Machine Learning in Finance: An Application of Predictive Models to Determine the Payment Probability of a Client

Andrés Buitrón, Celeste Rodríguez, María Baldeon Calisto
Universidad San Francisco de Quito, USFQ, Colegio de Ciencias e Ingeniería, Departamento de Ingeniería Industrial and Instituto de Innovación en Productividad y Logística CATENA-USFQ, Diego de Robles s/n y Vía Interoceánica, Quito, Ecuador 170901
abuitronc@estud.usfq.edu.ec, mcrodriguez@estud.usfq.edu.ec, mbaldeonc@usfq.edu.ec

Sebastián Bonilla
Universitat Politecnica de Catalunya- Barcelona Tech, Facultad de Matemáticas y Estadística, Carrer de Jordi Girona, 31, Barcelona España
sebastian.bonilla@estudiantat.upc.edu

Abstract

In the last decades, private debt collection has become an important industry in Ecuador. Debt collection Agencies are organizations that manage the process of collecting money from delinquent debts and their performance is often measured by the collection success rate. However, determining which clients will repay their liabilities is a complex process and many times subjectively judged. Machine learning (ML) models have been successfully implemented on the financing sector for various applications, however a limited amount of work has been published in the prediction of a client´s debt payment probability. In this study, ML models are trained to predict a client´s payment probability to an Ecuadorian Debt Collection Agency the first three months after signing a payment agreement. Specifically, a feedforward neural network, logistic regression, and gradient boosting ensemble models are implemented using the SEMMA data mining methodology, which comprises the steps of sample, explore, modify, model, and assess. Furthermore, by analyzing the results of the models, relevant features that determine whether a costumer will pay or not its debt are identified. The results show that neural networks perform better than the competing models in terms of classification accuracy.

Keywords
Neural Networks, Machine Learning, Client Payment Prediction, Debt Collection Agencies, Payment Probability.

1. Introduction
Financial institutions lend credit to costumers, when their current cash availability is not enough to meet their requirements. A debt then is generated, and it becomes delinquent when there is no payment made in the established agreement or billing notice (U.S. Treasury 2015). Delinquent debts can be a problem for both costumers and creditors, as the collection process involved is time consuming and resources should be focused to commit to the original agreement.

Creditors often have a well-defined procedure for collection, where they try to contact the debtors through written communications, telephone, or personal contact. Successful debt collection is very important to the profitability of the business (Rial 2008), so when the creditors are not able to collect the money, the debts are sent to third-party debt collection agencies (DCAs). DCAs are organizations that manage the process of collecting money from delinquent debts and their performance is often measured by the collection success rate. These companies have the option of managing the collection or making the debt theirs by buying it from financial institutions at a lower value (Beck et al. 2017).

In Ecuador, 7 million debt collection efforts were made from January to June of 2021 accounting for $1,773,000 total debt to financial institutions (SBE 2021). According to Rial (2008), several national factors may influence the delinquency rate. These include economic changes, employment rates, and currency inflation or deflation. The collection task needs to consider macroeconomic variables, as well as individual ones to develop a complete analysis and a successful strategy.

According to Beck et al. (2017), the main field of DCAs is the collection of past-due receivables via agreements with the client. Usually, the DCA and the client sign payment terms that have a mutual agreement for the collection of the debt through monthly fees, but these agreements are not always honored. Given that a DCA´s performance and profitability is often measured by the collection success rate, determining which clients will repay their liabilities is critical. However, the classification of clients is a complex process and many times subjectively done. Machine learning (ML) models have been successfully implemented on the financing sector for various applications, however a limited amount of work has been published in the prediction of a client´s debt payment probability. This research proposes an analytical approach to debt collection on delinquent debt. Specifically, a logistic regression, feedforward neural network, and gradient boosting ensemble are applied to predict the probability that a costumer will pay its debt the first three month after making a payment agreement to a DCA. The dataset used to train the machine learning models was provided by an Ecuadorian DCA and comprises of information from September 2020 to August 2021. The application of the ML models also provides insight that can help stablish strategies for a better debt collection strategy, which can benefit both creditors and debtors. Additionally, in the present study we analyze which relevant features play a role in whether a client pays a delinquent debt or not, which can lead the DCAs to canalize a better debt collection process. Hence, the contributions of our work are two-fold:

- We present a feedforward neural network model to predict whether a costumer will pay its debt after making a payment agreement to a third-party debt collection agency with 78% test sensitivity and 0.70 AUC. To the best our knowledge, we are the first work to propose a prediction model for this task in an Ecuadorian DCA.
- Based on our results, we analyze which relevant features determine if a customer will pay or not its debt, giving a relevant insight for the business.

1.1 Objectives

The study aims to propose a machine learning model to predict a client´s payment probability to a Debt Collection Agency the next three months after a payment agreement. Furthermore, an objective is to analyze which are the most relevant features that determine if a client will or will not pay its depth to improve the collection strategy.

2. Literature Review

Over last decades, machine learning has become a powerful tool for solving various complex problems as machine translation, natural language processing, image analysis, among others (Baldeon-Calisto and Lai-Yuen, 2020a, 2020b, 2021). Moreover, machine learning has been widely used for different applications in the finance sector. Matsumaru et al. (2019) proposed models for predicting bankruptcy risk in companies, making use of a multiple discriminant analysis, artificial neural networks and support vector machines. Data from 64,708 companies in Japan from the period between 1991 and 2015 were used to assess models for predicting bankruptcy from different types of industries. The conclusion states that SVM was more accurate in predicting risk in an aggregate (industry) and individual level (company). In the study of Sniegula, et al. (2019), client churn is predicted using various machine learning models like K-means, decision trees and neural networks; a public dataset from the platform “bigml” with 3333 records and 20 features was used. Decision trees had the best performance with 78% sensitivity and 98% recall scores. Kumar, et. al (2019) proposed a model based on decision trees and neural networks to predict customers likely to leave a banking company. The study was conducted using public data from the Kaggle platform, its conclusions allow to formulate customer retention strategies. Similarly, Bahrami, et. al (2020) used the supervised methods of logistic regression and support vector machine, and unsupervised learning models like DBSCAN to predict which consumers will or will not pay the next payment period agreed, with the use of customer data of a telecommunications company collected in 2014. In an investigation presented by Yeh & Lien (2009), the performance of various machine learning models is assessed for estimating customers prone to default on payment installments at a financial institution in Taiwan. Shoghi (2019) proposes an optimization procedure based on Markov chains and machine learning models like gradient boosting decision tree to prioritize debtors with highest marginal value of debt to collect more debt in smaller periods of time compared. Despite the several studies of machine learning applications in finance and client

classification, there has been few research in DCAs where overdue debts have been incurred and classification of clients can lead to operational savings.

3. Methods

SEMMA is a methodology developed by the SAS Institute for the implementation of data mining applications (SAS 2017). Ilyas et al. 2019 states that it is widely used in the development of machine learning models because it provides a framework to attain meaningful information from data. The acronym SEMMA stands for the sequential phases of the methodology which are Sample, Explore, Modify, Model, and Assess. The phases are shown in **Figure 1** and described below (Balkan and Goul 2010):

- Sample: In this phase the dataset for modeling is selected. The dataset should be a representative sample of the population and contain sufficient information to obtain reliable conclusions.
- Explore: In this phase the dataset is cleaned and explored. The aim is to understand and discover trends in the data.
- Modify: In this phase relevant features or variables are selected for input to the model. Furthermore, variables are transformed or engineered in preparation for the modelling step.
- Model: The machine learning models are selected and trained.
- Assess: In this phase the models are evaluated and validated. Also, models are compared based on the selected evaluation metrics, and the best is then selected.

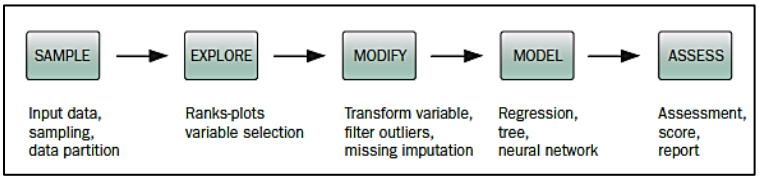


Figure 1. SEMMA Methodology Phases (Balkan and Goul 2010)

The SEMMA methodology is applied as follows in the proposed problem. First, in the Sample phase, data from a middle-sized DCA located in Ecuador was provided. The data is composed of four datasets from 11,918 clients from the period of September 2020 to August 2021. The data corresponds to unpaid debts from clients to financial and services institutions that the DCA bought. The datasets are briefly described in **Table 1**.

Table 1. Description of the DCA datasets for the study

Database Name	Description	Size
Sociodemographic	Educational, salary and location data from clients	11918 rows × 16 columns
Accounts	Purchase capital, type of debt and general details about debt transferor	14887 rows × 8 columns
Agreements	Details about payment agreements from clients to the DCA	12058 rows × 4 columns
Transactions	Date and number of transactions of clients regarding their debt	37691 rows × 4 columns

In the Explore phase an exploratory data analysis is conducted following the guide proposed by Denis (2016). Using multivariate visualization techniques, relevant features are analyzed and relationship between variables obtained. In the Modify phase, a final dataset is constructed by merging the most important variables of the four datasets and creating new features. Categorical variables are also transformed into dummy variables. In the Model phase, three machine learning models are applied to the dataset. Since the aim of the project is to classify costumers of the DCA regarding their payment probability and obtain insights about the classification rules, classification algorithms that have high interpretability or accuracy are selected. Hence, logistic regression, gradient boosting ensemble, and neural networks are considered. In these models the predictor variables are the variables selected from the four DCAs datasets and the response variable is dichotomous whose possible values are 1 if the client has fulfilled their credit obligations,

and 0 if not. Finally, in the Assess phase the models are compared using the evaluation metrics of accuracy, sensitivity, specificity, AUC, and ROC curve.

The following sections are structured as follows: In section 4 a description of the datasets utilized for the study are provided. In section 5 the implementation of the steps of the SEMMA methodology on the proposed case study are presented. Finally, in section 6 the conclusions of the study are shown.

4. Data Collection

The present work uses data collected from a middle-sized DCA from Ecuador. The data is comprised of four datasets obtained from September 2020 to August 2021. The first dataset contains "sociodemographic" information and has educational, work, credit, and demographic variables from the clients. **Table 2** presents the information collected in this dataset. The second dataset, named "accounts", stores data about the debt of each client and is shown in **Table 3**. A third dataset, named "agreements", contains data about the payment agreement and terms reached between the DCA and the clients. The features of this dataset are presented in **Table 4**. Finally, the fourth dataset "transactions" consist of the financial transactions between the client and the DCA prior to signing the payment agreement. The features of the dataset are shown in **Table 5.** All the debts considered for the analysis correspond to purchased overdue portfolio, thus this client database is considered as owned by the DCA and no third parties are involved in the debt collection.

Table 2. Sociodemographic features description.

"Sociodemographic" Features		
No.	**Name**	**Type**
1	Age	Demographic
2	Decease date	
3	Gender	
4	Civil status	
5	Province	
6	Region	
7	Education degree	Educational
8	Dependence	Work status
9	Salary	
10	Work Seniority	
11	Best Credit Qualification	Credit Information
12	Worst Credit Qualification	
13	Amount $ Best Qualification	
14	Amount $ Worst Qualification	
15	Risk Central operations	
16	Amount $ Risk Central	

Table 3. Accounts Features description.

"Accounts" Features		
No.	**Name**	**Type**
1	Account ID	Debt information
2	Transferor	
3	Portfolio	
4	Product	
5	Purchase capital	
6	Purchase date	
7	Overdue date	
8	Account Status	Payment condition

Table 4. Agreements Features description.

"Agreements" Features		
No.	**Name**	**Type**
1	Agreement date	Payment term
2	Total amount	
3	Number of fees	
4	Agreement Status	

Table 5. Transactions Features provided by the DCA.

"Transactions" Features		
No.	**Name**	**Type**
1	Amount	Payment/debt transaction
2	Status	
3	Date	
4	Concept	

5. Results and Discussion

5.1 Sample

The four datasets described in section 4 were merged into one dataset and used as input to develop the ML models. From the original dataset, which contained information about 11819 clients, only data from clients that had signed a payment agreement with the DCA is utilized. Hence, the dataset is reduced to 7289 observations. On the other hand, the unified dataset has a total of 56 predictive variables. The dataset is divided into 80% observations for training set and 20% observations for testing set. Hence information of 5832 clients are used for training the ML algorithms and information of 1457 clients solely for testing the models.

There are three response variables, which identify if a client has paid at least 70% of the agreed monthly fee the three months after signing the payment agreement. The response variable x_i , $i \in \{1,2,3\}$, is a dichotomous variable having a value of 1 if a payment was received in month i and 0 if not. The calculation of the response variable is presented in **Equation 1**.

$$x_1 = \begin{cases} 1 & if\ client\ payed\ at\ least\ 70\%\ of\ the\ fee\ the\ 1st\ month \\ 0 & if\ not \end{cases}$$

$$x_2 = \begin{cases} 1 & if\ client\ payed\ at\ least\ 70\%\ of\ the\ fee\ of\ the\ 1st\ and\ 2nd\ month \\ 0 & if\ not \end{cases}$$

$$x_3 = \begin{cases} 1 & if\ client\ payed\ at\ least\ 70\%\ of\ the\ fee\ of\ the\ 1st\ and\ 2nd\ and\ 3rd\ month \\ 0 & if\ not \end{cases}$$

Equation 1. Response variable definition for the study

It should be noted that clients classified as x_i = 1 are the class of interest, as they are expected to give a fast return on investment and the DCA should invest on efficiently signing a payment agreement with them.

5.2 Explore

The exploratory analysis is carried out according to the guide proposed by Denis (2016). First, the "accounts" dataset is analyzed. The assignor variable, which describes the institution to which the customer's debt belongs to, is evaluated. The classification of the assignor variable is shown in Table 6 and its distribution presented in Figure **2**.

Table 6. Type of transferor variable to which the institutions belong

Name of value	Institution type
Bank 1P, Bank 2P, Bank B, Bank I, Bank S, Bank G	Bank
Ori	Insurance company
Serv	Appliance company
Telephone C, Telephone M	Telecomunications Company

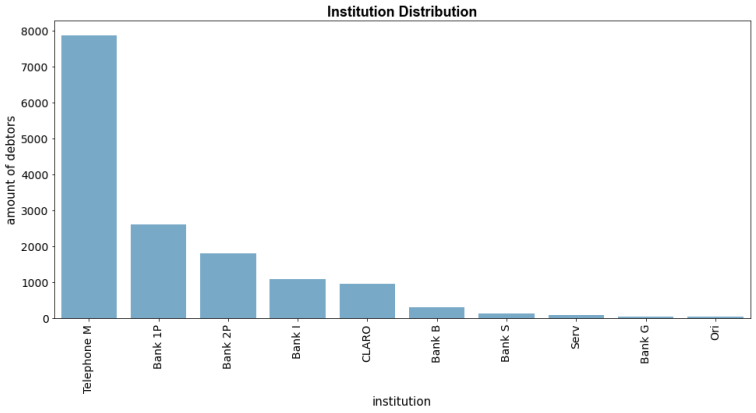


Figure 2. Distribution of the transferor variable

Telephone M, Bank 1P, and Bank 2P, are the institutions with the highest number of debtors. However, by considering the amount of the debt, Bank 1P, Bank I, and Bank 2P represent 80% of the total amount of the debt, as portrayed in **Table 7**. Hence showing that banking institutions accumulate most of the receivable debt.

Table 7. Percentage of main institutions based on purchase capital.

Transferor	Purchase capital (USD)	Percentage
Bank 1P	48.76529	30.03%
Bank I	39.29636	24.20%
Bank 2P	37.77617	23.63%
Telephone M	21.93224	13.51%

A new variable is created that quantifies the number of days past due for each customer. This variable is calculated as the difference between the current date and the due date of the debt. The empirical distribution of the variable is presented in **Figure 3.**

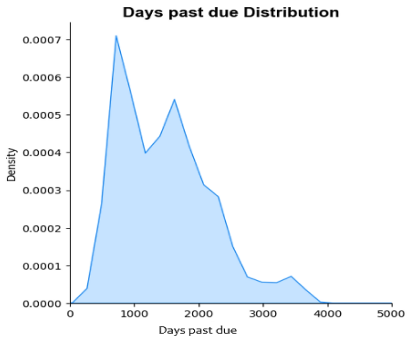


Figure 3. Distribution of the variable days past due

As shown, 80% of debtors are between 0 and 2068 days in debt, which equates to approximately 11 years. The mean is 1538.3 days, which is equivalent to 4 days, while the mode is of 692 days, which means approximately 2 years. For the Sociodemographic database, the distribution of each variable is analyzed as well as the correlation between them. The most important findings are the following. As shown in **Figure 4**, most debtors are men accounting for 59.6% of the observations. Furthermore, most defaulters have between 30 and 40 years old, followed by 40 to 50 years as portrayed in **Figure 5.** Salary is another deterministic variable. As presented in **Figure 6**, most clients earn from $400 to $600 monthly. Moreover, 80% of clients earn less than $4,600 monthly, while the mean is of $ 642.5.

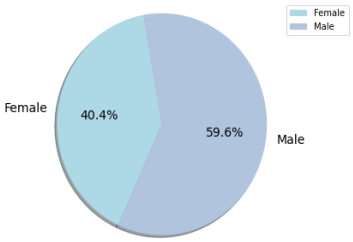


Figure 4. Percentage of debtors according to gender.

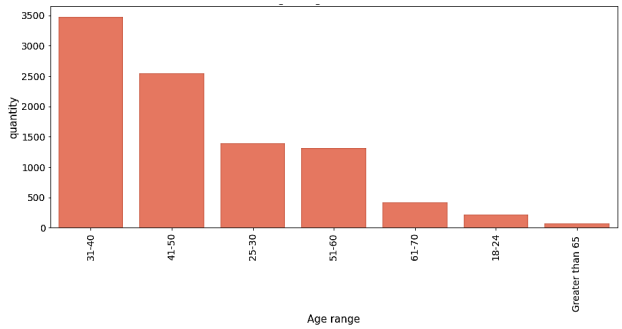


Figure 5. Distribution of the number of debtors by age range.

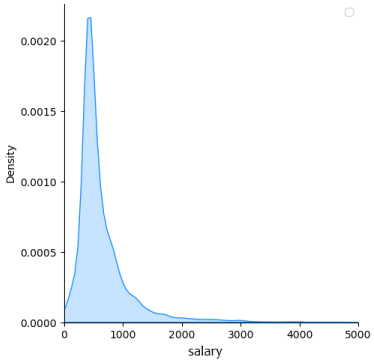


Figure 6. Distribution of the salary variable

The predictor variable named “dependent” describes if a client is affiliated to the Ecuadorian Social Security Institute (IESS), which provides a universal insurance and is mandatory for all working population. Only 67% of the debtors are affiliated while the others are independent, shown in **Table 8**.

Table 8. Percentage of beneficiaries and non-beneficiaries' of the IESS universal insurance

Dependent	Percentage
Beneficiaries	67.64%
Non beneficiaries	32.36%

Regarding the credit rating of each client, obtained through the credit bureau, most clients have a type E rating, which corresponds to the worst rating. The different qualifications are presented in **Figure 7**, with A1 being the best qualification and E the worst qualification that a client can obtain in the credit bureau.

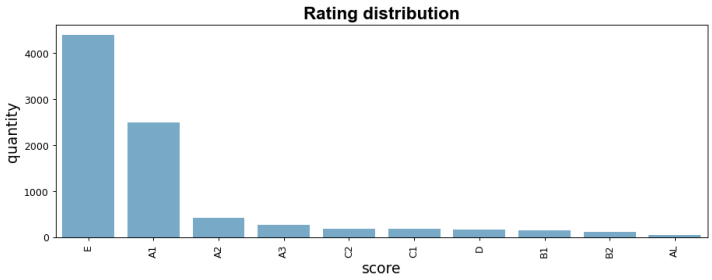


Figure 7. Debtor score Distribution

The variable "amounts" indicates the amount of money that each client owes. It has been divided into ranges for better understanding, and the most significant ranges are presented in **Table 9**. The range with the highest percentage is between 1,000 to 5,000 dollars, meanwhile it is found that 85% of the debts are between 0 and 10,000 dollars.

Table 9. Percentage of the amount of a client´s debt.

Amount of the Debt	Percentage
1000-5000	39.69%
0-1000	34.39%
5000-10000	11.13%

5.3 Modify

In the modify phase, first the null values of each variable are treated. All variables with missing values less than 35% are imputed using a K- Nearest Neighbor approach with three neighbors (N=3). An N=3 is selected because according to Beretta and Santianello (2016) this value helps conserve the original data structure. Secondly, categorical variables are transformed into dummy variables, creating a total of 203 predictive variables on the dataset. Finally, the most important variables are selected using a forward stepwise selection algorithm. This step determines the most significant predictive variables to avoid overfitting the training set and optimize the processing time. The forward stepwise selection algorithm is a partial search algorithm that finds the best combination of predictor variables by testing a subset of the possible combinations. Although it finds a local minimum, it reduces importantly the processing time and has shown to produce good results. The forward stepwise selection algorithm runs for 60 iterations and selects 56 variables from the dataset which included variables with information about the education, debt transferors, location, and credit history.

5. 4 Model

The models are implemented using Python 3.0.8 and the Sklearn library. One important characteristic of the dataset is that it is unbalanced, where only 13% of the clients made a payment according to the agreement signed with the DCA (corresponding to class 1). Therefore, by using the SMOTE technique proposed by Chawla et al. 2002 the minority class is oversampled. Models were selected given their specific characteristics and based on the literature review. The implementation of the models is presented next.

Logistic Regression: According to Shmueli (2020), the logistic regression (LR) is a multivariate statistical method analogous to a linear regression that returns the probability of a client pertaining to class 1. The regression´s parameters are calculated using the maximum likelihood technique, which maximizes the probability of obtaining the observed training dataset. This model is simple but provides insightful information about features. Specifically, the log odds from the logistic regression provides a good interpretation about which variables affect the classification of an observation as class 0 or 1. **Table 10** presents the odds ratios of this problem and the most important features that help to determine if a client will pay the next month after an agreement**.** As it can be seen, a client will pay the debt if the debt was acquired from Bank G, Telephone C, Telephone M, and Telephone M. Also, if a client is considered a corporate entity, is born in Province B, is an employee, and has a higher educational level will increase the odds of paying its debt.

Table 10. Odds ratio that determines if a client will pay its debt according to LR.

Variable	Odds ratio
Transferor: Bank G	25.59
Transferor: Telephone C	4.86
Transferor: Telephone M	3.69
Portfolio 52: Telephone M	3.68
Corporate Entity	3.17
Province B	2.31
Best Qualification Risk Center	1.99
Employee	1.29
Education level: Higher	1.22

Likewise, in **Table 11** the features that determine if a client will not make a payment after an agreement are portrayed. It should be noted that in this case the odds ratios are below 1. A client that has a risk qualification of C will tend not to pay the debt; having an initial educational and bigger purchase capital will increase the odds of not paying. As well if the debt is acquired from Bank B.

Table 11. Odds ratio that determines if a client will not pay its debt according to LR.

Variable	Odds ratio
Worst Qualification Risk Center: C	0.40
Portfolio 2: Bank B	0.24
Education level: Initial	0.17
Purchase capital	0.99

Gradient Boosting: The gradient boosting is an ensemble model that combines decision trees to obtain a stronger model that has a more accurate final prediction, and a "importance" is given to each feature. The model is trained in an iterative manner, in which each new tree is trained on a modified version of the original dataset. The gradient boosting ensemble has various hyperparameters, in this work the learning rate, maximum depth, and number of estimators are optimized to improve the model´s accuracy. The learning rate specifies how fast the model will learn. If the rate is too high, the optimal structure of the tree might be skipped. However, if the rate is too low, the model will learn slowly and can be inefficient (Bentéjac et al., 2021). The number of estimators refers to the number of trees that will be part of the ensemble. Finally, the maximum depth is the maximum number of nodes allowed for each tree. In general, implementing the hyperparameters separately can generate suboptimal configurations because information about the interaction between the hyperparameters is lost. A grid search technique with 5-fold cross-validation is used to select the most optimum values. The possible search ranges for each hyperparameter are shown in **Table 12**. The optimal values found are 1000 number of decision trees, learning rate of 0.01, and a maximum depth of 10.

Table 12. Hyperparameter value ranges for the Grid Search in the Gradient Boosting Ensemble.

Hyperparameter	Values
Number of trees	50, 100, 500,1000
Learning rate	0.0001, 0.001, 0.01, 0.1
Maximum depth	1, 5, 10, 20

This model also allows to determine which are the most important variables for the classification of the clients, although the interpretability is not the same as the one obtained with the logistic regression. In **Table 13** the most significant variables are presented, where the importance represents the decrease of the model´s impurity due to the participation of the predictor variable in the tree. It is observed that the most important variable is purchase capital with 16.83% (which represents the amount of the debt), followed by the salary with 10.63%, and customer age with 7.33%. Variables that have a value less than 1% importance have not been included due to their relative insignificance in the model.

Table 13. Importance of variables for the Gradient boosting model

Variable	Importance
Purchase Capital	16.83%
Salary	10.63%
Customer age	7.33%
Amount of the best grade	5.51%
Worst grade amount	5.46%
Total amount in credit registry	5.33%
Transferor: Telephone M	3.94%
Product M	3.56%
Labor Old	3.38%
Operational amount in risk center	1.95%

In comparison with the logistic regression model, there is certain discrepancy regarding the results of the feature importance. However, it is important to mention that the information provided by each model is distinct. The logistic regression analysis gives an insight about features that affect if a client will or will not pay, while the gradient boosting gives the relative importance of each feature to the model´s discriminative power. In the logistic regression, Bank G and Telephone C strongly affect the model, whereas they are unimportant for Gradient Boosting where Telephone M is the transferor which has an importance. Nevertheless, we can conclude in both models from whom the debt is acquired does have an effect in the payment probability. The variable education affects the logistic regression, where having an initial education reduces the probability of pay and having a higher education increases this probability, but it has no significant importance in gradient boosting. On the other hand, purchase capital shows to be an important variable for both models, being the probability of repay reduced as the value of the debt is increased.

Neural Networks: A neural network (NN) is a graph-based model inspired in the human brain that uses nonlinear functions to create a highly complex non-linear model. The basic structure of a NN consist of input, hidden, and output layers. Layers are made of nodes that give weights to input data to amplify or decrease its significance for the specific classification task (James et al. 2021)**.** The hyperparameters that can affect the performance of the model are the number of neurons that each layer has, which is set based on the complexity of the problem, the type of activation function that is usually set to Rectified Linear Unit (ReLU); and the learning rate that determines the speed at which the model learns. The neural network implemented has 3 hidden layers, with 15 neurons per layer, ReLU activation functions, and a learning rate of 0.01. The structure of the feedforward neural network applied is shown in **Figure 8.**

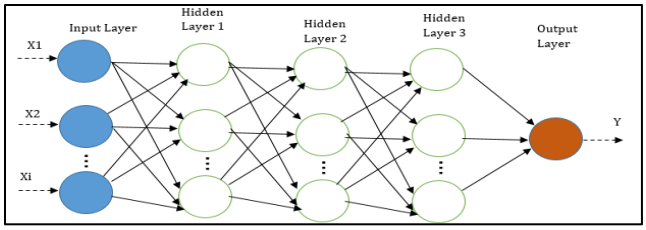


Figure 8. Neural Network structure applied for the classification of the DCA´s clients.

5.5 Assess

The models are evaluated using the accuracy, sensitivity, specificity, ROC curve, and AUC evaluation metrics. The accuracy is the ratio between the number of correct predictions made by the model and the total number of predictions, as presented in **Equation 2**. It provides significant initial information about the general performance of a model, however when the dataset is unbalanced the obtained values might hide a deficient prediction on the minority class (Al-jabery et al. 2019).

$$\text{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Equation 2. Classification accuracy metric

The sensitivity measures the rate of true positives, putting emphasis on correctly classifying class 1 or the clients that will repay their debt. The formula is shown in **Equation 3.** This is the most important evaluation metric for the problem in hand, as it is of interest to identify the clients that will pay their debt after signing the agreement.

$$Sensitivity = \frac{True\ positive}{(True\ positive + False\ negative)}$$

Equation 3. Sensibility metric

The specificity measures the rate of true negatives, as presented in **Equation 4.** It is the probability of calculating a prediction as negative when it is truly negative.

$$Specificity = \frac{True\ negative}{(True\ negative + False\ positive)}$$

Equation 4. Specificity metric

The relationship between sensitive and specificity metrics is determined by the separation limit between the two classes, which is known as the threshold. By varying the threshold, the values for sensitivity and specificity can be modified. In this work the appropriate threshold value is selected based on the desired performance on sensitivity and specificity. Finally, the ROC curve is a graph that determines the performance of a binary classification model for each possible threshold value. It measures how well a model correctly classifies the observations from the positive class and minimizes the false positive error (Gneiting and Vogel. 2021). The AUC is understood as the area under the ROC curve. The AUC ranges between 0 and 1, where a value of 1 indicates that a model perfectly classifies the dataset.

Table 14 depicts the accuracy, sensitivity and specificity for each model the three months after signing the payment agreement. The AUC and ROC curve are illustrated in **Figure 9**. The results demonstrate that the higher sensitivity and is achieved by the neural networks in month 1 and moth 3; while gradient boosting performs better in month 3. In reference to the AUC scores, there is a better performance during the first month and decrease for the second and third month for all the models. Logistic regression performs better in terms of accuracy in months 2 and 3, while Gradient Boosting does it so in month 1. Given that neural networks have the highest sensitivity in comparison to the other models for the first two months, which is the metric of interest because it measures how well the model identifies the clients that will honor their debt, it is recommended to be used in the future. Furthermore, it must be mentioned that predicting the client´s that will pay their debt the third month is very difficult due to the dataset imbalance as most clients do not pay their debt.

Table 14. Accuracy, sensitivity, and specificity of each month for the models.

Period	Accuracy	Sensitivity	Specificity	Model
Month 1	60.36	76.99	57.30	Logistic Regression
	70.44	63.070	**71.89**	Gradient Boosting
	60.56	**78.31**	57.30	Neural Networks
Month 2	**63.49**	62.79	**63.85**	Logistic Regression
	62.30	63.75	61.63	Gradient Boosting
	62.21	**70.17**	61.88	Neural Networks
Month 3	**68.25**	53.70	**72.22**	Logistic Regression
	64.19	**66.66**	62.50	Gradient Boosting
	60.28	45.95	60.66	Neural Networks

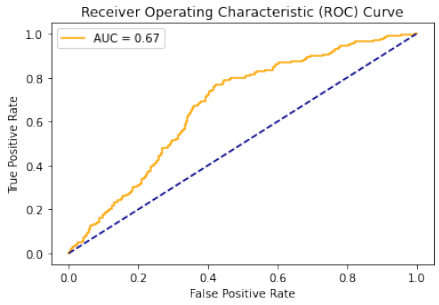
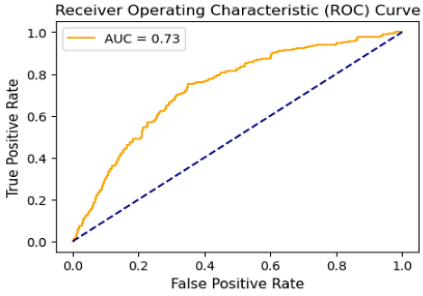
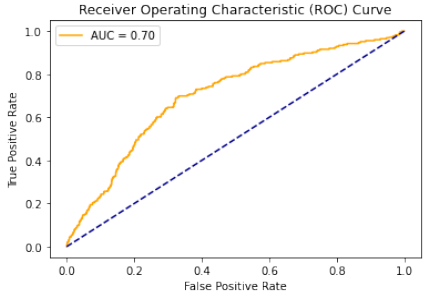
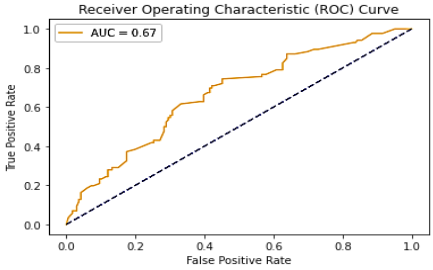
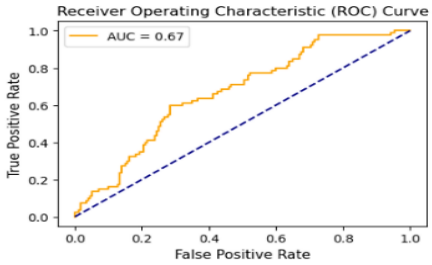
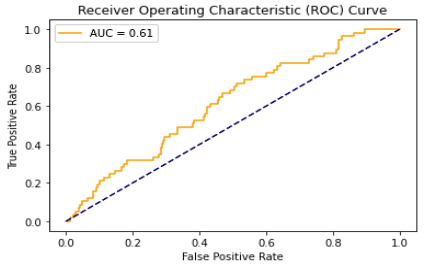
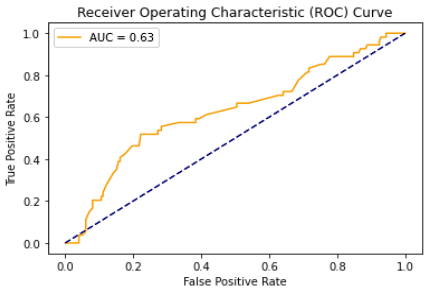
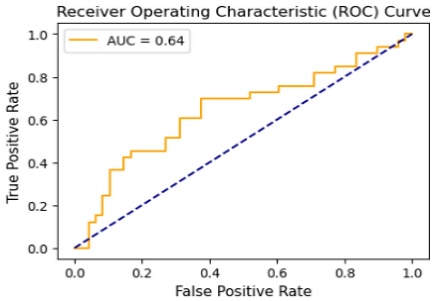
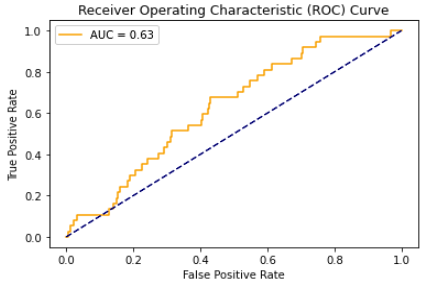
Month / Model	Logistic Regression	Gradient Boosting	Neural Networks
1			
2			
3			

Figure 9. ROC and AUC curve for each month for the logistic regression, gradient boosting, and neural network models.

6. Conclusion

In this work, a study was conducted in an Ecuadorian Debt Collection Agency to develop a ML model that predicts the clients that will pay their debt after signing a payment agreement. The problem was formulated as a supervised classification problem, following the SEMMA methodology for data mining projects. In the exploratory data analysis, the distribution of the variables was analyzed as well as the relationship between them. On the modify step, the null values were treated, and a forward stepwise selection method was used to reduce the dimensionality of the dataset and select the most important variables. On the model phase, the SMOTE oversampling technique was applied to the unbalanced data set to reduce the disproportionality of the minority class during training. A logistic regression, gradient boosting ensemble, and neural network model were implemented due to their interpretability and capability to model the response variable. The models were evaluated in the assess phase using the metrics of sensitivity, precision, specificity, ROC curve, and AUC. The results show that the neural network model outperformed the competing models in the prediction of the first- and second-month payment, while the gradient boosting performed best during the third month prediction in terms of sensitivity. Hence, it is recommended to use a neural network to classify customers based on the overall performance on the three months. Furthermore, it was found that the variables that affects if a client will pay the next months after an agreement are transferor, qualification in the risk center, education level, and purchase value. The DCA should put emphasizes in analyzing these variables before buying a debt and determining with which clients to sign first the payment agreement.

References

Al-jabery, K., Obafemi-Ajayi, T., Olbricht, G., and Wunsch, D., Computational Learning Approaches to Data Analytics in Biomedical Applications, 1st Edition, Academic Press, Missouri, 2019.

Balkan, S., and Goul, M., Advances in Predictive Modeling: How In-Database Analytics Will Evolve to Change the Game, Business Intelligence Journal, vol.15, pp. 17-25, 2010.

Bahrami, M., Bozkaya, B., and Balcisoy, S., Using Behavioral Analytics to Predict Customer Invoice Payment, Big Data, vol. 8, no. 1, pp. 25–37, 2020

Baldeon-Calisto, M., and Lai-Yuen, S., AdaResU-Net: Multiobjective adaptive convolutional neural network for medical image segmentation, Neurocomputing, vol. 392, pp. 325-340, 2020a.

Baldeon-Calisto, M., and Lai-Yuen, S., AdaEn-Net: An ensemble of adaptive 2D-3D Fully Convolutional Networks for medical image segmentation, Neural Networks, vol. 126, pp. 76-94, 2020b.

Baldeon-Calisto, M., and Lai-Yuen, S., EMONAS-Net: Efficient multiobjective neural architecture search using surrogate-assisted evolutionary algorithm for 3D medical image segmentation, Artificial Intelligence, vol. 119, 2021.

Beck, T., Grunert, J., Neus, W., and Walter, A., What Determines Collection Rates of Debt Collection Agencies? Financial Review, vol. 52, no. 2, pp. 259-279, 2017.

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. A., Comparative analysis of gradient boosting algorithms. Artificial Intelligence Review. Vol. 54, pp. 1937-1967, 2021.

Beretta, L., and Santaniello, A., Nearest neighbor imputation algorithms: a critical evaluation, BMC Medical Informatics and Decision Making, vol. 16, no. 3, pp. 197-208, 2016

Chawla N., Bowler K., Hall L., and Phillip W., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

Denis, J. Applied Univariate, Bivariate, and Multivariate Statistics, 1st Edition, John Wiley and Sons Ltd, 2016.

Yeh, I-C., and Lien C., The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Application, vol. 36, no.2 , pp. 2473–2480, 2009.

Gneiting, T., and Vogel, P., Receiver operating characteristic (ROC) curves: equivalences, beta model, and minimum distance estimation, Machine Learning, vol. 1, 2021

Ilyas, H., Sohail, K., and Aslam, U., Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA), Journal of Computational and Theoretical Nanoscience, vol. 16, pg. 3489–3503, 2019.

James, G., Witten, D., Hastie, T., and Tibshirani, R., An Introduction to Statistical Learning, 2nd Edition, Springer, 2021.

Kumar, G., Tirupathaiah, K., and Krishna Reddy, B., Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques, International Journal of Computer Sciences and Engineering, vol. 7, no. 6, pp.842-846, 2019.

Matsumaru, M., Kawanaka, T., Kaneko, S. and Katagiri, H., Bankruptcy Prediction for Japanese Corporations using Support Vector Machine, Artificial Neural Network, and Multivariate Discriminant Analysis, International Journal of Industrial Engineering and Operations Management, vol. 1, no.1, pp. 78 – 96, 2019

Rial, R., Best Practice in Consumer Collections, 2nd edition, VRL KnowledgeBank Ltd., 2008.
SAS Help Center, Available: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm, Accessed on October 20, 2021.
Shoghi, A. Debt Collection Industry: Machine Learning Approach, Journal of Money and Economy, vol. 14, no. 4, pp. 453-473. 2019
Sniegula, A., Poniszewska-Marańda, A., and Popovic, M., Study of machine learning methods for customer churn prediction in telecommunication company, Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. pp. 640-644, New York, United States, December 2-4, 2019.
Shmueli, G., Bruce, P., Gedeck, P., and Patel, N., Data Mining for Business Analytics, 1st Edition, John Wiley and Sons Ltd, New Jersey, 2020.
Superintendencia de Bancos del Ecuador SBE. Available: https://estadisticas.superbancos.gob.ec/portalestadistico/portalestudios/?page_id=1826. Accessed on November 14, 2021.
US Department of Treasury.Available: https://www.fiscal.treasury.gov/files/dms/chapter6.pdf. Accessed on October 20,2021.

Biographies

Andrés Buitrón is a student of the Industrial Engineering at San Francisco de Quito University. Throughout his student career he has participated in community outreach projects like MIT Genesys whose aim is to improve micro and small firms' operations. His areas of interest include data analytics, quality control, and supply chain. He is currently working in the clothing retailing sector where he aspires to gain knowledge in logistics applications.

Celeste Rodriguez is a student of Industrial Engineering at the Universidad San Francisco de Quito. During her student career she has been interested in topics such as operations research, statistics, data analytics, logistics and project planning. She currently works on sales, marketing of communication, and advertising media.

María Baldeon Calisto is a professor and researcher in the Industrial and Management Systems Engineering department at Universidad San Francisco de Quito, Ecuador. She earned a BS in Industrial Engineering at Universidad San Francisco de Quito, a master's degree in safety, health, and environment at Universidad San Francisco de Quito, a master's degree in industrial engineering at University of South Florida, and a PhD in Industrial Engineering at University of South Florida. Her research interests include data analytics, deep learning, medical image analysis, and hyperparameter optimization.

Sebastián Bonilla is a PhD Candidate at Universitat Politecnica de Catalunya with a focus on Models and Methods with mix data. He earned a BS in Industrial Engineer from Universidad San Francisco de Quito, Ecuador and holds a master's degree in Statistics and Operations Research from Universitat Politecnica de Catalunya and Universitat de Barcelona, Spain. His research interest includes statistical learning for Finance and Marketing, data analytics, and assemble methods. Currently, he works as a Senior Data Scientist at a Finance Institution.