# Application of Decision Tree to Banking Classification Model

*Juan Freire*[1], *Cesar Guevara*[2]

[1] Digital School, Universidad Internacional SEK
Quito, Ecuador

[2] Centrer of Mechatronics and Interactive Systems (MIST), Universidad Tecnológica Indoamérica
Ambato, Ecuador

## ABSTRACT

In this research, we will focus on INSOTEC NGO, an entity dedicated to granting microcredits to entrepreneurs with limited economic resources. This company is present in rural areas of Ecuador, increasing its income in recent years. The organization plans to become a bank in the long term and expand its operations to near countries such as Colombia and Peru. However, the entity's customer classification processes have had many drawbacks because it is currently a manual procedure that generates a high operational burden, slow response times to customers, huge inefficiency rates, and a great problem to continue growing. This project proposes to model an artificial intelligence algorithm that classifies the organization's clients based on the different variables that are considered convenient for the analysis. The method selected to meet this objective is a Random Forest, a supervised learning method that builds models that are easy to interpret. Its implementation complexity is very low, it allows continuous and categorical values, and it handles noise from data from different sources very well. This new process will guide the organization to implement these models in other areas such as risk, finance, auditing, and operations.

**Keywords:** Random Forest · decision tree · credit customer classification.

# 1. INTRODUCTION

In recent decades, the economy of the different productive sectors has developed thanks to the expansion of financial services, the loans provided by these entities offer the opportunity to obtain income to boost the different businesses, deferring payments for the following months (Lanzarini et al., 2017).

Credit classifiers that use statistics were introduced in the 1950s, are now used worldwide, and have become a fundamental part of the credit granting process. The main objective is to evaluate the payment behavior of the lenders, in order to reduce the percentage of error (Zhang et al., 2019).

There are investigations that implement algorithms for the classification of credit clients, in which neural networks, naive bayes, sensory vector machines and decision trees are used. However, in this study Random Forest (RF) was selected because it has been shown to have high precision in several investigations (Ziemba et al., 2020).

From several of the related works Pradhan et al. (2020) it compares some algorithms such as neural networks, decision trees, SVN and RF in a data set of credit clients, which has 4,600 records and 47 variables. First, it used the "feature selection" technique which reduced the set to 31 variables, in the experiment the RF algorithm obtained better results with a precision of 85% and an error rate of 10%.

A similar experiment was carried out by Arora & Kaur (2019) where several classifiers were also com-pared for a data set of credit clients from the "Len-ding Club" database. This dataset is found in kaggle, which has 42,530 records and 143 variables with a history from 2007 to 2011. Subsequently, the chi-square, gain ratio, relieff and Bolasso techniques were applied, reducing the set to 36 variables. In the application of classification models, Random Forest obtained an accuracy of 97.9%, an AUC of 93.4% and an error rate of approximately 2.1%.

In another study carried out by Siswanto et al. (2019), a model with C4.5 decision trees is proposed to a credit customer base, which has 1044 records and 8 attributes. The result of the model was an accuracy of 93.5%, with an error rate of 8.8% of which 92 were FP and 0 were FN.

Another important contribution was proposed by Subasi & Cankurt (2019), to evaluate the payment behavior of clients in one of their loan products. In this work some algorithms were compared as in Arora & Kaur (2019) and Pradhan et al. (2020), in the experiment the data set is composed of 25,000 records and 23 variables. After applying the different classification algorithms, RF obtained the highest percentage of accuracy with 89.01%, an AUC of 95% and an error rate of 11.68%.

Finally, the objective of this research is to create a customer classification model with a high

level of precision, for which the necessary data preprocessing techniques will be used and an algorithm based on Random Forest will be used for the model. Additionally, as a result of this research, the risk assessment processes, the credit rating of customers and the percentage of error in the classification of the organization's customers will be improved.

The research is structured as follows: in section 2, the methods and materials used in the study are presented. The database used in the analysis, the method for the selection of variables, which is chi-squared, and RF, which is the algorithm selected for the classification of the data set with Insotec clients, is described.

In section 3 the proposed model is explained, in addition, the data preprocessing that includes the elimination of atypical data and selection of attributes with the chi square technique is detailed.

Section 4 describes the model proposed in the research. Subsequently, in section 5 the results are analyzed to be compared with related works. Finally, section 6 contains the conclusions and future lines of this research.

## 2. METHODS AND MATERIALS

In this section, the methods and materials used in this document will be described theoretically, in the first point the data set used in the study is detailed, later, the chi-square technique is analyzed for the selection of variables and RF as an algorithm of prediction.

Table 1. Detail of the data set variables.

| **Variable** | **Datatype** | **Type of information** | **Example** |
|---|---|---|---|
| Loan type | Text | Type of loan delivered | Direct |
| Year | Entire | Year of last transaction | 2017 |
| Month | Text | Month of last transaction | 2 |
| Gender | Entire | Gender | Male |
| Civil status | Text | Civil Status | Single |
| Instruction | Text | Instruction Level | Primary |
| Atraso | Text | Maximum delay in days | 30 |
| Age | Entire | Current age | 20 |
| Patrimony | Text | Current equity in dollars | 10000 |
| Province | Text | Province of residence | Pichincha |
| Canton | Text | Canton of residence | Quito |
| Time of operation | Entire | Business uptime in days | 100 |
| Initial debt | Double | Initial debt of the last loan | 2000 |

| | | | |
|---|---|---|---|
| Economic activity | Text | Economic activity | Manufacture |
| Current rating | Text | Credit rating | A-1 |
| Profit | Entire | Business profit in dollars | 10000 |
| Income | Entire | Business income in dollars | 20000 |
| Expenses | Entire | Business expenses in dollars | 10000 |
| It is a good payer | Text | Whether it qualifies as a good payer or not is the kind that divides the dataset | GP y BP |

The customer data set is made up of two classes GP and BP, the absolute frequency of the GP is 58,472 instances while the BP have 5,424 records. On the other hand, the relative frequency for GP is 92%, while for BP it is only 8%, which is why it is concluded that the data set is unbalanced since there is a much higher percentage of GP.

## 2.2. RANDOM FOREST (RF)

Random Forest RF is an algorithm that arises from the combination of decision trees without correlation, subsequently averages its results, in this way it is an assembled method that combines the results of the different trees to obtain a value for the entire set of trees (Safari, 2020).

This technique introduced by Breiman (2001) can consider a data set $x$, where $x_i$ represents each of the iterations that are formed from a random sample obtained from the original set, $T_b$ represents each of the trees and $B$ the number of trees that are found in the forest regression, this regression is represented in equation 1 (Uthayakumar et al., 2020).

$$\hat{y}(x_i) = \frac{1}{B}\sum_{b=1}^{B} T_b(xi) \quad (1)$$

## 2.3. CHI SQUARE

Chi square is an independence test that helps determine the degree of correlation between the variables. It is used to reduce the number of columns in a data set, for which the variables that have a greater degree of dependency must be selected, when the value is greater there is greater dependency with the class Ramya & Kumaresan (2015). The formula proposed by Hidalgo et al. (2020) and (Guevara & Peñas, 2020) is shown in equation (2), in this case *c* are the degrees of freedom, *x* represents the values of the client data set and m the expected values. Additionally, the *m* observations are classified into *k* classes, in this case they are 2 corresponding to the GP and BP.

$$x_c^2 = \sum_{i-1}^{k} \frac{(x_i - m_i)^2}{m_i} \quad (2)$$

# 3. CREDIT CUSTOMER CLASSIFICATION MODEL

This section will briefly describe data preprocessing, the application of the chi-square method for the selection of variables, the description of the model and the configuration of the algorithm.

## 3.1. DATA PREPROCESSING

The data pre-processing stage has a very significant importance in data analysis, since it eliminates noise, instances and variables that do not add value to the study(Kotsiantis et al., 2006).

In the experiment, outliers were eliminated to avoid noise in the analysis, after carrying out this procedure the set was reduced to 58,641 records. Subsequently, the chi-square technique was applied for the selection of variables, where only 8 of the 19 variables have relevance in the analysis, in the results the variables with the highest correlation with the class are backwardness and current grade with values of 29,010 and 24,751 respectively. These variables have a higher level of importance to identify the GP and BP. On the other hand, the variables operating time, year, loan type, expense, income and economic activity obtained much lower values, but greater than 0, which is why they were also included within the selected fields. The chi square results in the data set are observed in figure 1.

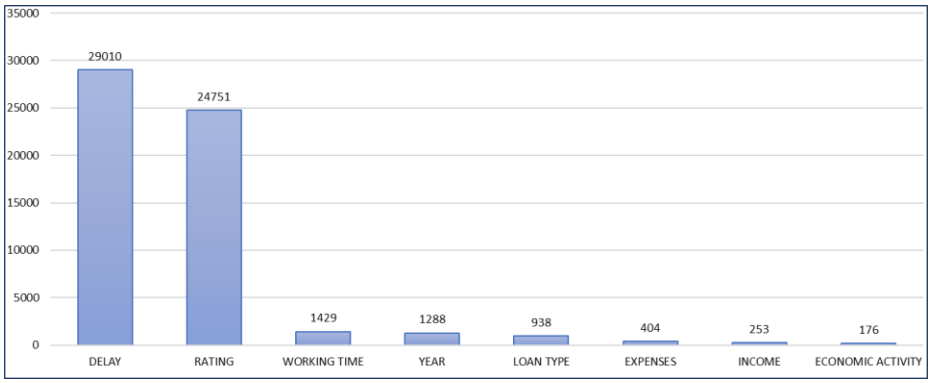


Figure. 1. Variable selection bar chart with chi square.

The vector is defined as follows $d_n = \{a, ca, tf, an, tp, g, i, ae\}$.. The variable $a$ represents

the arrears, $cr$ to the current rating, $ot$ is the business operating time, $yt$ to the year of the last transaction, $tp$ is the type of loan of the last transaction, $g$ to the business expense, $i$ to the income and $ae$ to the economic activity of the client.

# 4. MODEL WITH RANDOM FOREST

For the creation of the prediction model, Random Forest was selected since by literary review it has obtained the highest level of precision in similar cases. As presented by Kotsiantis et al. (2006) and Arora & Kaur (2019). In the aforementioned studies, the RF models obtained the best results compared to others that used different techniques such as neural networks and decision trees.

As mentioned in section 3.1, 8 variables and 58,641 instances were selected to be entered in the model, 50 trees were used for the configuration of the model as recommended (Mercadier & Lardy, 2019), for the number of variables and depth of the trees. individual, the maximum value was selected, which is 8 and 50, since some investigations have shown better results (Ahmad et al., 2017).

# 5. RESULTS

For the evaluation of the model, cross-validation was used with $k = 1\ to\ 10$, the results of the confusion matrix that were obtained for the training set was a precision of 97.7%, an accuracy of 97.4% and 2.6% error pertaining to 742 FP and 447 FN instances. On the other hand, that of the set of test data obtained a precision of 97.3%, an accuracy of 97.4% and a 2.6% error that belongs to 122 instances of FP and 31 of FN.

Additionally, in the research it was recommended that the indicated number of trees for a data set of credit clients be 50 trees, when conducting this experiment, the accuracy and precision was around 97% in both cases and the percentage of error is less than 3%, which is consistent with research and shows that the results of the algorithm are very good. However, the algorithm has greater difficulty in predicting the BPs because the percentage of error is 10%, while it is more precise to find the GPs since the percentage of error is 1.5%.

# 6. CONCLUSIONS AND FUTURE WORKS

In the research, it was identified that the key part for obtaining the results was the data preprocessing where the outliers were eliminated, and the most important variables for the study were selected with the chi-square technique.

In this work, we apply chi-square for select attributes most relevant to understand the most important variables from the database.

The Random Forest assembled method yielded very good results for the data set, with a precision of 97% and an error rate of 2.8%.

Additionally, the proper configuration of the model improved the results. However, the number of variables should be increased in order to create a more robust classification model.

For future research, the same analysis should be carried out with a greater number of variables and include some tests to evaluate the robustness of the client classification model.

# REFERENCES

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Arora, N., & Kaur, P. (2019). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936. https://doi.org/10.1016/j.asoc.2019.105936

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Guevara, C., & Peñas, M. S. (2020). Surveillance Routing of COVID-19 Infection Spread Using an Intelligent Infectious Diseases Algorithm. *IEEE Access*, *8*, 201925–201936. https://doi.org/10.1109/ACCESS.2020.3036347

Hidalgo, J., Guevara, C., & Yandún, M. (2020). Generation of User Profiles in UNIX Scripts Applying Evolutionary Neural Networks. In I. Corradini, E. Nardelli, & T. Ahram (Eds.), *Advances in Human Factors in Cybersecurity* (pp. 56–63). Springer International Publishing.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, *1*, 111–117.

Lanzarini, L., Villa Monte, A., Bariviera, A. F., & Jimbo Santana, P. (2017). Simplifying credit scoring rules using LVQ + PSO. *Kybernetes*, *46*, 8–16. https://doi.org/10.1108/K-06-2016-0158

Mercadier, M., & Lardy, J. P. (2019). Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research*, *277*(1), 351–365. https://doi.org/10.1016/j.ejor.2019.02.005

Pradhan, M. R., Akter, S., & Al Marouf, A. (2020). Performance Evaluation of Traditional

Classifiers on Prediction of Credit Recovery. In T. Sengodan, M. Murugappan, & S. Misra (Eds.), *Advances in Electrical and Computer Technologies* (pp. 541–551). Springer Singapore.

Ramya, R., & Kumaresan, S. (2015). *Analysis of feature selection techniques in credit risk assessment*. 1–6. https://doi.org/10.1109/ICACCS.2015.7324139

Safari, M. J. S. (2020). Hybridization of multivariate adaptive regression splines and random forest models with an empirical equation for sediment deposition prediction in open channel flow. *Journal of Hydrology*, *590*, 125392. https://doi.org/10.1016/j.jhydrol.2020.125392

Siswanto, Abdussomad, Gata, W., Wardhani, N. K., Gata, G., & Prasetvo, B. H. (2019). The Feasibility of Credit Using C4.5 Algorithm Based on Particle Swarm Optimization Prediction. *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 416–421. https://doi.org/10.23919/EECSI48112.2019.8977074

Subasi, A., & Cankurt, S. (2019). Prediction of default payment of credit card clients using Data Mining Techniques. *2019 International Engineering Conference (IEC)*, 115–120. https://doi.org/10.1109/IEC47844.2019.8950597

Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2020). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. *Journal of King Saud University - Computer and Information Sciences*, *32*(6), 647–657. https://doi.org/10.1016/j.jksuci.2017.10.007

Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, *121*, 221–232. https://doi.org/10.1016/j.eswa.2018.12.020

Ziemba, P., Radomska-Zalas, A., & Becker, J. (2020). Client evaluation decision models in the credit scoring tasks. *Procedia Computer Science*, *176*, 3301–3309. https://doi.org/10.1016/j.procs.2020.09.068