Expert Systems with Applications 39 (2012) 8071–8078

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data

Francisco Louzada [a,*], Paulo H. Ferreira-Silva [b], Carlos A.R. Diniz [b]

[a] Universidade de São Paulo, SME-ICMC, São Carlos, Brazil
[b] Universidade Federal de São Carlos, DEs, São Carlos, Brazil

ARTICLE INFO

Keywords:
Classification models
Naive logistic regression
Logistic regression with state-dependent sample selection
Performance measures
Credit scoring

ABSTRACT

Statistical methods have been widely employed to assess the capabilities of credit scoring classification models in order to reduce the risk of wrong decisions when granting credit facilities to clients. The predictive quality of a classification model can be evaluated based on measures such as sensitivity, specificity, predictive values, accuracy, correlation coefficients and information theoretical measures, such as relative entropy and mutual information. In this paper we analyze the performance of a naive logistic regression model (Hosmer & Lemeshow, 1989) and a logistic regression with state-dependent sample selection model (Cramer, 2004) applied to simulated data. Also, as a case study, the methodology is illustrated on a data set extracted from a Brazilian bank portfolio. Our simulation results so far revealed that there is no statistically significant difference in terms of predictive capacity between the naive logistic regression models and the logistic regression with state-dependent sample selection models. However, there is strong difference between the distributions of the estimated default probabilities from these two statistical modeling techniques, with the naive logistic regression models always underestimating such probabilities, particularly in the presence of balanced samples.

1. Introduction

The proper classification of applicants is of vital importance for determining the granting of credit facilities. Historically, statistical classification models have been used by financial institutions as a major tool to help on granting credit to clients.

The consolidation of the use of classification models occurred in the 90s, when changes in the world scene, such as deregulation of interest rates and exchange rates, increase in liquidity and in bank competition, made financial institutions more and more worried about credit risk, i.e., the risk they were running when accepting someone as their client. The granting of credit started to be more important in the profitability of companies in the financial sector, becoming one of the main sources of revenue for banks and financial institutions in general. Due to this fact, this sector of the economy realized that it was highly recommended to increase the amount of allocated resources without losing the agility and quality of credits, at which point the contribution of statistical modeling is essential.

Classification models for credit scoring are based on databases of relevant client information, with the financial performance of clients evaluated from the time when the client–company relationship began as a dichotomic classification. The goal of credit scoring models is to classify loan clients to either good credit or bad credit (Lee, Chiu, Lu, & Chen, 2002), predicting the bad payers (Lim & Sohn, 2007).

In this context, discriminant analysis, regression trees, logistic regression, logistic regression with state-dependent sample selection and neural networks are among the most widely used classification models. In fact, logistic regression is still very used in building and developing credit scoring models (Caouette, Altman, & Narayanan, 1998; Desai, Crook, & Overstreet, 1996; Hand & Henley, 1997; Sarlija, Bensic, & Bohacek, 2004). Generally, the best technique for all data sets does not exist but we can compare a set of methods using some statistical criteria. Therefore, the main thrust of this paper is to investigate and compare the performance of the naive logistic regression (Hosmer & Lemeshow, 1989) and the logistic regression with state-dependent sample selection (Cramer, 2004) using performance measures, in terms of a simulation study. The idea is to analyze the impact of disproportional samples on credit scoring models. Logistic regression with state-dependent sample selection is a statistical modeling technique used in cases where the sample considered to develop a model, i.e. the selected sample, contains only a portion, usually small, of the individuals who make up one of two study groups, in general the most frequent group. In credit scoring, for instance, the group of good payers is expected to be the predominant group. In short, this recent technique makes a correction in the estimated default probability from a naive logistic regression model (Cramer, 2004).

* Corresponding author. Tel.: +55 16 3373 6614.
E-mail address: louzada@icmc.usp.br (F. Louzada).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.134

1.1. Literature review

The first credit scoring models were developed around 1950 and 1960, and the methods applied in this kind of problem referred to methods of discrimination suggested by Fisher (1936), where the models were based on his discriminant function. As Thomas (2000) points out, David Durand, in 1941, was the first one which recognized that the discriminant analysis technique, invented by Fisher in 1936, could be used to separate good credits from the bad ones. According to Kang and Shin (2000), Durand presented a model which attributed weights for each variable using discriminant analysis. Thus Fisher's approach can be seen as the starting point for developments and modifications of the methodologies used for granting of credit until today, where statistical techniques, such as discriminant analysis, regression analysis, probit analysis and naive logistic regression, have been used and examined (Banasik, Crook, & Thomas, 2001; Boyes, Hoffman, & Low, 1989; Greene, 1998; Orgler, 1971; Sarlija et al., 2004; Steenackers & Goovaerts, 1989). Particularly, considering state-dependent sample selection (Cramer, 2004) in order to make a correction in the estimated default probability from a credit scoring model.

The predictive quality of a credit scoring model can be evaluated based on measures such as sensitivity, specificity, correlation coefficients and information measures, such as relative entropy and mutual information (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000).

Generally, there is no overall best statistical technique used in building credit scoring models, so that the choice of a particular technique depends on the details of the problem, such as the data structure, the features used, the extent to which it is possible to segregate the classes by using those features, and the purpose of the classification (Hand & Henley, 1997). Most studies that made a comparison between different techniques tried to discover that the most recent/advanced credit scoring techniques, such as neural networks and fuzzy algorithms are better than the traditional ones. Nevertheless, the more simple classification techniques, such as linear discriminant analysis and naive logistic regression, have a very good performance, which is in majority of the cases not statistically different from other techniques (Baesens et al., 2003; Hand, 2006).

1.2. Paper organization and main results

This paper is organized as follows. Section 2 describes the two commonly used statistical techniques in building credit scoring models: the naive logistic regression and the logistic regression with state-dependent sample selection. Section 3 presents some useful measures that are used to analyze the predictive capacity of a classification model. Section 4 describes the details of a simulation study performed in order to compare the techniques of interest. In Section 5 the methodology is illustrated on a real data set from a Brazilian bank portfolio. Finally, Section 6 concludes the paper with some final comments.

Our empirical results reveal that there is difference between the distributions of estimated default probabilities by the use of these two techniques, especially in the cases where the sample used for building the model, i.e. the training sample, is balanced. However, there is no significant difference in predictive capacity among the models adjusted using different fractions of individuals of the most frequent group.

2. Credit scoring models

In this Section, the two statistical techniques used for building credit scoring are described. The first model is the naive logistic regression model, which was proposed by Hosmer and Lemeshow (1989) and is widely used for credit scoring modeling. The second model is the logistic regression with state-dependent sample selection model (Cramer, 2004), which differently from the first one, takes into account the principles of sample selection and their application to a logistic model.

2.1. Naive logistic regression

Naive logistic regression is a widely used statistical modeling technique in which the response variable, i.e. the outcome is binary (0, 1) and can thus be used to describe the relationship between the occurrence of an event of interest and a set of potential predictor variables. In the context of credit scoring, the outcome corresponds to the credit performance of a client during a given period of time, usually 12 months. A set of individual characteristics, such as marital status, age and income, as well as information about his credit product in use, such as number of parcels, purpose and credit value, are observed at the time the clients apply for the credit.

Let us consider a large sample of observations with predictors x_i and binary (0, 1) outcomes Y_i. Here, the event $Y_i = 1$ represents a bad credit, while the complement $Y_i = 0$ represents a good credit. The model specifies that the probability of i being a bad credit as a function of the x_i is given by,

$$P(Y_i = 1|x_i) = p(\boldsymbol{\beta}, x_i) = p_i. \tag{1}$$

In the case that Eq. (1) is a naive logistic regression model, p_i is given by,

$$p_i = \frac{\exp(x_i'\boldsymbol{\beta})}{1 + \exp(x_i'\boldsymbol{\beta})}. \tag{2}$$

(see Hosmer & Lemeshow, 1989). Thus the objective of a naive logistic regression model in credit scoring is to determine the conditional probability of a specific client belonging to a class, for instance, the bad payers class, given the values of the independent variables of that credit applicant (Lee & Chen, 2005).

2.2. Logistic regression with state-dependent sample selection

Now let us consider the situation where the event $Y_i = 1$ represents a bad credit but it has a low incidence, while the complement $Y_i = 0$ represents a good credit but it is abundant.

Suppose we wish to estimate β from a selected sample, which is obtained by discarding a large part of the abundant zero observations for reasons of convenience. Assume also that the overall sample, hereafter full sample, is a random sample with sampling fraction α and that only a fraction γ of the zero observations, taken at random, is maintained. The probability that the element i has $Y_i = 1$ and it is included in the sample is given by αp_i, but for $Y_i = 0$ it is given by $\gamma\alpha(1 - p_i)$, where p_i is calculated from Eq. (2). Then, the probability that an element of the selected sample has $Y_i = 1$ is given by,

$$\tilde{p}_i = \frac{p_i}{p_i + \gamma(1 - p_i)}. \tag{3}$$

The sketch of the proof of Eq. (3) is given in Appendix A.

2.3. Estimation procedure

The likelihood of the observed sample can be written in terms of $\tilde{p}_i$ as follows,

$$\log L = \sum Y_i \log \tilde{p}_i(\boldsymbol{\beta}, x_i, \gamma) + (Y_i - 1)\log \tilde{p}_i(\boldsymbol{\beta}, x_i, \gamma). \tag{4}$$

If the selected sample is drawn from a known full sample (as here) γ is always known. Thus the parameters of any specification of p_i from Eq. (1) can be estimated from the selected sample by standard maximum likelihood methods. In the special case that Eq. (1) is a naive logistic regression model, $\tilde{p}_i$ is given by,