UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Machine Learning in Finance: Application of Predictive Models to Determine Payment Probability

.

Andrés Sebastián Buitrón Chang
María Celeste Rodríguez Villalva

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 20 de diciembre de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Machine Learning in Finance: Application of Predictive Models to Determine Payment Probability

Andrés Sebastián Buitrón Chang
María Celeste Rodríguez Villalva

Nombre del profesor, Título académico	**María Gabriela Baldeón, PhD.**

Quito, 20 de diciembre de 2021

© DERECHOS DE AUTOR

Nombres y apellidos: Andrés Sebastián Buitrón Chang

Código: 00202713

Cédula de identidad: 1724061815

Nombres y apellidos: María Celeste Rodríguez Villalva

Código: 00204027

Cédula de identidad: 1805368246

Lugar y fecha: Quito, 20 de diciembre de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

En las últimas décadas, la cobranza de deuda privada se ha convertido en una industria importante dentro del Ecuador. Las empresas de gestión de cobranza son organizaciones que manejan el proceso de recolección de efectivo de deudas vencidas y su rendimiento es medido por la tasa de éxito de recuperación. Sin embargo, determinar qué clientes van a pagar sus deudas es un proceso complicado y muchas veces juzgado subjetivamente. Los modelos de aprendizaje automático han sido exitosamente implementados en el sector financiero para varias aplicaciones pero existe una limitada cantidad de publicaciones en la predicción de probabilidades de pago de clientes. En este estudio, modelos de aprendizaje automático fueron entrenados para predecir la probabilidad de pago de clientes en una organización de gestión de cobranza ecuatoriana para los tres primeros meses después de un acuerdo de pago entre partes. Específicamente, los modelos de redes neuronales, regresión logística y métodos de potenciación de gradiente fueron implementados utilizando la metdología de minería de datos de SEMMA que sigue los pasos de Sample, Explore, Modify, Model y Assess. Se analizó los resltados de los modelos y se obtuvieron variables relevantes que determinan si un cliente pagará o no su deuda. Los resultados muestran que el modelo de redes neuronales tiene mejor rendimiento que los otros modelos evaluados en términos de precisión de clasificación.

Palabras clave: Redes Neuronales, Aprendizaje automático, Predicción de pago de cliente, Empresa de gestión de cobranza, Probabilidad de pago.

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 month after signing a payment agreement. Specifically, a neural network, logistic regression, and gradient boosting ensemble models are implemented using the SEMMA data mining methodology, which comprises of the steps of Sample, Explore, Modify, Model and Assess. Furthermore, 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 (NN) perform better than the competing models in terms of classification accuracy.

Key words: Neural Networks, Machine Learning, Client Payment Prediction, Debt Collection Agencies, Payment Probability.

TABLE OF CONTENTS

TABLES´ INDEX

FIGURES’ INDEX

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, 2005), so when the creditors are not able to collect the money, the debts are send 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 (2005), several factors from a nation 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 collection strategy.

According to (Beck et al. 2017), the main field of DCAs is the collection of past-due receivables via agreements with the client. Payment terms that have a mutual agreement for the collection of the debt through monthly fees are signed between the DCA and the client, but these agreements are not always honored. This research proposes an analytical approach to debt collection on delinquent debt. Specifically, three machine learning (ML) models 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 August 2020 to September 2021. The application of the ML models also provides insight that can help stablish strategies for a better debt collection system 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 machine learning model to predict whether a costumer will pay its debt after making a payment agreement to a third-party debt collection agency with an 82% accuracy. 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.

LITERATURE REVIEW

Machine learning has been widely used in studies on the finance sector for different applications. 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 Sniégula, 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 ,achine, 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 Cheng 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 (2020) 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.

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 machine learning development process 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 explore. 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 selected.

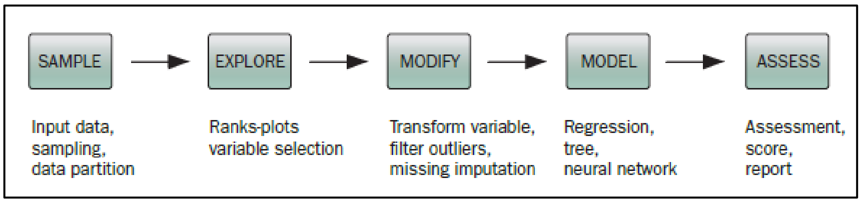

SAMPLE
EXPLORE
MODIFY
MODEL
ASSESS
Input data, sampling, data partition
Ranks-plots variable selection
Transform variable, filter outliers, missing imputation
Regression, tree, neural network
Assessment, score, report

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 to their debt	37691 rows × 4 columns

In the explore phase an exploratory data analysis is conducted following the guide proposed by Denis (2020). Using multivariate visualization techniques relevant features and relationship between variables are analyzed, obtaining important business insights about the clients. 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 also transformed to dummies. 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 interpretability or have a high accuracy are selected. In this way, 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 potentially risky and non-risky clients with respect to the fulfillment of their credit obligations. The ML models implemented are described next.

Logistic Regression: According to Shmueli (2019) it is a multivariate statistical method analogous to a linear regression, that returns the probability of a client paying or not paying the debt. 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 very powerful information. Specifically, the logit function from the logistic regression provides a good interpretation about which variables affect the classification of an observation as class 0 or 1, which is the reason this model is selected.

Gradient Boosting: It is an ensemble model that combines weak learners to obtain a stronger model that has a more accurate final prediction. 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, from which the learning rate, maximum depth, and number of estimators are optimized to improve the model´s accuracy. The learning specifies how fast the model will learn. If the learning is high, the optimal structure of the tree might be skipped. However, if the rate is low, the model will learn slowly and can be inefficient (Dash, 2020). The number of estimators refers to the number of trees that will be part of the ensemble. Finally, the maximum depth is the is the maximum number of levels allowed for each tree (Ippolito, 2019). In general, implementing the hyperparameters separately can generate suboptimal configurations because information about the interaction is lost. In this work, a grid search technique with cross-validation is used to select the most optimum values.

Neural Networks: Neural networks are a series of algorithms used to recognize relationships in data sets, inspired of neurons in a brain. They use nonlinear functions on variables to predict a response. The basic structure of NN consist of input, hidden and output layers. Layers are made of nodes that give weights or coefficients to input data and then amplify or decrease significance to that inputs regarding to the classification task. The weights pass through an activation function, which tells if these coefficients should pass through the network in order to be taken into account for the classification task, being an "activated" neuron. Finally, a linear regression with the number of K activations is estimated to classify further records. (James et al. 2021)**.** The hyperparameters that can affect the performance of the model are the number of neurons that each layer has, being decided according to the complexity of the problem; activation function usually being Rectified Linear Unit; and learning rate that determines the speed at which the model learns.

Finally, in the assess phase the models are compared using the evaluation metrics of sensitivity, specificity, AUC and ROC. The description of each evaluation metric is discussed below.

Accuracy: Is the ratio between the number of correct predictions made by the model and the total number of predictions, as presented in **Equation 1**. 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).

$$\text{Classification accuracy} = \frac{Correct\ predictions}{Total\ predictions}$$

Equation 1. Classification accuracy

Analyzed similarly in **Equation 2**, accuracy is presented based on the rate of true negatives and false positive.

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

Equation 2. Accuracy metric

Sensitivity: Measures the rate of true positives, putting emphasis on correctly classifying class 1. The formula is shown in **Equation 2.**

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

Equation 3. Sensibility metric

Specificity: Measures the rate of true negatives. it is the probability of calculating a prediction as negative, this being 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.

ROC curve and AUC: 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 et al. 2019). 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 model perfectly classifies the dataset.

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

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 August2021. 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", supplies 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. This dataset consists of information of 7.289 clients that have made 22.014 transactions regarding their debt to the DCA. 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

"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

“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

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

Table 5. Transactions Features

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

RESULTS AND DISCUSSION

SAMPLE

The four datasets described in section 4 are merged into one dataset and used as input to develop the ML models. From the original dataset that contained information about 11819 clients, only data from clients that had signed a payment agreement with the DCA is utilized. Hence, reducing the dataset 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 and 20% observations for testing. Hence information of 5832 clients are used for training the algorithms and information of 1457 clients solely for testing the models.

There are three response variables, which identifies 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 7**.

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

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

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

Equation 7. 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.

EXPLORE

The exploratory data analysis is carried out according to the guide proposed by Denis (2019). First, the Accounts data set is analyzed. The Assignor variable, which is the institution to which the customer's debt belongs, has been evaluated. The institutions are described in **Table 6**:

Table 6. Transferor type

Institution	Institution type
Bank 1 P	Banks
Bank 2P	Banks
Bank B	Banks
Bank I	Banks
Bank S	Banks
Bank G	Banks
Ori	Production
Serv	Production
Telephone C	Telephone Company
Telephone M	Telephone Company

Next, the distribution of institutions and the number of debtors.

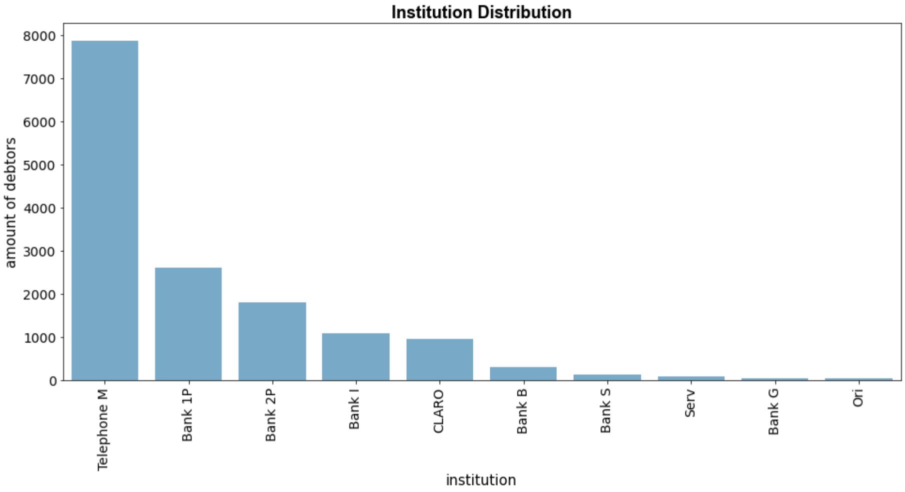


Figure 2. Distribution of the transferor variable

It is obtained that Telephone M, Bank 1P and Bank 2P, are the institutions with the highest number of debtors.

By counting the amount of the debt according to each institution, the **Table 7** is obtained:

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

Transferor	Purchase capital	Percent
Bank 1P	48.76529,39	30.03%
Bank I	39.29636,93	24.20%
Bank 2P	37.77617,94	23.63%
Telephone M	21.93224,66	13.51%

The entities Bank 1P, Bank I and Bank 2P represent 80% of the amount of the debt. These three institutions accumulate most of the debt receivable. Therefore, the difference between the amount of debt and the number of debts is noted.

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

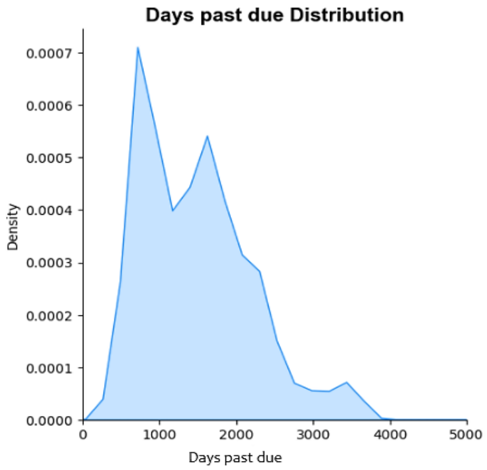


Figure 3. Distribution of the variable days past due

The statistical metrics in **Table 8** are calculated to corroborate the information in figure 2.

Table 8. Statistical metrics of the variable days past due.

Statistical metrics	Results
minimum value	195.0
maximum value	44530.0
quantile / 80	2068.0
upper limit	0
lower limit	4771.5

It is obtained that 80% of the debtors have a debt between 0 and 2068 days, which is equivalent to approximately 11 years.

For the Sociodemographic database, an analysis is made of each variable and the correlation between them. As shown in **Figure 4**, the largest number of debtors are men.

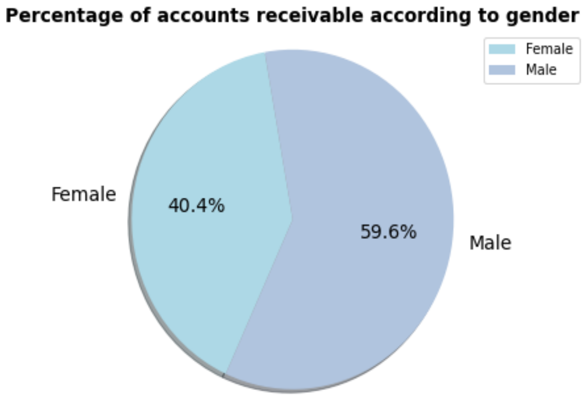


Figure 4. Percentage of debtors according to gender.

Likewise, most defaulters are between 30 and 40 years old, followed by 40 to 50 years as shown in **Figure 5**.

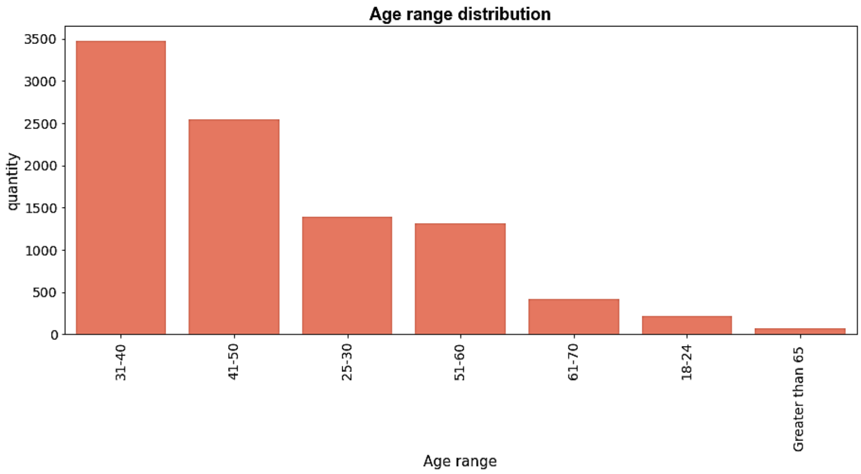


Figure 5. Distribution of the number of debtors by age ranges.

Salary is another deterministic variable. As shown in **Figure 6**, most clients earn around $ 400 to $ 600, this information is corroborated in **Table 9**, that according to statistical analysis, 80% of clients earn less than $ 4,600 in monthly salary and the mean is $ 642.5. The salary is 33% null, so they are completed using a KNN imputation method.

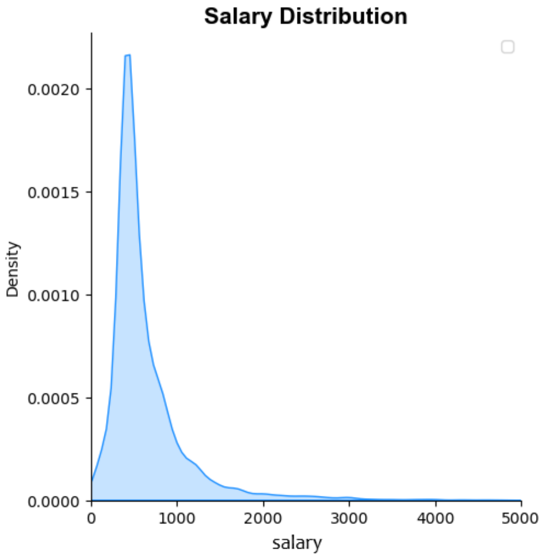


Figure 6. Distribution of the salary variable

Table 9. Statistical metrics of the salary variable.

Statistical metrics	Results
quantile / 80	2068.0
mean	642.5
upper limit	0
lower limit	4771.5

The dependent variable determines the number of debtors who are beneficiaries of the Ecuadorian Social Security Institute and who do not have this insurance. It is obtained that only 67% of the debtors are affiliated while the others are independent, shown in **Table 10**.

Table 10. Percentage of beneficiaries and non-beneficiaries' people

Dependent	Percent
Beneficiaries	67.640%
Non beneficiaries	32.36%

When correlating the dependency and salary variables in **Figure 7**, it turns out that the salary for non-beneficiaries does not appear in the database, therefore the percentage of missing data for the salary column is deducted.

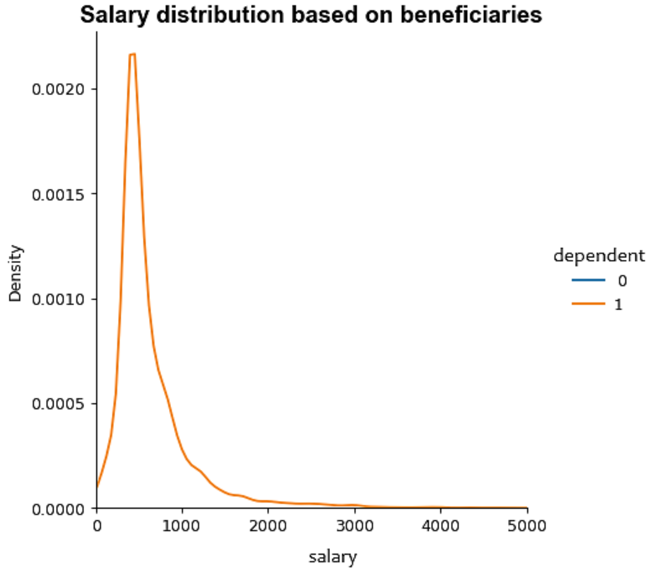


Figure 7. Distribution of the Salary variable with respect to the dependent variable.

Regarding the rating obtained through the credit bureau, according to **Figure 8**, most clients have a type E rating, which corresponds to the worst rating.

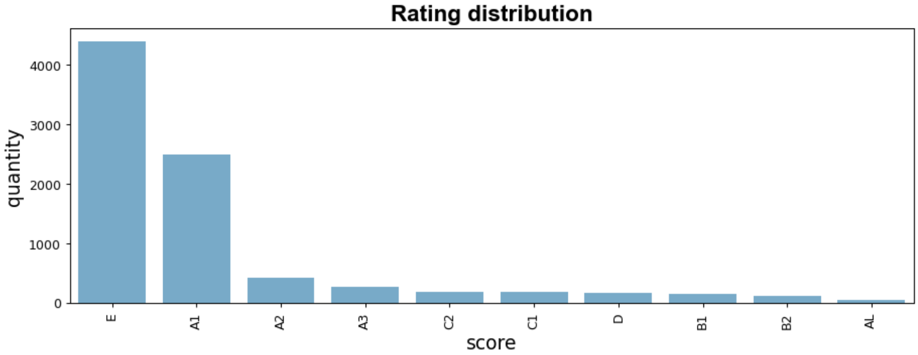


Figure 8. Debtor score Distribution

The information of the variable amounts in the risk center has been divided into ranges to better visualize and understand the data. It is had that 85% of this corresponds to debts between 0 to 10,000 dollars. Being the range of 1,000 to 5,000 dollars the amount with the highest percentage in **Table 11**.

Table 11. Percentage of the ranges of the amounts of the clients in the credit registry.

Amount range in Risk Center	Percent
1000-5000	39.69
0-1000	34.39
5000-10000	11.13

MODIFY

In the modify step, first the null values are imputed using a K- Nearest Neighbor approach with three neighbors (N=3). An N=3 is selected because according to Beretta and Santianello (2018) this value helps conserve the original data structure. Secondly, categorical variables are transformed into dummies. A total of 203 predictive variables are obtained with this technique. Finally, the most important variables are selected using a forward stepwise selection approach. This step determines the most significant predictive variables to avoid overfitting and optimize the processing time. The forward stepwise selection techniques are partial search algorithm that finds the best combination 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 run for 60 iterations and selected 56 variables from the dataset which included variables from education, debt transferors, location, and credit history.

MODEL

The models are implemented using Python 3.0.1 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 (corresponding to class 1). Therefore, using the SMOTE technique proposed by Chawla et al. 2002 the minority class is oversampled. The implementation of the models are presented next.

Logistic Regression: The odds ratio represent the most important features based on their influence on the response variable. **Table 12** represents the most important features that help to determine if a client will pay the next month after an agreement**.**

Table 12. Odds ratio that determine if a client will pay.

Variable	Odds ratio
Transferor: Bank G	25.59
Transferor: Telephone C	4.86
Transferor: Telephone M	3.69
Portfolio 52: Telephone M	3.68
Legal Person	3.17
Province B	2.31
Best Qualification Risk Center	1.99
Dependent (work)	1.29
Education level: superior	1.22

Likewise, **Table 13** represents the features that determine that a client will not make a payment after an agreement. It should be noted that in this case odds ratios are below the value of 1.

Table 13. Odds ratio that determine if a client will not pay.

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

Gradient Boosting: First, the optimal hyperparameters for the model are obtained. The possible search ranges are shown in **Table 14**. The optimal variables found are number of trees: 1000, Learning rate: 0.01, maximum depth: 10.

Table 14. Selection of GBM hyperparameters.

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

The analysis of the most important variables is not equal to the interpretability obtained in the logistic regression model, but it allows us to understand which are the most significant variables for the development of the In **Table 15** the most significant variables for month 1 are shown, month 2 and month 3. It is observed that the most important variable is Purchase Capital with 16,827%. followed by the Salary variable with 10.63% and Customer age.

Table 15. 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 Logistic Regression model, it is observed that there is a discrepancy with the GMB results. This may be since the analysis in this last model is carried out after performing dummy variables. The importance of the variables in this last model is less compared to the first. Variables that have a value less than 1% have not been considered due to their relative insignificance in the model.

Neural Networks. The neural network implemented has 3 hidden layers and ReLU activation functions. Also a lerning rate of 0.01 is applied. The structure of the neural network is shown in **Figure 9**

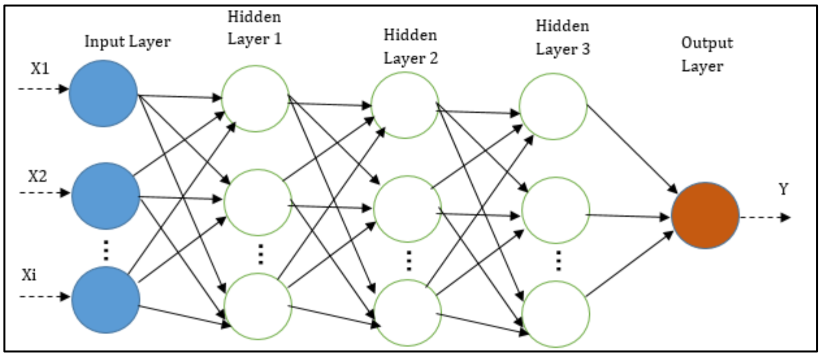


Figure 9. Neural Network diagram for the model.

ASSESS

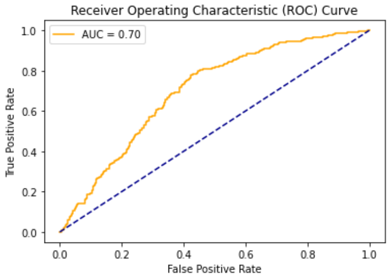
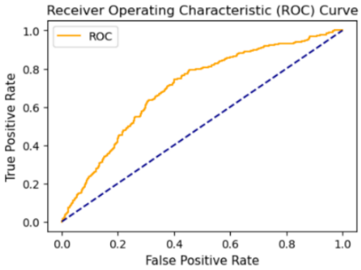
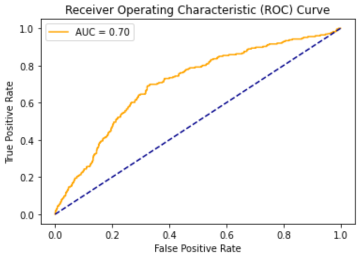
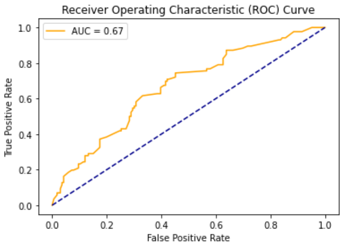
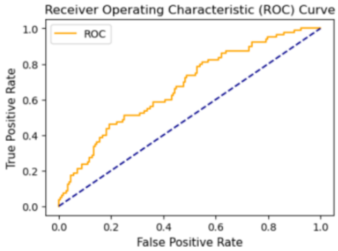
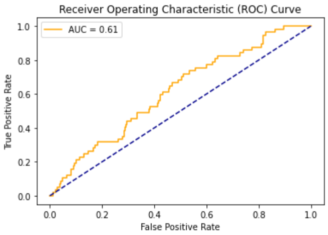
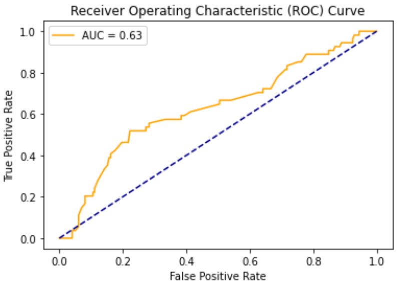
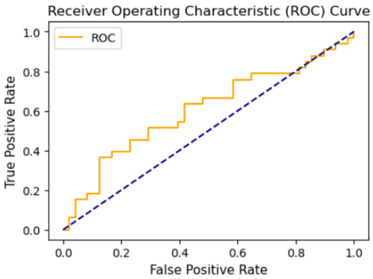
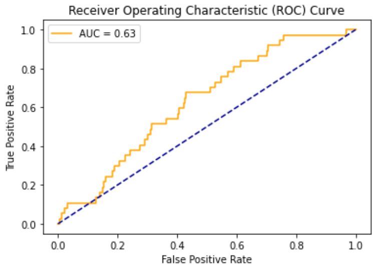
The models were evaluated using the evaluation metrics defined in the phase assess. **Table 16** represents the accuracy, sensitivity and specificity for each model and the three months after signing the payment agreement. Finally the AUC and ROC curve is illustrated in **Figure 10**. The results demonstrate the higher sensitivity and specificity is achieved by the neural networks in month 1 and 3. For month 2 the gradient boosting ensamble performed the best. In general, the AUC scores are better for the first month and decrease for the second and third months they decrease. The logistic regression is the model with the worse performance. Hence, the neural networks are recommended to be used in the future.

Table 16. Accuracy, sensitivity and specificity of each month for the models.

Period	Accuracy	Sensitivity	Specificity	Model
Month 1	60.91	75.66	58.19	Logistic Regression
Month 2	64.68	55.81	69.27	
Month 3	63.88	57.4	65.65	
Month 1	62.89	70.95	62.10	Gradient Boosting
Month 2	56.75	82.50	43.12	
Month 3	65.43	48.48	81.25	

Month 1	57.33	77.87	53.57	Neural Networks
Month 2	62.21	70.17	61.88	
Month 3	53.36	67.56	52.99	

Figure 10. ROC and AUC curve for each month for the logistic regression, gradient boosting, and neural network models.

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

CONCLUSIONS

In this work, a study was conducted in a Debt Collection Agency where the aim is to classify the potential payments of the clients in the first, second and third months after signing an agreement. The problem was formulated as a supervised classification problem, following the SEMMA methodology for data mining projects. In the exploratory data analysis, the relationship between the databases was understood and the process to join these databases was idealized to create the response variables for month 1, month 2 and month 3. A forward stepwise selection method was used to reduce the dimensionality of the dataset and select the most important variables. An "SMOTE" oversampling technique was also applied to the unbalanced data set and reduce the disproportionality of the class of clients paying the debt, which was also the class of interest. Three machine learning models were applied being Logistic Regression, Gradient Boosting and Neural Networks. The models were evaluated using the metrics of precision, sensitivity, specificity, and AUC. The results showed that neural networks outperformed the models during the first and third month of the prediction, while gradient boosting performed best during the second month of the prediction. The variables that affect that a client pay the next month prior to an agreement are Transferor, Qualification in Risk Center, Education Level, Purchase value. It is recommended to use a neural network to classify customers according to their payment probability of payment and optimize the company's collection processes.

REFERENCES

Allamsetty, A. (2018). A Brief Introduction to Gradient Boosting. Boston, United States: Northeastern University.

Al-jabery, K., Wunsch II, D., & Ajayi, T. (2019). Data analysis and machine learning tools in python. Missouri: Elsiever Inc.

Bahrami, M., Bozkaya, B., & Balcisoy, S. (2020). Using Behavioral Analytics to Predict Customer Invoice Payment. Big Data, 8(1), 25–37. doi:10.1089/big.2018.0116

Beck, T., Grunert, J., Neus, W., & Walter, A. (2017). What Determines Collection Rates of Debt Collection Agencies. Financial Review, 52(2), 259-279.

Benavides, A. R. (2018). Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones. Sevilla, Andalucía: Universidad de Sevilla.

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

Chen Yeh, I., Lien C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.   Expert Systems with Applications 36 (2009) 2473–2480

Dash, S. (2020, October 21). Machine Learning. Retrieved from https://www.machinelearningplus.com/machine-learning/gradient-boosting/

Gneiting, T., Vogel, P., & Walz, E.-M. (2019). Receiver Operating Characteristic (ROC) Curves. Baden-Wurtemberg, Germany: Heidelberg Institute for Theoretical Studies (HITS) and Karlsruhe Institute of Technology (KIT).

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

Ippolito, P. P. (2019). An introduction on how to fine-tune Machine and Deep Learning models using techniques such as: Random Search, Automated Hyperparameter Tuning and Artificial Neural Networks Tuning. Towards Data science, 12-19.

Kumar, G., Tirupathaiah, K., Krishna Reddy, B. (2019). Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques. International Journal of Computer Sciences and Engineering, 7(6), 842-846.

Matsumaru et al. (2019). Bankruptcy Prediction for Japanese Corporations using Support Vector Machine, Artificial Neural Network, and Multivariate Discriminant Analysis. Volume 1, No. 1, May 2019 pp. 78 - 96

Rial, R. (2005). Best Practice in Consumer Collections. Articles and Whitepapers on Collection & Recovery, VRL Publishing.

SAS Help Center. (2017). SEMMA Methodology. Retrieved from: https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.ht

Shoghi, A. (2020). Debt Collection Industry: Machine Learning Approach. Journal of Money and Economy Vol. 14, No. 4, Fall 2019 pp. 453-473

Superintendencia de Bancos del Ecuador SBE. (2021). Gestión de cobranza junio 2021. Retrieved from: https://estadisticas.superbancos.gob.ec/portalestadistico/portalestudios/?page_id=182

US Department of Treasury. (2015). Chapter 6: Delinquent Debt Collection. Retrieved from: https://www.fiscal.treasury.gov/files/dms/chapter6.pdf