Expert Systems with Applications 36 (2009) 4736–4744

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Consumer credit scoring models with limited data

Maja Šušteršič [a,*], Dušan Mramor [b], Jure Zupan [c]

[a] Petrol d.d., Ljubljana, Dunajska c. 50, 1000 Ljubljana, Slovenia
[b] Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia
[c] National Institute of Chemistry, Ljubljana, Hajdrihova ul. 19, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Consumer credit scoring
Neural networks
Genetic algorithm
Principle component analysis
Variable selection

ABSTRACT

In this paper we design the neural network consumer credit scoring models for financial institutions where data usually used in previous research are not available. We use extensive primarily accounting data set on transactions and account balances of clients available in each financial institution. As many of these numerous variables are correlated and have very questionable information content, we considered the issue of variable selection and the selection of training and testing sub-sets crucial in developing efficient scoring models. We used a genetic algorithm for variable selection. In dividing performing and nonperforming loans into training and testing sub-sets we replicated the distribution on Kohonen artificial neural network, however, when evaluating the efficiency of models, we used *k*-fold cross-validation. We developed consumer credit scoring models with error back-propagation artificial neural networks and checked their efficiency against models developed with logistic regression. Considering the dataset of questionable information content, the results were surprisingly good and one of the error back-propagation artificial neural network models has shown the best results. We showed that our variable selection method is well suited for the addressed problem.

1. Introduction

Financial institutions manage credit risks for businesses and consumers differently. In spite of the fact that procedures for granting loans to businesses are less universal, quantitative business credit scoring models were developed first (Altman, 1968; Beaver, 1966) mainly due to a wider availability of company data. There has been an impressive development from their first introduction to their later forms (Altman, 1993; Goonatilake & Treleaven, 1995; Hand, 1998; Trippi & Turban, 1996). In the past, due to the limited number of usually standardized types of consumer loans and scarce availability of data financial institutions predominantly used simple subjective qualitative methods to evaluate creditworthiness of consumer loan applicants (i.e. Sinkey, 1992).[1]. Quantitative consumer credit scoring models were developed much later than those for business credit mainly due to the problem of availability of data. In many countries legal (privacy protection) and other reasons prevented the buildup of publicly available databases. Data were limited to the own databases of financial institutions. Nowadays, some data are publicly available in several countries and financial institutions and researchers have developed many different quantitative credit scoring techniques.

Classical statistical methods that are used to develop credit scoring models are linear discriminant analysis, linear regression, logit, probit, tobit, binary tree and minimum method (Baesens, Setiono, Mues, & Vanthienen, 2003a; Baesens et al., 2003b; Thomas, 1998; West, 2000). The two most commonly used are linear discriminant method (LDA) and logistic regression (Baesens et al., 2003b; Desai, Crook, & Overstreet, 1996; Lee & Chen, 2005; Lee, Chiu, Lu, & Chen, 2002; Thomas, 2000; West, 2000). The weakness of the linear discriminant analysis is the assumption of linear relationship between variables, which is usually nonlinear and the sensitivity to the deviations from the multivariate normality assumption. The logistic regression is predicting dichotomous outcomes and linear relationship between variables in the exponent of the logistic function, but does not require the multivariate normality assumption. Because of the linear relationship between variables both LDA and logistical regression are reported to have a lack of accuracy (Thomas, 2000; West, 2000). On the other hand there are also studies showing (Baesens et al., 2003b), that most of the consumer credit scoring datasets are only weakly nonlinear and because of that LDA and logistical regression gave good performance.

There are also more sophisticated models known as artificial intelligence: expert systems, fuzzy systems, neural networks and genetic algorithms. Among these the neural networks are very

* Corresponding author. Tel.: +386 31 66 88 00/1 4714423; fax: +386 1 4366865.
E-mail address: maja.sustersic@petrol.si (M. Šušteršič).

[1] At this stage of the development the use of qualitative data was logical. It was shown that even for micro-companies their accounting data do not contain much of information, that could be used for bankruptcy prediction (see Mramor & Valentincic, 2003).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.016

promising (Goonatilake & Treleaven, 1995) and the alternative to the LDA and logistic regression, due to the possible complex nonlinear relationship between variables. In the literature in most cases of credit scoring problems the neural networks are more accurate than LDA and logistic regression (Desai et al., 1996; Jensen, 1996; Lee et al., 2002; Piramuthu, Shaw, & Gentry, 1994; Richeson, Zimmermann, & Barnett, 1996; West, 2000). The neural networks have their weaknesses in their long training process, and after obtaining the optimal network's architecture, the model acts as a "black box" and there is not easy to identify the relative importance of potential input variables. One can find also a few studies with genetic algorithms (Kim & Sohn, 2004; Walker, Haasdijk, & Gerrets, 1995), but in the last years the hybrid systems seem to be the most promising (Hsieh, 2005; Lee & Chen, 2005; Lee et al., 2002).

The datasets for the mentioned studies were usually collected by credit unions. They consisted of a relatively small number of variables: from 5 to 20. As these were the only available variables and as their selection was done by credit unions on the basis of past consumer loan experiences of financial institutions, researchers did not regard selection of variables as a crucial step of the model development. Because of the relative small number of variables they used all of them or their selection was based mainly on classical statistical methods like *t*-test or chi-square-test (Avery, Calem, & Canner, 2004; Kim & Sohn, 2004), multivariate adaptive regression splines (Lee & Chen, 2005) or artificial neural network (Glorfeld & Hardgrave, 1996; Hsieh, 2005; West, 2000). The weaknesses of the statistical methods usually appear when multicollinearity between a large number of variables exists and in the case of neural networks in their time consuming process especially when large number of variables exists. The highest number of variables that we found in the literature was 57 (Jacobson & Roszbach, 2003). The authors included publicly available or governmentally supplied variables, such as sex, citizenship, marital status, postal code, taxable income, taxable wealth, house ownership and variables reported by the Swedish banks like the total number of inquiries made about an individual and the number of unsecured loans and the total amount of unsecured loans. Most of the variables (41) were not used for the development of the model, because either they lacked a bivariate relation with dependent variable or displayed extremely high correlation with another variable that measured approximately the same thing but had greater explanatory power.

Contrary to the previous research, we developed consumer credit scoring models for financial institutions where data that were used in previous research are not available. We base our model selection primarily on accounting data on transactions and account balances of clients that are readily available in each financial institution. Therefore, the number of input variables is in our study larger than in other studies, many of the variables are highly correlated and for a great majority we do not know how much creditworthiness information (if any) they contain as they are currently not used in credit assessments. Hence, the issue of variable selection is a crucial and a challenging problem to solve before different credit scoring techniques are used to develop the best performing models. As it is known, different variable selection methods give different results on the same dataset. To increase the quality of variable selection we compare a statistical principal component analysis (PCA) with a nonstatistical genetic algorithm. For the genetic algorithm we divided performing and nonperforming loans into training and testing sub-sets randomly and in such a way, that both types of loans proportionally covered the whole Kohonen neural network, however, when evaluating the efficiency of models, we used *k*-fold cross-validation (Hsieh, 2005). The efficiency of models using only principle component variables was smaller.

We developed consumer credit scoring models with logistic regression and error back-propagation artificial neural networks.[2] Considering questionable information content of the dataset that we use, the results of the models are surprisingly good – prediction power of our models is approximately the same or even better than those of the latest studies. Error back-propagation neural network model using variables selected by genetic algorithm is showing the best results.

We start with short explanations of the methods, the research procedure and the data used in this study. The description of the selection of variables and the division of the master dataset into training and testing sub-sets follows. Different models and their results concerning efficiency of consumer loans classification are presented next and followed by the conclusion.

2. Principal component analysis, genetic algorithm and neural networks

Principal component analysis (PCA) is an effective transformation method for reduction of a large number of correlated variables where variable selection is hard to achieve. Namely, the result of PCA is a set of new independent variables that can be directly used by credit scoring techniques. PCA is a statistical method used frequently for reducing the dimensionality of a given dataset of correlated variables while maintaining as much of the variables' variability as possible. This efficient reduction of the number of variables is achieved by obtaining orthogonal linear combinations of the original variables – the so-called principal components (PCs). This is possible with a transformation of the co-ordinate system to a new one. The transformation is done by rotation of the old co-ordinate system into the new one in such a way that the most of the relevant information is collected around smaller number of new axes (PCs). The first principal component PC_1 preserves most of the remaining variability in the original variables, the second component PC_2 preserves the second most variability existing in the original variables, and so on. Each PC is an eigenvector of the variance–covariance matrix of the original variables. This analysis provides two important outputs: the percentage of variance explained by *i*th principal component PC_i and the correlations between each PC and the original variables. The first one is computed by dividing the eigenvalue associated to the corresponding PC by the total sum of the eigenvalues. The first output provides the importance of the component in the terms of the variability of the original variables (see Godoy & Stiglitz, 2006).

Genetic algorithm (GA) is an efficient optimization procedure. The basic principle of the genetic algorithm is inspired by the mechanisms of biological evolution. The main idea of a genetic algorithm is to start with a population of possible solutions to a given problem, and to continue by a production of series of new generations of many different solutions, assuming to find better and better ones. Genetic algorithm operates through a simple cycle consisting of the following four stages: creation of the population, evaluation, selection, and reproduction in which the last three stages are cycled until no more improvement in the evaluation stage is detected.[3]

The starting point of genetic algorithm is the creation of a population of "members" which represent candidate solutions to the problem being solved. The members (candidate solutions) are evaluated by the fitness function. This assesses the degree to which the solutions are good at solving the given problem. The value returned by this fitness function is used for the selection of members as "parents" for the production of the next generation (population)

[2] We decided to examine these two types of models, because they were most promising according to previous research (see Sustersic, 2001).

[3] For more details and applications for finance and business see Goonatilake and Treleaven (1995) and Hand (1998).