






Financial Credit Risk Measurement Using a Binary Classification Model

Oscar Chiluiza¹, Cathy Guevara-Vega² , Antonio Quiña-Mera² ,
Pablo Landeta-López² , and Javier Montaluísa¹

¹ Universidad de las Fuerzas Armadas ESPE, Quijano y Ordoñez Street, Latacunga, Ecuador
{owchiluiza, fjmontaluísa}@espe.edu.ec

² eCIER Research Group, Universidad Técnica del Norte, 17 de Julio Avenue, Ibarra, Ecuador
{cguevara, aquina, palandeta}@utn.edu.ec

Abstract. Currently, financial institutions have several problems in the analysis of information to grant a credit or a loan, causing losses that involve collection expenses, notifications, legal processes, among others. Thanks to the digital transformation and technological progress, Artificial Intelligence and especially Machine Learning can be used to analyze customer data and predict non-compliance with the payment of their obligations to institutions. The objective of this work is to apply CRISP-DM (Cross Industry Standard Process for Data Mining) methodology and the Random Forest (XGBoost), Logistic Regression and Neural Networks of supervised learning models, to implement a predictive model that allows evaluating credit risk. With the result of the application of the predictive model, it is concluded that the use of Machine Learning tools helps to optimize the evaluation of credit risk in financial entities. Once CRISP-DM methodology was used for the analysis, development, and evaluation of the models, it was concluded that the most efficient model is the Random Forest. Based on this experience, future work could implement this type of model in other areas such as: fraud detection, customer segmentation or a recommendation engine that can suggest financial products and services based on customer needs and behaviors.

Keywords: Predictive model · CRISP-DM · Random Forest · Logistic Regression · Neuronal Networks

1 Introduction

The risks are the possibility of occurrence of adverse events that can modify the normal progress of the established functions, which can generate delays in achieving the established objectives and losses in the financial entities [1]. In Ecuador a problem has been evidenced. SEPS (Superintendencia de Economía Popular y Solidaria) [2] as a control and regularization entity for the Savings and Credit Cooperatives of Ecuador, has not established regulations that allow the use of models for credit management, this has even caused some cooperatives to close due to the increase in late payments in their credit portfolios and the lack of liquidity. On the other hand, we have noticed that there is a limitation of technological tools that help control risk in financial entities; therefore, these

entities have been forced to invest more resources to find their own alternatives to mitigate credit risk. Considering this problem, we decided to apply Design Science Research (DSR) approach [3] as the research methodology for this study. The following research question is proposed: *How to assess credit risk in Savings and Credit Cooperatives by applying a prediction model.*

In addition to this, we propose the following hypothesis:

H₀: If a predictive model is developed, then the credit risk will be evaluated, optimizing the granting of credits in the financial institution.

This work aims to answer the research question with the creation of a predictive model for the evaluation of the credit risk of Savings and Credit Cooperatives using Machine Learning models and prediction techniques and the CRISP-DM methodology (Cross Industry Standard Process for Data Mining) [4] which contains six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. We decide to use CRISP-DM because it is a popular methodology for Data Mining projects [4, 5]. The purpose of the predictive model is to identify the most important aspects that characterize credit clients, and with an evaluation using Machine Learning techniques, determine the ideal predictive model that helps establish the client's credit status accurately and with a percentage of minimal error. The investigation was carried out in a Savings and Credit Cooperative of Ecuador, whose credit risk management process is carried out through credit rating, but it has been shown that this process is not sufficient to assess whether a loan can be granted to a client. For all this, there is a need to use technological tools and specifically prediction models for credit risk to better evaluate their clients.

The rest of the document is structured as follows: Sect. 2) Research design based on DSR, which shows the theoretical foundation, design, and construction of the artifact (predictive model). Section 3) Results of the evaluation of the designed artifact. Section 4) Discussion of the results with related works. Section 5) Conclusions of the study and future work.

2 Research Design

We have designed the research based on the guidelines of the Design Science Research approach [3], see Table 1.

2.1 Population and Sample

Total data is the sum of all clients with a credit history within the institution, giving a total of 68,164 records stored in its production and historical database and made up of 27 variables. For data protection due to banking secrecy, the names and personal identifications were hidden.

One of the most important points is the process of cleaning and eliminating unnecessary, inconsistent, redundant, or erroneous information in the extraction of data from the variables, it is observed that there is information with null values, which may be due to

Table 1. Research design methodology.

Activity	Components	Paper section
Problem diagnosis	Problem; Objective Population and sample	Introduction Research design
Theoretical foundation	Machine Learning Supervised learning Logistic Regression Model Random Forest Model Neuronal Network Model Confusion Matrix CRISP-DM	Research design
Artefact design: predictive model development	Based on CRISP-DM	Predictive model development
Artifact evaluation	Evaluation of the models with the classification algorithms	Results

failures in the filling in the client file, in addition there are inconsistencies or omissions in filling out the information.

2.2 Theoretical Foundation

A practical approach to Machine Learning with tools and techniques for Java implementations is discussed in the work of Witten et al. [6].

The survey conducted by Sebastiani, discusses the main approaches to text categorization that fall within the machine learning paradigm. The dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories [7]. There is an article about classification and regression trees, it gives an introduction to the subject by reviewing some widely available algorithms and comparing their capabilities, strengths, and weakness in two examples [8].

An important book as a guide to logistic regression modeling for health science and other applications is presented in [9]. It provides an easily accessible introduction to the logistic regression (LR) model and highlights the power of this model by examining the relationship between a dichotomous outcome and a set of covariables.

Random forests are an effective tool in prediction. Because of the Law of Large Numbers, they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. This es analyzed in the work of Breiman [10], it is a necessary article to understand Random Forests basics.

University of Washington researchers propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, we provide insights on cache access patterns, data compression and sharding to

build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems [11].

A review of deep supervised learning, unsupervised learning, reinforcement learning & evolutionary computation, and indirect search for short programs encoding deep and large networks is presented in [12], to assign credit to those who contributed to the state of the art about Deep Learning (DL). Finally, we must mention the work of Tharwat [13], who introduces a detailed overview of the classification assessment measures with the aim of providing the basics of these measures and to show how it works to serve as a comprehensive source for researchers who are interested in this field. This overview starts by highlighting the definition of the confusion matrix in binary and multi-class classification problems. Many classification measures are also explained in detail, and the influence of balanced and imbalanced data on each metric is presented.

3 Predictive Model Development

We use the CRIPS-DM.v3 (Cross-Industry Standard Process for Data Mining) methodology [4] to develop the predictive model, from business analysis to implementation and presentation of results.

Phase 1: Business Understanding

Determine Business Goals. According to the business situation of the financial entity, there is a database of current and even historical credits. However, there are no studies of customer behavior that can provide conclusions and patterns that make predictions about future customers who may or may not receive credit.

The Savings and Credit Cooperative is a financial entity of segment 2, which seeks to integrate itself into this new Data Mining technology, which projects an analysis of historical and current data that the institution has stored, with this process the institution seeks to improve the following objectives: i) Improve the evaluation of a client to know if he is suitable or not for a loan. ii) Streamline the credit rating and delivery process. Iii) Minimize the probability of non-compliance in credit payments.

From a business perspective, predicting that a new customer is trustworthy can reduce the rate of arrears of loans, this is considered a success criterion, another measure of success is the increase in the percentage of agility when granting a loan.

Evaluate the Initial Situation. The Savings and Credit Cooperative is a de facto company dedicated to productive microcredits. It currently has more than 40,000 members, mostly: merchants, farmers, artisans, public and private employees, carriers; its capital is made up of the contributions of all the members and savings of its depositors, who have been contributing based on the trust generated during all these years of work, the Savings and Credit Cooperative has 12 Agencies and its Head Office in the city of Latacunga.

For this project, the databases that store current and historical information of all the credits granted and whose statuses are canceled, active, reclassified, and expired are required. Due to banking secrecy, the presentation of data such as personal identification, names and surnames is restricted.

The existing risks that must be considered for this work are the following:

- The development time of the project, once the proposal was accepted by the General Management, all the credit information of the financial entity was downloaded and saved in another SQL Server 2016 database engine.
- Additional or outstanding costs: we will work with the initial data, so no additional costs are required by the entity or the researcher.
- The quality of data: the process of granting credit starts from the update of the account, that is, the member updates the data such as personal information, home, spouse, work, and financial status. With this, it can be considered that the data quality is good.

Set Data Mining Goals. The objectives set for the development of the project are the following: i) Generate patterns that help the evaluation of a client to verify if he is suitable or not for a loan. ii) Generate patterns that help with the problem of increased arrears and low profitability.

The data mining criterion was to predict with the variables or data already trained, if a partner is subject to credit or not with a percentage of 90% success and effectiveness.

Regarding the techniques for data extraction, the SQL programming language was used to generate the scripts that allow the information consulted and extracted. Classification tasks were used to generate predictive models, and the modeling techniques were Random Forest, Logistic Regression and Neural Networks.

Phase 2: Understanding the Data

Collect Initial Data. For the development of this project, the following data was collected: customer information, type of credit, amount, payment frequency, economic activity, etc., these data are stored in tables related to the credit granting process, also for reasons legal and banking secrecy identification data, names and surnames were omitted. Because the objective of the project is to make predictions, the client was classified good and bad according to the last credit rating they obtained at the time of canceling. Due to the large number of records that are necessary for the development of the project, it was decided to develop SQL statements that help generate all the information based on the data investigated in conjunction with the financial institution.

Describe Data. According to the analysis, the variables with which to work for the development of the model are disclosed. Each of them is detailed, classified into 27 input variables, and one output variable, which are described in [14].

Explore Data. With the variables obtained, a search was made to the database, to identify all the information that can be obtained and will be necessary for the realization of this research project, a statistic of each variable to be processed was made.

Verify Data Quality. To data quality verification, ISO/IEC 25012 standard - "Data Quality Model" [15] will be used, which specifies a general quality model for those data that are defined in a structured format within a computer system; through a matrix the good and bad data will be valued.

Characteristics that will be considered to data verification matrix: i) Accessibility (AC): Specifies the degree to which data can be accessed in a specific context. ii) Accordance (CO): Verifies that the corresponding data complies with current standards, conventions,

or regulations. iii) Confidentiality (CF) (associated with information security): ensures that data is only accessed and interpreted by specific authorized users. iv) Efficiency (EF): Analyzes the degree to which data can be processed and provided with expected performance levels. v) Precision (PR): The data requires exact values or with discernment in a specific context. vi) Traceability (TZ): Analyzes whether the data provides a record of the events that modify it. vii) Understandability (CP): The data is expressed using appropriate languages, symbols and units and can be read and interpreted by any type of user.

Phase 3: Data Preparation

Select Data. Data is made up of the sum of all clients with a credit history within the institution, giving a total of 68164 records stored in its production and historical database and made up of 27 variables.

Clean Data. An important process is cleaning and eliminating unnecessary, inconsistent, redundant, or erroneous information in the extraction of variable data. It is detected that there is information with null values, which may exist due to failures at the time of filling out the client file, in addition there are inconsistencies or omissions in filling out the information.

Build Data. A count of ranges was carried out where the minimum and maximum are specified, which will help transform the categorical data to numbers, facilitating the training of the algorithm and improving the interpretation of data.

Integrate Data. Only the data stored in the Core from the Informix database engine was used, so integration with other sources was not necessary.

Format Data. Once the categorization is done, as a result processed data with numerical values that reflect the value of each record are obtained. Data is available at: [14]. This type of processing is part of a data normalization. The order of the variables does not affect the development of the project, so it is not necessary to change, the normalization of variables is a great help when generating a model.

Phase 4: Modeling

Select Modeling Techniques. The selected models focused on predictive models aimed at solving problems in banking areas such as predictions of arrears, predictions of non-payment. The selected techniques are the following: Decision Trees (Random Forest), Neural Networks and Logistic Regression.

The three techniques are supported by Python as a tool to be used for model generation and evaluation to find the most accurate model.

Design Model Tests. Before the construction of the models for the research work, a test plan was considered, to determine the procedures to validate the quality and accuracy of each model. Two stages were contemplated for the design of the test plan. The first stage consisted in divide the data in training data that covers 70% and test data that covers 30%. The second stage was the validation of the models for which the Confusion

Matrix Evaluation Technique was used. The metrics to evaluate the models are listed: Accuracy, Error rate, Sensitivity, Specificity, Precision, and Negative predictive value. The selection of the metrics was based on the review of several research papers [16–18].

Build the Models. Random Forest: For the construction of the model, XGBoost tool was used, (Extreme Gradient Boosting) [11], the following parameters were defined: A bi-nary classification was used as the main objective of the model, a maximum depth of 150, a minimum observation weight of 25, a subsample of 0.85, the column per tree sample of 0.8, a minimum loss reduction needed of 5, job number 16, a learning rate of 0.025, and a speed of 1305. Additionally, the `scale_pos_weight` parameter was used to adjust the behavior of the algorithm with unbalanced classification problems. In Logistic Regression and Neural Network, the `class_weight = "balanced"` was used to balance the data in training.

For this model, its construction was in accordance with the definition of the initial parameters: import of the libraries, connection to the database, division of the data in training and testing. The most important variable for this model was the type of client, that is, if a member is new or has already had credits in the financial institution before (recurring).

Logistic regression: For the construction of the model, the parameters that are defined by default were used. It was based on the default parameters since they meet the objective of the project, which is to verify if a member is subject to credit or not, for this model the most important variable is age.

Neural network: For the creation of the neural network with scikit-learn, the `sklearn.neural_network.MLPClassifier` class was used for classification, the following parameters were defined: the number of hidden neurons (10, 10, 10), maximum number of iterations (500), regularization parameter (0.0001), Adam-based solution optimizer for large volumes of data, random number for the weights (21) and a tolerance of 0.000000001.

Evaluate Models. For the verification of the Confusion Matrix, `confusion_matrix` was imported from the sklearn library, and it was used together with the real data and those that have been previously predicted.

The results of the evaluation are displayed in the Results section.

Phase 5: Evaluation

For the development of this research work, three Machine Learning classification models were applied: Random Forest, Logistic Regression and Neural Networks.

The results of the evaluation are displayed in the Results section.

Phase 6: Implementation or Deployment

Schedule Deployment. To implement the model within the financial institution, four phases were created, which helped employees to use Machine Learning for credit risk assessment:

Phase1: After creating the model and reviewing which is the best option to achieve the business objectives, it is necessary to export the model, that is, serialize the trained model to be able to use it in a Web API.

Phase2: A web API is developed with Flask [15], which helps to create web applications with Python. Through this application, collaborators will be able to enter from any web browser and will be able to make the prediction of the client after entering some data.

Phase3: In this phase, a training in the use of the web API is carried out for the collaborators who work in credits and operations area, being these the ones in charge of generating the credit requests and the respective approvals.

Phase4: As a last phase, interviews or work meetings will be held with those responsible for the operational processes to see the results obtained when classifying the partner through the model and the web API that was carried out using Machine Learning.

Plan Monitoring and Maintenance. Monitoring and maintaining the implementation of the model is one of the important steps to carry out a good prediction of the client. It should be considered that there is a data update program that the financial entity carries out from time to time and with the refinement of the parameters, the model classification process can be improved. Also, each month there is cancellation of credits or pre-cancellations, these historical data are updated in the institution's database.

As a monitoring and maintenance plan, the following processes can be followed: i) Selection and extraction of data updated every six months, that is, a data mining process. ii) Generation of the model with the new data, without forgetting that 70% is needed for training and 30% for testing. iii) Exporting the model using the serialize and update process in the Web API. iv) Have a model update log and save a version for each model.

4 Results

4.1 Evaluation of the Models

For the verification of the Confusion Matrix, in the Python programming language, `confusion_matrix` is imported from the `sklearn` library, and it is used together with the real data and those that have been previously predicted.

Table 2 represents a comparison of the data with the metrics of our models, resulting in a general model with a very good accuracy percentage of 90%.

Decision Tree Model (Random Forest). The error matrix generally indicates that the degree of classification is quite good with 90% accuracy and an error or misclassification rate of 10%.

The model also indicates that it classifies positive cases with a probability of 99.6% and negative cases with a probability of 6%.

Also, if the classifier detects that a customer is good, it is with a 91% probability. And if he says he isn't, then the customer is bad with a probability of 58%.

Logistic Regression Model. The error matrix generally indicates that the degree of classification is quite good with 90% accuracy and an error or misclassification rate of 10%.

The model also indicates that it classifies positive cases with a probability of 99% and negative cases with a probability of 0.3%.

Table 2. Comparative table of the models

Model/Metric	Accuracy	Error Rate	Sensitivity	Specificity	Precision	Negative Prediction
Random Forest (ML1)	0.90	0.10	0.99	0.06	0.91	0.58
Logistic regression (ML2)	0.90	0.10	0.99	0.003	0.90	0.33
Neural network (ML3)	0.90	0.10	0.99	0.06	0.91	0.53

Also, if the classifier detects that a customer is good, it is with a 90% probability. And if he says he isn't, then the customer is bad with a probability of 33%.

Neural Network Model. The error matrix generally indicates that the degree of classification is quite good with 90% accuracy and an error or misclassification rate of 10%.

The model also indicates that it classifies positive cases with a probability of 99% and negative cases with a probability of 6%.

Also, if the classifier detects that a customer is good, it is with a 91% probability. And if he says he's not, then the customer is bad with a probability of 53%.

Now, the data of the models applied to the business objectives and the Data Mining objectives are analyzed.

Business Objectives:

- a) Improve the evaluation of a client to know if he is suitable or not for a loan: if the comparison table is observed, the ML1 and ML3 model predicts with an accuracy of 91% if a client is good, but only the ML1 model predicts that a client is bad with 58%.
- b) Streamline the credit qualification and delivery process: the three models meet this objective since when entering the data of the variables and executing the model there is a very short time in which the model solves and gives a good or bad result.
- c) Minimize the probability of default in credit payments: both the ML1 and ML3 models give this probability, but ML1 has the 58% higher of the two models.

Data Mining Objectives:

- a) Generate patterns that help the evaluation of a client to verify if he is suitable or not for a loan. The ML1 and ML2 models show a summary of the Important Variables of each model indicating which pattern is the most outstanding at the time of generating a model and with this verify if a client is suitable or not.

The pattern shows the client's behavior regarding direct obligations to the institution. To determine what these patterns are, the classification of variables by their

importance was generated. This indicates that a classification of clients can be made using these variables through an interview or survey to generate a pre-selection of the ideal clients to deliver a loan. In addition, with the selection of variables, a model can be generated, thus improving the effectiveness of the algorithm.

- b) Generate patterns that help with the problem of increased delinquency and low profitability: the ML1 and ML2 models show a summary of the Important Variables of each model, indicating which pattern is the most prominent when generating a model and with this verify the delinquency and low profitability.

Approved models

When reviewing the objectives of both the business and Data Mining, it is noted that the three most important factors in choosing the best model are Accuracy, Precision and Negative Prediction. Summarizing these metrics, the first two represent the percentage of success and quality of the model at the time of making the true predictions, while the third metric indicates the percentage of negative prediction, that is, the percentage of non-compliance in the payment of the credit.

It is concluded that the Random Forest model applying XGBoost with an accuracy of 90%, a precision of 91% and a negative prediction of 58%, is the accepted model for implementation.

4.2 Results Validation

The results of the models evaluated with the extracted data are presented below:

Random Forest. The first model is developed with XGBoost, one of the modules of the Scikit-learn library. The data of the confusion matrix are shown in the following graph:

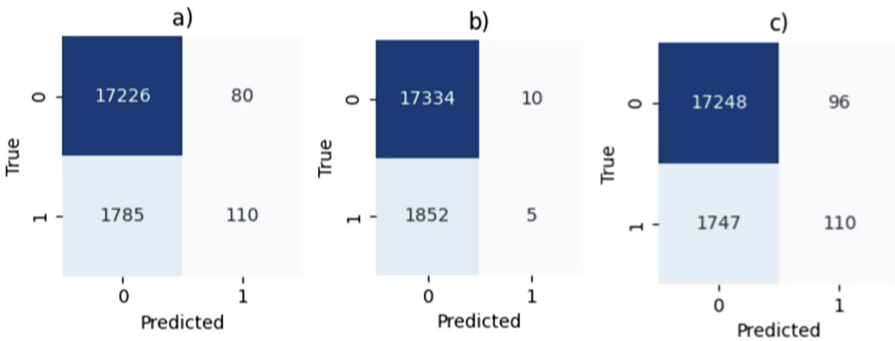


Fig. 1. a) Random Forest Confusion Matrix, b) Logistic Regression confusion matrix and c) Confusion matrix Neural Networks

Figure 1(a) indicates that there are 17,336 records that were classified correctly and that 1,865 were erroneously classified. In addition, the following can be detailed:

17,226 clients who are classified as subject to credit.
110 clients who were correctly classified as not subject to credit.
80 clients that were erroneously classified as not subject to credit.
1785 clients who were erroneously classified as subject to credit.

Logistic Regression. The second model developed with Logistic Regression in the same way a module of the Scikit-learn library, Fig. 1 (b) shows the data of the confusion matrix.

Figure 1(b) indicates that there are 17,339 records that were classified correctly and that 1,862 were erroneously classified. In addition, the following can be detailed:

17,334 clients who are classified as subject to credit.
5 clients who were correctly classified as not subject to credit.
10 clients who were wrongly classified as not subject to credit.
1,852 clients who were erroneously classified as subject to credit.

Neural Networks. The third model is developed with the MLPClassifier module of the Scikit-learn library, Fig. 1 (c) shows the confusion matrix data.

Figure 1(c) indicates that there are 17,358 records that were classified correctly and that 1,843 were erroneously classified. In addition, the following can be detailed:

17,248 clients who are classified as subject to credit.
110 clients who were correctly classified as not subject to credit.
96 clients who were erroneously classified as not subject to credit.
1747 clients who were erroneously classified as subject to credit.

5 Discussion

In the study by Li and Wang [19], they present the development of credit risk measurement models based on data mining, they do it in the same context of our research; however, they do not provide or do not have an appropriate deployment plan for the normal user or customer. This finding motivated the realization of this work to help the credit analyst by entering data, verify whether the client is subject to credit.

In another study found by Kruppa, Schwarz, Arminger y Ziegler [20], propose a framework to estimate credit risks for consumer loans, using Machine Learning. They refer to logistic regression as the main model construction technique, which is related to our research work. However, this technique has its limitations, being the easiest to use and its configuration is based on data by default. Faced with these difficulties, we present our contribution with the analysis and evaluation of the models with various machine learning techniques, Random Forest being the most effective.

The work proposed by Song and Wu [21] uses data mining to determine the risk of excessive financialization. They used the genetic algorithm (GA), neural network and principal component analysis (PCA) data collection and processing methods. The results suggest that the data mining technology based on back propagation neural network

(BPNN) can optimize the input variables and effectively extract the hidden information from the data.

Another study about financial risk prevention with data mining is proposed by Gao [22], who analyzed 21 companies with high trust. The results show that the financial risk evaluation index system of four dimensions of solvency, operation ability, profitability, growth ability and cash flow ability can affect the financial risk of enterprises. Compared with the traditional data mining algorithm, the algorithm of financial risk index evaluation model constructed in this study has the best performance.

Greek researchers Lappas and Yannacopoulos [23], affirm that in addition to the automatic processing of credit ratings, the opinion of experts is required. They propose a combination strategy of integrating soft computing methods with expert knowledge. The ability of interpretation of the predictive power of each feature in the credit dataset is strengthened by the engagement of experts in the credit scoring process. Testing instances on standard credit dataset were established to verify the effectiveness of the proposed methodology.

Among the limitations presented by our research work, it is considered that the development of the model is tailor-made for the financial institution. It would be advisable to take this model to another context, to evaluate and instantiate the performance of said model. In addition, the analysis of credits according to the manuals, policies and regulations of each financial entity is considered another limitation. This means that the variables for the model would not be the same and it would be necessary to reengineer the ideal variables for the construction of the model.

6 Conclusions and Future Work

Once the CRISP-DM methodology was used for the analysis, development and evaluation of the models as mentioned in phase 5 of said methodology, it was concluded that the most efficient model is the Random Forest.

In addition, to verify if the problem was solved and the solvency of the hypothesis, a web application was made, in which the data of the variables is entered and with the processing of the trained model it results in whether the client is good (the person is subject to credit) or bad (the person is not subject to credit) and with this analysis verify if the credit granting process was optimized.

Hypothesis H0 raised with 95% is accepted, in the development of the research work. It is verified that the use of binary classification techniques is an effective method for the evaluation of credit risk in the Savings and Credit Cooperative. The choice of the CRISP-DM methodology helped to generate objectives that meet the needs of the business line and focus on solving the hypothesis. It is clarified that the customer data used in the model to test the hypothesis were real and correspond to a financial institution in Ecuador.

For the elaboration of the models, an investigation of the Machine Learning techniques was carried out, which are within the predictive analysis and that focus on credit risk, with this analysis three classification-type models were used: Random Forest, Logistic Regression and Neural Networks.

With the development and evaluation of these Machine Learning models, it was confirmed that the Random Forest model using the XGBoost module is the most accurate to predict whether a client is subject to credit.

Through the evaluation of the models, performance patterns of a client were generated, based on the extraction of the important variables, this helped the operations personnel to carry out a brief validation of the partner in the field.

With the application of Machine Learning for the generation of the most effective model and the development of a Web API (data entry through a form and prediction process through the serialized model), the credit delivery process was improved, and the customer service and attention improved. Customer by 40% more, in reference to the manual process. As future work, it is proposed to apply empirical strategies to evaluate the proposed model. In addition, develop a predictive model for the collection area, to predict customer behavior when deciding to pre-cancel an investment. Finally, it is proposed to carry out an analysis with Machine Learning in the compliance area, verifying if there is or will be money laundering through the behavior of the transactions carried out by the client. Supplementary material for this research study is available at: [14].

References

1. Duffie, D., Singleton, K.: *Credit risk: Pricing, Measurement, and Management*. Princeton University Press (2012)
2. Superintendencia de Economía Popular y Solidaria: Inicio – Superintendencia de Economía Popular y Solidaria. <https://www.seps.gob.ec/> (2022). Accessed 21 Mar 2022
3. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in IS research. *Manag. Inf. Syst.* **28**(1), 75–105 (2004)
4. Chapman, P., et al.: *CRISP-DM 1.0: Step-by-step data mining guide* (2000)
5. Moro, S., Laureano, R.M.S., Cortez, P.: Using data mining for bank direct marketing: an application of the CRISP-DM methodology. In: *ESM 2011 – 2011 European Simulation Modelling Conference Modelling Simulation 2011*, no. Figure 1, pp. 117–121 (2011)
6. Witten, I., Frank, E., Hall, M., Pal, C.: *Data Mining: Practical Machine Learning Tools and Techniques* (2016)
7. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* **34**(1), 1–47 (2002). <https://doi.org/10.1145/505282.505283>
8. Wei-Yin, L.: Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**(1), 14–23 (2011). <https://doi.org/10.1002/widm.8>
9. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: *Applied Logistic Regression*, 3rd edn. Wiley (2013)
10. Breiman, L.: Random Forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
11. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
12. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015). <https://doi.org/10.1016/j.neunet.2014.09.003>
13. Tharwat, A.: Classification assessment methods. *Appl. Comput. Informatics* **17**(1), 168–192 (2018). <https://doi.org/10.1016/j.aci.2018.08.003>

14. Chiluiza, O., Guevara-Vega, C., Quiña-Mera, A., Landeta-López, P., Montaluisa, J.: Supplementary material: financial credit risk measurement using a binary classification model. Zenodo (2022). <https://doi.org/10.5281/zenodo.7274756>
15. Iso25000.com: ISO/IEC 25012. <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012> (2021). Accessed 26 Mar 2022
16. Buckland, M., Gey, F.: The relationship between Recall and Precision. *J. Am. Soc. Inf. Sci.* **45**(1), 12–19 (1994). [https://doi.org/10.1002/\(SICI\)1097-4571\(199401\)45:1%3c12::AID-ASI2%3e3.0.CO;2-L](https://doi.org/10.1002/(SICI)1097-4571(199401)45:1%3c12::AID-ASI2%3e3.0.CO;2-L)
17. Visa, S., Ramsay, B., Ralescu, A., Van Der Knaap, E.: Confusion matrix-based feature selection. In: *CEUR Workshop Proceedings*, vol. 710, pp. 120–127 (2011)
18. Marom, N.D., Rokach, L., Shmilovici, A.: Using the confusion matrix for improving ensemble classifiers. In: *2010 IEEE 26th Convention Electrical Electronics Engineers in Israel IEEEI*, pp. 555–559 (2010). <https://doi.org/10.1109/EEEI.2010.5662159>
19. Li, W., Wang, S.: Research on assessment method for credit risk in commercial banks of china based on data mining. *Appl. Mech. Mater.* **303–306**, 1361–1364 (2013). <https://doi.org/10.4028/www.scientific.net/AMM.303-306.1361>
20. Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A.: Consumer credit risk: Individual probability estimates using machine learning. *Expert Syst. Appl.* **40**(13), 5125–5131 (2013). <https://doi.org/10.1016/j.eswa.2013.03.019>
21. Song, Y., Wu, R.: The impact of financial enterprises' excessive financialization risk assessment for risk control based on data mining and machine learning. *Comput. Econ.* **60**, 1245–1267 (2021). <https://doi.org/10.1007/s10614-021-10135-4>
22. Gao, B.: The use of machine learning combined with data mining technology in financial risk prevention. *Comput. Econ.* **59**, 1385–1405 (2021). <https://doi.org/10.1007/s10614-021-10101-0>
23. Lappas, P.Z., Yannacopoulos, A.N.: A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Appl. Soft Comput.* **107**, 107391 (2021). <https://doi.org/10.1016/j.asoc.2021.107391>