Improvement in Bank Financing Access for Peruvian MSE using prediction and classification models

Alvaro Samaniego, Eng[1], Jonatán Rojas, Mg[1], Alexia Cáceres, Mg[1], Alessandro Gilardino, Eng[1], Felipe Viamonte, Eng[1] and Renzo Benavente, Mg[1]

[1]Pontificia Universidad Católica del Perú, Perú, alvaro.samaniego@pucp.edu.pe, jrojasp@pucp.pe, alexia.caceres@pucp.pe, agilardino@pucp.pe, felipe.viamonte@pucp.pe and renzo.benavente@pucp.pe

Abstract – Peruvian economic development has grown 149% in the last decade – 15% yearly – mainly because of micro and small enterprises (MSE) which stands for about 42% GdP thus MSE becomes a key driver for national development and growth.

However, as MSE can acquire various business know-how and transform into influential partners for big enterprises, they still face obstacles that limits their mid and long-term survival rate (avg. life cycle of 8.3 years, 6% of annual mortality rate), mainly because of the lack of access to ways to gain funds in order to invest in new machinery or upgrades in business infrastructure to be more productive and, in consequence, be more profitable and contribute to national growth by more tax payments.

The current research assesses this problem by understanding which are the main difficulties that MSE face when trying to apply to a loan for a financial institution. A dedicated study was performed to identify and prioritize which variables are statistically relevant to the approval or rejection of a loan application and then use data mining techniques such as logistic regression and neural networks to model the probability of approval of a loan application and in consequence give highlights of the most important parameters that MSE should improve in their business model.

Keywords-- Data Mining and MSE, predictive analysis of bank credit, Logistic regression on the likelihood of bank-issued credit.

I. Introduction

According to Law N° 28015 – 2003 of Promotion and Formalization of the MSE, the microbusiness, and small business is the economic unit established by a natural or legal person, under any form of business organization or management stated by current laws, that develops extraction, transformation, production, marketing of goods and service activities. The economic development in Perú depends greatly on MSE because they represent more than 99% of the formal productive units [1] and earn about 40% of the national Gross Domestic Product [2]. This is evidence of the MSE's role at both an economic and institutional level as they have the highest employment rate in Latin America [3]; however, they have weaknesses that limit their growth such as informality, low production levels, credit access difficulty, among others.

We will give two perspectives about their characteristics: The TUO perspective from the Productive Development and Business Growth Support Law [4] and the SBS Normative perspective [5].

TABLE 1
MSE definition according to Produce and SBS

Business Type	Produce	SBS
Microbusiness	Annual sales for a maximum of 150 UIT. (1UIT = S/. 4,150)	Total credit indebtness in the financial system less than S/. 20,000
Small Business	Annual sales between 150 and 1700 UIT.	Total credit indebtness in the financial system between S/. 20,000 and S/. 300,000

In many cases, the MSE are stable for long periods but most of that time is dedicated to surviving, deprioritizing and losing growth opportunities. This is mainly because of difficulties they have to get or apply for bank-issued credit. According to credit information central Sentinel, 63% of MSE have credits in the regulated financial system (2.23 million of business), 25% of these have a bad score because of late payments of any sort, according to SBS criteria [6].

On the other hand; in the last decade, Bank-issued credit emission has increased exponentially because of two fundamental reasons: A great diversification of credit products with customized requirements and the impulse of the entrepreneurship culture in our country mainly because of the MSE. This client increase rate has motivated financial institutions the need to manage non-payment probability of clients more efficiently and reliably, also known as the credit risk, switching from an interpersonal estimation (interviews or expert opinions) to a databased structured approach. This risk management process becomes a client risk sorting using external (by hiring risk sorting agents such as MOODY´s) or internal methods (actual bank methods, predictive or sorting mathematical models); in both cases, clients don´t have access to the followed criteria for risk category assignation.

This research performs an statistical analysis of significant variables that financial entities consider when they issue a loan to a MSE. From information collected by the bank entity, no name given, the sorting method will be replicated using internal methods or credit scoring to find the relevant factors for the approval of the bank-issued credit for a MSE. In this way, insights will be generated that will maximize the probability of bank-issued credit access. Common internal methods for

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.7
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

sorting have been used: logistic regression, logistic regression based on sorting trees, and machine learning in the form of neural networks. Results of each predictive model have been compared to determine the most efficient and most accurate for sorting.

The objective of our research is to determine which are the significant variables that impact the bank-issued credit authorization for a MSE, based on its size and business type, and improve on aspects that are related to those variables to become more competitive and raise their employment and sustainability level.

II. Literature Review

The state of the art and the conceptual definitions of the main tools that will support this research will be detailed:

A. *Data mining overview*

Data mining is the process of applying a specific algorithm to large data samples for extracting patterns and allowing the automated discovery of interesting knowledge [7]. Data mining has three main goals: description, prediction, and prescription, mainly divided into two main categories, supervised and unsupervised learning [8], and has been applied in many fields such as banking, finance, telecommunications, insurance, marketing, and several more [9, 10, 11].

Although there are many data mining algorithms developed, can be divided into three main stages: the first stage is determining the relationships and regularities in the data sample, the second stage refers to applying the rules observed in order to predict new objects and familiarities, and the third stage is related to the analysis of variance to process deviations and error analysis [8].

Due to the increasing development in data mining techniques, many researchers are applying them in diverse fields including performance prediction systems [12]. There is empirical evidence related to the determinant factors for deposit fee pricing using data mining methods and data provided by the Banks and highlighting the importance of considering client´s characteristics for the determination of the deposit rates [13]. Data mining techniques have many applications for customer profile discrimination, for example, on credit scoring models. The use of selection algorithms for characteristics and group sorting can improve bank performance when dealing with credit scoring problems. The simultaneous and hybrid use of many sorting and learning algorithms, with their parameter setup considered, is recommended and results in data mining hybrid models. For example, there are models that start with data gathering and preprocessing and then use algorithms for principal component analysis, genetic algorithms and finally, the resultant accuracy of the implementation of the sorting algorithm is evaluated [14]. Another reason for data mining usage is that nowadays number of databases managed by financing entities is so big that a massive financial data groups analysis for risk group sorting becomes necessary. Data mining techniques can be useful to reduce risk for the financial entities when predicting which clients will make their payments on time [15].

B. Credit Scoring

Bank performance forecasting is the main issue for managers, mainly due to poor control may cause a bankruptcy, generating an adverse influence on the country's economy [12]. In the last decade, banking systems couldn't satisfy the financial needs of small enterprises, highlighting inefficiencies in their general loan system processes, which includes acceptance or rejection of a credit application based on the knowledge, experience, and judgment of the evaluator [16]. After USA financial crisis in 2007-2008, was clearly stated that corporative credit qualification is a key role in credit risk management reflecting that credit risk evaluation for guaranteed loans is an important operation in bank systems to ensure that moneylenders will comply with their payments as scheduled [17].

There are several credit qualification models applied to datasets, such as logistic regression, multilayer perceptron, and vector support machines [18]. In the last years, artificial intelligence has been widely used to assess credit risk, achieving important improvements in credit scoring systems and bankruptcy forecasting, minimizing losses for non-payment. Recent research shows that a simple classifier based on unprecise probabilities and uncertainty measures may achieve a better compensation between several important aspects for this kind of study, such as precision and area under the ROC curve [19]. As can be observed, according to literature, most studies focus on developing an accurate qualification credit model to determine if the loan will be granted to new appliers [20]. As a result, rating and credit evaluation is a key analytic technique as support for managers to assess appliers [21, 22].

C. *Logistic regression*

Given the predictive variables $X_1, X_2, X_3, \dots, X_k$, an ordinary linear regression can be used (1).

$$E\{Y|X\} = X\beta \qquad (1)$$

The expected result of the binary variable is to find the probability of occurrence $Probability\{Y = 1\}$. Nevertheless, it cannot be adjusted to a linear model due to the values $Probability\{Y = 1\}$ can be higher than 1 or less than zero. For that reason, it would be convenient to use a statistic model that allows contemplating binary answers, being a good choice the binary logistic regression model. The model is expressed as $Y = 1$, given the values included in the matrix X that shows the independent variables (2).

$$Probability\{Y=1|X\} = \frac{exp^{(X\beta)}}{1+exp^{(X\beta)}} \quad (2)$$

The logistic regression model was developed mainly by Cox, Walker, and Duncan [23]. The success ratio over the failure, odds, can be determined from equation 2 in equation (3).

$$odds = \frac{Probability\{Y=1|X\}}{Probability\{Y=0|X\}} = exp^{(X\beta)} \quad (3)$$

Finally, to find the coefficient of the regressive variables, it is convenient to find the natural logarithm of the improvement ratio, obtaining the logit function, which is shown in equation (4).

$$logit = ln(odds) = X\beta \quad (4)$$

D. Neuronal or neural networks

Currently, the number of defaulted loans and the competition in the banking market has increased. Nevertheless, some commercial banks are reluctant to use data mining tools in order to support credit decisions. According to researches, the artificial neuronal networks represent a new family of statistic tools and promises of data mining tools that have been successfully used in classification problems in many domains allowing the evaluation of the loan application with improvements in the effectiveness of the credit decision, as well as the time saved for decision making [24]. During the last years, research has been developed on neuronal emotional networks, which have been successfully applied for the recognition of patterns. However, they differ from the learning of a basic neural network, either in the proportion of training data, as well as in the validation used during training and testing [21]. Despite this, multilayer perception neural networks (MLP) are widely used in automatic credit rating systems with high precision and efficiency, with different initial weights and instances of training [25].

III. Predictive Models Application

Based on the literature review in this section the predictive models proposed will be the conditional inference tree-based logistic regression model and neural network modeling, mainly because of (1) ease of use and interpretation (2) to assess the benefits of using tree inference method to improve logistic regression results. A case study of a local bank will be modeled based on these two methods using an extract of the historical loans applications (approved and rejected) to compare accuracy and performance measurements and finally perform statistical validation with test/control group.

A. Dataset description for MSE loan assessment scoring

Dataset for this case study consists mainly of sociodemographic and credit scoring features, such as economic sector, type of person (legal/natural), and geographic information about agency that managed loan applications according to bank clustering. It also includes features about loan applications, such as the amount of money requested (local /foreign currency), sale and evaluation channel, credit rate, and probability of default based on client. Finally, there are also variables related to marketing campaigns and in which month they were executed.

This dataset contains 31,157 cases of loan applications and a mix of 29 categorical and numerical variables that are shown in Table II. To do proper predictive models, some of these categorical variables will be treated as indicator variables and will be respectively transformed. Approved/denied loans in this dataset are labeled as 1 or 0, respectively. Dataset is divided into training and validation subsets, and the ratio of instances for each subset will be 80-20 for all predictive models.

TABLE II
Decision attributes used for evaluating credit risk in bank dataset

ID	Attribute	Description
1	IDCLIENT	Unique client identifier
2	MONTH	Month when loan application was submitted
3	TYPE_CAMP	Type of market campaign
4	STATUS	1 if loan application is approved, 0 otherwise
5	DENOMINATION	Type of exchange (local / foreign)
6	APPROVED_AMOUNT	Approved loan amount, in local exchange
7	AMOUNT_DOLLARS	Requested loan amount, in foreign exchange
8	AMOUNT_SOLES	Requested loan amount, in local exchange
9	RATE	Loan credit rate
10	SECTOR_ECONOMIC	Economic sector
11	CAMP	Market campaign description
12	TYPESOL	Guarantee inclusion variable
13	LNPER	Class of person (legal / natural)
14	EXCHANGE_SOL	Exchange rate
15	AREA	Geographic variable
16	REGION	Geographic variable
17	ZONE	Geographic variable
18	DESDEPART	State where loan was processed
19	S_CAN	Sales cannel
20	E_CAN	Evaluation cannel
21	DJUS	Level 3 justification on rejected applications
22	DJUS_CLUS	Level 2 justification on rejected applications
23	JUST_CLUST	Level 1 justification on rejected applications
24	ESTCANEV	Status of loan-Sales channel
25	LG_ OUTLAY	'S' for money outlay, 'N' otherwise
26	PBCLIENT	Probability of default
27	SEG_COM	Commercial clusterization, bank criteria
28	SEGMENT	Global clusterization, bank criteria
29	YEAR	Year when loan application was submitted

B. Conditional inference tree-based logistic regression

The decision tree-based logistic regression model is built on three steps. In first step, data will be preprocessed and cleaned, creating coded variables for categorical attributes. Then, a simple logistic regression model will be tested using a training subset to analyze significant variables and reduce dimensionality. Next, for these significant attributes, a conditional inference tree will be deployed to find cutoffs that could improve the model's predictive power compared to non-grouped variables. Finally, groups formed with these cutoffs will be coded and then tested with a logistic regression model to estimate model efficiency in training and test subset.

Data preprocessing and simple logistic regression model

After six iterations, using p-value criteria for variable selection, final significant predictors are shown in Table III. Even though some attributes appeared as non-significant, they could improve performance in the model with decision tree grouping.

Pseudocode

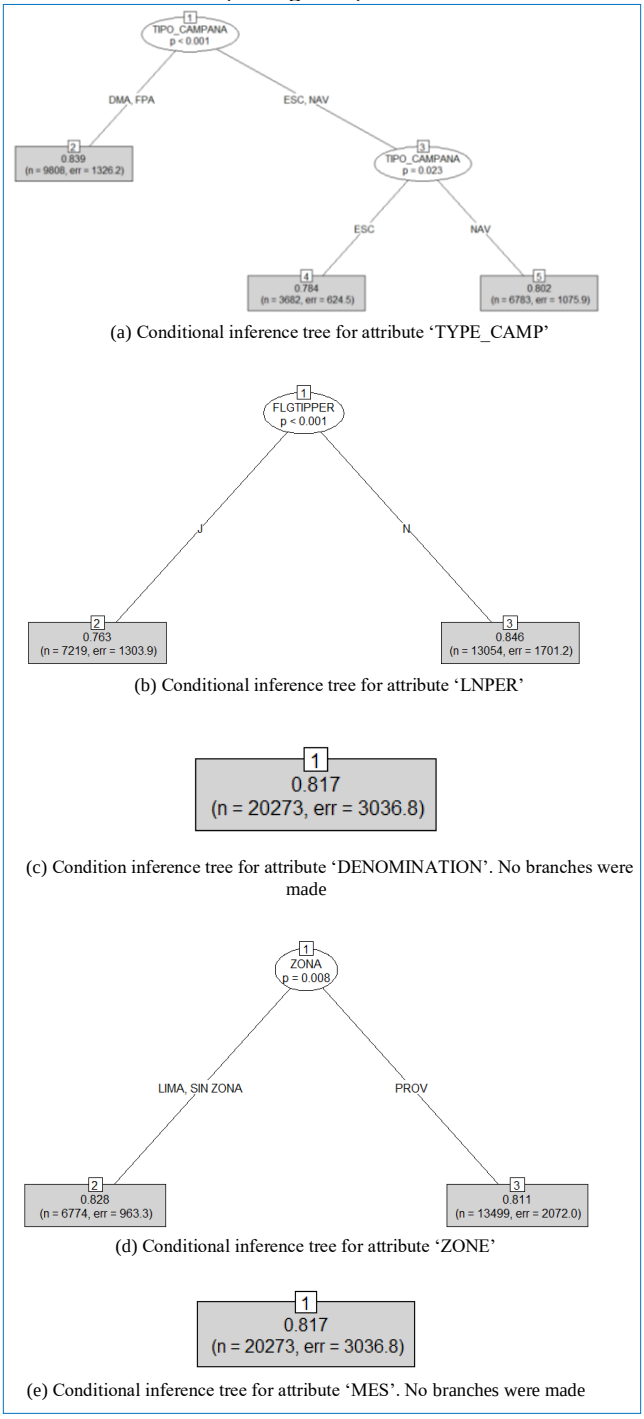
```
#first iteration: evaluate model
Modelo <- glm(formula = STATUS ~  TYPE_CAMP + DENOMINATION
              + APPROVED_AMOUNT + RATE + SECTOR_ECONOMIC +
              CAMP + TYPESOL + LNPER + AREA + REGION + ZONE
              + DESDEPART + S_CAN + E_CAN + PBCLIENT +
              SEG_COM + SEGMENT + YEAR,data = Base_Bank
              [ind_muestra==0,],family=binomial)
summary(modelo)
# errors in the variables E_CAN, REGION, ZONA, AREA, CAMP
remove them
#second iteration: evaluate model
Modelo <- glm(formula = STATUS ~ TYPE_CAMP + DENOMINATION +
              APPROVED_AMOUNT + RATE + SECTOR_ECONOMIC +
              TYPESOL + LNPER + DESDEPART + S_CAN + PBCLIENT
              + SEG_COM + SEGMENT + YEAR,data =
              Base_Bank[ind_muestra==0,],family=binomial)
summary(modelo)
#Iterated six times until obtaining the optimum model
# Final iteration. Evaluate efficiency measures with GINI
p <- predict(modelo, type="response")
pr <- prediction(p, Base_Bank$STATUS[ind_muestra==0])
auc <- performance(pr, measure = "auc");auc_yvalue <-
auc@y.values[[1]]
GINI <- if(auc_yvalue>0.5){(2* auc_yvalue-1)}else{(2*(1-
auc_yvalue)-1)}
auc2 <- performance(pr, "tpr", "fpr")
KS = max(attr(auc2,'y.values')[[1]]-
attr(auc2,'x.values')[[1]])
```

TABLE III
Coefficients for selected attributes with simple logistic regression, training subset

Coefficients	Estimate	Std.Error	Z value	P-value
(Intercept)	0.96612	0.63781	1.515	0.13
TYPE_CAMPESC***	-0.5457	0.07087	-7.7	1.36e-14
TYPE_CAMPFPA***	0.42885	0.07016	6.113	9.79e-10
TYPE_CAMPNAV***	0.48738	0.1013	4.811	1.50e-06
LNPER***	0.58507	0.03909	14.967	< 2e-16
DENOMINATION	0.80741	0.63238	1.277	0.202
ZONEPROV***	-0.21994	0.03987	-5.517	3.45e-08
ZONESINZONA	8.87198	119.46805	0.074	0.941
MONTH***	-0.11885	0.02088	-5.692	1.25E-08
Akaike Information Criterion (AIC) = 19001 Baseline selected automatically by software				

Conditional inference decision trees for selected variables

Categorical variables are coded as dummy variables with names concatenated as variable name and each unique value for each variable. After this, variables 'TYPE_CAMP', 'LNPER', 'DENOMINATION', 'ZONE' and 'MONTH' are modeled with conditional inference decision trees, giving the following splits for each attribute (see figure 1).



(a) Conditional inference tree for attribute 'TYPE_CAMP'

(b) Conditional inference tree for attribute 'LNPER'

(c) Condition inference tree for attribute 'DENOMINATION'. No branches were made

(d) Conditional inference tree for attribute 'ZONE'

(e) Conditional inference tree for attribute 'MES'. No branches were made

Fig. 1 Conditional inference decision trees

Groups formed for each variable will be labeled from left to right, according to each tree and leaf. For example, in variable 'TYPE_CAMP', first group will be labeled as 'TYPE_CAMP_g1'. This new variable will be binary, and it will be 1 if TYPE_CAMP = 'DMA' or 'FPA' and 0 otherwise. This method will be used for all variables in the new model. Variables that don't have splits will be the same as the original variables.

Conditional inference tree-based logistic regression

The new set of dummy variables created from the previous step will be tested with a logistic regression model. Because all variables are dummified, and according to logistic regression theory, a reference baseline must be established. This baseline will be the combination of all categorical values that gives the least approval rate within each group in each variable. Results are shown in Table IV.

TABLE IV
Coefficients for selected attributes with decision tree based logistic regression

Coefficients	Estimate	Std.Error	Z value	P-value
(Intercept)	0.03389	0.63215	0.054	0.9573
TYPE_CAMP_g1***	0.51237	0.07172	7.144	9.07 e-13
TYPE_CAMP_g3***	0.53624	0.11629	4.611	4 e-06
DENOMINATION_g1S	0.84828	0.63017	1.346	0.1783
LNPER_g2***	0.56617	0.03890	14.556	< 2 e-16
ZONE_g1***	0.21786	0.03984	5.468	4.55 e-8
MONTH_g1*	-0.04125	0.01653	-2.496	0.0126

Akaike Information Criterion (AIC) = 19035
Baseline: TYPE_CAMP_g2, DENOMINATION_g1N, LNPER_g1, ZONE_g2

Comparison of AIC leads to this modelling to be more efficient than the simple logistic regression model.

Performance results and classifier efficiency

Learning ratio used is 80-20 for training and test subsets, selected randomly but replicating general approved/rejected proportion. In Table V, performance results are shown. Threshold for positive/negative classification is 0.5.

Table V
Performance results.

Conditional inference tree-based logistic regression	Value
Training-Test subset ratio	80/20
Training time, seconds	0.09
Evaluation time, seconds	0.07
Training dataset accuracy	81.65%
Test dataset accuracy	81.98%
Overall acurracy	81.72%

C. Neural network-based model

The neural network-based credit assessment evaluation consists of two phases: first, where data preprocessing phase where each numerical value of the applicant's attributes in the dataset is normalized separately using the min-max approach, and categorical variables, on the other hand, will be transformed into a dummy or coded variables; and second, evaluating the applicant's attributes and deciding whether to accept or refuse the application using the neural network algorithm. Once the neural network converges to a set minimum error value, learning is defined as "accomplished" so the testing process for model efficiency can be done.

Loan application data preprocessing for neural network

In this phase, data is filtered, prepared, and separately processed to be prepared for neural network modeling. First, any duplicated instances that mean two loan evaluations of the same amount in the same moment, one rejected and one approved, will be deleted taking preference of those instances that are not approved. Then, attributes that only have values when the application is rejected will be excluded in the model, as this will perform as perfect predictors and could generate important bias (variables 1, 21, 22, 23, and 24). Then, categorical variables are coded. Next, all numeric variables are separately normalized to values between 0 and 1 using the min-max normalization method, which means finding the maximum and minimum value for each attribute and rescaling each value of this attribute. After all the preprocessing, the dataset reduces its size to 25,342 cases and 17 mixed numeric and coded variables.

Neural network application

A supervised neural network that relies on the backpropagation learning algorithm was used, due to its implementation simplicity and its easy understanding. Figure 2 explains the general construction of a neural network model with ***n*** input layers, l hidden layers, and h nodes in each hidden layer, and how the backpropagation learning algorithm works.

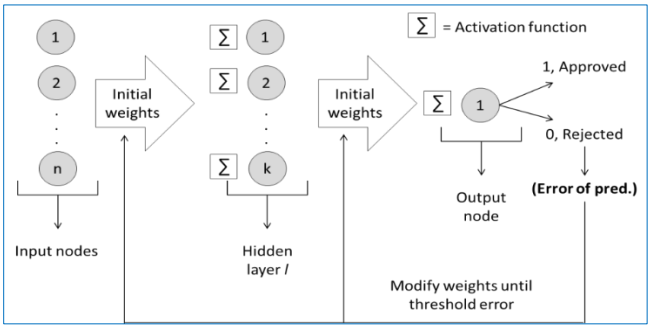


Fig. 2. General construction of an ANN. Initially random weights are selected for each hidden layer, and compounded variables are computed with the activation function to calculate input for hidden layer l, until final output node. Then these weights are adjusted with backpropagation algorithm, based on partial derivatives on prediction error to adjust initial weights.

This neural network input layer has 17 neurons, where each input neuron uses a normalized numerical value. It has only one hidden layer that contains 4 neurons or nodes. More layers of neurons per layer are not tested because of computational limitations.

For this type of prediction (binary), the activation function will be the sigmoid function, as this type of function can rescale outputs to 0-1 intervals. A threshold value of 0.5 is used to distinguish between predictive approved and rejected loan applications. If the output result from the output node is greater

than 0.5, the loan application is clasified as approved; otherwise, it is classified as rejected. All parameters defined for this neural network model are specified in Table VI.

Pseudocode

```
# Neural networks with 1 layer, 1,2,3 nodes and 2 layers,
            1,2,3 nodes
library(neuralnet)
#STEP 1: Standardize, categorical variables such as dummies
          and numerical variables. min-max normalization,
          we select some variables.
Base_NeuralNetwork = Base_BanK[ ,-(which(names(Base_Bank)
       %in% c("CAMP",
       "TYPESOL","AREA","REGION","YEAR","YEARMONTH","SEG_COM
       "))),with=F]
# numerical variables (i): MONTH, STATUS, APPROVED_AMOUNT,
AMOUNT_DOLLARS, AMOUNT_SOLES, RATE, PBCLIENT, SEGMENT
Base_NeuralNetwork$i = (Base_NeuralNetwork$i -
        min(Base_NeuralNetwork$i)) /
        (max(Base_NeuralNetwork$i)-
        min(Base_NeuralNetwork$i))
# categorical variables
for (j in
             which(sapply(Base_NeuralNetwork,class)=="chara
             cter")) {
  clases=unique(Base_NeuralNetwork[[j]])
  for (i in 1:length(clases)) {
   eval(parse(text=paste("Base_NeuralNetwork$",
             names(Base_NeuralNetwork)[j],"_g",i,"=as.numer
             ic(Base_NeuralNetwork$",names(Base_NeuralNetwo
             rk)[j],"=='",clases[i],"') ",sep="")))  }
}
# categorical variables (k): {TYPE, DENOMINATION,
   SECTOR_ECONOMIC, LNPER, ZONE, DESDEPART, S_CAN, E_CAN} -
   > NULL
Base_NeuralNetwork$k=NULL
# neural network model
fm<-
   as.formula(paste("STATUS~",paste(names(Base_NeuralNetwor
   k) [names(Base_NeuralNetwork)!=" STATUS"],collapse = " +
   "),sep= ""))
a<-neuralnet(fm,data =
   Base_NeuralNetwork[ind_muestra==0,],hidden = 1,
   linear.output = F,lifesign = "full")
output=compute(a,Base_NeuralNetwork[ind_muestra==0,-
   which(names(Base_ NeuralNetwork)==" STATUS"),with=F])
prediction_net <- ifelse(output$net.result>0.7,1,0)
table(prediction_net,Base_NeuralNetwork$STATUS[ind_muestra=
   =0])
```

TABLE VI
Neural network model parameters for training dataset

Neural network parameters	Value
Input layer nodes	17
Hidden layer nodes	4
Output layer nodes	1
Minimum required error	0.01
Obtained error	0.009
Maximum allowed iterations	100,000
Performed iterations	47,058

Performance results and classifier efficiency

As specified, the learning ratio used is 80-20 for training and test subsets, selected randomly but replicating general approved/rejected proportion. In Table VII, implementation results are shown, such as minimum obtained error, training time, evaluation time, the accuracy of the model in training and test dataset, and overall accuracy.

Pseudocode

```
# indicators
p_prueba <- predict(modelo,prueba, type="response")
pr_prueba <- prediction(p_prueba,
              Base_Bank$STATUS[ind_muestra==1])
auc <- performance(pr, measure = "auc");auc_yvalue <-
              auc@y.values[[1]]
GINI_prueba <- if(auc_yvalue>0.5) {(2* auc_yvalue-1)} else
   {(2*(1-auc_yvalue)-1)}
auc2_prueba <- performance(pr, "tpr", "fpr")
KS_prueba = max(attr(auc2,'y.values')[[1]]-
              attr(auc2,'x.values')[[1]])
```

Table VII
Performance results

Neural network performance results	Value
Training-Test subset ratio	80/20
Obtained final error	0.078
Training time, seconds	559.8
Evaluation time, seconds	0.06
Training dataset accuracy	82.67%
Test dataset accuracy	83.11%
Overall acurracy	82.77%

IV. Conclusions and Future Research

The use of logistic regression models combined with conditional inference trees can be useful to enhance the accuracy and performance of the logistic predictive model. Variables 'TYPE_CAM', 'LNPER', 'DENOMINATION', 'ZONE' and 'MONTH' are modeled with conditional inference decision trees, giving an overall accuracy of 81.72%. Results of implementing credit assessment with neural network models have an overall accuracy of 82.77% but required significant training time compared to the enhanced logistic model. As neural networks can provide more accuracy and performance, it requires more computational power that could be a limitation for similar models. However, as this type of information is very business-sensitive, additional data for this research was not reachable.

Based on overall results, both models have good statistical adjustment, so peruvian MSE could use this knowledge to be prepared to adjust some characteristics of its business model to improve their probability to apply and get the loan approved so that they can be more competitive and in consequence contribute to national economic growth and development.

Future research guidelines could assess (1) the statistical significance of interaction between these variables to know if results are sensitive to this type of interactions and (2) the use of survey data from MSE to determine the influence – if exists – of organizational structure in financial access to credit loans.

References

[1] Muñoz, J., Concha, M., Salazar, O.: Analizando el endeudamiento de las Micro y Pequeñas Empresas. Revista Moneda - Banco Central de Reserva del Perú, N°156, pp.19-24 (2013).

[2] Asociación de Emprendedores del Perú (ASEP).: MYPES aportan el 40% del PBI. https://asep.pe/index.php/mypes-aportan-el-40-del-pbi/, last accessed 2018/05/21.

[3] Organización Internacional del Trabajo (OIT).: Panorama Laboral Temático: Pequeñas Empresas, grandes brechas. Empleo y condiciones

de trabajo en las MYPE de América Latina y el Caribe. OIT: Lima, Perú (2015).
[4] Ministerio de la Producción (PRODUCE).: Decreto Supremo N° 013 – 2013: Texto Único Ordenado de la Ley de Impulso al Desarrollo Productivo y al Crecimiento Empresarial (2013).
[5] Superintendencia de Banca, Seguros y AFP.: Resolución N°11356 – 2008: Reglamento para la Evaluación y Clasificación del deudor y la exigencia de provisiones (2008).
[6] Diario Gestión.: Casi 900 mil Mypes tienen problemas con sus deudas. Publicado el 19 de junio (2013). https://gestion.pe/economia/empresas/900-mil-mypes-problemas-deudas-41210, last accessed 2018/05/21.
[7] Farnazeh, A., Fadlalla, A.: Data mining application in accounting: A review of the literature and organizing framework. International Journal of Accounting Information Systems 24, 35-58 (2017).
[8] Vadim, K. Overview of different approaches to solving problems of Data Mininn. Procedia Computer Science 123, 234-239 (2018)
[9] Abdou, H., Pointon, J., El-Masry, A. Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert System with Application 35, 1275-1292 (2008)
[10] Cho, V., Ngai, E. Data mining for selection of insurance sales agents, Expert systems 20(3), 123-132 (2003).
[11] Gschwind, M. Predicting late paymen: A study in tenant behavior using data mining techniques. Journal of Real State Portfolio Management 13(3), 269-288 (2007).
[12] Lin, S., Shiue, Y., Chen, S., Cheng, H.: Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks. Expert Systems with Applications (2009).
[13] Abellán, J., Castellano, J.: A comparative study on base classifiers in ensemble methods for credit scoring. Journal Expert Systems With Applications 73, 1–10 (2017) .
[14] Nemati, F., Sajedi, H., Khanbabaei, M.: A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. Journal of Retailing and Consumer Services, Volume 27, Pages 11-23 (2015).
[15] Pérez-Martín, A., Pérez-Torregrosa, A., Vaca, M.: Big Data techniques to measure credit banking risk in home equity loans. Journal of Business Research (2018).
[16] Martínez, J., Pérez, G.: Assessment of a credit scoring system for popular bank savings and credit. Contaduría y Administración, Volume 61, Pages 391-417, April - June (2016).
[17] Narindra, G., Badra, C., Rian, F.: Assessing Credit Risk: an Application of Data Mining in a Rural Bank. International Conference on Small and Medium Enterprises Development with a Theme "Innovation and sustainability in SME Development" - ICSMED, (2012).
[18] Luo C., Wu, D., Wu, D.: A deep learning approach for credit scoring using credit default swaps. Engineering Applications of Artificial Intelligence (2016).
[19] Batmaz, I., Danısoglu, S., Yazıcı, C., Kartal, E.: A data mining application to deposit pricing: main determinants and prediction models. Applied Soft Computing Journal (2017).
[20] Chen, W., Xiang, G., Liu, Y., Wang, K.: Credit risk Evaluation by hybrid data mining technique. Systems Engineering Procedia 3 194 – 200, (2012).
[21] Khashman, A.: Credit risk evaluation using neural networks: Emotional versus conventional models. Applied Soft Computing, Volume 11, Pages 5477-5484, December (2011).
[22] Huang, C., Chen, M., Wang, C.: Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications 33, 847 – 856 (2007).
[23] Harrell, F.: Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Second Edition. Springer Series in Statistics (2015).
[24] Ali, H., Kamel, S.: Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach. Review of Development Finance, Volume 4, Pages 20-28, January - March (2014).
[25] Zhao, Z., Xu, S., Ho Kang, B., Kabir, M., Liu, Y., Wasinger, R.: Investigation and improvement of multi-layer perception neural networks for credit scoring. Expert Systems with Applications, Volume 42, Issue 7, 1 May 2015, Pages 3508-3516, (2014).