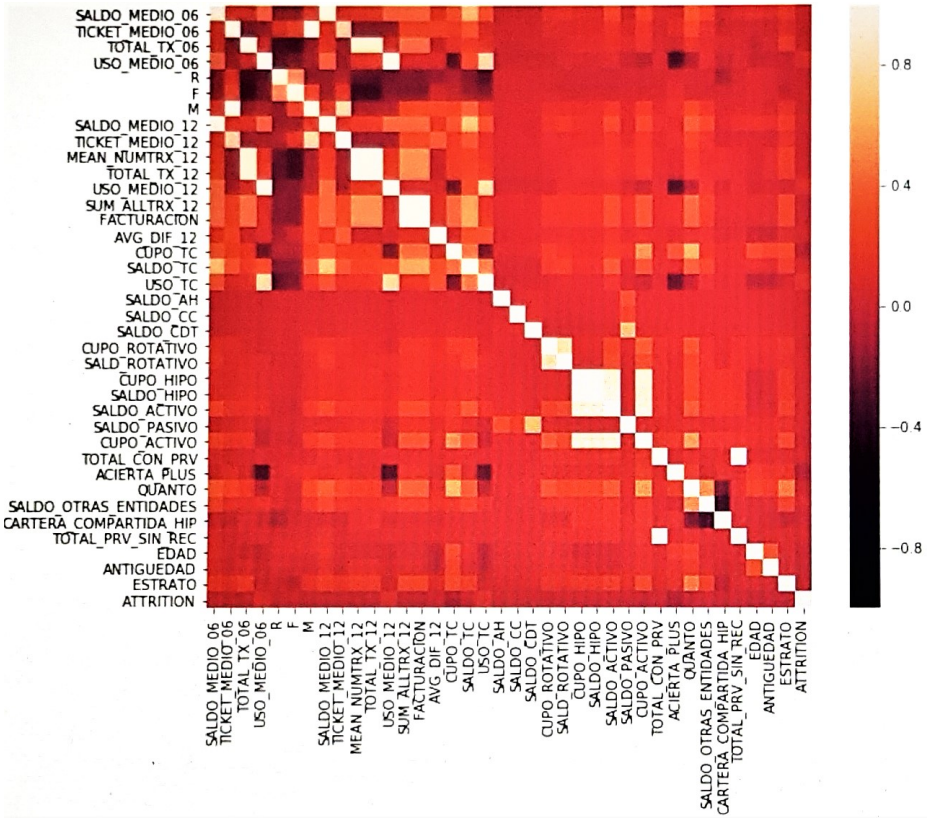


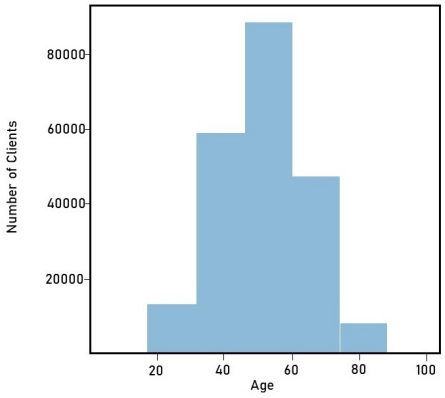
Fig. 2: Correlation Matrix

self-employed workers are the ones who most register this process. This is due to different factors that influence their economic activity such as having one or more products with the same financial institution and not being able to fulfill their obligations. Likewise, due to the nature of their work they are a more sensitive target to campaigns or offers of other banking entities.

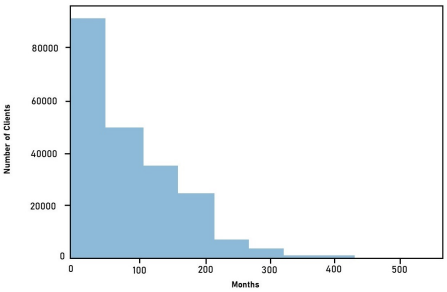
In Fig. 4 the analysis was performed on the banking variable Quanto, which represents the estimate of customer income in DataCrédito. It can be seen that there is a difference and that customers who cancel tend to have lower income.

After carrying out the exploration on some variables, the analysis made was conducted to know the number of null values to determine if any variables got discarded for having more than 20% of null records. However, as can be seen in Table 4, no variable exceeds this threshold; therefore none were discarded. The table shows the six variables with the highest percentage of null records.

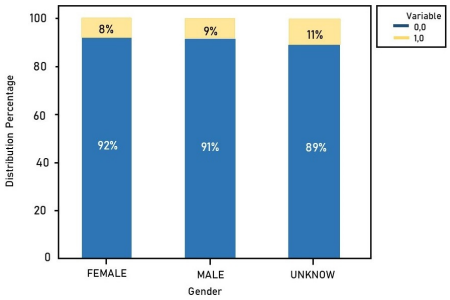
As a final part of the preparation, the selection of the variables to be used with the models is made to identify which ones are the best qualified. Two techniques were considered for the selection of variables. The first technique is



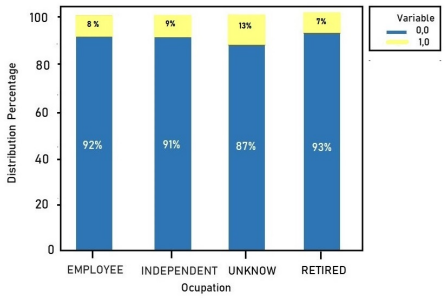
(a) Histogram showing distribution of client ages



(b) Length of relationship with client in months



(c) Gender vs Attrition Diagram



(d) Occupation vs Attrition Diagram

Fig. 3: Client attrition characteristics

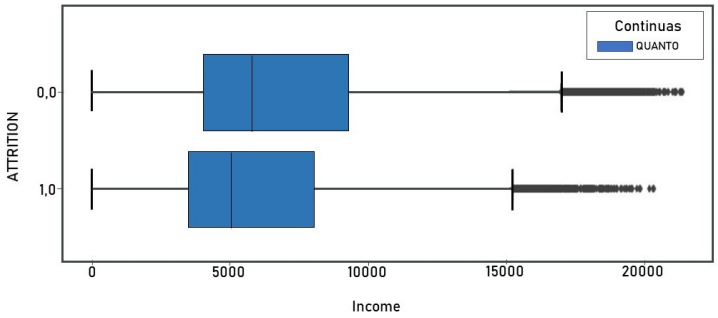


Fig. 4: Bloxpot Quanto vs Attrition.

Kolmogorov-Smirnov statistics (KS) [17]:

$$KS = \max_{a \in [L,H]} |F_{m,BAD}(a) - F_{n,GOOD}(a)| \tag{2}$$

Table 4: Percentage of Nulls by variable.

Variables	**Percentage of Nulls**
Total transaction last 6 months	12.04
Avg. ticket Transactions last 6 months	12.02
Avg. days between purchases or last 12 months	12.00
Shared portfolio with Mortgage 3 months	5.53
Mean transactions last 12 months	4.79
Avg. ticket Transactions last 12 months	4.79

where the main idea is to find the vertical difference in the distributions, then both functions are subtracted from the accumulated distribution, the maximum difference is obtained to see how different both distributions can be and the absolute value is used so that it does not matter that $F(x)$ has a negative sign.

The second technique is the Information Value (IV) [12]:

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right) \tag{3}$$

This technique was the one that allowed the analysis of all the variables, since with it the number of bins could be modified and thus all the records of these variables were analyzed. Three functions were created to handle the different types of data in the data base. In Fig. 5 the results obtained with the score given to each of the variables are observed, where the nine best qualified were chosen. With this result, the final data set obtained was used in the models.

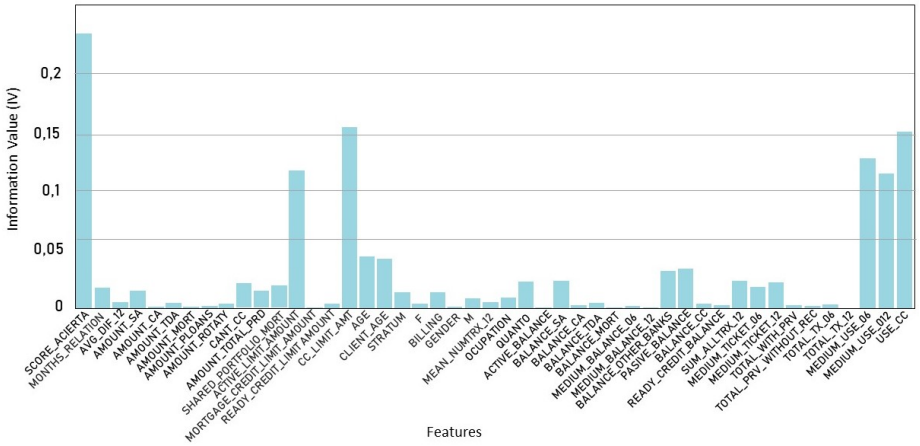


Fig. 5: Information Value

5 Modeling and evaluation

Machine learning techniques are data approaches based on prediction and the construction of analytical models, with the aim of identifying patterns that can reduce the risk of leakage and have reliable results.

The study was conducted entirely in the Anaconda 2020.02 tool using the Jupyter Notebook IDE version 6.03 and Python version 3.8.3. The libraries with the versions used for the execution of the models are listed in Table 5.

Table 5: libraries and versions.

Library	**Version**
woe	0.1.4
lightGBM	2.3.1
sklearn	0.23.2
XGBoost	1.2.0

For the analysis of the final data set, the following algorithms were taken into account in order to obtain the predictive model that best indicates which customers are going to cancel their credit cards and thus carry out marketing campaigns focused on customer's needs.

5.1 Random forests Model

Random Forests (RF) is an ensemble algorithm that uses decision trees as base classifiers, each contributing one vote for assigning the most frequent class to the input vector. The RF increases the diversity of decision trees by growing them from different subsets of data [4]. Unlike decision trees, RF do not test the entire feature space when deciding how to divide the tree. Only one random subset of the space features is considered in each division, through a two-stage process [8]:

- A considerable number of decision trees are generated with the data set. Each tree contains a random subset of m variables (predictors) such that $m < M$ (where M = total predictors).
- Each tree grows to its maximum extent to obtain better results in the model predictions.

This algorithm allows the analysis of the understanding of the model through an input-output process, in order to obtain a number of variables to be selected in the participation of each tree. To find the parameters, it is necessary to leave one of the two variables with the value determined by the algorithm and the other will increase, in order to obtain the possible iterations.

5.2 XGBoost Model

The Extreme Gradient Boosting (XGBoost) [6] algorithm is a technique that uses weak tree models in order to take these results to generate a stronger one, with better predictive power and greater stability in the results, implying: a loss function to optimize an algorithm based on learning to obtain the results and finally a model that minimizes the loss function.

The XGBoost algorithm is obtained through an initial tree $F0$ to predict the target variable y, which adjusts to the error of the previous step. The results of $F0$ and $h1$ are combined to obtain the tree $F1$. This process is iterative until the error is minimized as much as possible in the following way [10]:

$$\mathrm{Fm}(x) < -\mathrm{Fm} - 1(x) + \mathrm{hm}(x) \tag{4}$$

Using decision trees with low bias and high variance, this algorithm allows obtaining a real value score, independent of its classification and regression. This has the purpose of potentiating the results through the sequential process of the data with a loss function, which allows minimizing the error of iteration after iteration to become categories and to be able to build the next classifier.

5.3 Logistic Regression Model

Logistic regression [5] is a model that allows one to analyze whether or not one variable depends on the other, in order to minimize the sum of the error boxes, where the answers can only be two values: presence with probability P and / or absence with probability $1 - P$. The objective of this model is to analyze the probability of occurrence through the level of the same values, as well as to determine the variable that best fits or describes the relationship between the regressive variable and response variables.

The objective is to determine whether a variable or a set of explanatory variables have a coefficient equal to zero, in order to determine the parameters that must be estimated from the data to obtain the global adjustment of the model.

5.4 LightGBM Model

The Light Gradient Boosting Machine (LightGBM) [11], model uses the Gradient Boosting technique; in this way the trees are built in a more agile and sequential way and each tree that is added serves to refine the previous prediction. In other words, it starts with a constant value and each new tree is trained to predict the error in the sum of all the predictions of the previous trees. Once the process is finished, the predictions are calculated by adding the results of all the trees that were built. In this way every time a new tree is added it focuses on the samples of the model that are performing badly to improve them.

Two cross-validation methods were used, stratified k-Fold, which is an improved variant of k-fold, since when dividing the data, it keeps the classes balanced, which is very important. The classification is a response variable, and

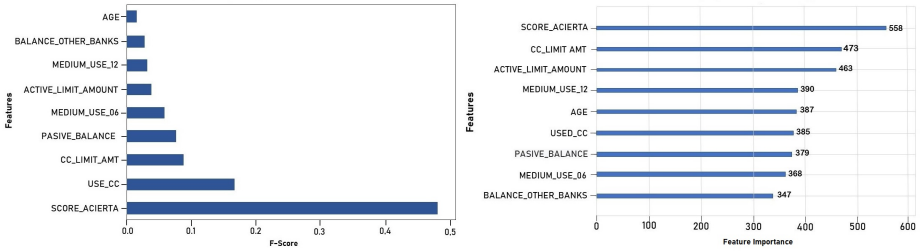
samples with only 1 o only 0, will not be obtained. This method was used for the XGBoost, Random Forest and Logistic Regression models, in which five folds were used.

For the LightGBM model, lightgbm.cv was used, which is an option for this algorithm, the idea is to obtain more information about the estimation of the generalization error by evaluating the performance in a KFold division with fixed parameters.

6 Results and Comparison of Models

As the last part of the study, the LightGBM, XGBoost, Random Forest and Logistic Regression models were applied to the training and test databases that were obtained by dividing the final data set with the selected variables.

The graphs shown in Fig. 6a and Fig. 6b show the importance of the variables for two of the four models executed. In other words, these are the most relevant predictors. In both the Random Forest and LightGBM models, it was observed that the "Score Acierta" variable is the one that has the most weight. This variable discriminates best and that contributes the most to the models and refers to the Customer rating in DataCrédito.

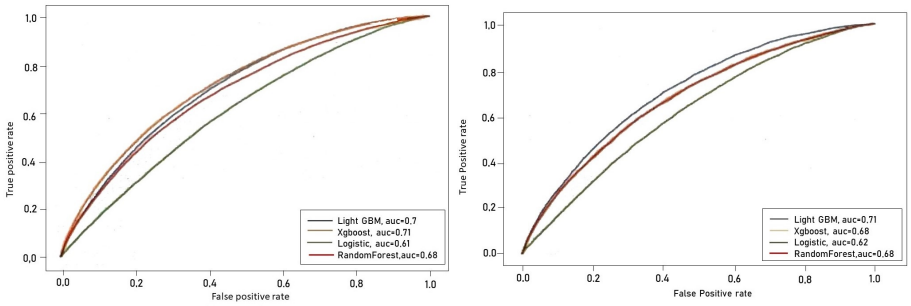


(a) Random Forests (b) LightGBM

Fig. 6: Feature Importance

On the other hand, the Receiver Operating Characteristics (ROC) curves were obtained with the area under curve (AUC) metric first for the training base. In Fig. 7a the results obtained with this data set are observed where the XGBoost model was the one with the best prediction, being slightly superior to the LightGBM model, but to a greater extent than the Random Forest and logistic regression models.

However, when performing the same procedure on the test data base, a change was observed Fig. 7b. as the LightGBM model had an AUC greater than that of the Random Forest, XGBoost and Logistic Regression models.



(a) Training data (b) Test data

Fig. 7: Receiving Operator Characteristic (ROC) curves

The ROC curves and the AUC metric obtained give the technical foundation to compare the obtained models, since their AUC on the test data was higher; the LightGBM model was chosen.

7 Conclusions and future work

Attrition prediction models are important to reduce the cancellation of credit card products and analyze which customers have a greater tendency to cancel. This work was carried out through a strategic segmentation focusing on the client to generate campaigns according to the profile of each one, to understand and anticipate their behavior. Likewise, productive resources were focused on high-value groups, as it is cheaper to retain a client than attract a new one.

To generate a strengthening of the relationship between the client and the banking institution, understanding and anticipating their needs; also make the customer less sensitive to the competition to identify and quantify the impact of a retention program to understand customer output.

Future work could usefully include: (1) The implementation of the selected with a more updated, standardized database and a wider time window; (2) Evaluation of other algorithms such as neural networks, genetic algorithms and SVMs to compare their performance with the results obtained; (3) Evaluation of attrition models for other products handled by the bank.

References

1. Amieva-Huerta, J., Urriza González, B.: Crisis Bancarias: causas, costos, duración, efectos y opciones de política. CEPAL (2000)
2. Armstrong, G., Adam, S., Denize, S., Kotler, P.: Principles of marketing. Pearson Australia (2014)
3. Athanassopoulos, A.D.: Customer satisfaction cues to support market segmentation and explain switching behavior. J. of Business Res. **47**(3), 191–207 (2000)

4. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
5. Castaño, H.F., Ramírez, F.O.P.: El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito. Revista Ingenierías Universidad de Medellín **4**(6), 55–75 (2005)
6. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: Xgboost: extreme gradient boosting. R package version 0.4-2 pp. 1–4 (2015)
7. Chitra, K., Subashini, B.: Customer retention in banking sector using predictive data mining technique. In: ICIT 2011 The 5th International Conference on Information Technology (2011)
8. Dfuf, I.A.: Análisis de Sensibilidad Mediante Random Forest. Ph.D. thesis, Universidad Politécnica de Madrid (2018)
9. He, B., Shi, Y., Wan, Q., Zhao, X.: Prediction of customer attrition of commercial banks based on svm model. Procedia Computer Science **31**, 423–430 (2014)
10. Jesús, E.Z.J., Gerencia, C.: Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito application of random forest and xgboost algorithms based on a credit card applications database. IngenIería InvestIgacIón y tecnología **21**(3), 1–16 (2020)
11. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems. pp. 3146–3154 (2017)
12. Krishnan, S.: Weight of evidence and information value using python (2018), `https://medium.com/@sundarstyles89/weight-of-evidence-and-information-value-using-python-6f05072e83eb`
13. LaMorte, W.W.: Cox proportional hazards regression analysis (2016), `https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html`
14. Manrai, L.A., Manrai, A.K.: A field study of customers' switching behavior for bank services. Journal of retailing and consumer services **14**(3), 208–215 (2007)
15. Mittal, V., Kamakura, W.A.: Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. Journal of marketing research **38**(1), 131–142 (2001)
16. Van den Poel, D., Lariviere, B.: Customer attrition analysis for financial services using proportional hazard models. European journal of operational research **157**(1), 196–217 (2004)
17. Řezáč, M., Řezáč, F.: How to measure the quality of credit scoring models. Finance a úvěr: Czech Journal of Economics and Finance **61**(5), 486–507 (2011)
18. Tang, L., Thomas, L., Fletcher, M., Pan, J., Marshall, A.: Assessing the impact of derived behavior information on customer attrition in the financial service industry. European Journal of Operational Research **236**(2), 624–633 (2014)
19. Verbeke, W., Martens, D., Mues, C., Baesens, B.: Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications **38**(3), 2354–2364 (2011)
20. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proc. of the 4th international conference on the practical applications of knowledge discovery and data mining. pp. 29–39. Springer-Verlag London, UK (2000)
21. Xie, Y., Li, X., Ngai, E., Ying, W.: Customer churn prediction using improved balanced random forests. Expert Systems with Applications **36**(3), 5445–5449 (2009)
22. Zhao, X., Shi, Y., Lee, J., Kim, H.K., Lee, H.: Customer churn prediction based on feature clustering and nonparallel support vector machine. International Journal of Information Technology & Decision Making **13**(05), 1013–1027 (2014)