COMPUTATIONAL EXPERIMENT TO COMPARE TECHNIQUES IN LARGE DATASETS TO MEASURE CREDIT BANKING RISK IN HOME EQUITY LOANS

A. PÉREZ-MARTÍN & M. VACA
Department of Economics and Finance. Miguel Hernandez University of Elche, Spain

ABSTRACT

In the 1960s, coinciding with the massive demand for credit cards, financial companies needed a method to know their exposure to risk insolvency. It began applying credit-scoring techniques. In the 1980s credit-scoring techniques were extended to loans due to the increased demand for credit and computational progress. In 2004, new recommendations of the Basel Committee (as called Basel II) on banking supervision appeared. With the ensuing global financial crisis, a new document, Basel III, appeared. It introduced more demanding changes on the control of borrowed capital.

Nowadays, one of the main problems not addressed is the presence of large datasets. This research is focused on calculating probabilities of default in home equity loans, and measuring the computational efficiency of some statistical and data mining methods. In order to do these, some Monte Carlo experiments with known techniques and algorithms have been developed.

These computational experiments reveal that large datasets need BigData techniques and algorithms that yield faster and unbiased estimators.

Keywords: BigData; Credit Scoring; Monte Carlo; Discriminant analysis; Support Vector Machine.

1 INTRODUCTION

There are a variety of methodologies available for assessing credit risk, from the personalized study of an expert in risk analysis, to different statistical and econometric methods of Credit Scoring. However, in a first step, it is not feasible to apply specific analyses in the study of home equity loans. Credit Scoring methods are more efficient and are more objective and consistent in their predictions and, so, can be analyzed and used to make decisions about a lot of credit applications quickly and inexpensively.

Credit Scoring can be considered, as observed by some authors, as a way to identify different groups within a population. One of the first proposals to solve this problem was introduced in statistics by [1] using discriminant analysis and multivariate statistical technique. He sought to distinguish three varieties of plants by physical measurements. [2] was the first to recognize that the same statistical techniques could be used to optimize the differentiation between good and bad loans.

It is called credit scoring and all credit rating systems allow for assessing the risk associated to a banking operation automatically. The risk may depend on the customers and credit characteristics, such as solvency, type of credit, maturity, loan amount and other features inherent to financial operations. It is an objective system in which approval of credit does not depend on the discretion of the analyst. This system must be automatic to reduce costs and processing time.

As for an automatic evaluation it is necessary to use fast and adaptive techniques like machine learning to calculate, in a reasonable period of time, there is a probability of default with historical massive datasets.

In the 1960s, the United States began to develop and apply the techniques of Credit Scoring for credit risk assessment to estimate probability of default [3]. From 1970, Credit Scoring

ISSN: 2046-0546 (paper format), ISSN: 2046-0554 (online), http://www.witpress.com/journals
DOI: 10.2495/CMEM-V5-N5-771-779

models were based on statistical techniques (in particular, discriminant analysis), but were then generalized in 1990 [4]. Best statistical resources were developed at the same time that technology progressed. It was necessary for financial institutions to make their risk assessment more effective and efficient.

The use of credit scoring models is not only due to the generalization of credit. Banking regulation and supervision has also encouraged its use in the past three decades. Financial and credit institutions are subject to the so-called 'prudential policy'. It means that the equity must be maintained to ensure smooth operation and to cover several risks to which they are subject, including credit risk [5].

During the late twentieth and early twentyfirst century, there has been economic growth and consumer credit has increased spectacularly. The need for financial institutions to increase the market share is a current reality; the larger the volume of credit granted by a company, the greater its potential benefits, but should be linked to an increase in the quality, because otherwise the end result would be a significant deterioration in the income. Statistical methods for assessing credit risk have become increasingly important [6].

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. Basel III introduces more demanding changes on the control of borrowed capital. An increase in reserves based on their risk occurs in financial institutions. The improvement in the accuracy of the assessment of credit risk is a potential benefit to the financial institutions, even if it is small. Over the past decades there were several different investigations that have compared different methods for measuring risk.

Today, credit scoring models are based on mathematics, econometric techniques and artificial intelligence [7, 8]. Empirical studies by various authors present alternative approaches that compare different techniques and algorithms (decision trees, logistic regressions, discriminant analysis, parametric and not parametric method, support vector machines, etc.), see [9 – 30].

All the methods suggested in the scientific literature referred are suitable for classifying good or bad credit. Each methodology analyzed by different authors has its own 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 [31, 32].

Therefore, scientific literature has not solved the problem efficiently. In addition, there is an increase in financial operations, with a consequent increase in the volume of databases. The volume of databases that manage financial companies is so great, and it is necessary to fit this problem. BigData techniques applied to massive financial datasets for segment risks groups is the solution. Big Data helps to extract the value of data and thus make better decisions without the runtime component, which involves high cost that makes the problem intractable. In this paper, two methods for solving the problem of credit scoring in home equity loans are proposed. First of all we measure how a loan can be classified and how cost impacts execution time. To evaluate this, different Monte Carlo simulation experiments are performed.

The execution time component may be important in deciding which method has to be applied, due to the massive volume of data. It can be much more competitive as a computationally efficient method provides advantages in terms of time expected in resolving requests. Our main goal is to compare credit scoring methods that can be both effective and efficient.

In Section 2, the methods used in our research are presented. In Section 3, simulation experiments are developed, and several efficient measures are shown. Finally, in Section 5, conclusions and recommendations are offered.

2 THE MODELS

Let us consider two methods supported by two different models. One of them is a classical statistical procedure, Quadratic Discriminant Analysis (QDA). The second one is a data mining class of procedures, Support Vector Machines (SVM).

Quadratic Discriminant Analysis is a technique more advanced than Linear Discriminant Analysis (LDA) formulated by Fisher [1]. LDA is a classifier that assumes homogeneous covariance matrix for each class. In QDA is not assumption that the covariance matrix of each class is homogeneous and better for classification [33]. QDA algorithm is more recommended than LDA in large datasets [34]. In our research we use MASS [35] package of R software [36] with two discriminant variables.

SVM algorithms are supervised models to analyze binary class labels of a response variable. In a SVM, a hyperplane has been built in order to separate observations for classification. Several SVM algorithms can be found in the literature. In this research we use an SVM with linear kernel (LSVM) that is very closely related to a linear programming problem. In our research we use the e1071 [37] package of R software [36].

For the sake of brevity we skip developed formulas, because it is very easy to find them in the literature.

3 SIMULATION EXPERIMENTS

The Monte Carlo simulation experiment is designed to compare the success rate of well-classified loans for QDA and LSVM techniques and the time that it involves.

In the web agustinperez.edu.umh.es/academia/research/papers/ numerical results of the simulation study are available under Big Data 2016 tag.

Two sets of random data are generated, to obtain training and testing datasets. Training dataset allows to obtain the models parameters (QDA and LSVM). These models parameters are used to predict target variables with testing dataset. These predictions will be used to calculate the success rate on a number of correct classifications. Each dataset is generated as 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: $x_{ij1} = (b_i - a_i)U_{ij} + a_\mathrm{i}$ with $U_{ij} = \dfrac{j}{n_i + 1} \cdot a_i = 1,\ b_i = 1 + \dfrac{1}{I}(I + i)$.
- Other explanatory variable: Generate as an uniform distribution from x_{ij2} to x_{ijp}.
- Random effects and errors: $ui \sim N(0, \sigma_1^2 = 1).\ eij \sim N(0,\ \sigma_1^2 = 1)$.
- Target variable: Calculate:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + u_i + \beta_p x_{ijp} + e_{ij}, \quad \text{with } \beta_0 = \cdots = \beta_p = 1$$

- Recategorize target variable to success and default cases:

$$y_{ij} \le median\ (y) \Rightarrow y_{ij} = 0$$

$$y_{ij} > median\ (y) \Rightarrow y_{ij} = 1$$

The simulation experiment follows the steps:

1. Repeat $K = 10^4$ times $(k = 1,..., K)$
   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 confusion matrix for QDA and SVM with the testing dataset.
   1.4. Calculate the success rate with the successes of the confusion matrix (elements of the main diagonal) and total elapsed time of both methods.
2. Calculate the average success rate and the average time for each method.

The simulations are carried out for the 4 combinations of sizes (records) presented in Table 1.

For each combination of table 1, 5 groups of explanatory variables x have been included. The number of explanatory variables are p = 1, 2,10, 50,100. With these values, we finally generated and analyzed 40 x 10^4 datasets belonging to two methods.

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 2.4GHz and 24GB Ddr3 RAM and implemented in R software [36].

4 RESULTS

Firstly, in the simulation experiment we focus our attention on the success rate for QDA and LSVM methods. In most of the combinations of datasets (16 of 20), the LSVM method arises as the best procedure to determine the success rate. See figure 1. The percentage of well-classified increases as does the number of explanatory variables, from 64.5% to 88.43%. For large datasets (5000 records) and a great number of explanatory variables, LSVM has better success rates. When the number of explanatory variables are equal or less than 10, differences are unnoticeable.

After searching for the best method in terms of efficiency in prediction, let us see the results of our computational problem on Big Data. It is well known that LSVM is one of the slower methods existing today. In our research we have tried to link with the increase of the explanatory variables. In figure 2, the average execution times for process are plotted. For the sake of better visualization of the execution times, only values for p = 1,2,10 have been plotted. It can be seen that the QDA method is the fastest. And LSVM is far slower than QDA when the number of record n increases.

In figure 2, it seems that increase in execution time grows exponentially. The relationship between increases in p and increases in execution time can be:

Table 1: Groups of datasets sizes.

g $I^{(g)}$	1 10	2 20	3 30	4 50
n_i	100	100	100	100
n	1000	2000	3000	5000

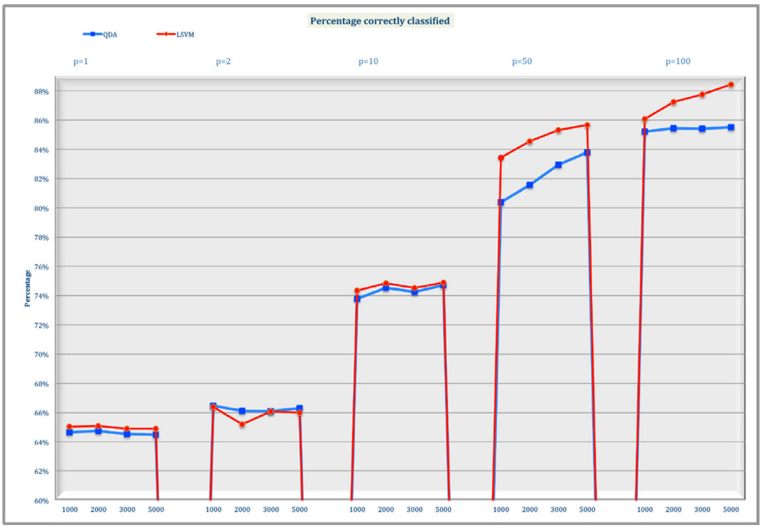

Figure 1: Success rate for QDA and LSVM.

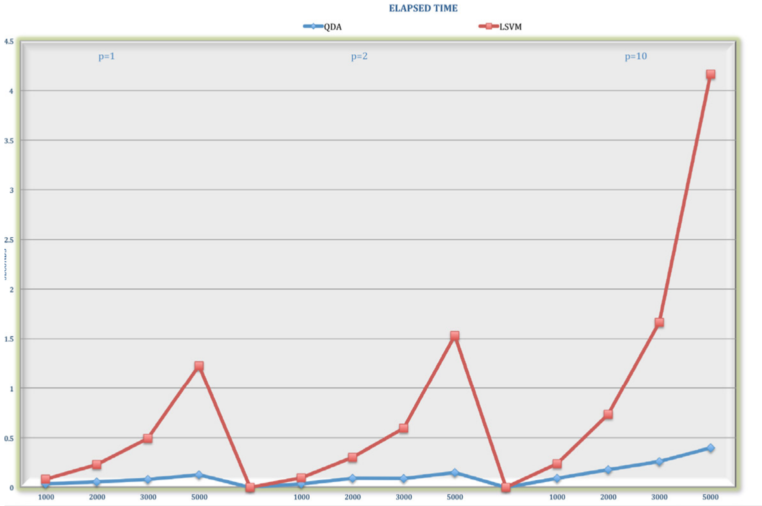

Figure 2: Elapsed time for QDA and LSVM for p = 1; 2; 10.

- *Constant.* This is impossible because in figure 2 the time grows.
- *Linear.* The ratio between times are constant, that is, it depends on the number of explanatory variables.
- *Exponential.* The ratio between groups of times grows in multiplicative way.

Figure 3 has been created in order to observe how the time increases with increasing number of explanatory variables. We have calculated the ratio between execution time for each

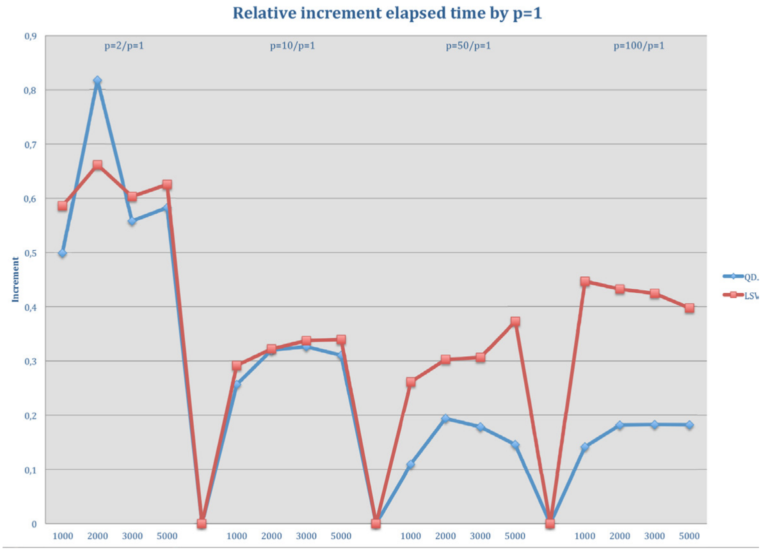

Figure 3: Comparative for execution time for QDA and LSVM for time values by p = 1.

p and execution time for p = 1. This ratio has been weighted by the increase that occurs concerning p = 1 as follows:

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

For example, the relative time increment for p = 50 is $\frac{time(p=50)}{50 \times time(p=1)}$. 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 Relative Increment reaches 1. As it can be seen in figure 3, the values of the Relative Increment of time are all less than 1. This means that the increase in the number of explanatory variables does not affect in the same way.

5 CONCLUSIONS

In our research experiment we attempted to create files that can represent loans for any bank branch. For this reason we simulated dataset from n = 1000 to n = 5000 record and from p = 1 to p = 100 explanatory variables.

Two methods have been proposed, QDA and LSVM. We calculated measures of effectiveness and efficiency. Regarding the effectiveness, we have found that, usually, LSVM is the best method for estimating credit risk. But in terms of computational efficiency, LSVM takes more time than QDA to solve the same problem. At the worst case, LSVM method takes 20 times more runtime than QDA.

A linear relationship between time and the number of explanatory variables has been encountered. It would be very productive to find a functional relationship between runtime and number of explanatory variables. It would also be very appropriate if it can be increased to a higher number of procedures.

REFERENCES

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

[2] Durand, D., *Risk Elements in Consumer Instalment Financing.* National Bureau of Economic Research: Massachusetts, 1941.

[3] Escalona Cortes, A., *Uso de los Modelos Credit Scoring en Microfinan-zas.* Ph.D. thesis, Institución de Enseñanza e Investigación en Ciencias Agrícolas, Montecillo, Texcoco, Mexico, 2011.

[4] Gutierrez, M., Modelos de credit scoring: que, cómo, cuando y para que. *MPRA Paper,* **16377,** pp. 1-30, 2007.

[5] Trias, R., Carrascosa, F., Fernandez, D., Pares, L. & Nebot, G., Riesgo de cróeditos: Conceptos para su medicióon, basilea ii, herramientas de apoyo a la gestióon. *AIS Group - Financial Decisions,* 2005.

[6] Hand, D. & Henley, W.E., Statistical classification methods in consumer credit scoring: a review. *Royal Statistical Society,* **160(3),** pp. 523-541, 1997.
http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x

[7] Ochoa, J.C., Galeano, W. & Agudelo, L., Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Coyuntura Económica,* **16,** pp. 191-222, 2010.

[8] Canton, S.R., Rubio, J.L. & Blasco, D.C., Un modelo de credit scoring para instituciones de microfinanzas en el marco de basilea ll. *Journal of Economics, Finance and Administrative Science,* **15(28),** 2010.

[9] Srinivasan, V. & Kim, Y.H., Credit granting: a comparative analysis of classification procedures. *Journal of Finance,* **42,** pp. 665-683, 1987.
http://dx.doi.org/10.1111/j.1540-6261.1987.tb04576.x

[10] Boj, E., Claramunt, M.M. & Fortiana, J., Selection of predictors in distance-based regression. *Communications in Statistics-Simulation and Computation,* **36(1),** pp. 87-98, 2007.
http://dx.doi.org/10.1080/03610910601096312

[11] Boj, E., Claramunt, M. & Esteve, J., A.and Fortiana, Criterios de se-leccióon de modelo en credit scoring, aplicacióon del anóalisis discriminante basado en distancias. *En Anales del Instituto de Actuarios Españoles,* **15,** pp. 833-869, 2009.

[12] Tam, K. & Kiang, M., Managerial applications of neural networks: The case of bank failure predictions. *Management Science,* **38,** pp. 926-947, 1992.
http://dx.doi.org/10.1287/mnsc.38.7.926

[13] Desai, V., Crook, J. & Overstreet, G., A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* **95,** pp. 24-37, 1996.
http://dx.doi.org/10.1016/0377-2217(95)00246-4

[14] Yobas, J., M.B.and Crook & Ross, P., Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics,* **11,** pp. 111-125, 2000.
http://dx.doi.org/10.1093/imaman/11.2.111

[15] Boj, E., Claramunt, M.M., Esteve, A. & Fortiana, J., Credit scoring basado en distancias: coeficientes de influencia de los predictores. *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009,* ed. F.M. Estudios, Cuadernos de la Fundacióon MAPFRE: Madrid, pp. 15-22, 2009.

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

[17] Boj, E., Claramunt, M.M., Granóe, A. & Fortiana, J., Projection error term in gower's interpolation. *Journal of Statistical Planning and Inference,* **139,** pp. 1867-1878, 2009. http://dx.doi.org/10.1016/j.jspi.2008.07.021

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

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

[20] Artóís, M., Guillóen, M. & Martóínez, J.M., A model for credit scoring: an application of discriminant anlysis. *QUESTIIO,* **18(3),** pp. 385-395, 1994.

[21] Gordy, M.B., A comparative anatomy of credit risk models. *Journal of Banking and Finance,* **24,** pp. 119-149, 2000.
http://dx.doi.org/10.1016/S0378-4266(99)00054-0

[22] Cheung, S., Provincial credit ratings in canada, an ordered probit analysis. *Bank of Canada,* **6,** 1996.

[23] West, D., Neural network credit scoring models. *Computers and Operations Research,* **27,** pp. 1131-1152, 2000.
http://dx.doi.org/10.1016/S0305-0548(99)00149-5

[24] Bonilla, M., Olmeda, I. & Puertas, R., Modelos paramóetricos y no paramóetricos en problemas de credit scoring. *Revista Espanñola de Fi-nanciacion y Contabilidad,* **32(118),** pp. 833-869, 2003.

[25] Liu, C., Frazier, P. & Kumar, L., Comparative assessment of the measures of thematic classification accuracy. *Remote Sens Environ,* **107,** pp. 606-616, 2007.
http://dx.doi.org/10.1016/j.rse.2006.10.010

[26] Falkenstein, E., Risk calc for private companies: Moody's default model. rating methodology. *Moody's Investor Service, Global Credit Research,,* 2000.

[27] Moses, D. & Liao, S., On developing models for failure prediction. *Journal of Commercial Bank Lending,* **69,** pp. 27-38, 1987.

[28] Wiginton, J.C., A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal ofFinancial and Quantitative Analysis,* **15(03),** pp. 757-770, 1980.

[29] Van Gestel, T., Baesens, B., Garcia, J. & Van Dijcke, P., A support vector machine approach to credit scoring. *Bank en Financiewezen,* **2,** pp. 73-82, 2003.

[30] Yu, L., Yao, X., Wang, S. & Lai, K.K., Credit risk evaluation using a weighted least squares svm classifier with design of experiment for parameter selection. *Expert Systems with Applications,* **38(12),** pp. 1539215399, 2011.
http://dx.doi.org/10.1016/j.eswa.2011.06.023

[31] Morales Gonzalez, D., Perez Martin, A. & Vaca Lamata, M., 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,** pp. 141-152, 2013.

[32] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, J., M. andSuykens & Vanthienen, J., Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal ofthe Operational Research Society,* **54(6),** pp. 1082-1088, 2003.

[33] Seber, G., *Multivariate Observations. Wiley series in probability and mathematical statistics.* John Wiley and Sons, Inc.: New-York, 1938.

[34] Marks, S. & Dunn, O.J., Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association,* **69(346),** pp. 555?-559, 1974.

[35] Venables, W.N. & Ripley, B.D., *Modern Applied Statistics with S.* Springer: New York, 4th edition, 2002. ISBN 0-387-95457-0. http://dx.doi.org/10.1007/978-0-387-21706-2

[36] R Core Team, *R: A Language and Environment for Statistical Comput*ing. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[37] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F., *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien,* 2014. R package version 1.6-4.

ACKNOWLEDGEMENT

The authors are grateful to the Conselleria the Educación, Generalitat Valenciana for the economical support under the grant GVA/2016/053. That has made possible this research