Application of Machine Learning algorithms for the prediction of payment by agreement in a debt collection company with the CRISP-DM methodology

Martín Calero Peréz, 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
mfcalero@estud.usfq.edu.ec, mbaldeonc@usfq.edu.ec

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

Daniel Riofrío
Universidad San Francisco de Quito USFQ, Colegio de Ciencias e Ingenierías, Carrera de Ingeniería en Ciencias de la Computación and The Applied Signal Processing and Machine Learning Research Group USFQ, Quito, Ecuador 170157
driofrioa@usfq.edu.ec

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 probability and 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, specificity of 0.93, and AUC of 0.98 on the validation set. This algorithm also allows to identify the most discriminative features for the prediction, 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.

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

1. 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 loans, 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 has increased its economical activity 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 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. Meanwhile 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 analytical 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 that modify the debt collection management to improve the response.

This project is developed in a DCA located in Quito, Ecuador that buys debt accounts from external companies and also provides a debt collection management service. 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 when 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. Hence, 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 cross industry data mining projects. 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 with a sensitivity score of 0.97, specificity of 0.93, and AUC of 0.98. At the deployment stage of the methodology, the latter model is utilized to identify the most significant variables for debt collection.

1.1 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 distinct ML models to achieve a 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, 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. In this problem, algorithms use the characteristics of the input space to split the data in categories 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 papers analyzed the top five models used are Decision Tree (DT), Support Vector Machine (SVM), Random Forrest Classifier (RFC), Naïve Bayes (NB) and JRip. Within ML algorithms, Artificial Neural Networks (ANNs) have gained notoriety due to their high performance with raw, high dimensional, and heterogenous datasets (Zhang et al., 2017) (Baldeon-Calisto and Lai-Yuen, 2020a, 2020b, 2021). ANNs have multiple hidden layers of neurons that transform representations of individual features of a dataset to make a classification.

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 ANNs, RFC, and SVM, where RFC had the best performance. Antulov-Fantulin et al (2021) compared the performance of Gradient Boosting Machine (GBM), RFC, Lasso Regression, and ANNs 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 in their implementation is 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 accuracy on 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, considering accuracy the best model was LR, and when taking in account the AUC the 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.

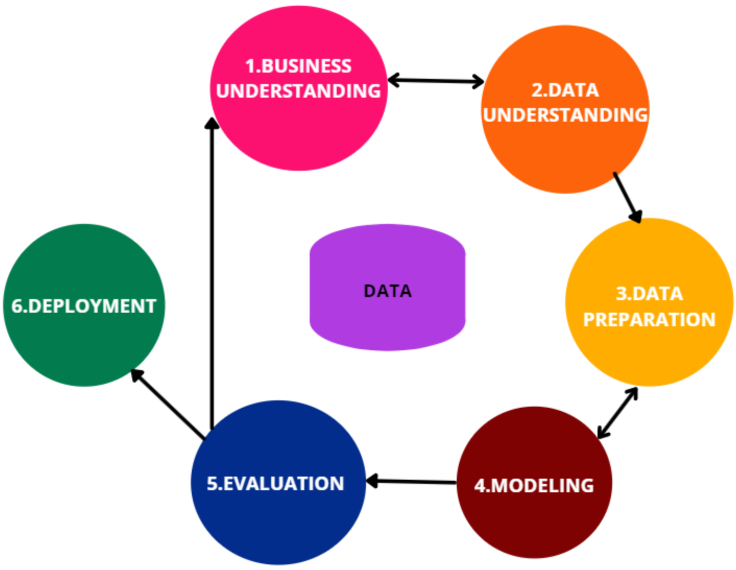

Figure 1: CRISP-DM methodology

As Schröer et al. (2021) states, 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 models are selected and trained, and success criteria established. 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. In phase 1 the problem is understood by interviewing the DCA`s project supervisor and obtaining an explanation of the collection strategies and dataset provided. In phase 2, an Exploratory Data Analysis (EDA) of the dataset is performed. 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). Following the steps presented by Nisbet et al. (2018b), input variables are categorized into continuous and discrete, the response variable is identified, an outlier and null value analysis implemented, correlation coefficients calculated, and an analysis of data bias and imbalance done. Given that the provided dataset is unbalanced, in phase 3 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 for 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. In phase 4, the ML 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 also applied for each model. In phase 5, 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. Finally, phase 6 will be performed by the DCA after providing the results of this work.

4. Results

4.1 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. 58 columns correspond to predictor variables and store information about the clients. The last column has the response variable that is binary. The latter indicates if a payment agreement was settled. From the 58 predictor variables, 9 features corresponding to the robot debt management from month 2 onwards are empty, so those variables are 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. Therefore, the final dataset has a total of 50 features as described in Table 1.

Table 1: Database Feature's Description

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 Information
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

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

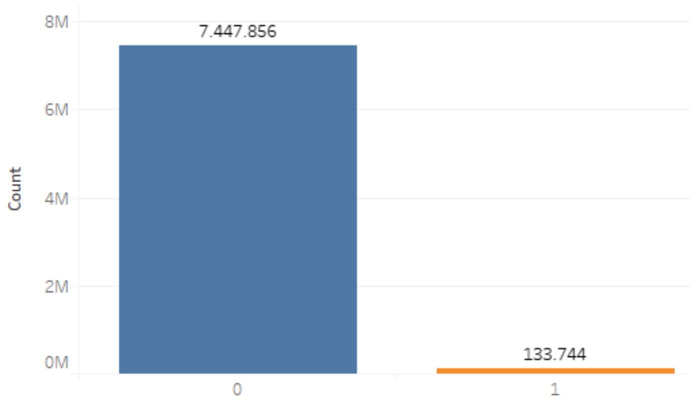

Figure 2: Agreement Class Imbalance

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 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.

For the numerical features, both a correlation matrix and an atypical value analysis is 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. In Figure 3 the correlation matrix is presented, where only features with a correlation above 0.5 are shown to reduce clutter. To avoid multicollinearity, only one variable is selected to remain in the dataset from all pair of variables that have a correlation higher than 0.8.

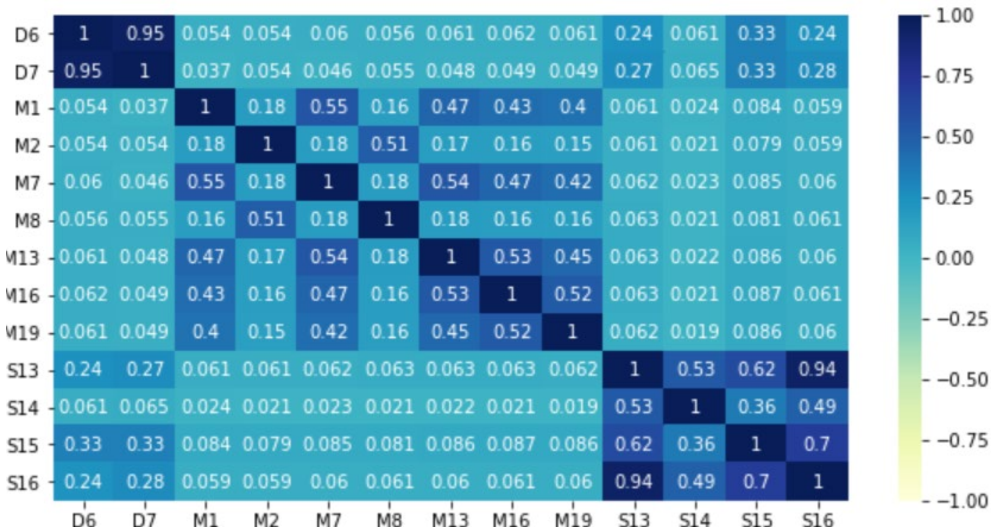

Figure 3: Correlation Matrix of predictor having a correlation higher than 0.5.

4.2 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. Noting that a sample of size of 16,604 observations provides a 99% confidence and 1% margin error on 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 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 is 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.

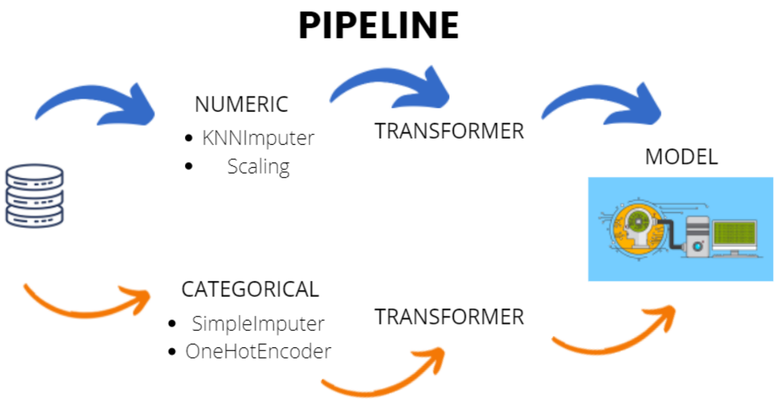

Figure 4: Pipeline Process

4.3 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).

4.3.1 Random Forrest Classifier

In this algorithm, multiple random tree classifiers are built on different subsets of the dataset 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, the Gini impurity measurement is applied to 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

4.3.2 Logistic Regression
This algorithm uses a probability function called logit to make a classification. LR "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.
4.3.3 Gradient boosting Machine

This model uses an initial function to create a first prediction, then this prediction is boosted "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.

4.3.4 Multilayer Perceptron
As Lord et al, (2021) describes, the MLP neural network is a graphical model where neurons from the same layer 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 is used for feature selection before implementing this model. 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 (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), how the learning rate is updated during training (constant, adaptive), and the activation function for the neuron (tanh or relu).

4.4 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´s prediction. Accuracy is the amount of correctly classified instances from the total instances and calculated as presented in equation 1:

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

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

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

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

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

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

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

The ROC is a graph that displays the sensitivity and the true negative rate with different thresholds. The AUC is then considered as the probability that the model will be able to distinguish between classes.

The best hyperparameters for the models after applying Grid Search are selected using the 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 3. Moreover, the ROC curves and AUC from the models are presented in Figure 5.

Table 3: Performance Metrics

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

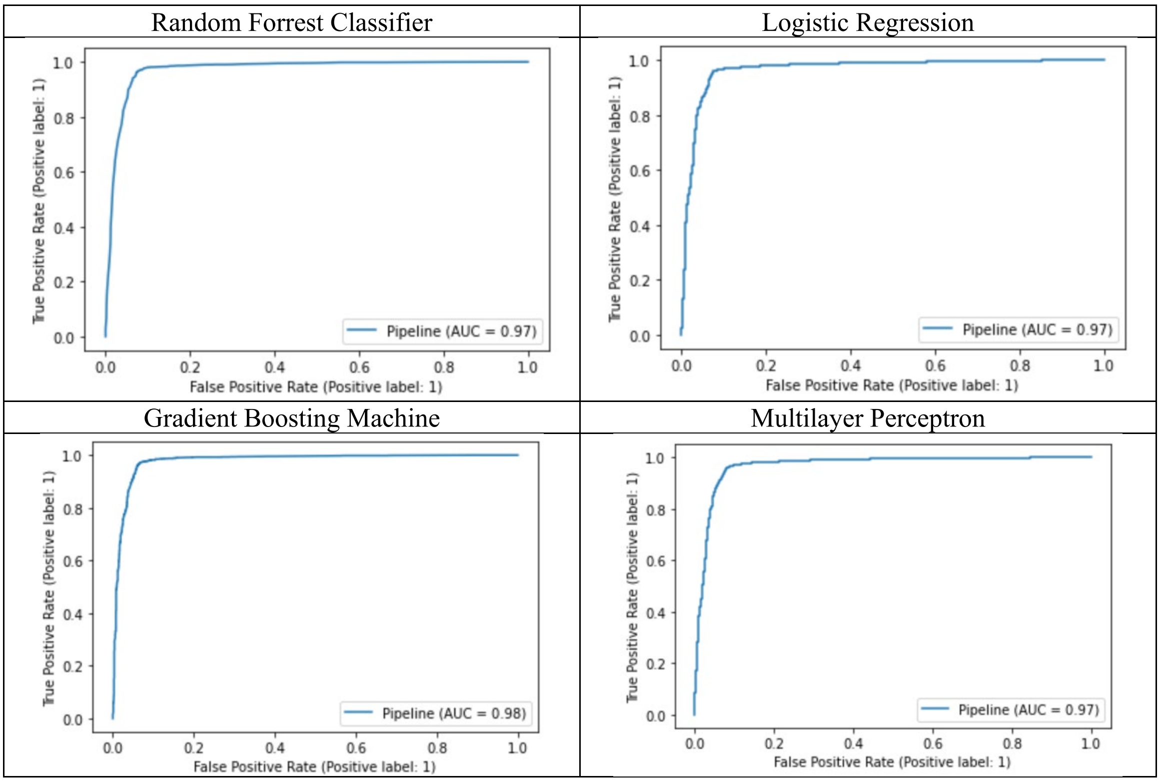

Figure 5: ROC Curves for each model

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 generalize to 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 outperforms 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 a higher value. Finally, the five most important features for the model obtained with the mean decrease over the Gini impurity are described in Table 4.

Table 4: 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.

4.4.1 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.

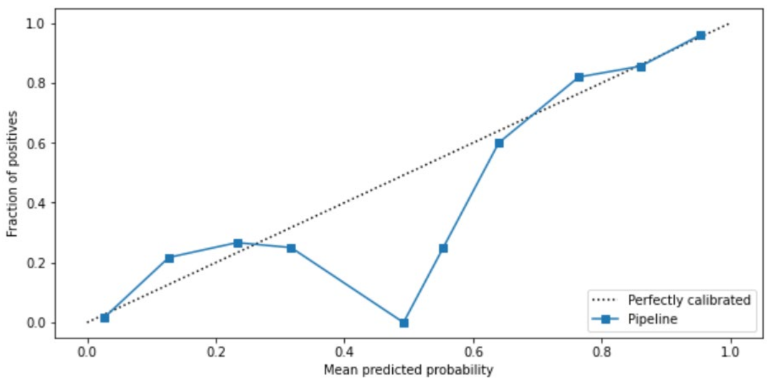

Figure 6: GBM reliability curve

As shown in Figure 6, 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, which is 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 7.

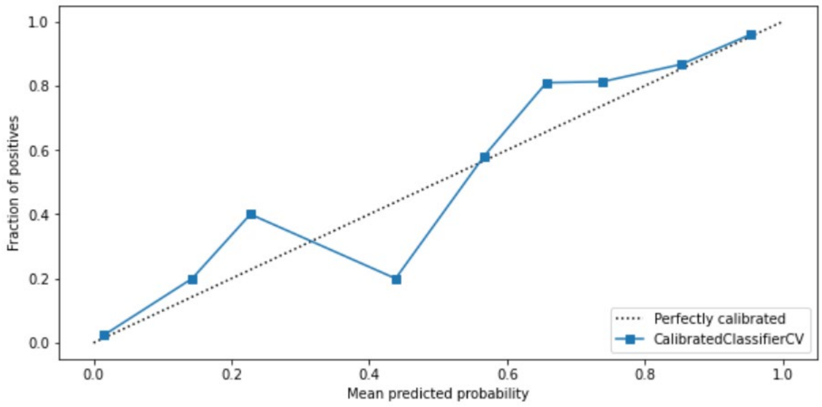

Figure 7: GBM calibrated reliability curve

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 Figure 8, 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.

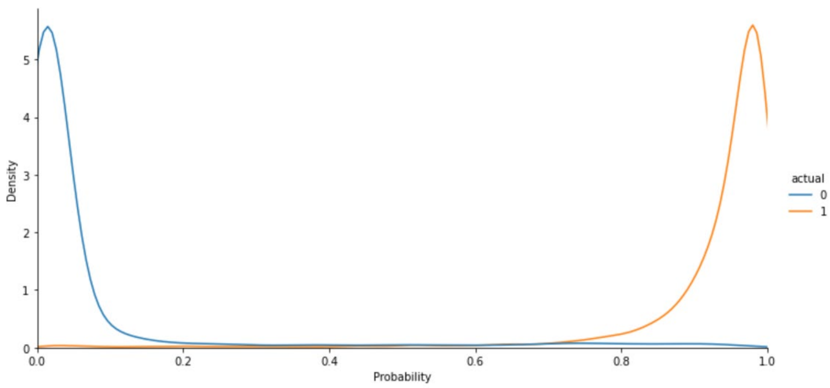

Figure 8: Prediction probability density function

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.

5. Conclusion

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 process is necessary. In this project a ML algorithm is developed for an Ecuadorian DCA to predict the payment probability of debtors using the CRISP-DM methodology comprised of 6 phases. 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 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, analysis of outliers, calculation of correlation coefficients, identification and management of missing values. For the data preparation, specific data manipulation was set in place depending on the models used. Due to the data imbalance on the training set, a random under sampling technique was applied. For Phase 4, the algorithms of RFC, GBM, LR, and MLP were selected for training. 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.

References

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

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

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

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

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

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

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. Machine Learning. Research Methods for Cyber Security, 153–173. (2017). https://doi.org/10.1016/B978-0-12-805349-2.00006-6

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

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

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

Laber, E., & Murtinho, L. 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. (2019). https://doi.org/10.1016/j.entcs.2019.08.050

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

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

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

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

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

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

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

Shobha, G., & Rangaswamy, S. Machine Learning. Handbook of Statistics, 38, 197–228. (2018).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. Machine learning algorithms , applications , and practices in data science. In Principles and Methods for Data Science (1st ed., Vol. 43). Elsevier B.V. (2020). https://doi.org/10.1016/bs.host.2020.01.002

Yu, B., Li, C., Mirza, N., & Umar, M. Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. Technological Forecasting and Social Change, 174(October 2021), 121255.

(2022).https://doi.org/10.1016/j.techfore.2021.121255
Zhang, L., Tan, J., Han, D., & Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today, 22(11), 1680–1685. (2017). https://doi.org/10.1016/J.DRUDIS.2017.08.010
Zhu, W., Zeng, N., & Wang, N. 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. (2010).

Acknowledgements
The authors thank the Applied Signal Processing and Machine Learning Research Group of Universidad San Francisco de Quito for providing the computing infrastructure (NVidia DGX workstation) to implement and execute the developed source code.

Biographies
Martín Calero is a student of Industrial Engineering at Universidad San Francisco de Quito, Ecuador. Throughout his student career he has developed interests in data analytics, simulation, operations research, logistics and supply chain management. He is currently working in a consulting company specialized in actuarial sciences and data analytics.

María Baldeon Calisto is a professor and researcher in the Industrial Engineering department at Universidad San Francisco de Quito, Ecuador. She is also director of CATENA Research Institute. She earned a BS in Industrial Engineering and a master's degree in safety, health, and environment at Universidad San Francisco de Quito. She also has a master's and Ph.D degree in Industrial Engineering from University of South Florida, USA. 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.

Daniel Riofrío works as a full-time professor at the Computer Science Engineering program at Universidad San Francisco de Quito USFQ where he is the coordinator of the program. He earned his Ph.D and MSc. at University of New Mexico, USA in Computer Science. His main research interests span novel applications of machine learning, computer vision, security and cybersecurity, and optimization techniques for radiotherapy and radiosurgery. He has supported the Open Technology Fund as a senior Information Controls Fellow during 2017 and holds a US patent 9,844,684 / 10,850,122 granted December 2017 / 2020.