





Article

Feature Selection to Optimize Credit Banking Risk Evaluation Decisions for the Example of Home Equity Loans

Agustin Pérez-Martín ^{1,*†} , Agustin Pérez-Torregrosa ^{1,†} , Alejandro Rabasa ^{2,†} 
and Marta Vaca ^{1,†} 

¹ Economic and Financial Studies Department, Miguel Hernández University of Elche, 03202 Elche, Spain; agustin.perez01@goumh.umh.es (A.P.-T.); mvaca@umh.es (M.V.)

² Operations Research Center, Miguel Hernández University of Elche, 03202 Elche, Spain; a.rabasa@umh.es

* Correspondence: agustin.perez@umh.es

† These authors contributed equally to this work.

Received: 21 July 2020; Accepted: 13 October 2020; Published: 6 November 2020



Abstract: Measuring credit risk is essential for financial institutions because there is a high risk level associated with incorrect credit decisions. The Basel II agreement recommended the use of advanced credit scoring methods in order to improve the efficiency of capital allocation. The latest Basel agreement (Basel III) states that the requirements for reserves based on risk have increased. Financial institutions currently have exhaustive datasets regarding their operations; this is a problem that can be addressed by applying a good feature selection method combined with big data techniques for data management. A comparative study of selection techniques is conducted in this work to find the selector that reduces the mean square error and requires the least execution time.

Keywords: credit scoring; feature selection; big data; data mining

1. Introduction

The granting of a loan implies a risk for the financial institution, which arises due to the possibility that the borrower does not comply with the repayment of the loan; that is, the possibility of delinquency. Financial institutions face a decision problem of whether to grant the loan to the client, which implies an assessment of the probability of each applicant presenting delinquency problems. Therefore, an institution attempts to calculate an uncertain event with data from the past. In other words, this is an estimation or prediction problem known as credit scoring. Any credit rating system that enables the automatic assessment of the risk associated with a banking operation is called credit scoring. This risk may depend on several customer and credit characteristics, such as solvency, type of credit, maturity, loan amounts and other features inherent in financial operations. It is an objective system for approving credit that does not depend on the analyst's discretion.

Today, this area of research represents a very important challenge for financial institutions for three fundamental reasons:

1. The credit risk represents 60% of the total risk of the financial institution.
2. The requirements by authorities to comply with Basel III, among other requirements, implies an increase in the entity's reserves based on a higher percentage of the calculation of the expected losses, with a consequent reduction of the benefits to be distributed to shareholders.
3. With the introduction of the Ninth International Financial Reporting Standard (IFRS 9) in January 2018, financial companies will have to calculate their expected losses due to default

during the 12 months after these financial instruments are implemented, transferring them as losses to their income statement, with a consequent reduction in profit.

Therefore, a good automatic method to select variables and optimize risk assessment decisions would mean lower expected losses and higher profits for the company and shareholders.

Financial institutions have exhaustive datasets for their operations; this allows them to gain an advantage (they have huge amounts of information related to creditors and debtors and can try to predict our behavior) and presents a problem (they need to obtain the best prediction without exceeding resources). Yu et al. [1] indicated that the credit industry requires quick decisions, and, in this sense, there should be a trade-off between computational performance and computational efficiency. Today, despite the fact that financial institutions have unlimited computing resources and computing power (Amazon Web Services (AWS), Azure, etc.), organizational reality sets limitations.

The processing of a model runs in parallel to many core business processes that are performed in batches during the night. Therefore, the data dump in the system is continuous at all levels and departments of the financial institution. This, together with the need for financial institutions to apply dynamic credit risk optimization models together with the selection of variables, leads us to consider this research topic. The models must allow the introduction of new variables at any moment in time, thus reflecting the economic, financial, social and political reality in order to anticipate both new predictions and a change in the importance of a variable. Therefore, the adjustments to which the model is subjected, together with the rest of the usual computer processes of the financial institution, will determine the need for a computationally efficient model.

According to Pérez-Martin et al. [2], computational efficiency decreases as the number of variables increases, while the number of records (potential borrowers) hardly influences this. Therefore, it is in these settings that a massive data problem appears, since the system must recalculate its credit scoring together with a new optimization of variables as its scenario has changed. This leads us to consider this research topic; i.e., to obtain a method for selecting variables and optimize risk assessment decisions in order to obtain results in a minimum time. The contribution made by this research is important in economic terms for financial institutions and at the same time for society. It is possible to give quick responses to risk managers and to be able to quickly correct a credit risk model to avoid a possible increase in expected losses, with a consequent loss of profits and increase in reserves for the financial institution, which represents a lower profit for distribution among shareholders.

Regarding the daily business of banks, they need to evaluate many loan applications and also be clear regarding the applications which must be rejected and accepted; in the past, for this task, employees engaged in personal study. Simon [3] concluded that this type of classical methodology has some problems, such as the non-reproducibility of the valuations (each analyst may have a different opinion) or a low level of confidence due to the subjectivity of the valuation, which can change every day. This method, as observed, has a significant degree of subjectivity, which together with its low efficiency and the difficulty of training new personnel makes it an inefficient method.

The general approach to this task is currently a model that is updated each night in a batch together with other processes that are executed by a bank's systems. The problem with a credit risk model batch is clear: each morning, the new model is deployed and ready to use for each user and must be the best model from the information available. Furthermore, banking institutions have limited computational capacity. The combination of these factors is a problem that a good feature selection method combined with big data techniques can solve. In fact, there are many research works in the literature that have sought the optimal method since Durand [4] introduced discriminant analysis. These works have employed approaches with different methodological methods, as listed below.

- **Linear Methods:** The authors of [5–9] considered this type of methodology to solve this problem, but the majority of authors did not obtain good results. For example, in the comparative study carried out by Yu et al. [5] in which the efficacy of 10 methods using three databases was compared to test the different methods: one referring to loans from an English financial company extracted

(England Credit) from [10], another referring to credit cards from a German bank (German Credit) and finally another database referring to credit cards in the Japanese market (Japanese Credit). The results show that the method ranks among the last classified, only surpassing neural networks with backwards propagation. However, subsequent research works (e.g., [11]) have shown the best percentages for this approach compared to other more complex methods, but the existing limitations due to the size of the data have been evident; in contrast, Xiao et al. [12] considered Logistic Regressions to be the best method if the information can be managed.

- Decision Trees: The authors of [9,13–16] used this methods in research comparing different methodologies. In none of these works, however, was this method the best of the methods used. For example, the results obtained in [9] can be observed, in which this approach lags far behind the rest of the methods in predictive precision. A predictive improvement can be observed in [17] where 11 methods are compared using the databases of German Credit and Australian Credit.
- Neural networks are commonly used to solve this problem in the literature, and successful results have been found, such as in the work of Malhotra and Malhotra [18] comparing the results with a discriminant analysis and obtaining good performance with the German Credit database. However, in general, there have been many research works that use neural networks without good or with insignificant results compared with the complexity of the models (e.g., [19–21]).
- Support Vector Machine obtains good results with classifications, as can be seen in [22], where the best accuracy results are obtained in comparison with four different methods. However, as Pérez-Martin and Vaca [23] concluded with a Monte Carlo simulation study, these best accuracy results have an impact on the efficiency (i.e., reducing it). Other authors have attempted to create or use another parametrization of the kernel, such as Xiao et al. [12], Bellotti and Crook [24] or Wang et al. [25], but the same problem remains regarding efficiency.

One important aspect is the method used when credit scoring must be performed, but it is possible to solve this problem by reducing the information used to model the behavior. To do this, it is necessary to select the variables that have the capacity to explain the best model. Managing large amounts of data poses a problem in terms of the server's efficiency and slows down its effectiveness. The number of considered features is called the data dimension, and, although high dimensionality in the feature space has advantages, it also has some serious shortcomings. In fact, as the number of features increases, more computation is required and the model accuracy and scoring interpretation efficiency are reduced [26,27]. Financial institutions need to provide a response to their potential borrowers, and so prediction models must meet two requirements: computational efficiency and predictive efficacy. We study this situation as a big data problem, and we consider a variable selection method for the model in order to reduce the volume of data, maintaining the model's efficiency with the aim of increasing its computational efficiency (i.e., achieving a lower execution time).

To reduce the computational time and maintain the effectiveness of the methods, one solution can be found through analytical methods of data mining ([28]). Data mining attempts to quantify the risk associated with a loan and facilitate credit decision-making, while allowing for and limiting the potential losses that a bank may suffer.

Therefore, variable selection is performed, and it is determined whether the prediction improves and if this improvement corresponds to a greater computational efficiency. It is expected that faster forecasts will be achieved, and information that does not result in any gain to the model must be eliminated. This involves applying a data mining algorithm that generates and reduces classification rule systems and achieves accurate predictions, but with a substantially lower temporal and memory consumption cost. The automatic selection of characteristics involves a set of analytical techniques that generate the combinations of variables that have a greater incidence of the behavior of a target variable. The selection of different variables can be carried out by applying different methodologies such as decision trees, rule systems and techniques such as principal components analysis with an incremental approach in which the algorithms employed seek to improve a certain metric of significance [29].

The importance of selecting variables will be determined if the contribution of the attributes is significant for the model. There are different approaches to this problem. One of the possible approaches is to use wrapper methods, and in this popular case, one methodology for selecting attributes to have been emphasized in the literature is genetic programming ([1,30,31]), or other models such as Linear Discriminant Analysis (LDA) ([14]) or different classes of artificial neural network ([30]), among other methodologies. However, if we consider the Basel principles regarding the model, all of these possible models breach these principles on the basis of their simplicity.

In this research, the impacts of two proposed feature selection techniques are measured—gain ratio and principal component analysis (PCA)—in terms of reducing dimensionality along with the impact of discretization in order to standardize the dataset values. To test these factors, we use a credit dataset from the Spanish Institute of Fiscal Studies (IEF) dataset. Later, we apply three credit scoring methods to predict when a home equity loan can be granted or not: the generalized linear model for binary data (GLMlogit), the linear mixed model (LMM) and the classification rules extraction algorithm-reduction based on significance (CREA-RBS). LMM and GLMlogit methods were chosen because they are the most efficient and effective methods ([2,32]). Section 2 gives an overview of the datasets used and the stages of our experimental study with a description of the methods used. In Section 3, the obtained results are shown. In Section 4, the conclusions are presented.

2. Materials and Methods

In this section, the stages of this study and all the progress are described, including the performance metrics used. We subsequently outline the experimental process used to assess the performances of the proposed models.

In summary, this study has four stages, as can be seen in Figure 1:

1. Construction of training and testing datasets—the details of the procedure are in Section 2.1.
2. Execution of the process of attribute selection and discretization—the entire process is explained in Section 2.2.
3. Training of classification models—three algorithms are selected for this study (GLMlogit, LMM and CREA-RBS), and the methods are explained in Section 2.3.
4. Collection of results—at the end of the process, two indicators are obtained to evaluate the effect of the feature selection methods on credit scoring models.

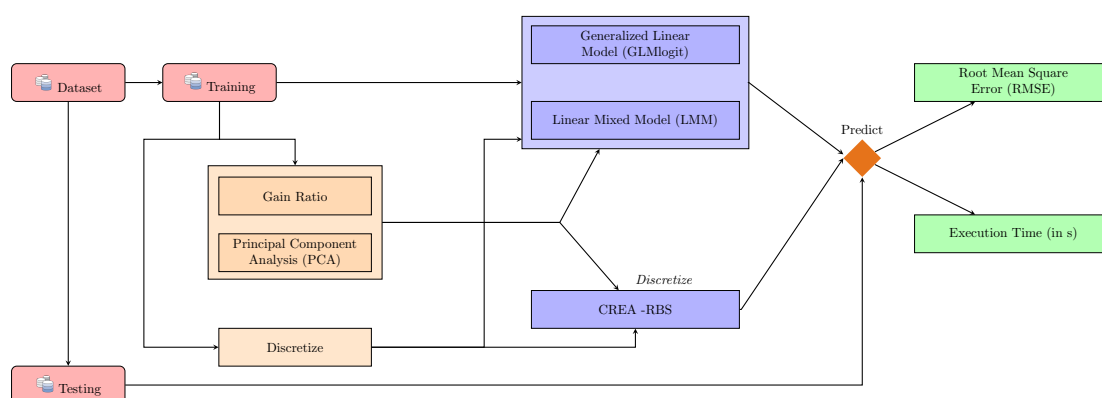


Figure 1. Summary of study. CREA-RBS, classification rules extraction algorithm-reduction based on significance.

All procedures were developed with a dedicated Intel Xeon E5-2420 server with the Linux Debian Jessie operating system (64 bits, 12 CPUs at 1.9 GHz and 32 GB Ddr3 RAM) and implemented in R software ([33]).

2.1. Datasets

To prove the efficacy of our four-stage procedure, two different datasets are used:

1. The German Credit dataset: The reason for choosing this database is because it is one of the most used in the existing literature on credit scoring research ([34], UCI Machine learning Repository (<http://archive.ics.uci.edu/ml/datasets.html>)).
2. The IEF semi-synthetic dataset: The IEF semi-synthetic dataset comes from the IEF dataset from the Spanish Institute of Fiscal Studies (IEF) and is obtained from its statistical repository ([32,35–39]). This subset was generated and transformed in order to achieve four necessary conditions: to homogenize diverse variables from different years, to remove missing data, to choose or transform the interest variables for our research and to reduce the size for computational reasons.

The variables used in the research were chosen based on the different articles and works consulted regarding the variables that financial institutions take into account in the decision to grant a mortgage loan ([7,18,40–51]). It is observed that there is unanimity in the choice of the explanatory variables and, at the same time, they are relevant in the time window of the study. The factors to take into account for the selection of the explanatory variables are the following [32]:

- (a) Social and personal factors: Variables related to the client's social and personal environment are selected, such as sex, age, marital status, number of dependents, population of the habitual residence, habitual residence and other dwellings.
- (b) Economic factors :Variables related to the availability of income are selected, such as salaries, the economic sector in which they carry out their activity and the size of the loan requested.

2.1.1. German Credit Dataset

The German Credit Dataset has 1000 instances: it has 700 instances corresponding to creditworthy applicants and 300 other instances corresponding to applicants to which credit should not be extended. The number of initial variables is 20: three of them are numeric variables, and there are 17 categorical variables ([34]). The features may be classified as follows:

- social attributes, such as marital status, age, sex; and
- economic attributes, such as business or employment and being a property owner.

A 70% random sample was extracted for training purposes, and the rest was used for testing the model. The distribution of observations, payers and defaulters is presented in Table 1.

Table 1. Distribution of observations in the German Credit dataset.

Class	Training	Testing	Total Cases
Payers	485	215	700
Defaulters	215	85	300
Total	700	300	1000

It should be noted that the German Credit database, although it comes from a financial institution, does not correspond to the number of real observations that exist in the databases of any financial institution (it is a very small database). Consequently, the conclusions that can be derived are approximate and limited. In any case, it is verified that the conclusions obtained in the investigation are consistent with the results in the two databases.

2.1.2. IEF Semi-Synthetic Dataset

The IEF semi-synthetic dataset comes from the IEF dataset from the Spanish Institute of Fiscal Studies (IEF) and is obtained from its statistical repository ([32,35–39]). This subset was generated and transformed to achieve four necessary conditions: to homogenize diverse variables from different years, remove missing data, choose or transform the interest variables for our research and reduce the size for computational reasons. The number of dataset entries is 5,101,260 observations; regarding the number of initial variables, nine of them are numeric and the rest are categorical. The features used may be classified as follows:

- social attributes, such as marital status, age, province and number of family members;
- economic attributes, such as business or employment, family income, properties and amount borrowed;
- statistical attributes based on adjustments such as corrected income—that is, the total income dichotomized by the minimum inter-professional salary (SMI) for further data processing; and
- synthetic variables, such as the response variable y_i , which was simulated using a linear regression model with several attributes of the dataset and a normal perturbation vector \mathbf{e} , with $e \sim N_n(0, 0.15)$. Then, the response variable was recategorized as binary (pay and default cases), as follows:

$$p_i = e^{(y_i)} / (1 + e^{(y_i)}) * 100$$

This expression can be considered as a probability of default by dividing into 100.

The original dataset was divided into training and testing sets as a function of the time attribute (Year), considering the last temporary period in those observations as the testing set and the rest as the training set. The distribution of observations, payers and defaulters is presented in Table 2.

Table 2. Distribution of observations in the Spanish Institute of Fiscal Studies (IEF) dataset.

Class	Training	Testing	Total Cases
Payers	2,131,601	811,375	2,942,976
Defaulters	1,461,779	696,505	2,158,284
Total	3,593,380	1,507,880	5,101,260

2.2. Methods and Algorithms

Two techniques are chosen to reduce the dimension, as explained below.

2.2.1. Gain Ratio

This technique is a version of the well-known information gain approach. It is a measure that gives the relationship between explanatory variables and the response variable. It measures how much information is communicated in comparison with the total:

$$\frac{H(Class) + H(Attribute) - H(Class, Attribute)}{H(Attribute)}$$

where H represents the entropy and uncertainty of data. Shannon [52] defined the entropy (H) of one discrete variable (X) with a probability function $P(X)$

$$H(X) = E[I(X)] = E[-\ln(P(X))],$$

where E is the expected value of an operator and I is the information function.

2.2.2. Principal Component Analysis (PCA)

PCA [53] is a multivariate statistical technique used to reduce the number of features with redundant information in a dataset into a smaller number of uncorrelated variables, called dimensions. These dimensions are a linear weighted combination of original features. For example, in mathematical terms, Dimension 1 from a set of n variables (X_1, \dots, X_n) is

$$Dim_1 = w_{11} * X_1 + w_{12} * X_2 + w_{13} * X_3 + \dots + w_{1n} * X_n,$$

where $w_{11}, w_{12}, w_{13}, \dots, w_{1n}$ are the weights for the initial variable to obtain Dimension 1.

2.2.3. Discretization Process

Datasets usually come in mixed formats: numeric and discrete. Discrete values are intervals in a continuous spectrum of values; while the number of numeric values for an attribute can be very big, the number of discrete values is often small or finite. The two types of values make a difference in learning classification models. Thus, we propose a discretization for the datasets in order to compare the effect [54,55]. For this experiment, two datasets are used. We consider the discretized criteria which appear in Tables 3 and 4 for the IEF dataset [32] and German Credit dataset, respectively.

Table 3. Discretized criteria for the IEF semi-synthetic dataset.

Original Feature	Discretized Criteria	Category Name
Percentage of ownership of the residence (ptvh)	Equal to 0%	ptvh-0
	Equal to 50%	ptvh-50
	Equal to 100%	ptvh-100
	Between 0% and 50%	ptvh- <50
	Between 50% and 100%	ptvh- >50
Percentage of ownership of the residence of spouse (pctvh)	Equal to 0%	pctvh-0
	Equal to 50%	pctvh-50
	Equal to 100%	pctvh-100
	Between 0% and 50%	pctvh- <50
	Between 50% and 100%	pctvh- >50
Average ownership percentage (Pmedowner)	Equal to 0%	PMO-0
	Equal to 50%	PMO-50
	Equal to 100%	PMO-100
	Between 0% and 50%	PMO- <50
	Between 50% and 100%	PMO- >50
Property tax deduction (PTD)	Does not have	PTD-0
	Has deduction	PTD-Dist0
Property income (PI)	Does not have	PI-0
	Has property income	PI-Dist0
Family Income (FI)	Negative	FI- <0
	Equal to 0 €	FI-0
	Between 0 and 3701 € (First Quartile)	FI-1st Quartile
	Between 3701 € and 10,730 €	FI- <Media
Amount (A)	>10,730 €	FI- + Media
	First Quartile	A-1stQ
	Second Quartile	A-2ndQ
	Third Quartile	A-3thQ
	Fourth Quartile	A-4thQ

Table 4. Discretized criteria for the German Credit dataset.

Original Feature	Discretized Criteria	Category Name
credit_amount	First Quartile	A-1stQ
	Second Quartile	A-2ndQ
	Third Quartile	A-3thQ
	Fourth Quartile	A-4thQ
duration_in_month	Under 12	duration-1
	Between 12 and 18	duration-2
	Between 18 and 24	duration-3
	Over 24	duration-4
age	Under 28	Age- <28
	Between 28 and 32	Age-28–32
	Between 32 and 36	Age-32–36
	Between 36 and 41	Age-36–41
	Between 41 and 49	Age-41–49
	Over 50	Age-50+

2.3. Credit Scoring Methods

The main goal of this research is to find the best selector that has the maximum efficiency and effectiveness possible in the scoring procedure of a home equity loan. For this purpose, we use the LMM and GLMlogit scoring methods because they are the most efficient and effective methods [2,32].

2.3.1. Generalized Linear Model (GLMlogit)

GLMlogit is an extension of the linear model that allows for other relations between response and explanatory variables. In this case, a binomial distribution with a canonical logit link was chosen. In general, this can be formulated as

$$E(Y) = \mu = g^{-1}(X\beta),$$

where $E(Y)$ is the expected value of Y , $X\beta$ is the linear predictor and g is the link function, which can be formulated as

$$g(p) = \ln\left(\frac{p}{1-p}\right)$$

2.3.2. Linear Mixed Model (LMM)

LMM is a linear model with mixed effects (fixed effect and random effect) that was developed by Fisher [56]. In matrix notation, a mixed model can be represented as

$$y = X\beta + Zu + e,$$

where y is a response variable vector, β is a vector of fixed effects, u is a vector of random effects and e is a vector of random errors, with $u \succ N(\bar{\mu}_u^2)$ and $e \succ N(\bar{\mu}_e^2)$. We consider regions as a random variable in both datasets.

2.3.3. Reduction Based on Significance (RBS)

The RBS algorithm [57] is an ensemble rule-based system based on the ID3 method which allows the rule set to be reduced and organized by the support and confidence of the facts.

This algorithm reduces the rules into a given rule set when each of them is associated with a rule significance value and is allocated into a specific significance value called a region. To determine each region [58], two metrics are used—rule support values and rule frequency values—to assign each rule its rule significance and its significance region (REG), respectively.

For a given rule $r_k^{\overline{AC}}$, the rule support or antecedent frequency $r_k^{\overline{AC}} \times fr_{\overline{A}}$ is the proportion of tuples in the data set containing an antecedent, such that

$$r_k^{\overline{AC}} \times fr_{\overline{A}} = \frac{|\overline{A}|}{N}$$

where $fr_{\overline{A}}$ is only based on each instance’s left-hand side (antecedent). All rules with the same left-hand side will have the same $fr_{\overline{A}}$ value. On the other hand, the rule confidence or rule frequency is defined as the proportion of tuples with the antecedent \overline{A} in D that also contain \overline{C} as a consequent, such that

$$r_k^{\overline{AC}} \times fr_{\overline{A}} = \frac{|\overline{A} \rightarrow \overline{C}|}{|\overline{A}|},$$

where $|\overline{A} \rightarrow \overline{C}|$ is the number of rules with the antecedent \overline{A} and consequent \overline{C} . The number of rules with the antecedent \overline{A} is labeled as $|\overline{A}|$.

Both the support and the confidence, as defined above, are used by the RBS algorithm to perform the reduction and classification of the rule system.

The work presented by Rabasa Dolado [58] contains the whole formalization of the RBS methodology (see Figure 2). Each rule in the rule system belongs to only one of the four defined exclusive regions:

- Region 2: This group of regions contains rules with high support and confidence regarding facts. Rabasa Dolado [58] called these direct rules, because these combinations are very reliable.
- Region 1: The region groups rules with high support and low-confidence rules.
- Region 3: These rules have a very low support. It should be understood that any conclusion can be reached.
- Region 0: The rules inside this region are discarded because this region is composed of rules with medium support and confidence of facts, and they may arise from randomness.

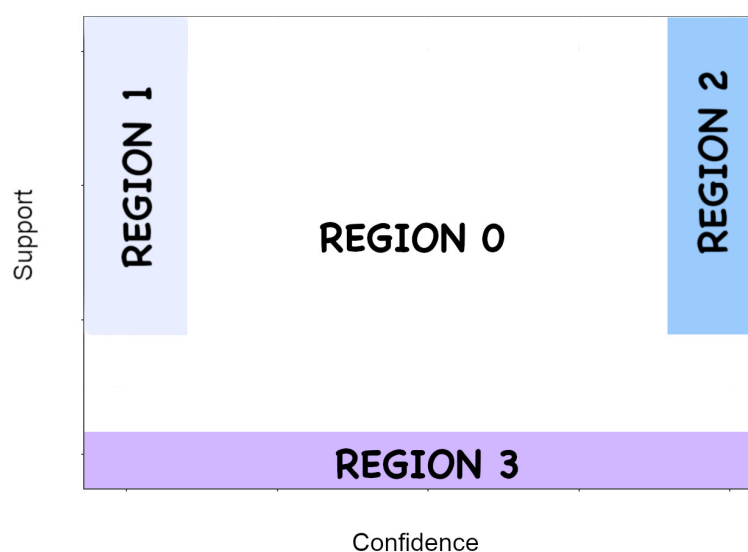


Figure 2. Structure of reduction based on significance (RBS) regions.

2.4. Model Evaluation

The model performances are compared using the Root Mean Square Error (RMSE) of fit to measure the accuracy of the models, along with the computer execution time for efficient measurement. RMSE is a good measure of effectiveness because it does not depend on a prediction threshold, instead relying on some ratios of the confusion matrix. RMSE assesses the quality of an estimator, not the result, and it could be defined as

$$RMSE = \sum_{i=1}^{n_i} \sum_{j=1}^I (\hat{p}_{ij} - p_{ij})^2 / n_{ij},$$

where \hat{p}_{ij} is the vector of predictions and p_{ij} is the observed vector values corresponding to the response variable of the model.

The computer execution time may be important in the decision to apply one method or another, due to the massive volume of the dataset. A computationally efficient method can be very competitive given the advantages in terms of the time expected for resolving requests. In our case, this time is measured between the start and end of the model fitting.

3. Results

In this work, we attempted to determine the effect of different processes in order to reduce the complexity of this problem. We calculated the gain ratio and PCA and performed discretization in order to rank the variables, create dimensions and create aggregated groups, respectively.

In Figures 3 and 4, we can see the most influential variable in the IEF semi-synthetic and German Credit datasets, respectively. As we can see in Figure 3, the most influential variable in the IEF dataset is Correctorincome. This is a feature which aims to discriminate between people with higher income or lower than the minimum inter-professional salary (SMI); the lower is the income compared to SMI, the more complicated it is to grant the payment of a bank credit. The other two most relevant variables are Familyincome and total tax deduction related to main residence (buying or improvements). The rest of the features have marginal importance. Moreover, economic attributes are more relevant than social variables according to the gain ratio.

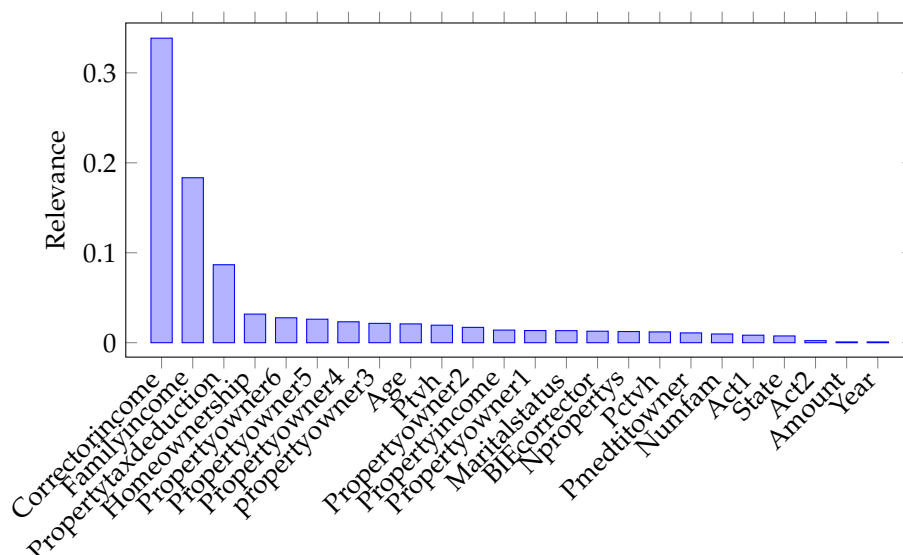


Figure 3. Results of the gain ratio for the IEF dataset.

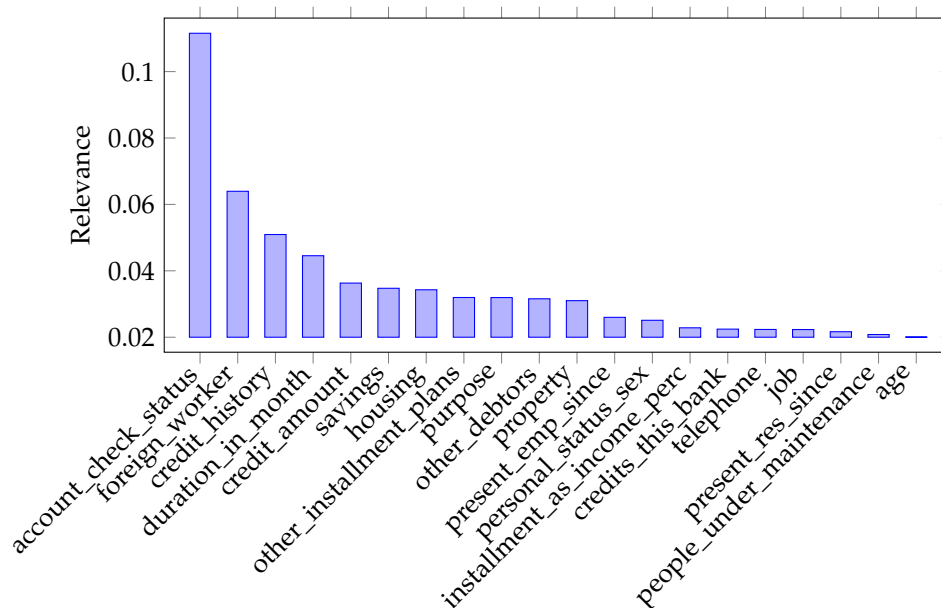


Figure 4. Results of the gain ratio for the German Credit dataset.

In the case of German Credit dataset (Figure 4), the most influential variable is `account_check_status`. This feature aims to determine the current amount of the account. The other four most relevant variables are `foreign_worker`, `credit_history`, `duration` and `credit_amount`. The rest of the features have marginal importance. Again, economic attributes are more relevant than social variables.

PCA is a technique used for feature extraction; thus, the output is a new dataset with new variables that are a combination of the input variables. It is very difficult to explain the meaning of the new variables. In the case of the IEF semi-synthetic dataset, the first 10 dimensions of PCA explain 25.34% of the total variability. This percentage is relatively low, but the next variables grow in very small increments of less than 5% of the previous explained variability.

In the German Credit case, the first dimension of PCA expresses 99.997% of the total dataset inertia. It is possible that this situation was created by the small size of the dataset. We decided to selected the five first dimensions because there is a small value of variance, but the possible interpretation of the tested models would not be logical except for the GLMlogit models.

Then, the dataset for the three proposed methods is adjusted. The execution time and the root mean square error (RMSE) are obtained by entering the features according to their contribution to the model. The results of these experiments are collected in Table 5 for RMSE and in Table 6 for execution times.

Table 5. Root mean square error (RMSE).

Dataset	GLM	LMM	RBS	
IEF Dataset	Raw	0.3635	0.3660	-
	Discretize	0.3527	0.3572	0
	Gain Ratio	0.3680	0.3960	0.3184
	PCA	0.3966	0.4181	0.1809
German Credit	Raw	0.3955	0.4037	-
	Discretize	0.4048	0.4120	0
	Gain Ratio	0.4045	0.4082	0.4303
	PCA	0.4275	0.4414	0.4082

Table 6. Execution times (in s).

Dataset		GLM	LMM	RBS
IEF Dataset	Raw	533.18	347.77	-
	Discretize	693.17	521.96	15.02
	Gain Ratio	17.687	32.092	0.4980
	PCA	23.184	53.952	0.928
German Credit	Raw	0.042	0.179	-
	Discretize	0.042	0.107	0.085
	Gain Ratio	0.015	0.095	0.041
	PCA	0.011	0.089	0.039

As we can see from the results in Table 5, the version of the dataset that performs worst is the selected dimensions of PCA, except for the RBS, which we analyze below. Moreover, the best version depends on the dataset used, although it is clear that it is a completed version. The difference between the two datasets is the individual’s total variability. In the case of the IEF dataset, there is a large number of cases and the discretization process reduces this number. In contrast, the German Credit Dataset contains a limited number of cases and the discretization process does not add value but homogenizes users who should not be equal but statistically cannot constitute an individual class.

RBS is the best method in most cases (it only shows worse results using gain ratio selection for the German Credit dataset, as seen in Figure 5). However, this model is not dynamic, because it is not able to adapt and predict new combinations of variables. For this reason, we must consider that this model does not assume that groups of observations may contain different combinations than the initial ones. RBS is a method that does not explain all combinations because it is focused on those of greater value. We can see the execution time of training models in Table 6, but they do not offer us relevant information on their own, beyond the fact that with a greater number of variables, there is more computational effort. In fact, if we consider those cases not collected by the model, whether positive or negative, we find that they perform worse except in two particular cases (gain ratio and PCA for German Credit considering accepting credit).

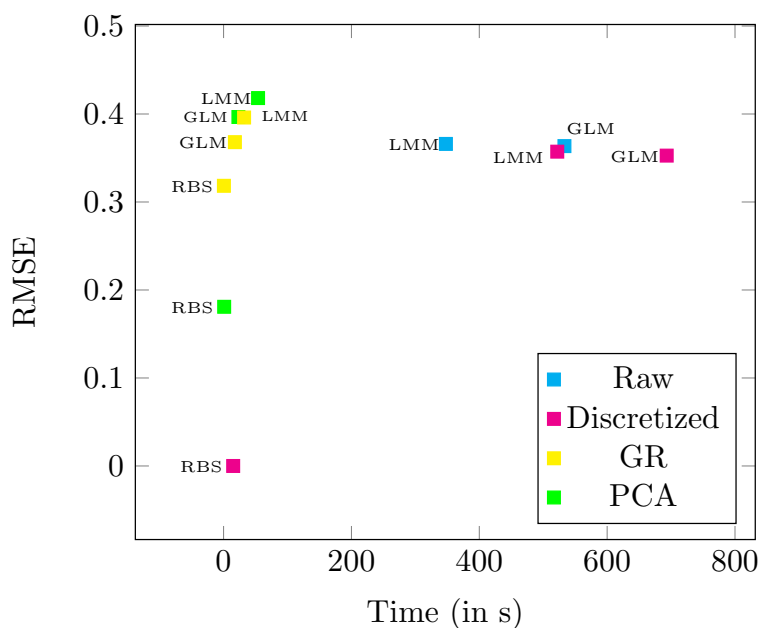


Figure 5. Relation between RMSE and time.

The mixture of the model precision (RMSE) with computational cost (time) is particularly interesting and is illustrated in Figure 5. The combination of a low time requirement and low value of RMSE shows the best methods. Three groups can be identified based on the results:

- The first group is formed by a complete dataset trained with a linear method. This group has the highest execution times and the second lowest RMSE. The best relationship within this group is LMM, because it saves more than 30%, losing less than 2% compared to GLMLogit.
- The second group is formed by a selected dataset with two proposed methods. This group has the second-lowest execution times and the worst RMSE. In general, GLMlogit behaves well; it obtains a good RMSE and shorter execution times. Within the methodologies used, the gain ratio is more effective than the PCA in view of the results.
- The last group is formed by RBS models regardless of whether there is a selection of variables or not. This stands out due to its low execution times, whatever the volume of data, and the best precision results. In this case, it is observed that the method of removing information makes the model lose precision.

4. Conclusions

We attempted to find efficient methods, discarding those with the lowest efficiency, for use with a banking dataset. For this reason, we used our own design with the IEF dataset, using some synthetic variables such as the target variable and borrowed amount.

Two feature selection methods were proposed using the elbow method, gain ratio and PCA. We calculated the measures of effectiveness (RMSE) and efficiency for the models adjusted with LMM, GLMlogit and RBS.

It is important to note that the method proposed in this variable selection investigation in its first step calculates the credit score with all the variables entered in the model. Once calculated, a selection is made using the most efficient proposed methods that offer more relevant information for making credit decisions. These variables are those that will be taken into account when deciding whether or not to grant a loan to a potential borrower, without having to enter the rest of the information, since it is irrelevant. As it is a dynamic model—i.e., new risk factors are added that are not constant over time and depend on economic, financial, political and even health changes (for example, COVID-19)—there are changes both in the calculation of the credit scoring and in the selection of variables. Therefore, it is important to interpret the results obtained in terms of variables that are part of the model; moreover, there is the possibility that new variables will appear and that they will have a greater weight, as well as that others will disappear or influence the decision less.

In the temporal analysis proposed in this research, the most important features with respect to the gain ratio in the IEF dataset used to predicting default probability were “Correctorincome” and “Familyincome”. These variables have a 51% weight, with respect to the total features. In the case of German Credit, the most important variables are “account_check_status” and “foreign_worker”, but these variables have only a 13.55% weight, with respect to the total features. When applying PCA, it is observed that, up to the 10th dimension, the inertia grows slowly in the IEF dataset. In the case of German Credit, it is observed that, up to the third dimension, practically all the inertia is represented.

RBS is the most effective method regarding accuracy, except that it is not an expert system and does not create responses for all cases. This situation may mean that, in stable environments, it can be used to estimate provisions for credit losses with customers where extensive knowledge is available for a financial institution or to provide a first filter for access to credit.

On the other hand, LMM has proven to be very efficient, practically equaling GLMlogit, and so this method could be a candidate to replace the logistic regressions (GLMlogit) used in the industry, since it remains a linear method, allows a simple interpretation but allows—through the random effect—the simplification of the adjustment of variables that are not particularly important but that should be introduced in the model.

For these reasons, RBS and LMM are proposed to evaluate credit risk in the presence of big data problems that need prior feature selection to reduce the dimensionality of a dataset. This method is applicable to any database and time horizon, because it selects the most important variables within the analyzed database while taking into account all of the information.

Author Contributions: Conceptualization, A.P.-M. and M.V.; methodology, A.P.-M.; software, A.R.; validation, A.P.-M, M.V. and A.R.; writing—original draft preparation, A.P.-T.; writing—review and editing, A.P.-M. and A.P.-T.; supervision, A.P.-M.; project administration, M.V.; funding acquisition, M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital, Generalitat Valenciana grant number GVA/2019/046.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Yu, L.; Yao, X.; Wang, S.; Lai, K. Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Syst. Appl.* **2011**, *38*, 15392–15399. [[CrossRef](#)]
2. Pérez-Martin, A.; Pérez-Torregrosa, A.; Vaca, M. Big data techniques to measure credit banking risk in home equity loans. *J. Bus. Res.* **2018**, *89*, 448–454. [[CrossRef](#)]
3. Simon, H. *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*; Wiley: Hoboken, NJ, USA, 1957.
4. Durand, D. *Risk Elements in Consumer Instalment Financing*; National Bureau of Economic Research: Cambridge, MA, USA, 1941.
5. Yu, L.; Wang, S.; Lai, K.K. An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *Eur. J. Oper. Res.* **2009**, *195*, 942–959. [[CrossRef](#)]
6. Loterman, G.; Brown, I.; Martens, D.; Mues, C.; Baesens, B. Benchmarking regression algorithms for loss given default modeling. *Int. J. Forecast.* **2012**, *28*, 161–170. [[CrossRef](#)]
7. Yu, L. Credit Risk Evaluation with a Least Squares Fuzzy Support Vector Machines Classifier. *Discret. Dyn. Nat. Soc.* **2014**, *2014*, 564213. [[CrossRef](#)]
8. Baesens, B.; Van Gestel, T.; Viaene, S.; Stepanova, M.; Suykens, J.; Vanthienen, J. Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc.* **2003**, *54*, 627–635. [[CrossRef](#)]
9. Sinha, A.P.; May, J.H. Evaluating and Tuning Predictive Data Mining Models Using Receiver Operating Characteristic Curves. *J. Manag. Inf. Syst.* **2004**, *21*, 249–280. [[CrossRef](#)]
10. Thomas, L.; Edelman, D.; Crook, J. *Credit Scoring and Its Applications*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2002. [[CrossRef](#)]
11. Alaraj, M.; Abbod, M.; Al-Hnaity, B. Evaluation of Consumer Credit in Jordanian Banks: A Credit Scoring Approach. In Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation (UKSIM '15), Cambridge, MA, USA, 25–27 March 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 125–130. [[CrossRef](#)]
12. Xiao, W.; Zhao, Q.; Fei, Q. A comparative study of data mining methods in consumer loans credit scoring management. *J. Syst. Sci. Syst. Eng.* **2006**, *15*, 419–435. [[CrossRef](#)]
13. Ong, C.S.; Huang, J.J.; Tzeng, G.H. Building credit scoring models using genetic programming. *Expert Syst. Appl.* **2005**, *29*, 41–47. [[CrossRef](#)]
14. Chen, W.; Ma, C.; Ma, L. Mining the customer credit using hybrid support vector machine technique. *Expert Syst. Appl.* **2009**, *36*, 7611–7616. [[CrossRef](#)]
15. Zhou, L.; Lai, K.K.; Yu, L. Least squares support vector machines ensemble models for credit scoring. *Expert Syst. Appl.* **2010**, *37*, 127–133. [[CrossRef](#)]
16. Tsai, C.F. Combining cluster analysis with classifier ensembles to predict financial distress. *Inf. Fusion* **2014**, *16*, 46–58. [[CrossRef](#)]
17. Li, J.; Wei, L.; Li, G.; Xu, W. An evolution strategy-based multiple kernels multi-criteria programming approach: The case of credit decision making. *Decis. Support Syst.* **2011**, *51*, 292–298. [[CrossRef](#)]

18. Malhotra, R.; Malhotra, D. Evaluating consumer loans using neural networks. *Omega* **2003**, *31*, 83–96. [[CrossRef](#)]
19. Twala, B. Multiple classifier application to credit risk assessment. *Expert Syst. Appl.* **2010**, *37*, 3326–3336. [[CrossRef](#)]
20. Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *Eur. J. Oper. Res.* **2011**, *210*, 368–378. [[CrossRef](#)]
21. Yao, P.; Lu, Y. Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Syst. Appl.* **2011**, *38*, 11300–11304. [[CrossRef](#)]
22. Purohit, S.; Kulkarni, A. Credit evaluation model of loan proposals for Indian Banks. In Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, 11–14 December 2011; pp. 868–873. [[CrossRef](#)]
23. Pérez-Martin, A.; Vaca, M. Compare Techniques In Large Datasets To Measure Credit Banking Risk In Home Equity Loans. *Int. J. Comput. Methods Exp. Meas.* **2017**, *5*, 771–779. [[CrossRef](#)]
24. Bellotti, T.; Crook, J. Support vector machines for credit scoring and discovery of significant features. *Expert Syst. Appl.* **2009**, *36*, 3302–3308. [[CrossRef](#)]
25. Wang, G.; Hao, J.; Ma, J.; Jiang, H. A comparative assessment of ensemble learning for credit scoring. *Expert Syst. Appl.* **2011**, *38*, 223–230. [[CrossRef](#)]
26. Liu, Y.; Schumann, M. Data mining feature selection for credit scoring models. *J. Oper. Res. Soc.* **2005**, *56*, 1099–1108. [[CrossRef](#)]
27. Howley, T.; Madden, M.G.; O’Connell, M.L.; Ryder, A.G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl. Based Syst.* **2006**, *19*, 363–370. [[CrossRef](#)]
28. Latimore, D. *Artificial Intelligence in Banking*; Oliver Wyman: Boston, MA, USA, 2018.
29. Martínez-Murcia, F.; Górriz, J.; Ramírez, J.; Illán, I.; Ortiz, A. Automatic detection of Parkinsonism using significance measures and component analysis in DaTSCAN imaging. *Neurocomputing* **2014**, *126*, 58–70. [[CrossRef](#)]
30. Zhou, L.; Lai, K.K.; Yu, L. Credit scoring using support vector machines with direct search for parameters selection. *Soft Comput.* **2009**, *13*, 149–155. [[CrossRef](#)]
31. Wang, X.; Xu, M.; Pustli, Ö.T. A Survey of Applying Machine Learning Techniques for Credit Rating: Existing Models and Open Issues. In *Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 122–132.
32. Vaca, M. Evaluación de Estimadores Basados en Modelos Para el Cálculo del Riesgo de Crédito Bancario en Entidades Financieras. Ph.D. Thesis, Universidad Miguel Hernández, Elche, Spain, 2017.
33. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
34. Dua, D.; Graff, C. *UCI Machine Learning Repository*; School of Information and Computer Sciences, University of California: Irvine, CA, USA, 2017.
35. Picos Sánchez, F.; Pérez López, C.; Gallego Vieco, C.; Huete Vázquez, S. *La Muestra de Declarantes del IRPF de 2008: Descripción General y Principales Magnitudes*; Documentos de trabajo 14/2011; Instituto de Estudios Fiscales: Madrid, Spain, 2011.
36. Pérez López, C.; Burgos Prieto, M.; Huete Vázquez, S.; Gallego Vieco, C. *La Muestra de Declarantes del IRPF de 2009: Descripción General y Principales Magnitudes*; Documentos de trabajo 11/2012; Instituto de Estudios Fiscales: Madrid, Spain, 2012.
37. Pérez López, C.; Burgos Prieto, M.; Huete Vázquez, S.; Pradell Huete, E. *La Muestra de Declarantes del IRPF de 2010: Descripción General y Principales Magnitudes*; Documentos de trabajo 22/2013; Instituto de Estudios Fiscales: Madrid, Spain, 2013.
38. Pérez López, C.; Villanueva García, J.; Burgos Prieto, M.; Pradell Huete, E.; Moreno Pastor, A. *La Muestra de Declarantes del IRPF de 2011: Descripción General y Principales Magnitudes*; Documentos de trabajo 17/2014; Instituto de Estudios Fiscales: Madrid, Spain, 2014.
39. Pérez López, C.; Villanueva García, J.; Burgos Prieto, M.; Bermejo Rubio, E.; Khalifi Chairi El Kammel, L. *La Muestra de Declarantes del IRPF de 2012: Descripción General y Principales Magnitudes*; Documentos de trabajo 18/2015; Instituto de Estudios Fiscales: Madrid, Spain, 2015.

40. Mylonakis, J.; Diacogiannis, G. Evaluating the Likelihood of Using Linear Discriminant Analysis as a Commercial Bank Card Owners Credit Scoring Model. *Int. Bus. Res.* **2010**, *3*. [[CrossRef](#)]
41. Hand, D.; Henley, W. Statistical Classification Methods in Consumer Credit Scoring: A Review. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1997**, *160*, 523–541. [[CrossRef](#)]
42. Boj, E.; Claramunt, M.M.; Esteve, A.; Fortiana, J. Credit Scoring basado en distancias: Coeficientes de influencia de los predictores. In *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009*; Estudios, F.M., Ed.; Cuadernos de la Fundación MAPFRE: Madrid, Spain, 2009; pp. 15–22.
43. Ochoa P, J.C.; Galeano M, W.; Agudelo V, L.G. Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Conyuntura Económica* **2010**, *16*, 191–222.
44. Cabrera Cruz, A. Diseño de Credit Scoring Para Evaluar el Riesgo Crediticio en una Entidad de Ahorro y crédito popular. Ph.D. Thesis, Universidad tecnológica de la Mixteca, Oaxaca, Mexico, 2014.
45. Moreno Valencia, S. El Modelo Logit Mixto Para la Construcción de un Scoring de Crédito. Ph.D. Thesis, Universidad Nacional de Colombia, Bogotá, Colombia, 2014.
46. Salinas Flores, J. Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART. *Rev. Fac. Ing. Ind. UNMSM* **2005**, *8*, 29–36.
47. Gomes Goncalves, O. Estudio Comparativo de Técnicas de Calificación Crediticia. Ph.D. Thesis, Universidad Simón Bolívar, Barranquilla, Colombia, 2009.
48. Lee, T.; Chen, I. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.* **2005**, *28*, 743–752. [[CrossRef](#)]
49. Lee, T.; Chiu, C.; Lu, C.; Chen, I. Credit Scoring Using the Hybrid Neural Discriminant Technique. *Expert Syst. Appl.* **2002**, *23*, 245–254. [[CrossRef](#)]
50. Steenackers, A.; Goovaerts, M. A credit scoring model for personal loans. *Insur. Math. Econ.* **1989**, *8*, 31–34. [[CrossRef](#)]
51. Quintana, M.J.M.; Gallego, A.G.; Pascual, M.E.V. Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras: comparación de resultados. *Pecunia Rev. Fac. Cienc. Econ. Empres. Univ. Leon* **2005**, *1*, 175–199. [[CrossRef](#)]
52. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [[CrossRef](#)]
53. Dunteman, G.H. *Principal Components Analysis (Quantitative Applications in the Social Sciences) Issue 69*; Quantitative Applications in the Social Sciences; Sage Publications, Inc.: Thousand Oaks, CA, USA, 1989.
54. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 47–58.
55. Zaidi, N.; Du, Y.; Webb, G. On the Effectiveness of Discretizing Quantitative Attributes in Linear Classifiers. *arXiv* **2017**, arXiv:1701.07114.
56. Fisher, R.A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **1919**, *52*, 399–433. [[CrossRef](#)]
57. Almi nana, M.; Escudero, L.F.; Pérez-Martín, A.; Rabasa, A.; Santamaría, L. A classification rule reduction algorithm based on significance domains. *TOP* **2014**, *22*, 397–418. [[CrossRef](#)]
58. Rabasa Dolado, A. Método Para la Reducción de Sistemas de Reglas de Clasificación por Dominios de Significancia. Ph.D. Thesis, Universidad Miguel Hernández, Elche, Spain, 2009.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).