

Better Efficiency on Non-performing Loans Debt Recovery and Portfolio Valuation Using Machine Learning Techniques



Jose Tupayachi and Luciano Silva

Abstract The following research is based on a portfolio of non-performing loans (NPLs), which was previously acquired and managed by a collection agency, the company under study is one of the owners of the portfolio. The study compares the efficiency and performance of several machine learning algorithms to develop and implement a forecasting tool to estimate the recovery rate of NPL portfolios. These models help to enhance and support the debt collection operation, allowing to forecast the number of debtors that will be recovered in the lifetime of the portfolio, as well as to efficiently manage resources (recovery task force) by reducing costs and expenses. The application aims to support the valuation process at the time of portfolio purchase. The study shows that the application using a binary ranking approach based on the XGBoost model outperforms other techniques, offering good results. It is also evident that product type was one of the most influential variables among the different models. The model using this algorithm could serve as a decision support tool, precisely in the operation of purchasing a portfolio of unprofitable debts, as it allows the quantification of the client's debt to be recovered by identifying the group of potential debtors with the highest probability of compliance, which would result in a faster and more efficient debt collection process.

Key words Non-performing loans · Machine learning · XGBoost

J. Tupayachi (✉)

Pontifical Catholic University of Peru, Lima, Peru

The University of Tennessee, Knoxville, TN, USA

e-mail: jtupayachi@pucp.edu.pe; jtupayac@vols.utk.edu

L. Silva

Pontifical Catholic University of Peru, Lima, Peru

e-mail: lsilva@pucp.pe

1 Introduction

1.1 Context and Overview

According to McKinsey non-performing debt recovery is defined as a form of business that generates income. Peru enjoys a well-positioned and credit-towards regulators grating credit penetration which avoids narrowing of the field of application as seen in Eduardo Lizarzaburu and Jesús del Río [15] study. In the social sphere, the so-called unprofitable debts affect both the financial institution (collection agency and the bank or investor) as well as the debtor. On one side, the base of the revenue-yielding activities of these companies depends on the generation and collection of the amounts owed. To maintain its profitability, the business must collect what is owed in the shortest possible time, but inadequate customer forecasting and filtering techniques can affect the business's operations.

After a defined period, financial institutions reschedule defaulted payments. Still, some borrowers do not meet their payment obligations, but entities such as the European Bank [10] encourage banks to dispose of NPLs after three years of management. Many banks or financial institutions encounter obstacles [2] to performing such tasks which include materializing costs in their equilibrium schemes [9]. As Bellotti and Brigo [5] point out, to decrease their defacement, lowering the losses and financial stability concerns, regulators suggest banks cluster their NPLs and sell them to specialized investors [6], called debt collection agencies.

Similarly, debt collection agencies aim to maximize their profits by offering the lowest price at the time of purchase. However, banks are not the only sellers of portfolios, collection agencies are also part of this space and partially worked portfolios are bought by other collection agencies. This research addresses and bases its results on a portfolio previously worked by a collection agency. It should be noted that the non-performing debt portfolio understudy has only one group of debts.

The process of purchasing and evaluating the portfolio is shown in Fig. 1. The diagram shows the NPL portfolio purchase operation as the starting point of the research. The developed solution supports the purchasing decision based on historical information and the methodology addresses how the study is conducted to meet the objectives and proposals. Therefore, a more accurate data-backed forecast of the future revenues of the operation can be obtained. It is important to emphasize that the recovery rate is analyzed from the position of the buyer of the non-performing debt portfolio. That is, from the point of view of the collection agency (Fig. 2).

1.2 Justification

Assessing the correct portfolio pricing is a poorly performed task as the current literature in the field reveals. The disposal of NPLs is hampered by the large bid-ask

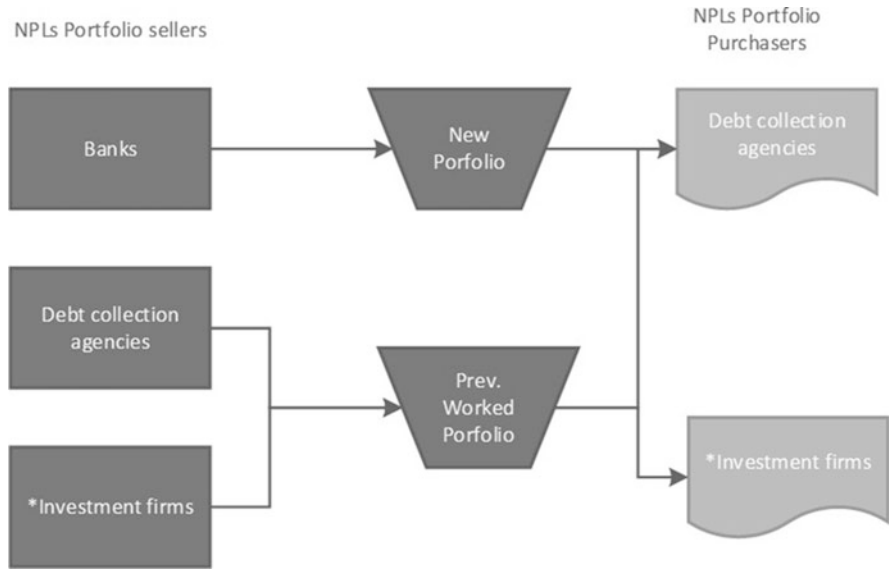


Fig. 1 Portfolio buy and sell operation

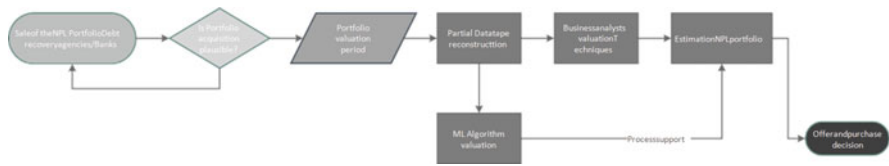


Fig. 2 Portfolio business model purchase operation diagram, showing the interaction of the solution presented for the operation

spreads characterizing their market, determined by discrepancies in data availability between banks and investors, and by poor valuation methodologies as indicated by Ye and Bellotti [4]. From this point, the following inquires emerged: Are these empirical algorithms powerful enough to consider the different interactions that occur between variables? Can they scan the different records and learn from historical data? On top of that, as researched by McKinsey [1] Latin American banks and debt collection agencies are behind on holistic digital transformation, which creates an opportunity for the implementation of new technologies to improve the decision support process. Data related to NPLs are stored by banks in tape backups, which may contain large amounts of information. These are transferred to the appraisers, who have to process the information to make forecasts and estimate their performance but given the complexity and quantity of the data, how efficient are traditional methods of evaluating datasets? According to the ESRB [22] companies still rely on poor valuation methodologies to forecast the recovery rate, therefore, the implementation of a tool to support the management and valuation of a portfolio is an essential task.

The use of machine learning algorithms in source information addressing this problem is still scarce. As Ye and Bellotti [4] explain more recently, machine learning techniques also started to be successfully applied to this field of research. However, most of the existing studies focus on corporate bonds or loans. Models focusing on retail credit products such as mortgages and credit cards still largely need to be investigated. Factors such as [5] the added privacy protection norms and confidentiality policies promoted by banks - did not allow previous references to identify many potential predictors of recovery rates for retail loans. Moreover, as Loterman [3] points out nowadays scholarly resources focusing on loan recovery, have a few outdated machine learning methodologies applied to this problem. The lack of literature in the field of non-performing loans debt recovery shows few and mostly unsuccessful investigations on NPLs. Additionally, the apogee of Machine Learning algorithms in the field of forecasting financial revenues, not yet debt collections, allows companies to base their decisions on such techniques as Gilles Loupe [14] states. While Garrigues [16] confirms an even higher trend noting that retail credits went from a delinquency rate of 3.41% in March to 5.79% in November 2020 [21], retail credits were directly affected by the economic crisis. Correct estimation of the recovery rate of the non-performing debt portfolio is an important component in the debt portfolio purchase transaction, which represents the performance and profit of the collection agency. The use of Machine Learning has not yet been fully developed in this field, leading to new research initiatives and the development of models to improve decision making.

1.3 Purpose and Tools

The purpose of this article is to show the development of a machine-learning algorithm to estimate the recovery rate of unprofitable portfolios [17]. As required by the business, the algorithm must offer a good ability to identify false positives, since these translate into an overestimation of the portfolio in the purchase operation. In other words, more value is invested in the purchase than can be recovered after handling by the collection company. The development compares different machine learning algorithms through the approaches proposed in the study, contributing results to the application area.

A specific field of machine learning supervised learning, is used, in which two approaches are utilized. The regression approach of the target variable (value to be predicted) is continuous and the classification approach in which the target variable is of discrete type. Nine supervised learning models [11] are used to forecast the recovery rates of the mentioned portfolios: Random Forest [13], XGBoost [18], Logistic Regression [20], and Unbalanced Random Forest [19]. For each model, the two aforementioned approaches are evaluated.

1.4 Methodology

A methodology with different approaches is applied to find the algorithm that best fits the business rules. The methodology used supports the way the problem is questioned and allows the conversion of the objectives into applicable approaches used in the experimentation. The classification capabilities of the models are assessed through the customer group approach using the scheme shown in Fig. 7. When an approach does not meet the business needs, a new step in the methodological cascade is taken. Customer prioritization enables the recovery of customer debt using the collection force (dedicated customer contact companies) efficiently and rapidly. The approach is realized through the methodological scheme, which gives a clear picture of what the research can achieve (Fig. 3).

The methodological cascade shows each of the steps followed to produce a model with the ability to differentiate false positives and to predict target variables. The algorithm is based on supervised models, as seen in items six, seven, and eight. The constructed methodology starts from the general requirement of the improvement of the purchasing operation (item one). Then, the reconstruction of the tape backup is performed. The first approach (item three) aims at forecasting the recovery rate. If a second approach (item four) is not possible, it uses multi-class classifiers to group the recovery amount. If the latter fails, it evaluates the binary classification that is expected to recover better scores than the other approaches analyzed. With these premises, the following experimental hypotheses are proposed.

1.5 Hypothesis

Based on the proposed methodology, it is necessary to transform the operational requirement into the desired outcome, i.e. to establish the independent variables and the target variable. The target variable is the result obtained after processing a supervised learning model [8], with dichotomous values.

The four approaches explained in the methodology are used as the basis for the construction of the datasets. Machine learning algorithms allow the formulation of models that can be understood as hypotheses to be tested. It should be noted that the use of classifier algorithms is aimed at discrete variables, while regression algorithms are aimed at continuous variables. In particular, binary classificatory models are used, where the alternative hypothesis assumes that the model can correctly discern between two categories: yes or no (or zeros and ones). The estimation of the amount of debt to be recovered will be covered by regression and classification algorithms; with the latter of the categorical type (more than two categories) and the binary type which, as noted in the previous chapter, prevails in this research.

It is emphasized that the algorithms presented are adapted to the need of each hypothesis. The use of models such as XGBoost and Random Forest together with a binary classification methodology give, through experimentation, superior results compared to the binary classificatory models produced by ImbalanceRFC and

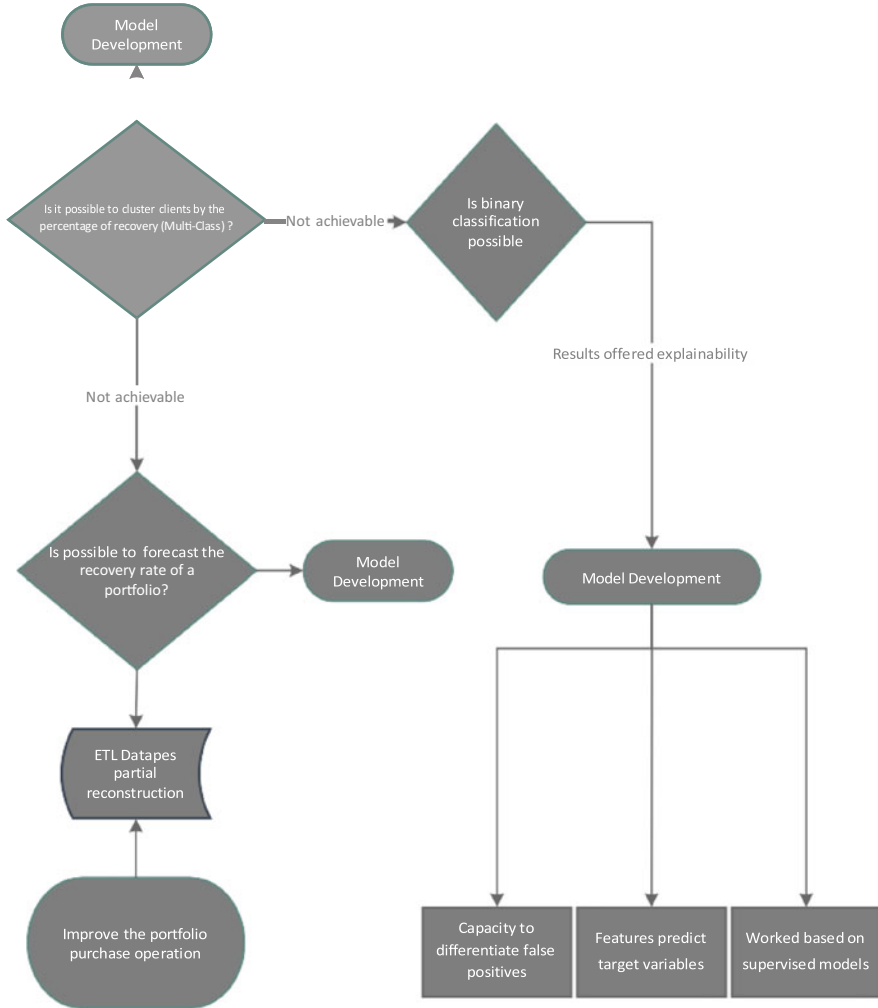


Fig. 3 Methodological cascade. It shows how the problem is approached from the general requirement “Improve the purchase transaction” to “The binary classification of the debtor”

Logarithmic regression. Meanwhile, the regression-based approach for the same models does not explain the target variable, which is usually validated with the coefficient of determination or R^2 , a metric used for continuous variables.

2 Data

The characteristics that explain the target variable, according to each of the hypotheses put forward, are extracted from the information provided by the seller of the portfolio in the tape backup. This process is carried out during the portfolio

evaluation period (limited time given to buyers to evaluate the tape backup to be purchased). During this period, the partial reconstruction of the portfolio is carried out, which corresponds to the second link in the methodological cascade presented in Fig. 2. These qualities make it possible to explain, after experimentation, the target variable.

The model's input variables include: (a) whether the type of product purchased by the borrower is commercial; (b) whether the credit was purchased by a company and; (c) whether the client is found to have active status with the national tax agency.

It should be noted that the above-mentioned variables belong to the set of categorical variables obtained from the dataset "clients". The aforementioned variables are identified and analyzed in the exploratory evaluation process to finally transform them into data that explain the target variable, after the "partial" reconstruction of the portfolio. Employing variable engineering, the datasets "customers", "payment" and "contact management" are worked on. From these, the 24 input variables for the developed model are obtained. The first dataset contains the inherent information of the debtor and the debt. The last two datasets allow the assessment of the quality of the debt and the opportunity to contact the customer again.

The three datasets arise from the "partial" reconstruction of the portfolio. It is called the "reconstruction process" since the information contained in the tape backups corresponds to flat files with a length of 1000 records each. Given this, it is required that the specialist area gathers and dumps the information in a company repository. It is called "partial" due to the limited time available to make the purchase decision and the non-inclusion of other sources of information that could contain data to complement the information provided.

2.1 Dataset Clients

It provides 61 predictive features of personal information containing socio-demographic variables, reported credit rate, and judicial status. This dataset contains 69,683 debtor records from major financial institutions in Peru. Overall, 96.99% of them are retail customers and from those 71.62% represent credit card debts. Each record corresponds to a loan and the identification number of each debtor is used as the primary key in the dataset.

2.2 Dataset Contact Management

This dataset contains the communications made by the recovery working group (third-party companies). It contains nine predictor variables that show the historical capacity to contact the debtor, through indicators such as (a) the number of calls; (b) contacts through digital channels, and; (c) visits made by the recovery task force, to collect the debt. It also includes information on customer response, grouped by

category. Finally, the quality of the contact is rated by linear coding following business criteria. The dataset contains information on 53,367 customers which is matched with information on each customer using a similar treatment of variables to that used in the previous dataset. The information, which contains both positive and negative responses, captured through contact, allows for an assessment of the willingness of customers to meet their payment obligation. This dataset contains information on debt collection in the time range from January 2015 to October 2019.

2.3 Dataset Payments

The dataset “payments” consists of a dataset showing the recoveries made by the previous owners of the non-performing debt portfolio. This information joins the payments recorded after the purchase transaction occurred dating back to August 2018.

There is a total of 69,335 payment records in the “Payments” dataset. It should be noted that for customers without any payments, a value of zero is assigned. In the portfolio, worked by the previous collection agency, 16,717 customers had not fulfilled their payment obligations. After the purchase transaction, only 3,437 had made at least one payment. This value is part of the target variable which, divided with the capital owed at the time of purchase, forms the target variable. The debt collection period recorded in this dataset is made up of payments made from November 2011 to the first day of August 2020 (Fig. 4).

2.4 Variable Selection and Mapping

2.4.1 Variable Interaction

Figure 5 shows that the variable related to the days that have passed from the date of purchase of the portfolio to the last payment made by the client presents a strong

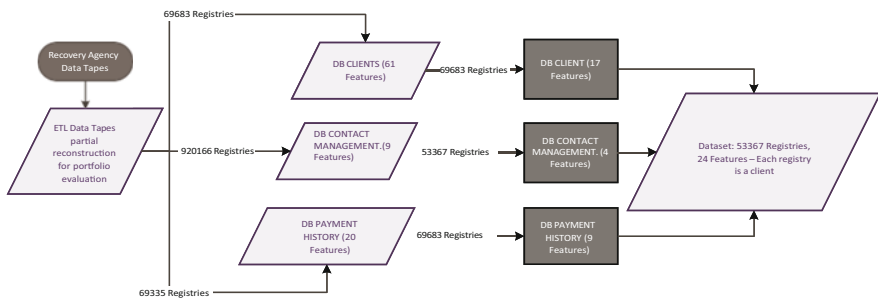


Fig. 4 Joining and transformation tree – partially reconstructed data tapes

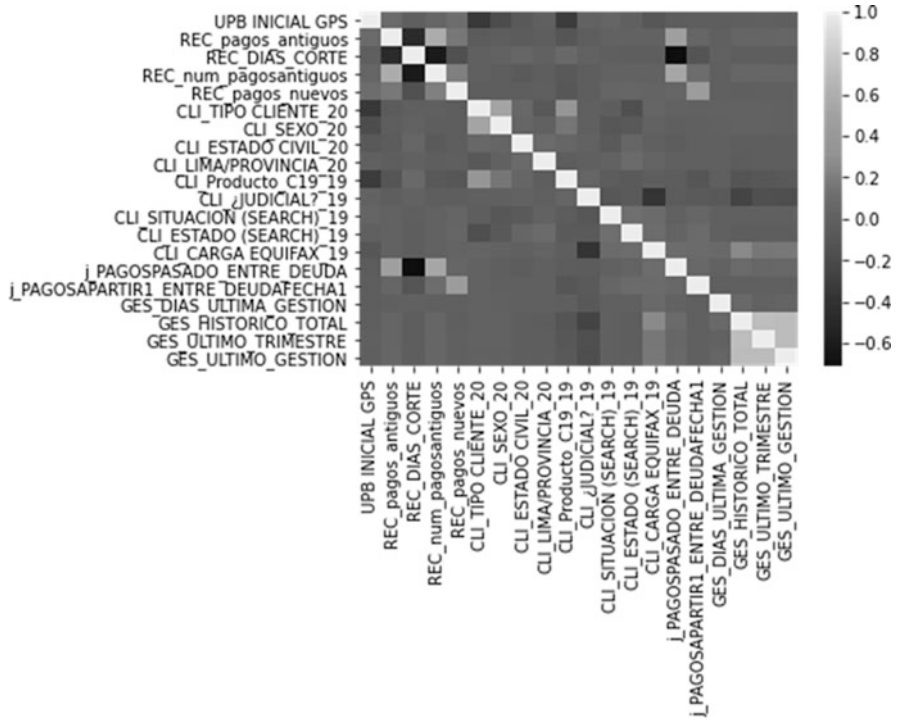


Fig. 5 Variable correlation heat map showing the interactions between variables

inverse correlation with the number of payments made up to the date of purchase. This characteristic shows the clients who have made at least one payment on at least one obligation up to the cut-off date, which is August 9th, 2018. This variable represents the date of the last payment issued by the debtor. It should be remembered that the proportion of customers who have made at least one payment represents only 31.32% of the total portfolio. In addition, it should be noted that this variable has a direct correlation with the variable representing the past payments made concerning the total capital of the debt, the variable is calculated according to the script in (1).

$$\begin{aligned} \frac{df[j_PAGOSPASADO_ENTRE_DEUDA]}{df[REC_pagos_antiguos]} &= \frac{df[UPB_INICIAL_COVINOC]}{df[UPB_INICIAL_GPS]} \end{aligned} \tag{1}$$

$$\begin{aligned} \frac{df[j_PAGOSAPARTIR1_ENTRE_DEUDAFECHA1]}{df[REC_pagos_nuevos]} &= \frac{df[UPBINICIALGPS]}{df[UPB_INICIAL_GPS]} \end{aligned} \tag{2}$$

In order to model the forecasting algorithm, the variable that measures the number of payments made, up to the date of purchase is removed from the dataset given its correlation with the variables shown in Fig. 5.

2.4.2 Variable Mapping

It is validated that each record in the dataset relates to one and only one customer. The dataset used as input data for the model consists of 53,367 records. This dataset represents the input for the algorithms tested in Table 1 where the target variable calculated in the script (2) is included.

As explained in Sect. 2.4, the cutoff date represents the portfolio purchase transaction and diverges from the equations in scripts (1) and (2). A function needs to be applied to map the classes into the target continuous variable, which corresponds to the output of the script (2). The target variable mapping allows each of the hypotheses set out in Table 1 to be established by transforming the continuous variable to a dichotomous variable.

Table 1 Hypothesis grouped by its corresponding algorithm

Met	Hypothesis description	XBG Class	Imb RFC	Log Reg	RFC	RFC	RFR Lasso	Linear Reg	XGB Reg
Binary	No payment will be received Does the client pay at least a quantity	x	x	x	x				
3 Multiclass	Does the client will pay more than 50 % Does the client will pay less than 50 % No payment will be received	x			x				
4 Multiclass	Does the client will pay more than one third does the client will pay between one third up to two thirds Does the client will pay more than two third No payment will be received	x			x				
Reg 1	What is the recovery rate of those who have paid in the past				x	x	x	x	x
Reg 2	What are the recovery rate of the whole dataset those never payed				x				x

3 Algorithm Selection, Results, and Deployment

3.1 Structure

The following development shown in the flowchart in Fig. 6 shows the different six stages of model development after “partial” portfolio reconstruction: Data processing (also known as extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system) allows to reconstruct the data tapes given by the seller.

The data flow shows the modeling process behind the execution of each of the approaches shown in the methodological cascade (Fig. 7).

The data flow shows the modeling process behind the execution. The first link consists of the extraction of the information and loading from the partially reconstructed tape backups. Through the engineering of variables and the transformation of the data, the variables that can describe the target variable are extracted, as explained in Sect. 2. The Implementation Consists of the Iteration of Processes That Follow a Certain methodology until obtaining the model that provides the best

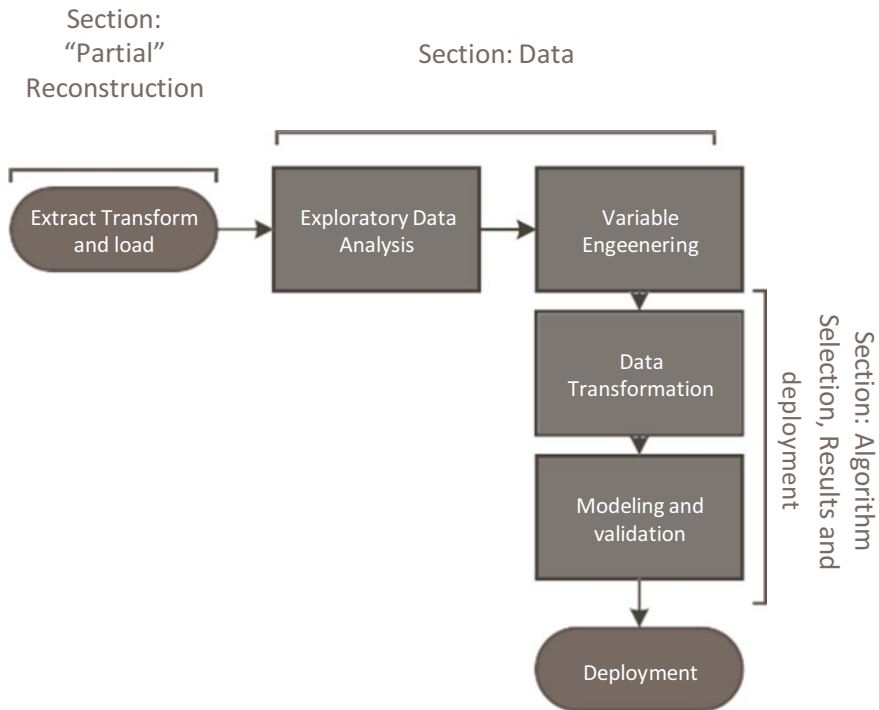


Fig. 6 Data processing scheme

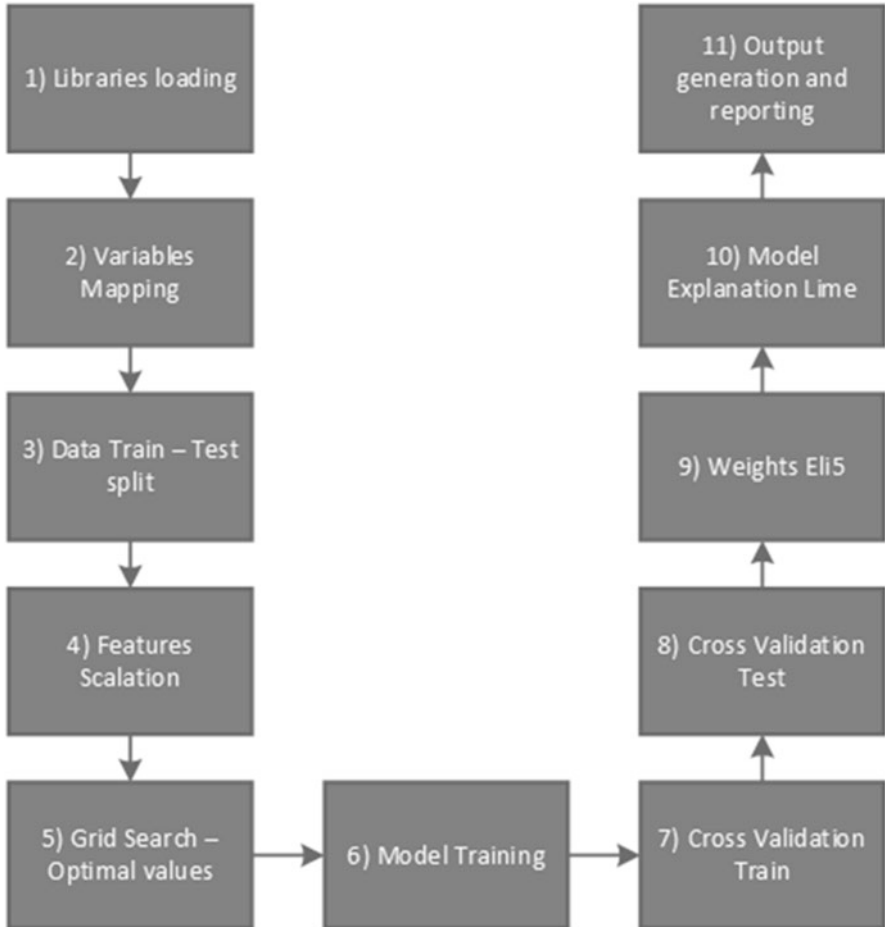


Fig. 7 Modeling scheme

metric, such as the F1- Score; this, after discarding the use of regression-based models.

The algorithms are shown in Table 1 are used inspired by the research carried out by Brigo and Bellotti [5] and the results offered, found that rule-based algorithms of ensemble type, especially random forests and the newly added boosted trees and Cubist, displayed the best forecasting performances. Due to the lack of research in the field of non-performing loans using classifiers, the approach taken by the two authors serves as a starting point. This research becomes the starting point for the experimentation and validation of the worked models, which is subsequently confirmed through the results. It should be noted that the approach taken by Brigo and Bellotti is based on the use of regressors, while the present research is based on the use of classifiers, as the approach taken by the two authors failed to achieve the

minimum threshold required by the methodology, therefore the classification approach provides new results for the field under study.

The scheme presented above shows the working route to obtain the results of the experiment. The first link illustrates the loading of the necessary libraries and packages (Python programming language). The mapping of variables allows masking the target variable, obtained from the preprocessing. The “data train-test split” allows the dataset to be split in a proportion of 70% (training) and 30% (testing). After the split, the information is scaled, where the minimum and maximum scaler is used given the different scales in the units presented by the variables. The “grid search CV” package is used to obtain the optimal configuration of hyperparameters of the base model. The information is loaded to the elaborated model and the training is carried out to obtain the results. These mainly consist of obtaining the confusion matrix from the cross-validation process as well as the F1-Score and R^2 performance metrics. The ability to explain the calculations made by the model is sought, together with the ranking of the most relevant variables. Finally, the output file of the information is generated.

3.2 Modeling Algorithms

The methodologies in Table 1 are addressed by categorizing them according to the method used and the hypothesis put forward for each of the approaches. The best-performing algorithms are then evaluated.

3.3 Metrics

The following metrics are used to compare performance between models (Table 2):

3.4 Data Pre-processing and Development

The development follows the scheme shown in Table 1.

Table 2 Used metrics to compare performance between models

Algorithm	Metrics
Regression	R^2
Binary classification	F1 Score (Macro)
Multiclass classification	F1 Score (Macro)

3.4.1 Binary Classification

Base Model Experimentation

This classification methodology offers the best results obtained from the four algorithms used: XGBoost, Random Forest, Logistic regression, and Imbalanced Learning. There is a clear recognition of false positives, which can cause an overestimation of the portfolio, as those defaulters would be counted as future payers. This results in good discrimination in the confusion matrix, as the business rule requires that the number of false positives, which are related to type 1 error, should be minimized. It is also worth mentioning that by expanding a group of customers (cluster) there is an improvement in the performance of the model.

It is evident from Fig. 8 that both tree models give high relevance to the variable that explains the capital owed at the time of the portfolio purchase and the binary variable that explains whether the product is a credit card or not; however, the model with the best F1-score is based on the XGBoost algorithm (Fig. 9).

The algorithm that best filters out false positives and is selected by the company is binary XGBoost due to the low overestimation that ensures the forecast. True positives make up the group of defaulting customers with the highest probability of recovering the principal owed in a shorter time. Therefore, the recovery task force can be targeted to this group of customers to benefit the company's performance. Unlike

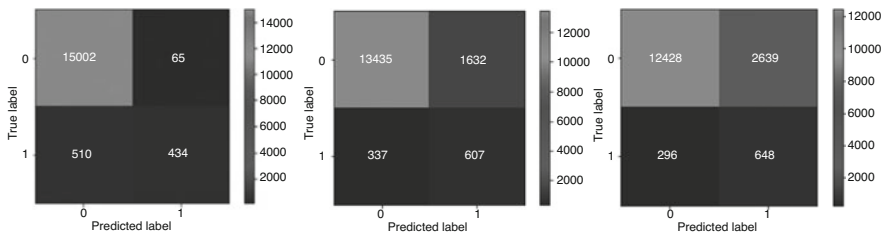


Fig. 8 Binary XGBoost, RFC, ImbalancedRFC Confusion Matrix

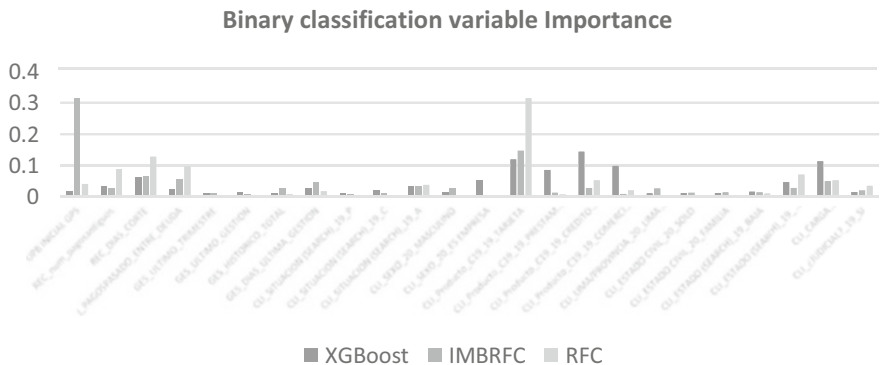


Fig. 9 Variable importance on binary classification algorithm

the XGBoost model, the Random Forest and Imbalanced Learning algorithms show higher type one error, but this generates an increase in positive true values. While this model may appear to have sifted out a better number of customers who meet at least a percentage of their debts, the increase in error type one may decrease and obliterate the profitability of the firm, since a higher bid price may be offered at the time of portfolio purchase; that is, a larger amount of money may have been paid, but never recovered. On the other hand, in the Logistic regression model, which was treated as the base model for the described methodology, the type one error is similar to the two models mentioned above. The unbalanced RFC model in Fig. 8 shows a true identification of 648 but has 2,639 records corresponding to type one error. Therefore, these records are misclassified being this number higher compared to the other binary classification algorithms to which the same analysis is applied (Fig. 10).

The graph shows the performance metrics obtained after the experimentation of the third approach (binary classification) of the methodology. The need for a macro F1- score metric of more than 65% is highlighted as an indispensable requirement. Three algorithms were capable of what was requested, the XGBoost model the one that gave the best performance in the indicated metric among the three algorithms that meet the criterion (Fig. 11).

The ROC AUC curve shows a performance measurement for the classification problems at various threshold settings, the ROC AUC curve achieves a value of 0.842. A model with perfect skill is depicted as a point at (1,1). A skillful model is represented by a curve that bows towards (1,1) above the flat line of no skill. The other two approaches within the classification algorithms used are the three clustered and four clustered, both representing part of the methodology proposed. These approaches offered less performance compared to the binary classification approach.

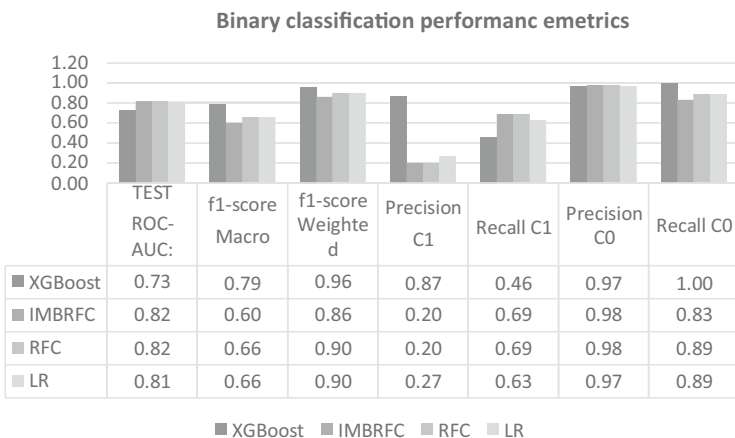


Fig. 10 Performance metrics binary classification

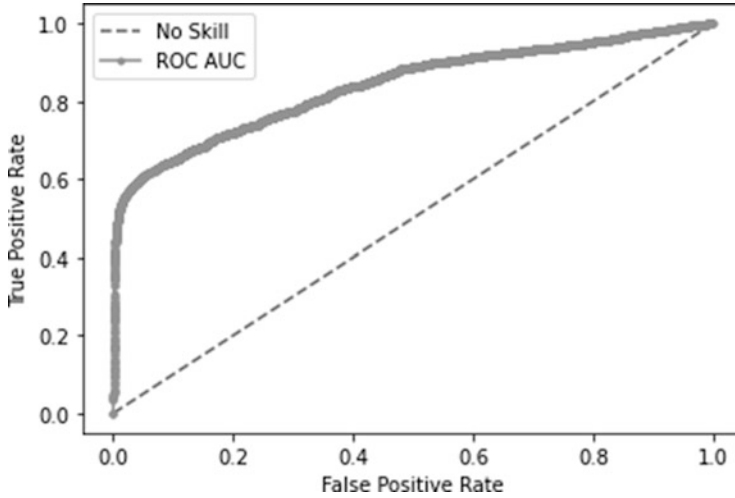


Fig. 11 ROC AUC curve XGBoost binary classification

Further Experimentation (Fine Tuning)

To improve the performance of the presented model and allow a greater generalization using the binary classification approach, the following hyperparameter testing is proposed for the XGBoost model. It should be noted that the hyperparameters presented are specific to each model, library, and programming language in which they are executed. We start with a base reference model to compare 3 Grid Search Method methods (a method that exhaustively searches, from a pre-established list of hyperparameters, to obtain the best performance). The first one, Coordinate Descent, optimizes one parameter at a time, and in the case of having several hyperparameters, there could be several local minima. This method is not optimal as there could be cases where a combined modification of two or more hyperparameters would produce a better score. The second method, Randomized Search, selects hyperparameter values from the search space randomly; and finally, the third method. The technique that calculates a posterior distribution on the objective function based on the data and then selects good test points against this distribution. For the experimentation of the whole data set, it was separated into two data sets (test and training) in a ratio of 3 to 7. The training set is again divided in a ratio of 3 to 7 to obtain a second level of train and test data sets (Fig. 12).

Selected hyperparameters are iterated following a discrete list of values. The graphic shows how a specific parameter (γ - Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger γ is, the more conservative the algorithm will be) varies according to its value change. After the experimentation process, the following values of the area under the curve of the AUC curve are obtained. This allows obtaining the discernment capacity of the model. It is important to clarify that the curve shows the performance through the

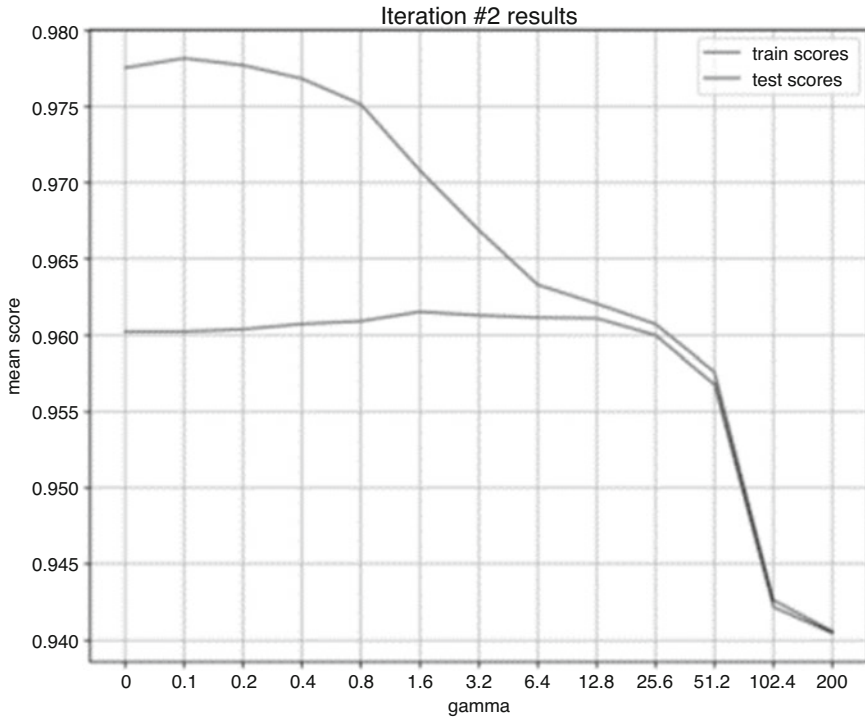


Fig. 12 Iteration parameter

Table 3 AUC fine tuning

	AUC Test	AUC (Never seen)
Benchmark	0.8224	0.8160
Coordinate	0.8277	0.8225
Randomized	0.8424	0.8344
Bayes	0.8408	0.8324

different methods used in the Grid Search CV to find the hyperparameters that give the best performance of the model (Table 3).

It is observed that the tuning of the hyperparameters using the Randomized Search and Bayesian Search methods results in a lower value of type 1 errors, as well as an increase in the AUC of approximately 2%.

3.4.2 Hypothesis Testing

The comparative table with the performance metrics for the models is shown in Table 4 is presented. Regression models do not appear because they do not offer significant results.

Table 4 Results and performance metrics

Classification		Regression							
Appr.	XGB Class	Imb RFC	Log Reg	RFC	Appr.	RFR	Lasso	Linear Reg	XGB Reg
Binary	0.79*	0.6*	0.66*	0.66*	Reg 1	0.3**	0.26**	0.21**	0.28**
3 Multiclass	0.59*			0.51*	Reg 2	0.3**			0.4**
4 Multiclass	0.48*			0.34*					

(*) The values obtained correspond to the F1-score (Macro) metric. The values are comparable to each other

(**) The values obtained correspond to the correlation metric R^2 . The values are comparable with each other

(Appr.) Denotes the selected approach

4 Conclusions

It is evident that the preprocessing of the information requires the use of a window of time in which two moments are denoted “pre and post-purchase”, this allows the construction of the target variable.

The methodological cascade starts from the specific hypothesis of predicting the recovery rate. After experimentation, it is concluded that the results do not meet the need of the business. To address a new casuistry, the use of the continuous target variable is ended and it is proposed three new approaches that have as their mission to respond to dichotomous target variables (groupings by percentages), culminating as fourth experimentation the approach of binary segmentation.

The approach based on classification models of dichotomous variables allows obtaining better performance metrics. Furthermore, within the classifier models, the binary classification approach is more favorable than the other approaches compared. Within the binary classification approach, the model based on “Boosted Trees” [12] (XGBoost Classifier) had the best performance among the other algorithms (0.79 in the F1-Score macro metric) for the prediction of the binary target variable.

The XGBoost-based binary classifier model obtained the best false positive differentiation rate (Type one error). A value of 0.41% (65/16011) corresponding to type one error was obtained. Thus, it could be concluded that the overvaluation of the portfolio, a requirement requested by the business, is avoided. The ROC AUC shows a value of 0.842 reveals the degree of discerning offered by the model.

The implementation presented allowed the classification of customers who would pay at least some amount of their debts so that the resources of the recovery force could be directed to that group. With this, it is expected that the expense of the debt collection operation will be minimized.

5 Recommendations

To improve the performance of the model, new variables and data transformations can be incorporated so that the model can have a greater capacity to explain the target variable, thus obtaining a better predictive capacity.

The inclusion of variables in the regression models containing data on willingness to pay and ability to pay can allow for better consistency in the model. That is, implementing, for Random Forest-based models, variables targeting characteristics that better detail customers who do not make any debt payments. For XGBoost-based models [7], the focus should be on features that detail customers who pay on time during portfolio operation.

In addition, manual refinement of hyper-parameters can improve predictability along with experimentation with new models that could improve predictability and the ability to predict the debt recovery rate. Plus, by using ensemble techniques, better performance could be obtained from two or three models. The inclusion of different data sources can reduce the bias of the model input data.

References

1. McKinsey & Company. (2019, July). Lessons from leaders in Latin America's retail banking market. <https://www.mckinsey.com/~media/mckinsey/industries/financial%20services/our%20insights/lessons%20from%20the%20leaders%20in%20latin%20americas%20retail%20banking%20market/lessons-from-leaders-in-latin-americas-retail-banking-market.pdf>
2. European Systemic Risk Board. (2019, March). Annual Report 2018. <https://doi.org/10.2849/042348>
3. Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 161–170. <https://doi.org/10.1016/j.ijforecast.2011.01.006>
4. Ye, H., & Bellotti, A. (2019). Modelling Recovery Rates for Non-Performing Loans. *Risks*. <https://doi.org/10.3390/risks7010019>
5. Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. (2019). Forecasting recovery rates on non-performing loans with machine learning. *Credit Scoring and Credit Control XVI*. <https://doi.org/10.3390/risks7010019>
6. Deloitte Hungary. (2019). What's beyond the peak? CEE loan markets still offer new opportunities. <https://www2.deloitte.com/content/dam/Deloitte/ce/Documents/about-deloitte/non-performing-bank-loans-npl-study-2019.pdf>
7. Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. <https://doi.org/10.1214/aos/1013203451>
8. Shalizi, C. (2008). Statistics 36–350: Data Mining. Carnegie Mellon University. <https://www.stat.cmu.edu/~cshalizi/350/2008/>
9. ESRB. (2019). The impact of uncertainty on activity in the euro area. European Union: ESRB. <https://doi.org/10.2849/224570> (pdf)
10. European Central Bank. (2018). *Guidance to banks on non-performing loans*. European Central Bank. <https://doi.org/10.2861/96204>
11. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
12. Coadou, Y. (2013). Boosted Decision Trees and Applications. EPJ Web of Conferences, 55, 02004. <https://doi.org/10.1051/epjconf/20135502004>
13. Statistics Department University of California Berkeley, & Breiman, L. (2001, January). RANDOM FORESTS. <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>
14. Gilles, L. (2014). Understanding Random Forests: From Theory to Practice. *Department of Electrical Engineering & Computer Science*. University of Liège. <https://doi.org/10.13140/2.1.1570.5928>
15. Lizarzaburu, E., & del Brío, J. (2016). Evolución del sistema financiero peruano y su reputación bajo el índice Merco. Período: 2010–2014. *Suma de Negocios*, (págs. 94–112). Lima. <https://doi.org/10.1016/j.sumneg.2016.06.001>
16. Garrigues. (2020). Transacciones con carteras de deuda (NPLs) y activos tóxicos (REOs) LatAm & Iberia – NPLs Task Force (4T 2020). https://www.garrigues.com/sites/default/files/documentos/transacciones_con_carteras_de_deuda_npls_y_activos_toxicos_reos_situacion_a_noviembre_de_2020.pdf
17. Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. <https://doi.org/10.1007/s42979-021-00592-x>
18. Chen, T., & Guestrin, C. (s.f.). XGBoost: A Scalable Tree Boosting System. <https://doi.org/10.1145/2939672.2939785>
19. Lemaitre, G., Nogueira, F., & Aridas, C. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18, 1–5. <https://www.jmlr.org/papers/volume18/16-365/16-365.pdf>

20. Shen, Aihua & Tong, Rencheng & Deng, Yaochen. (2007). Application of Classification Models on Credit Card Fraud Detection. International Conference on Service Systems and Service Management. 1 - 4. <https://doi.org/10.1109/ICSSSM.2007.4280163>.
21. Garrigues. (2021). Transacciones con carteras de deuda (NPLs) y activos tóxicos (REOs) LatAm & Iberia – NPLs Task Force (3T 2021). https://www.garrigues.com/sites/default/files/documents/transacciones_con_carteras_de_deuda_npls_y_activos_toxicos_reos_situacion_a_octubre_de_2021.pdf
22. European Systemic Risk Board, Suárez, J., & Sánchez Serrano, A. (Eds.). (2018). Reports of the Advisory Scientific Committee (N. 7). <https://doi.org/10.2489/617721>