Expert Systems with Applications 40 (2013) 6910–6917

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques

María-Dolores Cubiles-De-La-Vega [a], Antonio Blanco-Oliver [b,*], Rafael Pino-Mejías [a], Juan Lara-Rubio [c]

[a] Department of Statistics and Operational Research, Faculty of Mathematics, University of Seville, Avda. Reina Mercedes, s/n, 41012 Seville, Spain
[b] Department of Financial Economics and Operations Management, Faculty of Economics and Business Studies, University of Seville, Avda. Ramon y Cajal, 1, 41018 Seville, Spain
[c] Department of Financial Economics and Accounting, Faculty of Economics and Business Studies, University of Granada, Campus Cartuja, s/n, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Decision support systems
Microfinance institutions
Credit scoring
Efficiency
Statistical Learning
Data mining

ABSTRACT

A wide range of supervised classification algorithms have been successfully applied for credit scoring in non-microfinance environments according to recent literature. However, credit scoring in the microfinance industry is a relatively recent application, and current research is based, to the best of our knowledge, on classical statistical methods. This lack is surprising since the implementation of credit scoring based on supervised classification algorithms should contribute towards the efficiency of microfinance institutions, thereby improving their competitiveness in an increasingly constrained environment. This paper explores an extensive list of Statistical Learning techniques as microfinance credit scoring tools from an empirical viewpoint. A data set of microcredits belonging to a Peruvian Microfinance Institution is considered, and the following models are applied to decide between default and non-default credits: linear and quadratic discriminant analysis, logistic regression, multilayer perceptron, support vector machines, classification trees, and ensemble methods based on bagging and boosting algorithm. The obtained results suggest the use of a multilayer perceptron trained in the R statistical system with a second order algorithm. Moreover, our findings show that, with the implementation of this MLP-based model, the MFIś misclassification costs could be reduced to 13.7% with respect to the application of other classic models.

1. Introduction

Microfinance institutions (hereafter, MFIs) offer saving services and small loans (namely microcredits) to those sectors of the population with a very limited access to financial resources. For this reason, the goals and management criteria of the MFIs have less business component and higher social component than those used by new competitors (international commercial banks). The microfinance sector has rapidly grown in the last years, turning into a booming industry. As an example, the number of microfinance institutions grew by 474% in the period 1998–2008, while the number of customers grew by 1048%. This phenomenon has moved a large number of international commercial banks to operate in the microfinance sector. This reinforced interest has increased the competition between the players in this industry, but it is negatively affecting the MFIs. Therefore, the MFIs need to increase their efficiency in all their processes, minimize their costs and control their credit risk if they want to survival a long-term. In particular, credit scoring models may improve this efficiency. Their objective is to assign credit applicants to one of two groups: a 'good credit' group that is likely to repay the financial obligation or a 'bad credit' group that should be denied credit because of a high likelihood of defaulting on the financial obligation (Henley & Hand, 1996).

An appropriate automatic evaluation of the credit applicants offers several important advantages: the cost of credit analysis is reduced, cash flow is improved, faster credit decisions are enabled, the losses are reduced, a closer monitoring of existing accounts is possible, and prioritizing collections are allowed (West, 2000). In this sense, Rhyne and Christen (1999) suggest that credit scoring is one of the most important uses of technology that may affect management of MFIs, and Schreiner (2004) claims that experiments made in Bolivia and Colombia showed that the implementation of credit scoring improved the judgment of credit risk and thus cut, in more than $75,000 per year, costs of MFIs. Nevertheless, and unlike modeling in financial institutions, credit scoring algorithms in microfinance sector have been mostly based on classical statistical techniques, mainly linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression (LR). Some references in this sense are Vigano (1993), Sharma and Zeller (1997), Reinke (1998), Zeller (1998), Vogelgesang (2003), Kleimeier and Dinh (2007), Rayo, Lara, and Camino (2010). However, several authors, for example, Reichert, Cho, and Wagner (1983) and Karels

* Corresponding author. Tel.: +34 954 559 875; fax: +34 954 557 570.
E-mail address: aj_blanco@us.es (A. Blanco-Oliver).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.031

and Prakash (1987), point out that basic assumptions of LDA and QDA are often violated when applied to credit scoring problems. Other problems usually appearing in credit scoring data sets are the mixed nature of the data (quantitative and qualitative) and the high non-linearity in the association between the target variable and the predictors.

These problems can be faced with Statistical Learning algorithms. Statistical Learning is a framework for machine learning with a strong statistical basis. As it is remarked by Hastie, Tibshirani, and Friedman (2001), a related topic, data mining, is an important element of Statistical Learning. Both terms can be considered as parts of a wider process that was termed Knowledge Discovery from Data (KDD) by Fayad, Piatetsky-Shapiro, and Smith (1996), oriented to identify patterns in data sets.

There are many papers providing empirical evidences supporting these alternative algorithms in credit scoring. West (2000) developed several models applying various kinds of neural networks and he compared them with the classical statistical models (LDA and LR) and some non-parametric methods, such as k-nearest neighbor, kernel density and classification and regression trees (CART). Lee, Chiu, Lu, and Chen (2002) developed a two-stage hybrid credit scoring model using multilayer perceptrons (MLP) and multivariate adaptive regression splines (MARS). Multiple discriminant analysis (MDA) and MLP were compared in Malhotra and Malhotra (2003) to identify potential loans, revealing a better performance for MLP model. Kim and Sohn (2010) implemented support vector machine (SVM) models to predict the default of funded SMEs, comparing their performance with the MLP and LR models. Ince and Aktan (2009) compare the performance of several credit scoring models applied to credit card data set from a Turkish bank. These authors use four statistical methods: multiple discriminant analysis, logistic regression, artificial neural networks (ANN) and classification and regression trees, and suggest that CART obtain the best accuracy performance, following of LR, MDA and ANN.

However, similar works in the microfinance field are still expected to be developed. Following this research line, the main goal of this paper is precisely to build a wide set of credit scoring models for the microfinance institutions inside Statistical Learning framework. An empirical scheme has been adopted for this research, accessing to information of almost 5500 microcredits from a Peruvian Microfinance Institution. This data set was used to build and compare the following supervised classification rules to decide between default and non-default categories: linear and quadratic discriminant analysis, logistic regression, multilayer perceptron, support vector machines, classification trees, and three ensemble methods (bagging, random forests and boosting). According to Witten and Frank (2005), the different data mining methods correspond to different concept description spaces searched with different schemes. Thus, different description languages and search procedures serve some problems well and other problems badly, and that is the cause of the necessity to perform a careful comparison of different data mining techniques.

These classification models are freely available in the R system (R Development Core Team., 2012) which also provides the user with a powerful statistical programming language. Ihaka and Gentleman (1996) present an introduction to the main characteristics of the R system.

The remainder of the paper proceeds as follows. In Section 2, details of the analyzed data set are presented, including a detailed examination of the available variables. Classification models are presented from the point of view of the currently available R implementations in Section 3, where several practical questions associated with their use are also analyzed. In Section 4, the results for the different models are presented and a comparison of them is made. Finally, Section 5 provides the main conclusions of this study.

2. Data description

2.1. The data set

A data set of microcredits from a Peruvian Microfinance Institutions (*Edpyme Proempresa*) has been analyzed. It contains customer information during the period 2003–2008 related to: (a) personal characteristics of borrowers (marital status, sex, etc.); (b) economic and financial ratios of their microenterprise; (c) characteristics of the current financial operation (type interest, amount, etc.); (d) variables related to the macroeconomic context; and (e) any delays in the payment of a fee of microcredit. A previous cleaning of the data set was performed to improve its quality, and therefore abnormal cases, which had the top 1% and the bottom 1% of each variable, were removed. After eliminating missing and abnormal cases, 5451 cases remained. Among them, 2673 (49.03%) were default cases and 2778 (50.97%) were non-default cases. In line with other studies (for example, Schreiner, 2004), a credit is defined as default when it shows a delay in the payment of at least fifteen days.

To perform an appropriate comparison of the classification models the final data set was randomly split into two subsets; a training set of 75% and a test set of 25%. The test sample contains a total of 1363 cases (51.80% failed and 48.20% non-failed). The configuration of parameters of each model was performed through a 10-fold cross-validation procedure, as it will be described in Section 3. Our paper follows the extensive discussion in Hastie et al. (2001) regarding the mechanisms for an appropriate fitting and comparison of classification rules in the Statistical Learning framework.

2.2. Description of input variables

Tables 1–3 show the input variables used in this study. These tables also show the expected sign of the relationship between each input variable and the probability of default. Numerous qualitative variables are considered, following suggestions as Schreiner (2004), who claims that the inclusion of qualitative variables improves the prediction power of models. Moreover, since the default of borrowers has a close relationship with the general economic situation, variables linked to the macroeconomic context are also considered as input variables.

The absence of variables with information about the economic cycle has historically implied a major limitation of financial distress models. Furthermore, the macroeconomic environment is a key factor that directly affects the payment behavior of any borrower. For this reason, the following macroeconomic indicators were computed (Table 3): $\Delta VM_{i,j} = (VM_{i+j} - VM_i)/VM_i$, where $\Delta VM_{i,j}$ is the variation rate of the considered macroeconomic variable VM, i is the moment of the granting of the loan and j is microcredit duration.

Table 1
Description of predictor variables: financial ratios.

Variable	Description	Expected sign
R1	Assets rotation: income sales/total assets	–
R2	Productivity: gross utility/operating costs	–
R3	Liquidity: cash/total asset liquidity	–
R4	Liquidity rotations: cash/income sales × 360	+
R5	Leverage1: total liabilities/(total liabilities + shareholders' total equity)	+
R6	Leverage2: total liabilities/shareholders' equity	+
R7	ROA: Net income/total assets	–
R8	ROE: Net income/shareholders' equity	–