Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Big Data techniques to measure credit banking risk in home equity loans[☆]

A. Pérez-Martín[*], A. Pérez-Torregrosa, M. Vaca

Department of Economic and Financial Studies, Miguel Hernández University of Elche, Spain

ARTICLE INFO

Keywords:
Credit scoring
Big Data
Monte Carlo
Data mining

ABSTRACT

Nowadays, the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in Big Data techniques applied to massive financial datasets for segmenting risk groups. In this paper, the presence of large datasets is approached through the development of some Monte Carlo experiments using known techniques and algorithms. In addition, a linear mixed model (LMM) has been implemented as a new incremental contribution to calculate the credit risk of financial companies. These computational experiments are developed with several combinations of dataset sizes and forms to cover a wide variety of cases. Results reveal that large datasets need Big Data techniques and algorithms that yield faster and unbiased estimators. Big Data can help to extract the value of data and thus better decisions can be made without the runtime component. Through these techniques, there would be less risk for financial companies when predicting which clients will be successful in their payments. Consequently, more people could have access to credit loans.

1. Introduction

Any credit rating system that enables the automatic assessment of the risk associated to 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 amount, and other features inherent in financial operations. It is an objective system for approving credit that does not depend on the analyst's discretion.

In the 1960s, coinciding with the massive demand for credit cards, financial companies began applying credit scoring techniques as a means of assessing their exposure to risk insolvency (Altman, 1998). At the same time, the United States also began to develop and apply credit scoring techniques to assess credit risk assessment and to estimate the probability of default (Escalona Cortés, 2011).

Since 1970, credit scoring models had been based on statistical techniques, and particularly on discriminant analysis, which was generalized in 1990 (Gutierrez, 2007). However, with the development of better statistical resources and new advances in technology, it became necessary for financial institutions to carry out their risk assessments more effectively and efficiently.

Since the 1980s, due to the increase in credit demand and computational progress, credit scoring techniques have been extended to loans. In this context, some financial companies started to use different statistical techniques to optimize the differentiation between good and bad loans (Durand, 1941; Reichert, Cho, & Wagner, 1983).

In 2004, the recommendations of the Basel Committee (known as Basel II) on banking supervision appeared. Since then, the use of advanced methods of credit scoring have become a regulatory requirement for banks and financial institutions to improve the efficiency of capital allocation. In response to the global financial crisis, a new document (Basel III) appeared. This document introduced more changes and demands for financial companies regarding the control of borrowed capital and the ensuing increase in reserves based on their risk. As a result, improving the accuracy of credit risk evaluations has become a potential benefit to financial institutions.

In recent decades, although several different investigations have compared different methods for measuring risk, scientific literature has not solved the problem efficiently. As well as this, there has been a rise in financial operations, which has led to an increase in the volume of databases. All the methods analyzed in the scientific literature are suitable for the classification of good or bad credit, but all of them have advantages and disadvantages.

Nowadays, the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in Big Data techniques applied to massive financial datasets for segmenting risk groups. The presence of large datasets is approached through the development of some Monte Carlo experiments using known techniques and algorithms. These computational experiments reveal that large datasets need Big Data techniques and algorithms that yield faster and unbiased estimators.

[☆] The authors acknowledge financial support provided by Conselleria de Educación, Generalitat Valenciana under grant GVA/2016/053.

[*] Corresponding author.
E-mail addresses: agustin.perez@umh.es (A. Pérez-Martín), agustin.perez01@goumh.umh.es (A. Pérez-Torregrosa), m.vaca@umh.es (M. Vaca).

https://doi.org/10.1016/j.jbusres.2018.02.008
Received 20 June 2017; Received in revised form 1 February 2018; Accepted 5 February 2018

They are developed with several combinations of dataset sizes and forms to cover a wide variety of cases.

Big Data can help to extract the value of data and thus better decisions can be made. Then, the high costs of the runtime, which would make the problem intractable, can be avoided. Through these techniques, financial companies would have less risk when predicting which clients will be successful in their payments. Consequently, more people could have access to credit loans.

This research area is very important for financial institutions because credit risk is 60% of a company's total risk. With the introduction of the 9th International Financial Reporting Standards (IFRS 9) in January 2018, financial companies will have to calculate expected losses from defaults over the 12 months following these financial instruments. In this case, a good method for estimating risk would mean lower expected losses and a higher profit for the company and maybe more credit loans.

There are various methods available for assessing credit risk. These range from a personalized study by an expert in risk analysis to different statistical and econometric methods of credit scoring. Nowadays, in a first step, it is not feasible to apply specific analyses to the study of home equity loans. Credit scoring methods are more efficient, objective, and consistent in their predictions, since they can be used to analyze and make quick and inexpensive decisions about many credit applications.

In line with some authors, credit scoring can be considered as a way to identify different groups within a population. One of the first proposals to solve this problem was introduced in statistics by Fisher (1936) using discriminant analysis and a multivariate statistical technique. Durand (1941) was the first to recognize that the same statistical techniques can be used to optimize the differentiation between good and bad loans.

The use of credit scoring models is not only the result of the generalization of credit, but also the result of the banking regulations and supervision introduced in the past three decades. Financial and credit institutions are subject to what is known as “prudential policy”, which means the amount of equity must be maintained to ensure a smooth operation and to cover several risks which may arise, including credit risk (Trias, Carrascosa, Fernández, París, & Nebot, 2005).

Between the late 20th century and early 21st century, due to economic growth, consumer credit increased spectacularly. The need for financial institutions to increase their market share has become a reality today; the larger the volume of credit granted by a company, the greater its potential profits. However, this should be linked to an increase in quality, because otherwise the end result would be a significant deterioration in the income statement. Consequently, statistical methods for assessing credit risk have become increasingly important (Hand & Henley, 1997).

Since Basel II, the use of advanced methods of credit scoring has become a regulatory requirement for banks and financial institutions in order to improve the efficiency of capital allocation. Nevertheless, Basel III introduced stricter changes for controlling borrowed capital, where an increase in reserves occurs in financial institutions based on their risk. Improving the accuracy of credit risk evaluation is a potential benefit to financial institutions, even if it is only slight. Over the past decades, there have been different methods for measuring risk.

Nowadays, credit scoring models are based on mathematics, econometric techniques, and artificial intelligence (Ochoa, Galeano, & Agudelo, 2010; Canton, Rubio, & Blasco, 2010). Empirical studies by various authors present alternative approaches and compare different techniques and algorithms with the problem that they present different conclusions. These approaches include the following: decision trees used by Srinivasan and Kim (1987), Hand and Henley (1997), Galindo and Tamayo (2000), Huang, Hung, and Jiau (2006) or Lee, Chiu, Chou, and Lu (2006); the logistic regression technique described by Thomas (2000), Boj, Claramunt, Esteve, and Fortiana (2009b), and Alaraj and Abbod (2016); discriminant analysis used by Altman (1998), Yobas, Crook, and Ross (2000), and Boj, Claramunt, Esteve, and Fortiana (2009a); or the use of support vector machines by authors like Van Gestel, Baesens, Garcia, and Van Dijcke (2003), Liu, Frazier, and Kumar (2007) or Yu (2014). In our research, we try to address the problem by using several datasets with some Monte Carlo experiments.

All the methods analyzed in the scientific literature are suitable for the classification of good or bad credit, but all of them have advantages and disadvantages. The method or algorithm used depends on the structure of the data, the features used, the possibility of separating the classes by using these features, and the purpose of the classification of the data structure (Morales, Pérez-Martín, & Vaca, 2013, Baesens et al., 2003). Baesens et al. (2003) compare sixteen methods for credit risk evaluation based on eight datasets of different sizes and origin. They concluded that the experiments also indicated that many classification techniques yield performances which are quite competitive with each other. Only a few classification techniques were clearly inferior to the others, but they did not mention their computational efficiency. Yu, Yao, Wang, and Lai (2011) compare different methods for credit risk evaluation with two datasets (German and Australian UCI credit datasets). They concluded that weighted least squares support vector machine (LSSVM) classifier is the best for credit risk evaluation. They also indicated that the credit industry requires quick decisions, and in this sense, there should be a trade-off between computational performance and computational efficiency. In our paper, we carried out a computational measure as an incremental contribution.

Pérez-Martín and Vaca (2017) analyze quadratic discriminant analysis (QDA) and support vector machine with linear kernel (LSVM). With respect to effectiveness, LSVM is found to be the best method for estimating credit risk, but in terms of computational efficiency, LSVM takes longer than QDA to solve the same problem. For large datasets (5000 records) and a large number of explanatory variables, LSVM has better success rates. When the number of explanatory variables are equal or less than 10, differences are unnoticeable.

Scientific literature therefore has not solved the problem efficiently. As well as this, the increase in financial operations has led to an increase in the volume of databases. As a result, financial companies are faced with the problem of managing a huge volume of databases and the need to address this situation. Big Data techniques applied to massive financial datasets for segmenting risk groups is the solution. Big Data can help to extract the value of data and thus better decisions can be made. Then, the high costs of the runtime, which would make the problem intractable, can be avoided. Therefore, an automatic evaluation is necessary, using fast and adaptive techniques like machine learning, where the probability of default can be calculated with historical massive datasets in a reasonable period of time.

In this paper, eight methods for solving the problem of credit scoring in home equity loans are proposed. Firstly, measures are made of how a loan can be classified and how cost influences execution time. To evaluate this, different Monte Carlo simulation experiments are performed. Several (72×10^4) random datasets with different sizes are generated so that the result of a method does not depend on the data, and a linear mixed model (LMM) method is proposed as a new and better contribution.

The execution time component may be important when deciding whether to apply one method or another due to the massive volume of data. A computationally efficient method can be much more competitive, since it provides advantages in terms of time expected in resolving requests. The main goal in this study is to present and compare credit scoring methods that are effective and efficient.

In Section 2, the methods used in this research are outlined. In Section 3, simulation experiments are developed and several efficient measures are presented. Finally, in Sections 4 and 5, results, conclusions, and recommendations are given.

2. The models

The eight methods to be considered are as follows: 1) a classical statistical procedure called quadratic discriminant analysis (QDA); 2) classification and regression trees (CART); 3) prune decision tree (PRUNECART); 4) a linear regression model (LM); 5) a linear mixed model (LMM); 6) a data mining classification procedure, support vector machine (SVM); 7) neural network (NN); 8) a generalized linear model with logit link (GLMLOGIT).

A comparison is made of the different methods for credit risk evaluation with different techniques: traditional statistical methods, parametric and non-parametric methods (QDA, CART, PRUNECART), statistical methods (LM, LMM, GLMLOGIT), and artificial intelligence techniques (NN, LSVM).

Quadratic discriminant analysis is a more advanced technique than the linear discriminant analysis (LDA) formulated by Fisher (1936). LDA is a classifier that assumes homogeneous covariance matrix for each class, whereas QDA does not assume that the covariance matrix of each class is homogeneous, and is therefore better for classification (Seber, 1984). Furthermore, the QDA algorithm is better recommended than LDA in the presence of large datasets (Marks & Dunn, 1974).

In contrast to other decision trees (ID3, C4.5, C5.0, J48, Quinlan, 1993), CART decision trees (Breiman, Friedman, Olshen, & Stone, 1984) employ the mean square error or number of incorrect classifications (partition criterion of the Gini index) as optimality criterion. For the purposes of this study, it is preferable to reduce the error. Other methods maximize gaining information, but they can lead to over-adjustment and to obtaining insignificant branches (Bonilla, Olmeda, & Puertas, 2003). In CART, it is possible to prune the tree when there is irrelevant information that does not improve the optimum. The execution time is reduced by keeping the optimal prediction.

LM, LMM, and GLMLOGIT are included because LM is a simple and basic regression model, and LMM is a linear model with mixed effect (Dobson, 1990). It is considered important to observe the effects of introducing mixed effect in the model and to compare both. These methods are very easy to interpret in credit risk (Liu & Schumann, 2005; Piramuthu, 2006). As well as this, in credit risk, there is some flexibility in the basic hypothesis of the model thanks to its robustness (Morales et al., 2013).

In the GLMLOGIT method (McCullagh & Nelder, 1989), there is a response variable associated with co-variables, but unlike the linear model, it is not necessary to satisfy fundamental principles that are fulfilled in the linear model such as

1. Additive effects of covariables.
2. Normality of the response.
3. Homocedasticity.

A linear predictor based on a linear combination of explanatory variables is obtained. The variables can be continuous, categorical or a mixture of the two. The link function option in this research is the binomial logit function because the response is true or false. This method is used by several financial institutions.

Support vector machines (SVM) for binary classification (Van Gestel et al., 2003), is an important new methodology that has emerged in the area of machine learning and neural networks. The kernel based representation of SVMs makes it possible to formulate the classifier problem as a convex optimization problem, usually a quadratic programming problem. SVM are supervised models for analyzing binary class labels of a response variable. In a SVM, a hyperplane has been built in order to separate observations for classification. We use SVM with a linear kernel (LSVM), which is very closely related to a linear programming problem in operations research.

Neural network (NN) provides a particular advantage because the model does not require pre-specification (Ripley, 1996).

For the sake of brevity, the development of formulas has been omitted because they can be easily be found in literature.

3. Simulation experiments

This Monte Carlo simulation experiment is designed to compare the mean square error and the success rate of well-classified loans, for the eight methods proposed (QDA, CART, PRUNECART, LM, LMM, LSVM, GLMLOGIT, and NN techniques) and the time involved. This experiment is carried out to avoid the relationship between methods and data as found in the literature. For this reason, several random datasets have been generated, which would prevent patterns.

Two sets of random data are generated to obtain training and testing datasets. Training datasets make it possible to obtain the model parameters (QDA, CART, PRUNECART, LM, LMM, LSVM, GLMLOGIT, and NN). These model parameters are used to predict target variables with the testing dataset. The resulting predictions are then used to calculate the mean square error and the success rate as the number of correct classifications of the total.

Each dataset is generated as a mixed regression model (a fixed effect and a random effect) as follows:

For $i = 1,\ldots,I$, $j = 1,\ldots,n_i$:

- First explanatory variable: $x_{ij1} = (b_i - a_i)U_{ij} + a_i$ with $U_{ij} = \frac{j}{n_i+1}$. $a_i = 1$, $b_i = 1 + \frac{1}{I}(I + i)$.
- Another explanatory variables: Generated as a uniform distribution from x_{ij2} to x_{ijp}.
- Random effects and errors: $u_i \sim N(0, \sigma_1^2 = 1)$. $e_{ij} \sim N(0, \sigma_1^2 = 1)$.
- Target variable: Calculate:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + u_i + e_{ij},$$

with $\beta_0 = \ldots = \beta_p = (-0.95,1)$

- Recategorize target variable to successful and default cases:

$$p_{ij} = (e^{y_{ij}})/(1 + e^{y_{ij}})*100$$
$$P_{ij} = floor(p_{ij})$$

The simulation experiment follows the steps:

1. Repeat $K = 10^4$ times ($k = 1,\ldots,K$)
   1.1. Generate training and testing datasets of size $n = \sum_{i=1}^{I} n_i$.
   1.2. Calculate the models' parameters with the training dataset.
   1.3. Calculate the mean square error (RMSE) with the training datasets.
   1.4. Calculate average success rate and total elapsed time for the eight methods.
2. Calculate the average time for each method.

The simulations are carried out for the 6 combinations of sizes (records) presented in Table 1.

For each combination in Table 1, six groups of explanatory variables $\boldsymbol{x}$ have been included in order to verify the methods of wide datasets. The number of explanatory variables are $p = 1, 2, 10, 50, 100$, and 250. With these values, we finally generated and analyzed 720,000 datasets belonging to the eight methods.

Table 1
Groups of datasets sizes.

g	1	2	3	4	5	6
$I^{(g)}$	2	5	10	25	50	100
n_i	10	10	10	10	10	10
n	200	500	1000	2500	5000	10,000

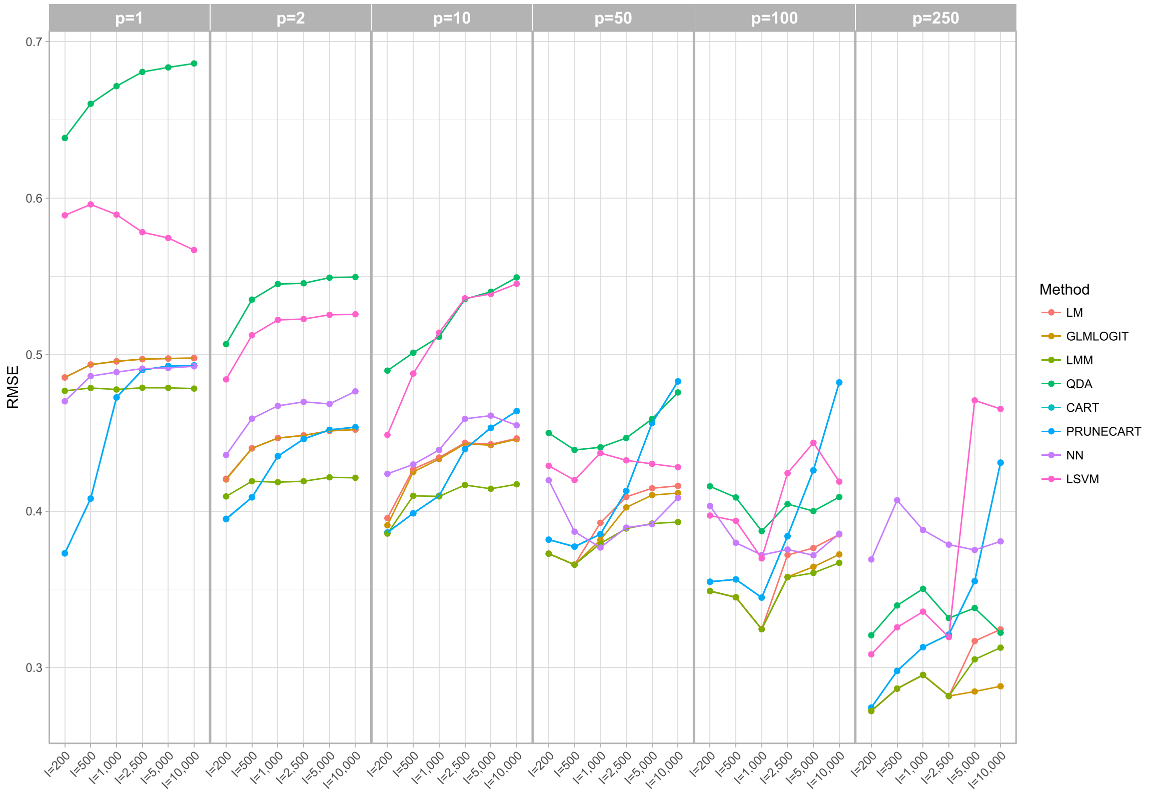

Fig. 1. RMSE.

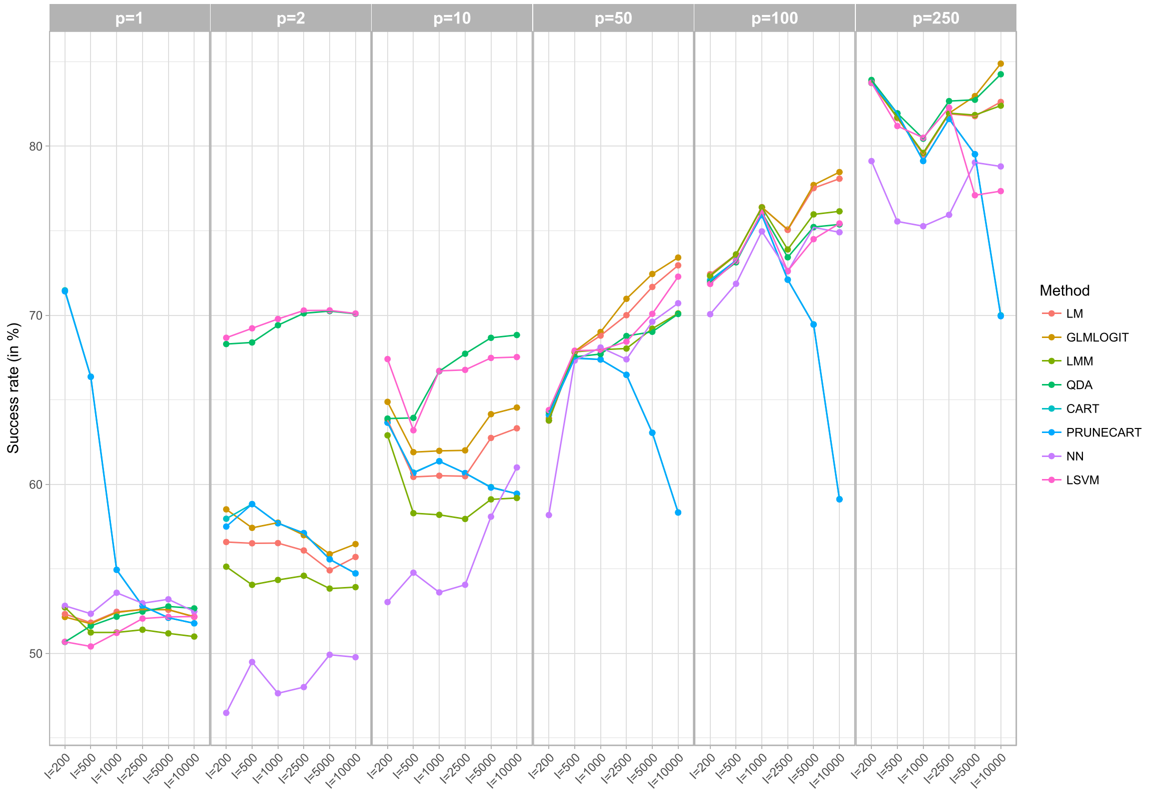

Fig. 2. Success rate.

All the simulations and procedures have been developed in a dedicated Intel Xeon E5620 server with Linux Debian squeeze operating system 64 bits, 8 CPUs at 2.4 GHz, and 24 GB Ddr3 RAM, and implemented in R software (R Core Team, 2015).

Table 2
Average total times for methods for p = 1, 2, 10, 50, 100, and 250.

		GLMLOGIT	CART	LM	PRUNECART	LMM	NN	QDA	LSVM
p = 1	I = 200	0.1170	0.1218	0.1140	0.1214	0.2251	0.3118	0.1197	0.5130
	I = 500	0.3678	0.3624	0.2832	0.3674	0.5124	0.7669	0.3620	3.0841
	I = 1000	0.8181	0.9048	0.7831	0.9674	1.1071	1.9144	0.9011	10.6799
	I = 2500	4.1835	4.2455	3.8588	4.4066	4.9294	7.0657	4.2303	63.5062
	I = 5000	15.5245	17.2117	16.2301	15.6821	17.3725	23.0798	17.2363	290.6021
	I = 10,000	70.5019	77.0731	73.6256	78.1395	75.6749	87.0670	84.3481	1924.4622
p = 2	I = 200	0.1607	0.1369	0.1207	0.1504	0.2228	0.2924	0.1475	0.4801
	I = 500	0.4675	0.4596	0.4444	0.4198	0.5638	0.8796	0.4370	2.4302
	I = 1000	1.1895	1.1214	1.0602	1.0955	1.4588	1.9720	1.1111	10.4809
	I = 2500	4.8048	4.9219	4.8926	5.1973	5.5394	7.6443	5.1935	50.0054
	I = 5000	18.4454	21.7528	19.7911	21.6193	19.5715	25.9201	21.7798	183.2284
	I = 10,000	94.4783	99.9641	91.4835	93.8493	80.4374	104.5231	98.0936	876.7011
p = 10	I = 200	0.2321	0.2186	0.2378	0.2510	0.3935	0.3846	0.2072	2.6114
	I = 500	0.7331	0.7599	0.7243	0.8641	1.0451	1.5696	0.8215	10.2398
	I = 1000	2.2530	2.6275	2.0933	2.4943	2.8878	3.5614	2.2876	31.7912
	I = 2500	11.3491	11.8600	12.1806	12.8664	11.6912	14.3573	12.3573	160.0460
	I = 5000	45.3098	51.6486	49.9349	47.6057	40.9611	58.1122	45.2535	901.3708
	I = 10,000	195.2906	198.7309	203.5109	210.8288	173.8847	212.4299	194.7537	3596.4383
p = 50	I = 200	0.7853	0.7150	0.6293	0.7405	0.9046	1.1734	0.6714	2.5142
	I=500	2.7525	2.4500	2.4922	2.7442	3.1189	6.1435	2.7249	22.5239
	I = 1000	7.6141	9.6485	8.1806	9.9753	8.6294	18.2833	8.5589	711.0637
	I = 2500	39.9701	47.8098	44.9164	49.7263	42.9063	74.4545	44.0226	70.4027
	I = 5000	146.5676	162.7924	160.9713	168.8385	154.8492	193.9680	159.4337	5072.2406
	I = 10,000	587.4827	556.9257	569.5652	599.1148	575.7265	645.5945	539.5353	35,913.1534
p = 100	I = 200	1.6177	1.4738	1.2519	1.5345	1.3778	2.1787	1.1935	3.8189
	I = 500	4.8899	5.9617	5.4510	6.0015	5.2912	11.0276	4.9508	21.7457
	I = 1000	14.8006	18.2913	15.3024	18.4230	17.2756	33.4146	14.4943	137.7739
	I = 2500	78.5340	84.5582	79.0578	86.1021	82.7671	137.0645	79.8543	3877.1001
	I = 5000	255.9099	253.3885	263.1605	300.7562	283.1745	393.0413	281.6063	10,943.7820
	I = 10,000	1034.5412	1042.5802	1105.8297	1103.1804	1063.1180	1208.5294	982.9024	203,937.4895
p = 250	I = 200	6.4747	3.5405	3.2845	3.5694	3.4186	4.2644	3.6577	6.8599
	I = 500	20.0630	13.4357	11.5822	12.8630	12.0170	20.1990	12.7286	40.3384
	I = 1000	46.7467	41.5634	36.1897	38.5789	39.5268	71.2766	36.1131	161.1132
	I = 2500	184.0902	195.8078	177.9616	198.8621	208.4802	304.2916	176.1228	3434.5314
	I = 5000	601.6096	666.8014	638.2155	692.9703	641.9674	877.8700	640.6941	142,661.0333
	I = 10,000	2450.8075	2432.2963	2578.1551	2507.6977	2438.3594	3185.6393	2275.8200	532,888.1111

4. Results

In the simulation experiment, the first focus of attention was on mean square error (RMSE) for all methods.

As observed in Fig. 1, the RMSE decreases as the number of variables increases for all the methods studied, and increases as the groups of dataset size ($I^{(g)}$) increase.

All methods obtain a bad RMSE with 1 variable, the highest being 0.686087 and the dataset size 10,000, which corresponds to the QDA method. In six of the eight methods, the highest RMSE occurred with 1 variable and the dataset size 10,000. The QDA method improves greatly as the number of variables increases. As previously mentioned, it has the highest RMSE value for 1 variable, which increases to a value of 0.320632188 with 250 variables and the dataset size 200.

The LMM, GLMLOGIT, and CART methods have lower RMSE values in 17, 11, and 6 times respectively out of a total of 36. CART obtains a smaller RMSE, but only where there are few variables (1 or 2 variables) and maximum dataset sizes 1000. LMM is closest to CART in the same events.

The difference between LMM and GLMLOGIT is very slight, with LMM maintaining better results for up to 10 variables and dataset size 10,000. From 50 variables, GLMLOGIT's RMSE is equal to that of LMM for dataset sizes of 200 and 500, for the rest of the dataset sizes LMM results are lower. As the number of variables increases, GLMLOGIT's RMSE equals LMM's RMSE for larger dataset sizes, with 100 variables matching dataset sizes 200, 500, and 1000. Finally, with respect to 250 variables, LMM's RMSE is equal to that of GLMLOGIT for dataset sizes 200, 500, 1000, and 2500.

The lowest RMSE methods are CART and PRUNECART. There is only a slight difference between them with 1 and 2 variables, and dataset sizes smaller than 1000. From 2 variables and dataset sizes greater than 1000, they are the methods that reach higher RMSE values.

In 24 of the 36 experiments, QDA is the method that obtains the worst results with respect to RMSE. CART, PRUNECART, QDA, and LSVM have parallel RMSE values. In relation to QDA, LSVM's RMSE is higher from 100 variables and above dataset size 1000, and from 250 variables and above dataset size 2500.

As can be observed in Fig. 2 (success rate), all methods, except for NN, behave the same; as the number of variables increases, the percentage of hits increases. With respect to NN, this percentage decreases with 1 to 2 variables, although success rates rise as data set sizes increase ($I^{(g)}$). In the cases of CART and PRUNECART, they follow the same pattern of an increasing number of variables, but as dataset sizes increase, the percentage of hits decreases.

According to the results obtained in percentage of hits, LSVM is the best method. It obtained the maximum in 11 of the 36 experiments, followed by GLMLOGIT with 9 cases and QDA with 8 cases.

The GLMLOGIT method obtained a maximum success rate of 84.87% with 250 variables and 10,000 dataset sizes.

In batch operations, response time is an important factor for financial institutions, since potential clients require a quick answer. Therefore, a good method is defined in terms of a reduction in the time

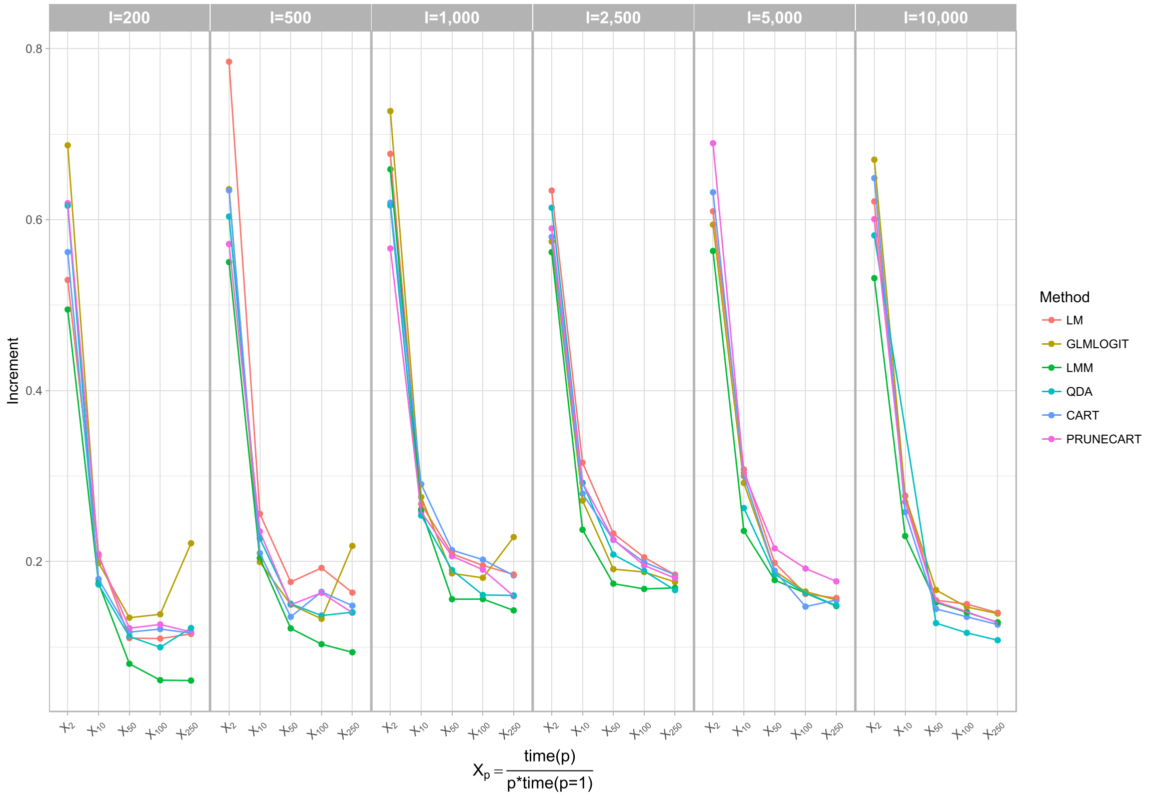

Fig. 3. Comparison of runtimes for $p = 1$ by time values.

taken to obtain an answer for that possible client (Yu et al., 2011).

In all methods, time increases as the number of variables and the dataset sizes increase. Table 2 shows how time increases exponentially as the variables and the size of the dataset increase.

The two most computationally efficient methods in 31% of the cases studied is LM and GLMLOGIT, followed by QDA, LMM, and CART. The results show that the difference in time between GLMLOGIT and LMM is not very significant, even in the majority of explanatory variables (2, 10, 50, and 250) and the dataset sizes 100,000.

The slowest computational method is LSVM, since in all cases it has the worst elapsed time, with an exponential growth from 10 variables and dataset size equal to 1000. Thus, with respect to runtime, LSVM becomes an inefficient method, followed by NN.

In Table 2, the relationship between increases in p and increases in execution time could be

- *Constant*. This is impossible because in Table 2 time increases.
- *Linear*. The ratio between times is constant; i.e., it depends on the number of explanatory variables.
- *Exponential*. The ratio between groups of times grows in multiplicative way.

In Fig. 3, elapsed time is studied as a relative increment (*OX* axis). It has been created to illustrate how time increases with the increase in the number of explanatory variables. A normalization has been created through the ratio between total time for each p and total time for $p = 1$. This ratio has been weighted by the increase that occurs regarding $p = 1$ as follows:

$$\text{Relative increment for } p = RI(p) = \frac{time(p)}{p \times time(p = 1)}$$

For example, the relative time increment for $p = 250$ is $\frac{time(p = 250)}{250 \times time(p = 1)}$. The value of p in the denominator acts as a modulator. If the ratio between times depends on the number of explanatory variables, when p appears in the denominator, the value of the $RI(p)$ goes to 1.

Here, the results for the LSVM and NN methods are also omitted from the graph.

It can be observed that all methods behave in approximately the same way with respect to relative increments. Time increases according to number of datasets sizes ($I^{(g)}$) up to $I^{(g)} = 2500$, after which it decreases. In all methods, the highest point is X_2 and the lowest is X_{100}, except for GLMLOGIT with respect to 250 variables in datasets less than $I^{(g)} = 2500$. The LMM method has the minimum total with 100 variables and in 2 dataset sizes. The slope from X_2 to X_5 is bigger than the slope for the rest of the cases and for all dataset sizes. In 75% of cases, the minimum time of dataset sizes is in X_{250} for all methods.

5. Conclusions

In this research experiment, the intention was to make files that can represent loans for any bank branch, without depending on a specific country or particular institution. For this reason, we simulated random datasets from $N = 2000$ to $N = 100{,}000$ records and from $p = 1$ to $p = 250$ explanatory variables.

Eight methods were proposed: QDA, CART, PRUNECART, LM, LMM, LSVM, GLMLOGIT, and NN. For each method, measures of effectiveness and efficiency were calculated. The most effective methods are GLMLOGIT and LMM. LM, LMM, and GLMLOGIT are the most computationally efficient methods, followed by QDA and CART. For large datasets, LMM takes less elapsed time, and so does GLMLOGIT for short datasets.

Among the target methods of this research, the ones that obtained the best results in terms of RMSE and total elapsed time are LM, GLMLOGIT, and LMM. In comparison with GLMLOGIT, LM could be ruled out because despite being more efficient, as indicated in the literature, the linear model is a particular case of the generalized linear model. For our investigation GLMLOGIT is more suitable than LM, since the response variable is binary categorical. In addition, in comparison with LMM, QDA is ruled out because LMM is more efficient, even though the QDA method is better in terms of computational efficiency. LMM is similar to GLMLOGIT, although for big dataset sizes LMM is better. Hence, GLMLOGIT and LMM are proposed for evaluating credit risk with simulated data.

This has several implications for financial institutions. When an asset is acquired, the impairment allowance is measured as the present value of credit losses from default events projected over the next 12 months. The allowance remains based on the expected losses from defaults over the next 12 months unless there is a significant increase in credit risk. If there is a significant increase in credit risk, the allowance is measured as the present value of all credit losses projected for the instrument over its full lifetime. If the credit risk recovers, the allowance can once again be limited to the projected credit losses over the next 12 months. The only way to calculate the expected losses from defaults is by the estimation of credit risk with the optimal evaluated method presented in this research.

Annex

All the proposed methods have been carried out through several of the R statistical software packages (R Core Team, 2015):

- QDA by MASS package (Venables & Ripley, 2002).
- CART and PRUNECART by rpart package (Therneau, Atkinson, & Ripley, 2014).
- LM by stats package (R Core Team, 2015).
- LMM by lme4 package (Bates et al., 2015).
- LSVM by e1071 package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2014).
- GLMLOGIT by stats and Matrix packages (Bates & Maechler, 2015).
- NN by nnet package (Ripley & Venables, 2015).

The different measures were developed by the authors.

References

Alaraj, M., & Abbod, M. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications, 64*, 36–55. http://dx.doi.org/10.1016/j.eswa.2016.07.017.

Altman, E. (1998). The important and subtlety of credit rating migration. *Journal of Banking and Finance, 22*, 1231–1247. http://dx.doi.org/10.1016/S0378-4266(98)00066-1.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635. http://dx.doi.org/10.1057/palgrave.jors.2601545.

Bates, D., & Maechler, M. (2015). *Linear mixed-effects models using 'Eigen' and S4*http://Matrix.R-forge.R-project.org/ (1.2–3).

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Grothendieck, G. (2015). *Linear mixed-effects models using 'Eigen' and S4*http://lme4.r-forge.r-project.org/ (1.1–10).

Boj, E., Claramunt, M. M., Esteve, A., & Fortiana, J. (2009a). Criterios de selección de modelo en credit scoring, aplicación del análisis discriminante basado en distancias. *En Anales del Instituto de Actuarios Españoles, 15*, 833–869.

Boj, E., Claramunt, M. M., Esteve, A., & Fortiana, J. (2009b). Credit scoring basado en distancias: Coeficientes de influencia de los predictores. In F. M. Estudios (Ed.). *Investigaciones en seguros y gestión de riesgos: RIESGO 2009* (pp. 15–22). Madrid: Cuadernos de la Fundación MAPFRE.

Bonilla, M., Olmeda, I., & Puertas, R. (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring. *Revista Española de Financiación y Contabilidad, 32*(118), 833–869. http://dx.doi.org/10.1080/02102412.2003.10779502.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth: Chapman and Hall.

Canton, S. R., Rubio, J. L., & Blasco, D. C. (2010). Un modelo de credit scoring para instituciones de microfinanzas en el marco de Basilea ii. *Journal of Economics, Finance and Administrative Science, 15*.

Dobson, A. (1990). *An introduction to generalized linear models*. London: Chapman and Hall.

Durand, D. (1941). *Risk elements in consumer instalment financing*. Massachusetts: National Bureau of Economic Research.

Escalona Cortés, A. (2011). (Ph.D. thesis)*Uso de los Modelos Credit Scoring en Microfinanzas*. Montecillo, Texcoco, México: Institución de Enseñanza e Investigación en Ciencias Agrícolas.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics, 15*, 107–143. http://dx.doi.org/10.1023/A:1008699112516.

Gutierrez, M. (2007). Modelos de credit scoring: Qué, cómo, cuándo y para qué. *MPRA Paper, 16377*, 1–30.

Hand, D., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Royal Statistical Society, 160*, 523–541. http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x.

Huang, Y.-M., Hung, C.-M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications, 7*, 720–747. http://dx.doi.org/10.1016/j.nonrwa.2005.04.006.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 50*, 1113–1130. http://dx.doi.org/10.1016/j.csda.2004.11.006.

Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment, 107*, 606–616. http://dx.doi.org/10.1016/j.rse.2006.10.010.

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society, 56*, 1099–1108. http://dx.doi.org/10.1057/palgrave.jors.2601976.

Marks, S., & Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association, 69*, 555–559. http://dx.doi.org/10.1080/01621459.1974.10482992.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall/CRC.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). *e1071: Misc functions of the Department of Statistics (e1071), TU Wien*http://CRAN.R-project.org/package=e1071 (R package version 1.6-4).

Morales, D., Pérez-Martín, A., & Vaca, M. (2013). Monte Carlo simulation study under regression models to estimate credit banking risk in home equity loan. *Data Management and Security Applications in Medicine, Science and Engineering, 45*, 141–152. http://dx.doi.org/10.2495/DATA130131.

Ochoa, J. C., Galeano, W., & Agudelo, L. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Coyuntura Económica, 16*, 191–222.

Pérez-Martín, A., & Vaca, M. (2017). Computational experiment to compare techniques in large data sets to measure credit banking risk in home equity loans. *Data Management and Security Applications in Medicine, Science and Engineering, 5*, 771–779.

Piramuthu, S. (2006). On preprocessing data for financial credit risk evaluation. *Expert Systems with Applications, 30*(3), 489–497. http://dx.doi.org/10.1016/j.eswa.2005.10.006.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. California, EUA: Morgan Kaufmann Publishers, Inc.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Reichert, A., Cho, C., & Wagner, G. (1983). An examination of conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics, 1*(2), 101–114.

Ripley, B., & Venables, W. (2015). *Feed-forward neural networks and multinomial log-linear models*http://www.stats.ox.ac.uk/pub/MASS4/ (7.3–10).

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Inglaterra: Cambridge University Press.

Seber, G. (1984). *Multivariate observations. wiley series in probability and mathematical statistics*. New-York: John Wiley and Sons, Inc.

Srinivasan, V., & Kim, Y. H. (1987). Credit granting: A comparative analysis of classification procedures. *Journal of Finance, 42*, 665–683. http://dx.doi.org/10.1111/j.1540-6261.1987.tb04576.x.

Therneau, T., Atkinson, B., & Ripley, B. (2014). *Recursive partitioning and regression trees*https://cran.r-project.org/package=rpart.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*, 149–172. http://dx.doi.org/10.1016/S0169-2070(00)00034-0.

Trias, R., Carrascosa, F., Fernández, D., París, L., & Nebot, G. (2005). Riesgo de créditos: Conceptos para su medición, basilea ii, herramientas de apoyo a la gestión. *AIS Group - Financial Decisions*.

Van Gestel, T., Baesens, B., Garcia, J., & Van Dijcke, P. (2003). A support vector machine approach to credit scoring. *Bank en Financiewezen, 2*, 73–82.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer0-387-95457-0http://www.stats.ox.ac.uk/pub/MASS4.

Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics, 11*(2), 111–125. http://dx.doi.org/10.1093/imaman/11.2.111.

Yu, L. (2014). Credit risk evaluation with a least squares fuzzy support vector machines classifier. *Discrete Dynamics in Nature and Society, 2014*, 1–9. http://dx.doi.org/10.1155/2014/564213.

Yu, L., Yao, X., Wang, S., & Lai, K. K. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications, 38*, 15392–15399. http://dx.doi.org/10.1016/j.eswa.2011.06.023.