UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Aplicación de Machine Learning a través de la metodología CRISP-DM para la predicción de pago por acuerdo en una empresa de cobro de deudas.

.

Martín Fernando Calero Pérez

Ingeniería Industrial

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

Quito, 17 de mayo de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias es Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Aplicación de Machine Learning a través de la metodología CRISP-DM para la predicción de pago por acuerdo en una empresa de cobro de deudas.

Martín Fernando Calero Pérez

Nombre del profesor, Título académico **María Gabriela Baldeón Calisto, Ph.D.**

Quito, 17 de mayo de 2022

© DERECHOS DE AUTOR

Nombres y apellidos: Martín Fernando Calero Pérez

Código: 202430

Cédula de identidad: 1723535579

Lugar y fecha: Quito, 17 de Mayo de 2022

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UNPUBLISHED DOCUMENT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

El cobro de deudas de una agencia de cobro de deudas (DCA) se ha vuelto más difícil debido a la pandemia. No obstante, en el último año la población se ha endeudado más, mientras que ha habido una disminución en la morosidad. Esto ha creado una oportunidad para que las DCA establezcan estrategias para mejorar el proceso de cobro de deudas. En este trabajo se implementa la metodología CRISP-DM en una DCA ecuatoriana para desarrollar un algoritmo de aprendizaje automático que prediga la probabilidad de acuerdo de pago de un deudor para establecer una estrategia de cobro de deudas. Se recopila, limpia y preprocesa un conjunto de datos desequilibrado con 7 447 856 registros para entrenar un Random Forrest Classifier, Gradient Boosting Machine, Regresión logística y un Multilayer Perceptron, utilizando una técnica de submuestreo aleatorio. El rendimiento de los modelos se compara utilizando las métricas de evaluación de sensibilidad, especificidad y AUC. El algoritmo de mejor rendimiento es Gradient Boosting Machine con una puntuación de sensibilidad de 0,97, especificidad de 0,93 y AUC de 0,98 en el conjunto de validación. Este algoritmo también permite identificar las características más importantes que contribuyen a la clasificación, siendo estas los días de mora, el día entre la adquisición de la cuenta y la fecha de incumplimiento, el nombre de la categoría comercial, el nombre del titular anterior de la cuenta, y el número de contactos directos realizados por un robot.

Palabras clave: Predicción de probabilidad de pago, Machine Learning, Random Forrest Classifier, Gradient Bosting Machine, Logistic Regression, Multi-Layer Perceptron, CRISP-DM.

ABSTRACT

Debt collection from a debt collection agency (DCA) has become more difficult due to the pandemic. Nevertheless, in the last year the population has incurred in more debts, while there has been a decrease in default loans. This has created an opportunity for DCAs to stablish strategies to improve the debt collection process. In this work, the CRISP-DM methodology is implemented in an Ecuadorian DCA to develop a machine learning algorithm that predicts a debtor´s payment agreement probability to stablish a debt collection strategy. An unbalanced dataset with 7,447,856 registers is gathered, cleaned, and preprocessed to train a Random Forrest Classifier, Gradient Boosting Machine, Logistic Regression, and Multi-Layer Perceptron using a random under-sampling technique. The models' performance is compared using the sensitivity, specificity, and AUC evaluation metrics. The best performing algorithm is the Gradient Boosting Machine with a sensitivity score of 0.97, 0.93 specificity, and 0.98 AUC on the validation set. This algorithm also allows to identify the most important features that contribute to the classification, these being the days past due, the day between the acquisition of the account and the default date, the name of the business category, the name of the prior account owner, and the number of direct contacts performed by a robot.

Key words: Payment Probability Prediction, Machine Learning, Random Forrest Classifier, Gradient Bosting Machine, Logistic Regression, Multi-Layer Perceptron, CRISP-DM..

TABLE OF CONTENTS

TABLE INDEX

FIGURE INDEX

INTRODUCTION

Debt management and collection is an integral process for the financial well-being of companies. In the last decades there has been an increase in the usage of loan services, such as credit card debts, which has made debt recollection management extremely important (Arora et al, 2022). Ecuador's credit portfolio, which records the credit amounts incurred by people in the country, indicates that the total amount grew at a rate of 18.1% between February 2021 and February 2022, and continues to show a sustained recovery after the contraction of the year 2020 (BCE, 2022). This means that the population is more economically active after the first year of the pandemic and can acquire more debt. On the other hand, the delinquency rate has decreased from 3.1% in February 2021 to 2.8% in February 2022 (BCE, 2022). This shows that even after a considerable increase in the quantity of loans, more debts are being paid. Nevertheless, the percentage increase in credit loans implies that more debt collection processes will be performed.

As Fedaseyeu (2020) states, when a credit card loan goes into default, the lender usually starts a debt collection process. It involves two types of approaches: in-house debt collection and third-party debt collection. In the in-house debt collection, creditors attempt to collect the debt on their own, while for third-party debt collection the creditors outsource debt collection to third-party agencies. As explained by Buitrón et al (2022), the performance and profitability of a third-party debt collection agency (DCA) is often measured by the collection success rate. Many times, the debt collection strategies used in Ecuadorian DCAs are subjective and much of the existing information is not considered. Therefore, an objective approach that maximizes debt collection return is needed.

Machine Learning (ML) has been used as a tool to analyze different characteristics of debt accounts and classification of the debt. Such is the case of predicting if a debt will be paid or

not by a client, or generating models that make a classification for credit card default (Arora et al., 2022; Louzada et al., 2016). Additionally, ML algorithms can be used to identify the most significant variables that contribute to the classification of a loan. Hence, companies can utilize this information to stablish diverse strategies to modify the debt collection management variables to improve the response. This project is developed in a DCA located in Quito, Ecuador that buys debt accounts from other companies and also provides a debt collection management service for external companies. The DCA uses three methods for debt collection, namely direct contact, indirect contact, and no contact. Direct contact is when the collection agency calls and has a conversation with the debtor. Indirect contact is when the DCA approaches the debtor via relatives, and no contact where no approach takes place. The intensity of the collection strategy is given by the amount and type of contacts that are produced in a month. So, an intensity increase could be to change from an indirect to a direct contact, or an increase in a type of contact. Also, a contact may be done by a person or with the help of a robot. The debt collection process has become more difficult to the company because of strict lockdown measurements and economic recession.

In this work, ML models are applied to predict a debtor's behavior and their probability of payment after signing a payment agreement. The CRISP-DM methodology is chosen to structure the project, which is specifically designed for a cross industry data mining project. This methodology contains six steps, where the dataset provided by the company with information of the last year is cleaned and preprocessed. Then, a comparison between a Random Forrest Classifier, Gradient Boosting Machine, Logistic Regression, and Multi-Layer Perceptron is performed using the sensitivity, specificity, and AUC evaluation metrics to select the best performing model. The Gradient Boosting Machine is chosen as the best performer. At the deployment stage of the methodology, the best model is utilized to identify the most

significant variables for debt collection, and calibrated to provide a confident prediction of payment probability.

DEVELOPEMENT OF THE TOPIC

1. Objectives

Main Objective:

- Use machine learning algorithms in an Ecuadorian debt collection agency to identify the most relevant characteristics of debtors and predict their debt payment probability.

Specific objectives:

- Apply the CRISP-DM methodology to obtain relevant results that allow the optimization of a debt collection strategy.
- Compare the performance of ML models' to achieve high prediction performance.

2. Literature Review

The first conception of what it is now known as Machine Learning came from Frank Rosenblatt, when he developed the Perceptron. The latter was an algorithm based upon the nervous system function that could recognize letters of the alphabet (Fradkov, 2020). Since then, ML has seen a major reappearance due to advances in computing power and data availability (Edgar & Manz, 2017). ML is a subset of Artificial Intelligence (AI), and is composed by algorithms that recognize patterns in historical data, acquire experience thought repetition, and make calculations that automate decision-making processes (Shobha & Rangaswamy, 2018). ML is generally classified as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. As Luxburg & Schölkopf (2011) describe, supervised learning consists of a training set of data X_i, correspondent to an input space or space of instances, and a set denoted as Y_i with the output or label space. In Supervised Learning, classification is one of the most employed and studied problems. These types of algorithms use the characteristics of the space of instances to split them in categories or classes

and label them as an output. There are many options to choose from for the ML application in this project. As an example presented by Azeem et al (2019) in a systematic literature review on ML algorithms applied for classification in code smell detection, from 15 final papers analyzed, the top five models used are Decision Tree (DT), Support Vector Machine (SVM), Random Forrest Classifier (RFC), Naïve Bayes (NB) and JRip; The best performance was attained with JRip, RFC, and tree-based classifiers. Within ML algorithms, there is a branch called Deep Learning that is composed of Artificial Neural Network (ANN) based methods. ANNs have multiple hidden layers of neurons (nonlinear individual processing units) that transform representations of individual features of a dataset to make a classification. ANNs are particularly useful with raw, high dimensional, and heterogenous datasets (Zhang et al., 2017) (Baldeon-Calisto and Lai-Yuen, 2020a, 2020b, 2021).

Machine learning has been successfully applied in the area of finance and banking. Yu et al., (2022) found that Classification and Regression Trees (CRT) have a good performance when classifying credit rating categories for decarbonized firms. The model achieved a score over 0.9 in F1-score, specificity, and accuracy. The authors also make a comparison between ANN, RFC, and SVM, where RFC had the second best performance. Antulov-Fantulin et al (2021) compared the performance of Gradient Boosting Machine (GBM), RFC, Lasso Regression (LASSO) and ANN, with 4 different approaches to mitigate unbalanced datasets for predicting bankruptcy in Italy. They found that GBM had the best performance, with a ROC of 0.98 Louzada et al., (2016) compared various classification methods for predicting credit scores and found that Logistic Regression (LR), ANN and DT were the most used across several papers, but the best performers were Fuzzy Logic and Genetic algorithms; Nevertheless, a big challenge is their implementation given the high computational cost. On a similar analysis, Arora et al (2022) performed a comparison between KNN, CRT, RFC, LR, SVM and NB when predicting credit card defaults. They achieved similar results between models, ranging between

0.76 and 0.82 on accuracy for the test sets. The best performing algorithms in order were SVM, LR and RFC. Buitrón et al (2022) tested LR, GBM and ANN with various error measurements to predict which clients will pay after a maximum period of three months. Making an emphasis in the specificity, they concluded that the best performer overall is ANN. Yet, on accuracy the best model was LR, and, when taking in account the AUC, GBM had higher results.

3. Methods

In this work the Cross Industry Standard Process for Data Mining (CRISP-DM) is implemented, which is an iterative methodology composed by six phases as presented in *Figure 1*. CRISP-DM was specifically designed for ML, but appliable in any type of project.

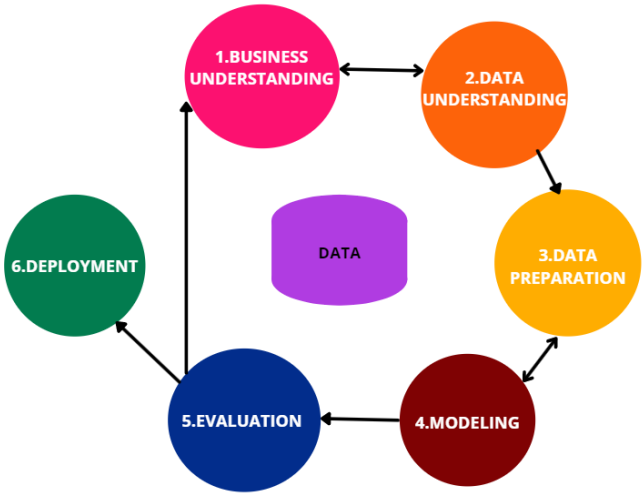


Figure 1: CRISP-DM methodology

As Schröer et al. (2021) presents, the first phase is business understanding, where the aim is to determine the data mining goal, type of problem, success criteria, and a compulsory project plan. The second phase is data understanding, where the data gathered is explored, described, and its quality examined. The third phase is data preparation, where inclusion and exclusion criteria are set for data features, in conjuncture with a preprocessing and cleaning of the data. The fourth phase is modeling, where the model selection is explained and constructed, stablishing their specific parameters and evaluation with the chosen success criteria. The fifth

phase is evaluation, where the results of the model are reviewed considering the business objectives. The final phase is deployment, where a formal integration of the prior work is instituted.

Each of the CRISP-DM phases are implemented in the proposed work as follows:

Phase 1: The business understanding begins with an approach to the DCA project supervisor for an explanation of the collection strategies and the dataset provided for the analysis. Also, a project charter was made to set the relevant characteristics of this study.

Phase 2: For the Data Understanding phase an Exploratory Data Analysis (EDA) is done. EDA's purpose is to "summarize the collected data in a meaningful way (…) this includes quantifying qualitative thoughts into quantitative data summaries" (Metcalf & Casey, 2016). With the steps presented by Nisbet et al. (2018b) the following calculations were performed: categorization of input variables into continuous and discrete, identification/definition of the target variable(s) in terms of available data elements, analysis of outliers and their management, calculation of correlation coefficients, identification and management of missing values, analysis of data bias and imbalance, and determining if the usage of data samples.

Phase 3: The data preparation begins with the EDA performed in phase 2. The provided dataset is unbalanced, hence, to improve the prediction power a randomized under-sampling solution is applied to level the quantity of observations from all classes. Missing values are treated by using a K-nearest-neighbors imputation is on the numerical data, and a simple imputation for categorical data. To avoid scaling problems, a normalization procedure is applied in the numerical data, and all categorical features are dummy encoded.

Phase 4: For the modeling phase the models selected are RFC, GBM, LR, and Multi-layer perceptron. The code is implemented with the application of a pipeline which ensures the preprocessing and imputation is performed for each dataset (training, validation, and test)

avoiding an interaction between them. A feature selection process, called Recursive Feature Elimination, and a Grid search hyperparameter optimization process is applied for each model.

Phase 5: For the evaluation, the sensitivity, precision, and AUC evaluation metrics are used to select the best performing algorithm. Also, the most relevant characteristics for the classification of clients are identified.

Phase 6: For the deployment, the results gathered from this project will help to stablish the new collection strategy for the company, and the models will be applied in other datasets.

4. Results

4.1.Business Understanding

In this step, three key elements are defined:

1.- Business objective: "To analyze and learn the most important features that predict a loan payment and stablish strategies that optimize the collection of debts."

2.- Determine data mining goals:

- Analyze the characteristics of the debtors
- Create agreement predictions.

3.- Generate a project plan.

4.2.Data Understanding

The data gathered from the company corresponds to the full 2021 year of debt management. It contains 59 columns, and 7.742.758 registers, from which 58 columns correspond to relevant variables from the clients and there is one response variable. Nine features corresponding to the robot debt management from the month 2 onwards were empty, so those features were dropped. In addition, a new variable is introduced that measures the difference between the date of the beginning of the debt and the date that the loan was acquired. The new feature is created to better represent the information from those two features that contained dates instead

of numerical values. Finally, the response variable is binary. It indicates if an agreement took place with the debtor where the company will receive only a part of the original debt or an installment payment. Therefore, the final dataset had 50 total features described in *Table 1*.

Table 1: Feature's Descriptions

Number	Code	Description	Type of information
1-12	D1 - D12	Information about the account's owner, the default date, original amount, actual amount, type of account of a prior owner, business category, and complementary data.	Debt
13-33	M1 - M21	The amount and type of contact used in a month	Debt Management
34-49	S1 - S16	Information about the debtor's socioeconomic status	Socioeconomic
50	RV	If a payment agreement was settled	Response Variable

There is a big unbalance between class 0 (no agreement) and class 1 (agreement) in the response variable, because the 0 class represents nearly 98% of the observations as shown in *Figure 2*. That is why a balancing solution must be considered. In this case, a random sampling approach is used.

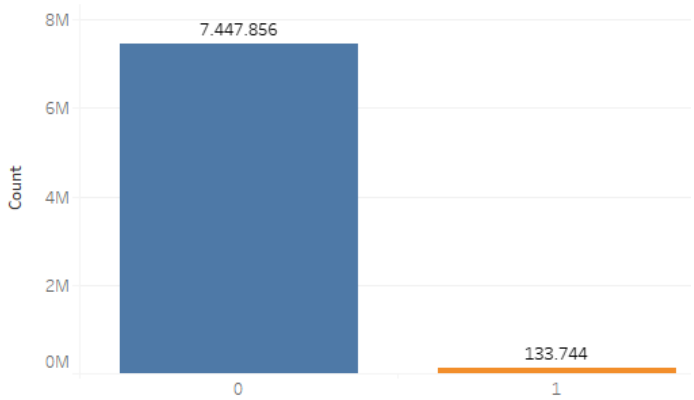


Figure 2: Agreement Class Imbalance

The predictor variables are cleaned. First, incoherent data is identified and treated in different manners. For instance, if numerical values are populating categorical features, the numbers are

erased, and converted into null values for a later imputation. Another important aspect for this step is the null value analysis. For each feature a count of null values was performed to identify a level of completeness. 82,765 registers had null values in almost all features, since the information contributed is negligent the observations are eliminated. The nine features eliminated correspondent to the robot debt management from the month 2 onwards that were empty increased the overall completeness of the dataset from 67.41% to 79.55%.

For the numerical features, both a correlation matrix and an atypical value analysis are done. An atypical value is set as the one that has a magnitude higher by 3 times the interquartile range. Outlier values are replaced by nan values to be later imputed in the preprocessing phase. The correlation matrix presented in *Figure 3*, where only correlation above 0.5 are shown to reduce clutter. To avoid multicollinearity, from all pair of variables that have a correlation higher than 0.8 only one variable is selected to remain in the dataset.

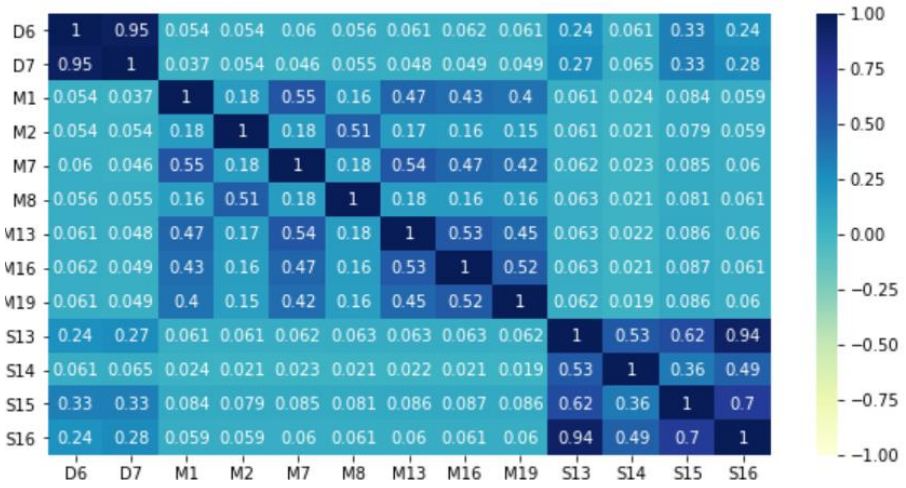


Figure 3: Correlation Matrix of predictor having a correlation of at 0.5.

4.3.Data Preparation

The class imbalance provokes a statistical prejudice to the ruling party skewing the findings of the report (Goyal et al., 2021). Therefore, to counteract the class imbalance in the response variable, a randomized under-sampling procedure is applied where all the observations of the

minority class are considering for training and the same number of observations of the majority class are sampled in a randomized manner. Considering that a sample of size of 16,604 observations provides a 99% confidence and 1% margin error in the current dataset, the balanced sample obtained using the under-sampling approach of 267,488 observations is statistically representative.

For this phase, a pre-processing is pipeline is implemented using the python sklearn library. The pipeline ensures that the training, validation, and test datasets do not have an interaction when being imputed. The objective of this step is to manage missing values, scale numerical variables, and dummy encode categorical values. There is a distinction in the procedure that was applied for the categorical and numerical features. For the missing values, instead of dropping all the registers which contained null values, an imputation of the mean was the strategy chosen. For the numerical features the type of imputation that was performed is KNN imputation, where the algorithm identifies the most similar registers and impute the mean from them. This is an improvement from a Simple Imputer, because instead of calculating the mean from all the values of a row, it does it from only those registers that are similar, taking in account the other features from the dataset. To avoid bias problems between scales of the numerical data, a normalization process is performed, were all the values are transformed to a value between 0 and 1. For the categorical data the simple imputation method was used because it has a feature that creates a category for missingness itself, and that provides more information to the model. Categorical data is also necessary to dummy-encode with OneHotEncoder, which transforms each class in each feature to a binary feature so that it can be processed by the algorithms which all have a mathematical basis. Finally, a column transformer helps to apply these processes to the original dataset. In *Figure 4* the pipeline process is shown.

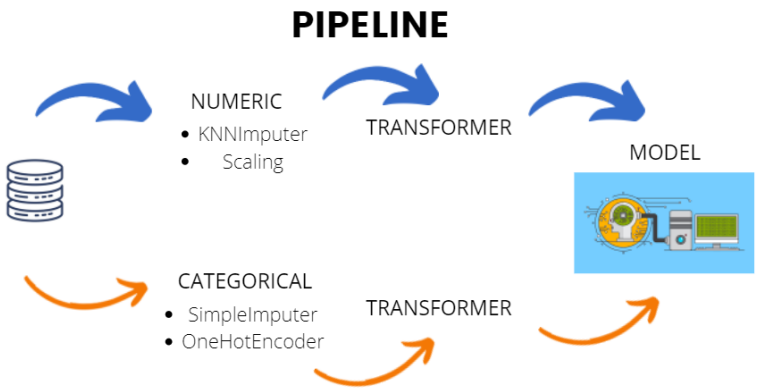


Figure 4: Pipeline Process

4.4.Modeling

For the modeling phase the RFC, LR, GBM and the Multilayer Perceptron (MLP) algorithms were selected because they have shown to provide excellent results in classification problems for banking, credit card scoring and defaults (Arora et al., 2022; Azeem et al., 2019; Buitrón et al., 2022; Louzada et al., 2016).

Random Forrest Classifier

This algorithm is a technique where multiple random tree classifiers are built on different subsets of data to reduce bias and variance (Yeturu, 2020). RFC is first used to evaluate each feature's importance. The results showed that only 18 variables had an importance over 1%, and 8 of them an importance over 2% as presented in *Table 2*. The feature importance is measured by a mean decrease over the impurity when a feature is taken out (Boulesteix et al., 2011). In this case is used Gini impurity that measures the probability of misclassification when an object is assigned to a class (Laber & Murtinho, 2019).

Table 2: Feature importance's for RF

Order	Feature	Importance
1	Days Past Due	0.272553
2	Days between the acquisition of the account and the default date	0.092945

3	No Contact Management of the month 0	0.066359
4	Type of account in a prior owner	0.057458
5	No Contact Management after 1 month	0.048700
6	No Contact Management after 2 months	0.039315
7	No Contact Management after 1 month with the robot	0.032100
8	Name of the prior account owner	0.029009

Logistic Regression

This algorithm uses a probability function called logit to make a classification, it "is one of the fundamental classification algorithms where a log odds in favor of one of the classes is defined and maximized" (Yeturu, 2020). For this model a Recursive Feature Elimination (RFE) is performed to select the most important features. Also, features with multicollinearity are removed.

Gradient boosting Machine

This model uses an initial function to create a first prediction, then it boosts this prediction "where incrementally, over steps, several weak classifiers are combined so as to reduce error" (Yeturu, 2020). For the application of this model a RFE was also implemented.

Multilayer Perceptron

As Lord et al, (2021) describes, the MLP neural network uses neurons of the same layer, which are not connected to each other, but connected to the neurons of the preceding and successive layers in the form of an activation function of a weighted summation of the outputs of the last hidden layer. The weights can be determined by solving an optimization problem. RFE was used for feature selection.

For each model a Grid Search was applied, were different values for important hyperparameters were tested. For the RFC the hyperparameter chosen were the impurity criterion (Gini or

Entropy), the number of tree estimators (between 10, 100, 1000, or 2000), and the maximum number of features used (the square root of the total number of features or the log2 of the total number of features. For the LR the hyperparameters chosen were the optimization solver (Liblinear or Saga), the inverse regularization strength "C" (0.1, 1, 10, 100) that controls the penalty to avoid overfitting, and the type of penalty itself was used (l1- Lasso Regression or l2 – Ridge Regression). In the case of the GBM, the hyperparameters chosen to be optimized are the learning rate (between 0.05, 0.1, and 0.2) and the number of estimators (between 100, 1000, and 5000). Finally, for MLP, the hyperparameters compared were the weight optimization solver (lbfgs, sgd, adam), the learning rate (constant, adaptive) that controls weather the learning rate value used is the same throughout the training, or if it adapts when the training loss stops decreasing, and the activation function for the neuron (tanh or relu).

4.5.Evaluation

The performance evaluation metrics calculated are the Accuracy, Precision, Specificity, Sensitivity and the area under the curve (AUC) of the Receiving Operating Characteristic curve. As Shobha & Rangaswamy (2018) and Zhu et al. (2010) explains, these metrics are calculated using the true positives (TP), true negatives(TN), false positives (FP), and false negatives (FN) from the model prediction. Accuracy is the amount of correctly classified instances from the total instances and calculated as presented in equation 1:

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (1)$$

Precision is the fraction of relevant instances among the retrieved instances as presented in equation 2:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Sensitivity is the fraction of relevant instances among the positive instances as presented in equation 3:

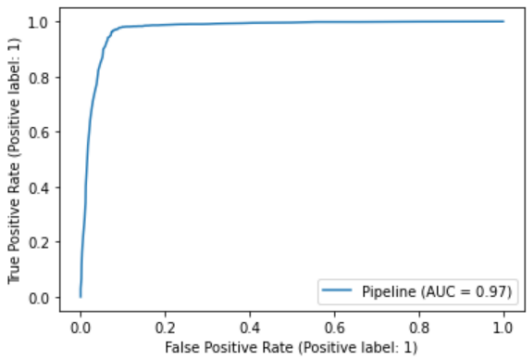
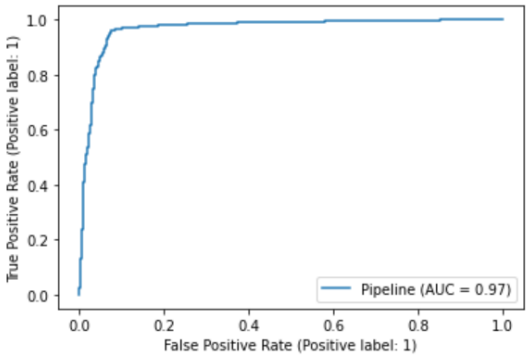
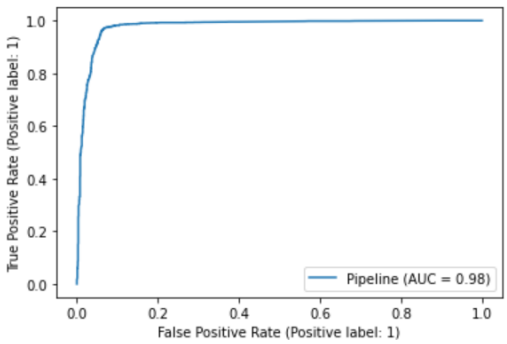
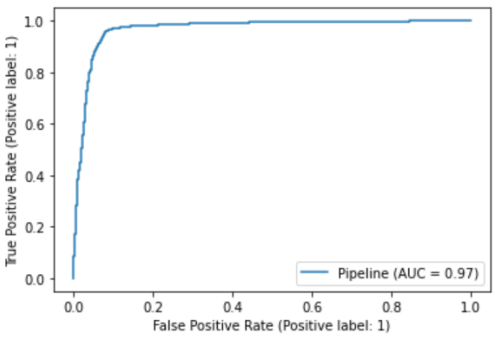
$$Sensitivity = \frac{TP}{TP+FN} \qquad (3)$$

Specificity is the fraction of non-relevant instances among negative instances as presented in equation 4:

$$Specificity = \frac{TN}{TN+FP} \qquad (4)$$

The ROC is a graph of the sensitivity and the true negative rate. The AUC is then considered as the probability that the model will be able to distinguish between classes. The ROC curves from the models are presented as follows in *Table 3*.

Table 3: ROC Curves for each model

Random Forrest Classifier	Logistic Regression
	
Gradient Boosting Machine	**Multilayer Perceptron**
	

For the parameter optimization performed with Grid Search the performance metric used is accuracy because it summarizes how well the model predicts for both classes. The parameters

that maximize accuracy for RFC are the entropy criterion, the square root of the number of features, and a 100 number of estimators. The best parameters for LR are the "Saga" solver, an inverse regularization strength of 0.1, and the l1 penalty. The best parameters for GBM are a learning rate of 0.05, and 100 estimators. For MLP the best parameters are the sgd solver, a constant learning rate, and the relu activation function. Once the best parameters were selected the metrics for each model were obtained from the training and validation sets as follows in *Table 4*.

Table 4: Performance Metrics

RFC			**LR**		
Metric	Training	Validation	Metric	Training	Validation
Accuracy	1.0	0.944938	Accuracy	0.932628	0.920275
Precision	1.0	0.926790	Precision	0.942393	0.937133
Specificity	1.0	0.921557	Specificity	0.944102	0.937827
Sensitivity	1.0	0.967724	Sensitivity	0.921071	0.903171
AUC	1.0	0.972202	AUC	0.972465	0.967777
GBM			**MLP**		
Metric	Training	Validation	Metric	Training	Validation
Accuracy	0.953106	0.950961	Accuracy	0.947858	0.931746
Precision	0.934548	0.932249	Precision	0.936952	0.929213
Specificity	0.932270	0.927368	Specificity	0.935871	0.926787
Sensitivity	0.974093	0.973952	Sensitivity	0.959931	0.936580
AUC	0.984171	0.978338	AUC	0.979618	0.968574

The RFC presents overfitting, where the metrics for the training set are much higher than the validation set. The LR and GBM models are balanced, which is desirable as they are generalizable for unseen data. To select the best model, the metrics of sensitivity and specificity are considered as per the company´s goal Sensitivity measures how well the model identifies the debtors that will settle an agreement, while specificity identifies the clients that

will not settle an agreement. In this case the GBM model outperformed RFC, LR and MLP for specificity, and had a minor decrease in sensitivity compared to LR and MLP. In terms of the Accuracy and AUC, the GBM has higher value. Finally, the five most important features for the model obtained with the mean decrease over the Gini impurity are described in *Table 5*:

Table 5: Feature Importance for GBM

Number	Feature	Importance
1	Days past due	0.852823
2	Days between the acquisition of the account and the default date	0.120480
3	Business category	0.005426
4	Name of the prior account owner	0.004941
5	Number of direct contacts performed by the robot	0.002457

As presented above the model highly relies on the Days past due, and Days between the acquisition of the account, and the default date to make a prediction. It is also important the Business category and the name of the prior account owner. On the other hand, the first management feature to appear is the number of direct contacts performed by the robot on the first month with an importance of 0.25%, which means that the DCA management is not contributing to the payment agreement as much as the prior descriptive characteristics of the clients and accounts bought. Additionally, feature importance provides a reference on those features that contribute the most to identify if a payment agreement will be done; so, it helps the DCA in various situations such as an early detection of possible payment agreements. A better analysis before buying a new account of debts or to focus the descriptive statistical analysis since there is a big amount of data available to process.

4.6.Model Calibration

The DCA wants to focus on the probability predicted by the model because they want to compare the effect in the probability prediction when they change the intensity of the debt management. This comparison is impossible to make only with the prediction of the classification, but a calibration of the probability's prediction is needed. For a model such as GBM the prediction outcome is binary (either 1 or 0), so the probability it predicts is calculated as the probability of the model getting the class right. As Vaicenavicius et al. (2020) explains, for the classifiers there is a real questioning of weather the model is making trustworthy probability predictions that can be interpreted as real-world probabilities. To assess this problem a graph called reliability curve is used, where the mean predicted probability of the classifier is compared to the real fraction of positives from the dataset:

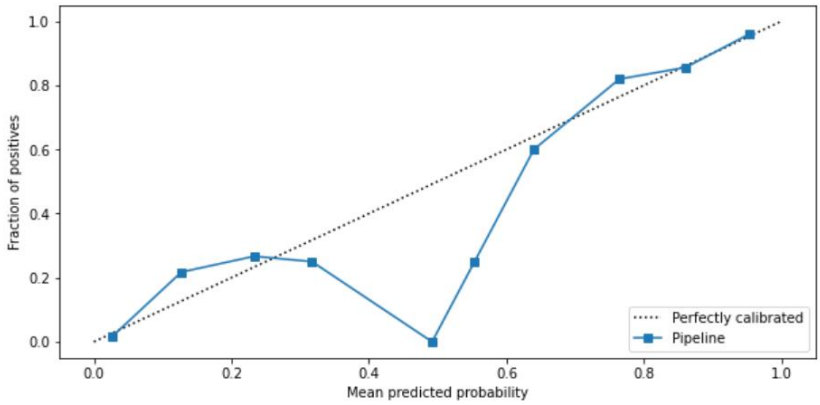


Figure 4: GBM reliability curve

As shown in Figure 4, there is a difference between the predicted and real probabilities of the model. The desired outcome would look like a diagonal straight line because that would mean that the real and predicted probabilities are the same. The model also presents a Brier Score, that measures the mean square difference between these probabilities, of 0.49. This score is better the closer to 0 it is. For the calibration of this model a sigmoid function stablished by Platt (1999) that transforms the predicted probabilities, assuming that the output of the model

is proportional to the log odds of a positive example. With the calibration, the reliability curve is shown as follows in Figure 5:

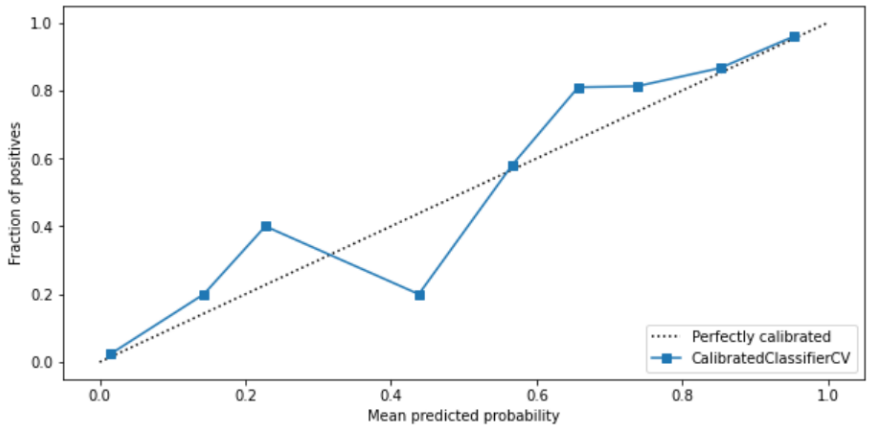


Figure 5: GBM calibrated reliability curve

Now the model has a better fit to a straight line, and presents a Brier Score of 0.425, so the model makes better predictions for probabilities. In **Error! Reference source not found.**, a graph that shows the probability density distribution of the prediction while differentiating the actual values is built to visualize how well the model is making the probability predictions.

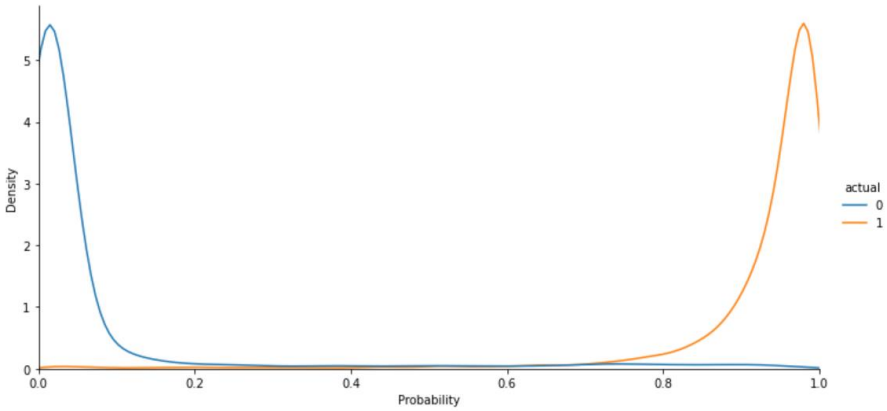


Figure 6: Prediction probability density function

The model makes good probability predictions since there is only a little portion of the probability density function for each class that overlaps with each other, and it is reflected in the high values obtained for the performance metrics.

CONCLUSIONS

After the pandemic, the amount of debt collection, and the ratio of people that perform payments has increased. So, for a DCA the analysis of the debt management is needed. The business understanding phase begun with an approach to the project supervisor in the company for an explanation of the collection strategies and the dataset that provided for the analysis. Also, the business objectives for the project were proposed. For the Data Understanding phase an Exploratory Data Analysis (EDA) was performed with the categorization of input variables into continuous and discrete, identification of the target variable(s) in terms of available data elements, analysis of outliers, calculation of correlation coefficients, identification and management of missing values. For the data preparation, depending on the models used, specific data manipulation were set in place. Due to the data imbalance on the training set, a random under sampling was performed. Furthermore, a pipeline of procedures for categorial and numerical data was implemented to avoid interactions between datasets when imputed. For Phase 4 the algorithms selected were RFC, GBM, LR, and MLP. For each model a hyperparameter tunning was performed to improve the default performance of the model. Once this step is done, in the evaluation phase the GBM was selected as the best model because it had the highest performance metrics. This model showed the most important features for the classification of the algorithm. Finally, for the deployment phase the conclusions of this projects are presented, and the code of the trained model will be used for the behavior prediction of the clients.

BIBLIOGREPHIC REFERENCES

Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior and Organization*, *183*, 681–699. https://doi.org/10.1016/j.jebo.2021.01.014

Arora, S., Bindra, S., Singh, S., & Kumar Nassa, V. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. *Materials Today: Proceedings*, *51*, 110–117. https://doi.org/10.1016/j.matpr.2021.04.588

Azeem, M. I., Palomba, F., Shi, L., & Wang, Q. (2019). Machine learning techniques for code smell detection: A systematic literature review and meta-analysis. *Information and Software Technology*, *108*(4), 115–138. https://doi.org/10.1016/j.infsof.2018.12.009

BCE. (2022). *Monitoreo de los principales indicadores internacionales*. https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Presentacion_Mar22.pdf

Boulesteix, A., Bender, A., Bermejo, J. L., & Strobl, C. (2011). Random forest Gini importance favors SNPs with large minor allele frequency. *Statistics*, *106*.

Buitrón, A., Rodríguez, C., & Calisto, M. B. (2022). *Machine Learning in Finance : An Application of Predictive Models to Determine the Payment Probability of a Client*.

Edgar, T. W., & Manz, D. O. (2017). Machine Learning. *Research Methods for Cyber Security*, 153–173. https://doi.org/10.1016/B978-0-12-805349-2.00006-6

Fedaseyeu, V. (2020). Debt collection agencies and the supply of consumer credit. *Journal of Financial Economics*, *138*(1), 193–221. https://doi.org/10.1016/j.jfineco.2020.05.002

Fradkov, A. L. (2020). Early History of Machine Learning. *IFAC-PapersOnLine*, *53*(2), 1385–1390. https://doi.org/10.1016/J.IFACOL.2020.12.1888

Goyal, A., Rathore, L., & Sharma, A. (2021). SMO-RF:A machine learning approach by random forest for predicting class imbalancing followed by SMOTE. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.12.891

Laber, E., & Murtinho, L. (2019). Minimization of Gini Impurity: NP-completeness and Approximation Algorithm via Connections with the k-means Problem. *Electronic Notes in Theoretical Computer Science*, *346*, 567–576. https://doi.org/10.1016/j.entcs.2019.08.050

Lord, D., Qin, X., & Geedipally, S. R. (2021). Data mining and machine learning techniques. *Highway Safety Analytics and Modeling*, 399–428. https://doi.org/10.1016/b978-0-12-816818-9.00016-0

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Luxburg, U. von, & Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic*, *10*, 651–706. https://doi.org/10.1016/B978-0-444-52936-7.50016-1

Metcalf, L., & Casey, W. (2016). Introduction to data analysis. *Cybersecurity and Applied Mathematics*, 43–65. https://doi.org/10.1016/B978-0-12-804452-0.00004-X

Nisbet, R., Miner, G., & Yale, K. (2018). A Data Preparation Cookbook. *Handbook of Statistical Analysis and Data Mining Applications*, 727–740. https://doi.org/10.1016/B978-0-12-416632-5.00018-9

Platt. (1999). Probablistic outputs for svm and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, *10*(3), 61–74.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Shobha, G., & Rangaswamy, S. (2018). Machine Learning. *Handbook of Statistics*, *38*, 197–228. https://doi.org/10.1016/BS.HOST.2018.07.004

Vaicenavicius, J., Lindsten, F., Widmann, D., Roll, J., Andersson, C., & Schön, T. B. (2020). Evaluating model calibration in classification. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*, *89*.

Yeturu, K. (2020). Machine learning algorithms , applications , and practices in data science. In *Principles and Methods for Data Science* (1st ed., Vol. 43). Elsevier B.V. https://doi.org/10.1016/bs.host.2020.01.002

Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change*, *174*(October 2021), 121255. https://doi.org/10.1016/j.techfore.2021.121255

Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, *22*(11), 1680–1685. https://doi.org/10.1016/J.DRUDIS.2017.08.010

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *Northeast SAS Users Group 2010: Health Care and Life Sciences*, 1–9.