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,

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

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

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

(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,

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

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,

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

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,

$$\tilde{p}_i = \frac{\exp(x_i'\boldsymbol{\beta})}{\exp(x_i'\boldsymbol{\beta}) + \gamma} = \frac{1/\gamma \cdot \exp(x_i'\boldsymbol{\beta})}{1 + 1/\gamma \cdot \exp(x_i'\boldsymbol{\beta})} = \frac{\exp(x_i'\boldsymbol{\beta} - \ln\gamma)}{1 + \exp(x_i'\boldsymbol{\beta} - \ln\gamma)}. \quad (5)$$

Thus the $\tilde{p}_i$ of the selected sample also obey a naive logistic regression model, and besides the intercept the same parameters β apply as in the full sample. If it is needed, the full sample intercept can be easily recovered by adding ln γ to the intercept of the selected sample.

3. Model validation

After building a statistical model, it is important we evaluate it. Particularly, we can say that a good model is one which produces scores that can distinguish good from bad credits, since the objective is to previously identify such groups and treat them differently considering distinct relationship policies (Thomas, Edelman, & Crook, 2002).

The evaluation performance of a model can be done by a comparison between its prediction and the real classification of a client. The true condition of a client is generally known and is present as basic information in the database. There are therefore four possible results for each client: (a) The result given by the classification model is a true-positive (*TP*). In other words, the model indicates that the client is a bad payer, which, in fact, he actually is; (b) The result given by the classification model is a false-positive (*FP*). In other words, the model accuses the client of being a bad payer, which, in fact, he actually is not; (c) The result given by the classification model is a false-negative (*FN*). In other words, the model classifies him as a good payer, which, in fact, he actually is not; (d) The result given by the classification model is a true-negative (*TN*). In other words, the model classifies the client as a good payer, which, in fact, he actually is.

These four possible situations above are summarized in Table 1. Two of these results, *TP* and *TN*, can be seen as indicating that the proposed model "got it right" and are important measures of financial behavior. Of the other results that can be considered to indicate that the proposed model "got it wrong", the one that gives a *FN* classification is often the most important and deserves the most attention. If the classification given by the proposed model is negative, and the client is, in fact, a bad payer, granting to this client the credit requested can end up with the client defaulting, which, added to other cases of defaults by bad payers, can result in very high overall default rates for the company, and a real threat to its financial sustainability. A *FP* result can also be negative for the company, since it may lose out by not approving credit for a client that would not, in fact, cause any default.

3.1. Some performance measures

The predictive capacity of a classification model is related to its performance measures, which can be calculated from Table 1. Among them we can cite sensitivity, specificity, positive and negative predictive values, accuracy, Matthews correlation coefficient, approximate correlation, relative entropy and mutual information.

3.1.1. Sensitivity, specificity and accuracy

Sensitivity (*SEN*) is defined as the probability that the classification model will produce a positive result, given that the client is a defaulter. In other words, sensitivity corresponds to the proportion of bad payers that are correctly classified by the classification model, and is given by,

$$SEN = P(M_+|D_+) = \frac{TP}{TP + FN}. \quad (6)$$

Specificity (*SPE*) is defined as the probability that a classification model will produce a negative result for a client who is not a defaulter. In other words, specificity represents the proportion of good payers that are correctly classified by the classification model, and is calculated as follows,

$$SPE = P(M_-|D_-) = \frac{TN}{TN + FP}. \quad (7)$$

Thus, a model with high sensitivity rarely fails to detect clients with the characteristic of interest (default) and a model with high specificity rarely classifies a client without this feature as a bad payer.

The Accuracy (*ACC*) is defined as the proportion of successes of a model, i.e. the proportion of true-positives and true-negatives in relation to all possible outcomes. Accuracy is then given by,

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}. \quad (8)$$

Table 1
Definitions used to validate classification models that produce dichotomized responses.

Result	Real	
From classification model	Positive (D_+)	Negative (D_-)
Positive (M_+)	True-positive (*TP*)	False-positive (*FP*)
Negative (M_-)	False-negative (*FN*)	True-negative (*TN*)

3.1.2. Positive and negative predictive values

The positive predictive value (*PPV*) of a model is defined as the proportion of bad payers identified as such by the model, while the negative predictive value (*NPV*) of a model is defined as the proportion of good payers identified as such by the model.

The positive and negative predictive values can be estimated directly from the sample, using the equations,

$$PPV = P(D_+|M_+) = \frac{TP}{TP + FP} \quad (9)$$

and

$$NPV = P(D_-|M_-) = \frac{TN}{FN + TN}, \quad (10)$$

respectively (Dunn & Everitt, 1995).

The more sensitive the model, the greater its negative predictive value. That is, the greater is the assurance that a client with a negative result is not a bad payer. Likewise, the more specific the model, the greater its positive predictive value. That is, the greater the guarantee of an individual with a positive result being a bad payer.

3.1.3. Matthews correlation coefficient

The correlation coefficient proposed by Matthews (1975) uses all four values (*TP*, *FP*, *TN*, *VN*) of a confusion matrix in its calculation. In general, it is regarded as a balanced measure which can be used even if the studied classes are of very different sizes. The Matthews correlation coefficient (*MCC*) returns a value between -1 and +1. A value of +1 represents a perfect prediction, i.e. total agreement, 0 a completely random prediction and -1 an inverse prediction, i.e. total disagreement. The *MCC* is given by,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (11)$$

(see Baldi et al., 2000).

If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; which results in a *MCC* of zero. There are situations, however, where the *MCC* is not a reliable performance measure. For instance, the *MCC* will be relatively high in cases where a classification model gives very few or no false-positives, but at the same time very few true-positives.

Table 2
95% confidence intervals for the performance measures, when we use the naive logistic regression technique.

Measures	K = 1	K = 3	K = 9
SEN	[0.8071; 0.8250]	[0.5877; 0.6008]	[0.3249; 0.3307]
SPE	[0.8187; 0.8334]	[0.9331; 0.9366]	[0.9768; 0.9777]
ACC	[0.8177; 0.8242]	[0.8123; 0.8194]	[0.8101; 0.8155]
PPV	[0.8179; 0.8400]	[0.8247; 0.8359]	[0.8258; 0.8341]
NPV	[0.8004; 0.8250]	[0.8047; 0.8170]	[0.8075; 0.8145]
MCC	[0.6354; 0.6485]	[0.5787; 0.5866]	[0.4404; 0.4439]
I	[0.3149; 0.3294]	[0.2419; 0.2475]	[0.1206; 0.1214]
H	[0.9989; 1.0000]	[0.9281; 0.9385]	[0.8105; 0.8217]
IC	[0.3149; 0.3295]	[0.2585; 0.2661]	[0.1469; 0.1493]
ACP	[0.8177; 0.8243]	[0.7908; 0.7945]	[0.7359; 0.7371]
AC	[0.6354; 0.6485]	[0.5815; 0.5891]	[0.4718; 0.4742]

Table 3
95% confidence intervals for the performance measures, when we investigate the performance of the model in a balanced test sample (50% good payers and 50% bad payers).

Measures	K = 1	K = 3	K = 9
SEN	[0.8255; 0.8499]	[0.8332; 0.8456]	[0.8343; 0.8431]
SPE	[0.7983; 0.8182]	[0.8022; 0.8134]	[0.8045; 0.8122]
ACC	[0.8219; 0.8243]	[0.8226; 0.8242]	[0.8232; 0.8239]
PPV	[0.8082; 0.8196]	[0.8104; 0.8172]	[0.8118; 0.8163]
NPV	[0.8242; 0.8416]	[0.8297; 0.8387]	[0.8306; 0.8368]
MCC	[0.6438; 0.6493]	[0.6453; 0.6490]	[0.6466; 0.6482]
I	[0.3241; 0.3305]	[0.3258; 0.3301]	[0.3273; 0.3292]
H	1	1	1
IC	[0.3241; 0.3305]	[0.3258; 0.3301]	[0.3273; 0.3292]
ACP	[0.8219; 0.8246]	[0.8226; 0.8245]	[0.8233; 0.8241]
AC	[0.6438; 0.6493]	[0.6453; 0.6490]	[0.6466; 0.6482]

Table 4
95% confidence intervals for the performance measures, when we use the logistic regression with state-dependent sample selection technique.

Measures	K = 1	K = 3	K = 9
SEN	[0.8061; 0.8221]	[0.5870; 0.6008]	[0.3258; 0.3278]
SPE	[0.8206; 0.8333]	[0.9330; 0.9366]	[0.9773; 0.9775]
ACC	[0.8173; 0.8241]	[0.8120; 0.8193]	[0.8111; 0.8127]
PPV	[0.8225; 0.8392]	[0.8237; 0.8365]	[0.8306; 0.8321]
NPV	[0.7989; 0.8211]	[0.8045; 0.8180]	[0.8088; 0.8106]
MCC	[0.6348; 0.6484]	[0.5779; 0.5859]	[0.4407; 0.4426]
I	[0.3143; 0.3294]	[0.2419; 0.2473]	[0.1205; 0.1212]
H	[0.9990; 1.0000]	[0.9271; 0.9389]	[0.8168; 0.8195]
IC	[0.3143; 0.3295]	[0.2578; 0.2655]	[0.1471; 0.1484]
ACP	[0.8174; 0.8242]	[0.7904; 0.7942]	[0.7359; 0.7367]
AC	[0.6348; 0.6484]	[0.5808; 0.5884]	[0.4718; 0.4735]

3.1.4. Approximate correlation

Burset and Guigó (1996) defined an approximate correlation measure to compensate for a declared problem with the *MCC*. That is, it is not defined if any of the four sums *TP* + *FN*, *TP* + *FP*, *TN* + *FP*, or *TN* + *FN* is zero, e.g. when there are no positive predictions. They use the average conditional probability (*ACP*) which is defined as,

$$ACP = \frac{1}{4}\left[\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right], \tag{12}$$

if all the sums are non-zero; otherwise, it is the average over only those conditional probabilities that are defined. Approximate correlation (*AC*) is a simple transformation of the *ACP* given by,

$$AC = 2 \times (ACP - 0.5). \tag{13}$$

It returns a value between −1 and +1 and has the same interpretation as the *MCC*. Burset and Guigó (1996) observe that the *AC* is close to the real correlation value. However, some authors, like Baldi et al. (2000), do not encourage its use, since there is no simple geometrical interpretation for *AC*.

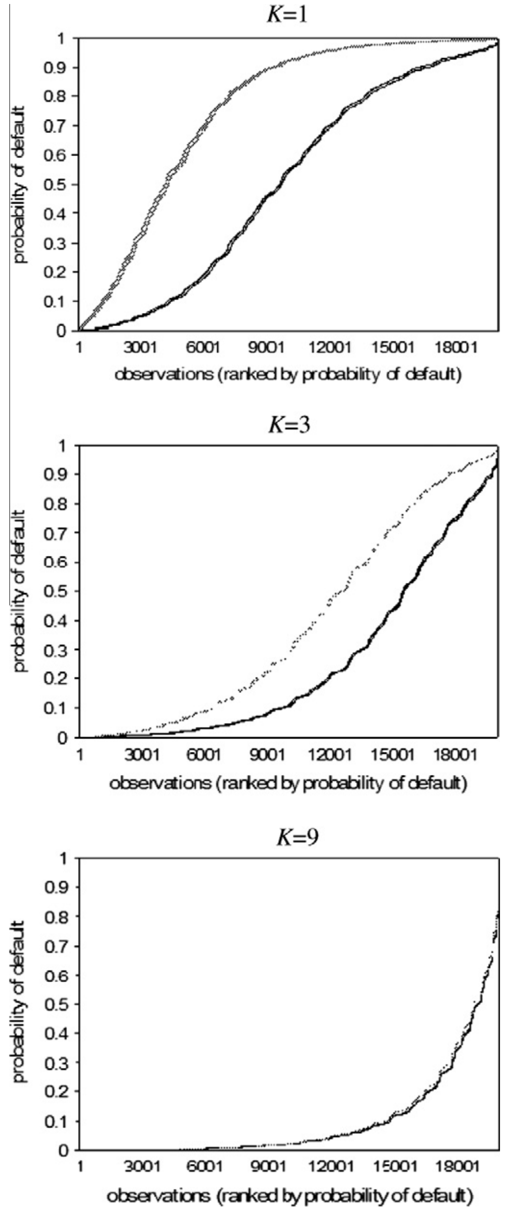

Fig. 1. Estimated cumulative probabilities, where ——— represents the 90% confidence limits for the original probability and ········· the 90% confidence limits for the adjusted one.

3.1.5. Relative entropy and mutual information

Suppose that $\mathbf{D}' = (\mathbf{d_1}, \ldots, \mathbf{d_N})$ is a vector of true conditions of clients and $\mathbf{M}' = (\mathbf{m_1}, \ldots, \mathbf{m_N})$ a vector of predictions from a classification model, both binary (0, 1). The d_i s and m_i s are then equal to 0 or 1, for instance, 0 indicates a good payer and 1 a bad payer. The mutual information between **D** and **M** is measured by,

$$I(\mathbf{D},\mathbf{M}) = -H\left(\frac{TP}{N},\frac{TN}{N},\frac{FP}{N},\frac{FN}{N}\right) - \frac{TP}{N}\log(\bar{d}\bar{m}) - \frac{FN}{N}\log(\bar{d}(1-\bar{m})) - \frac{FP}{N}\log((1-\bar{d})\bar{m}) - \frac{TN}{N}\log((1-\bar{d})(1-\bar{m})) \tag{14}$$

(see Wang, 1994), where N is the sample size, $\bar{d} = (TP + FN)/N, \bar{m} = (TP + FP)/N$ and

$$H\left(\frac{TP}{N},\frac{TN}{N},\frac{FP}{N},\frac{FN}{N}\right) = -\frac{TP}{N}\log\left(\frac{TP}{N}\right) - \frac{TN}{N}\log\left(\frac{TN}{N}\right) - \frac{FP}{N} \times \log\left(\frac{FP}{N}\right) - \frac{FN}{N}\log\left(\frac{FN}{N}\right)$$

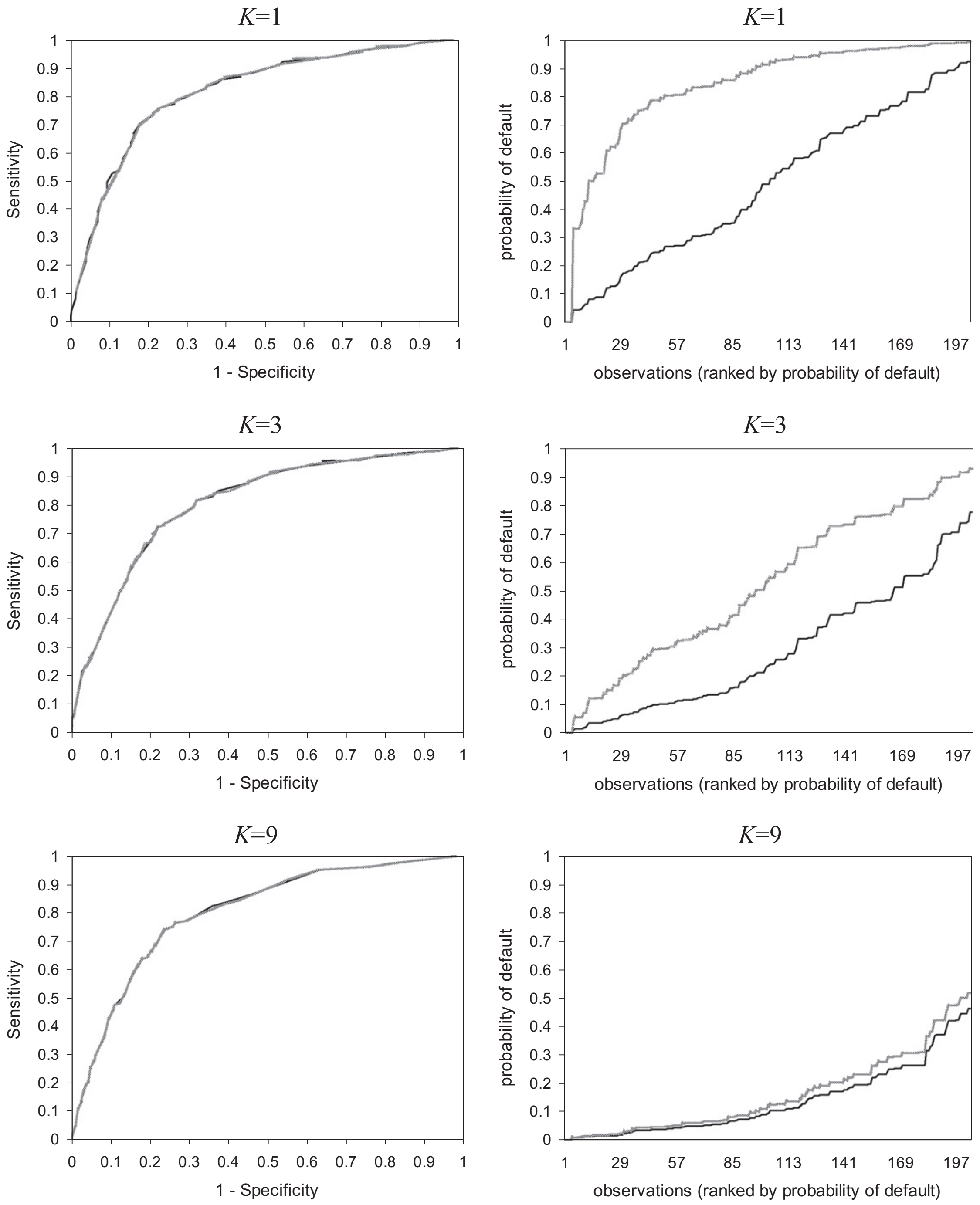

Fig. 2. The ROC curves constructed from the training sample of a bank's actual portfolio (panels on the left) and the corresponding distributions of estimated probabilities in a balanced test sample (panels on the right). To the panels on the left, ——— represents the naive logistic regression model and ········· the logistic regression with state-dependent sample selection one. To the panels on the right, ——— represents the original probability and ········· the adjusted one.

is the usual entropy, whose roots are in information theory (Baldi & Brunak, 1998; Kullback, 1959; Kullback & Leibler, 1986).

The mutual information always satisfies $0 \leqslant I(\mathbf{D}, \mathbf{M}) \leqslant H(\mathbf{D})$, where $H(\mathbf{D}) = -\bar{m}\log\bar{m} - (1-\bar{m})\log(1-\bar{m})$. Thus in the assessment of performance of a classification model, it is habitual to use the normalized mutual information coefficient (Rost & Sander, 1993; Rost, Sander, & Schneider, 1994), which is given by,

$$IC(\mathbf{D}, \mathbf{M}) = \frac{I(\mathbf{D}, \mathbf{M})}{H(\mathbf{D})}. \tag{15}$$

The normalized mutual information satisfies $0 \leqslant IC(\mathbf{D}, \mathbf{M}) \leqslant 1$. If $IC(\mathbf{D}, \mathbf{M}) = 0$, then $I(\mathbf{D}, \mathbf{M}) = 0$ and the prediction is completely random ($\mathbf{D}$ and $\mathbf{M}$ are independent). If $IC(\mathbf{D}, \mathbf{M}) = 1$, then $I(\mathbf{D}, \mathbf{M}) = H(\mathbf{D}) = H(\mathbf{M})$ and the prediction is perfect.

4. Simulation study

In this Section, we present results of the performed simulation study. The simulation study was conducted to compare the performance of the naive logistic regression and logistic regression with

Table 5
The evaluation (performance measures) of the adjusted models in a balanced test sample (50% good payers and 50% bad payers), for different values of *K*, where NLR is the naive logistic regression model and LRSD refers to the logistic regression with state-dependent sample selection model.

Measures	*K* = 1		*K* = 3		*K* = 9	
	NLR	LRSD	NLR	LRSD	NLR	LRSD
SEN	0.7451	0.7451	0.7255	0.7255	0.7745	0.7255
SPE	0.7843	0.7549	0.8137	0.8039	0.7745	0.8137
ACC	0.7647	0.7500	0.7696	0.7647	0.7745	0.7696
PPV	0.7755	0.7525	0.7957	0.7872	0.7745	0.7957
NPV	0.7547	0.7476	0.7477	0.7455	0.7745	0.7477
MCC	0.5298	0.5000	0.5413	0.5310	0.5490	0.5413
I	0.2133	0.1887	0.2236	0.2146	0.2299	0.2236
H	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
IC	0.2133	0.1887	0.2236	0.2146	0.2299	0.2236
ACP	0.7649	0.7500	0.7707	0.7655	0.7745	0.7707
AC	0.5298	0.5000	0.5413	0.5311	0.5490	0.5413

Table 6
The evaluation (performance measures) of the adjusted models in an unbalanced test sample (75% good payers and 25% bad payers), for different values of *K*, where NLR is the naive logistic regression model and LRSD refers to the logistic regression with state-dependent sample selection model.

Measures	*K* = 1		*K* = 3		*K* = 9	
	NLR	LRSD	NLR	LRSD	NLR	LRSD
SEN	0.7451	0.7451	0.7255	0.7255	0.7745	0.7255
SPE	0.7582	0.7451	0.7778	0.7778	0.7418	0.7614
ACC	0.7549	0.7451	0.7647	0.7647	0.7500	0.7525
PPV	0.5067	0.4935	0.5211	0.5211	0.5000	0.5034
NPV	0.8992	0.8976	0.8947	0.8947	0.9080	0.8927
MCC	0.4520	0.4379	0.4575	0.4575	0.4590	0.4392
I	0.1456	0.1373	0.1472	0.1472	0.1525	0.1365
H	0.8113	0.8113	0.8113	0.8113	0.8113	0.8113
IC	0.1794	0.1692	0.1814	0.1814	0.1880	0.1682
ACP	0.7273	0.7203	0.7298	0.7298	0.7311	0.7208
AC	0.4546	0.4407	0.4596	0.4596	0.4622	0.4415

Table 7
The evaluation (performance measures) of the adjusted models in an unbalanced test sample (90% good payers and 10% bad payers), for different values of *K*, where NLR is the naive logistic regression model and LRSD refers to the logistic regression with state-dependent sample selection model.

Measures	*K* = 1		*K* = 3		*K* = 9	
	NLR	LRSD	NLR	LRSD	NLR	LRSD
SEN	0.7451	0.7451	0.7255	0.7255	0.7745	0.7255
SPE	0.7505	0.7353	0.7712	0.7702	0.7331	0.7614
ACC	0.7500	0.7363	0.7667	0.7657	0.7373	0.7578
PPV	0.2492	0.2382	0.2606	0.2596	0.2438	0.2526
NPV	0.9636	0.9629	0.9620	0.9619	0.9670	0.9615
MCC	0.3248	0.3109	0.3325	0.3314	0.3271	0.3228
I	0.0688	0.0640	0.0703	0.0699	0.0715	0.0670
H	0.4690	0.4690	0.4690	0.4690	0.4690	0.4690
IC	0.1467	0.1365	0.1498	0.1490	0.1525	0.1428
ACP	0.6771	0.6704	0.6798	0.6793	0.6796	0.6752
AC	0.3542	0.3408	0.3596	0.3586	0.3592	0.3505

state-dependent sample selection. Here, we considered some circumstances that may arise in the development of credit scoring models, and which involve the number of clients in the training sample and its degree of unbalance.

4.1. General specifications

First, we generated a population of clients of a hypothetical credit institution, with the following composition: 1,000,000 good payers and 100,000 bad payers. The data relating to good payers were generated from a six-dimensional multivariate normal distribution, with mean vector $\boldsymbol{\mu}'_B = (0, \ldots, 0)$ and covariance matrix $4 \times I_6$, where I_6 is the identity matrix of order 6; while the data for bad payers were generated from a six-dimensional multivariate normal distribution, with mean $\boldsymbol{\mu}'_M = \left(1/\sqrt{6}, \ldots, 1/\sqrt{6}\right)$ and covariance matrix I_6 (Breiman, 1998). We categorized the six observed continuous covariates using the quartiles. Afterwards, we took out a stratified random sample (full sample) of 100,000 good payers (10% of the population size of this group) and 10,000 bad payers (10% of the population size of this class) of the generated population. Then the selected samples were obtained by keeping all bad payers of the full sample, plus 10,000 $*$ *K* good payers taken out randomly from the full sample. Here, we studied only the cases where *K* = 1, 3 and 9, which correspond to 10,000, 30,000 and 90,000 good payers in the selected samples, respectively. For each *K*, we performed 1000 simulations, i.e. 1000 samples were obtained for each *K*. In each one of them, the good payers were selected from their group in the full sample via simple random sampling without replacement.

For each obtained sample and for each *K* we applied the following procedures: a naive logistic regression model (Hosmer & Lemeshow, 1989) and a logistic regression with state-dependent sample selection model (Cramer, 2004) were fitted to the data and their predictive capacity was investigated in the training sample by the calculation of their performance measures. However, the evaluation of a model in the training sample produces better results than if assessed in an independent sample, since a model incorporates peculiarities of the sample used for its construction (Abreu, 2004). Thus, we also used a balanced test sample with 10,000 good payers and 10,000 bad payers, taken out randomly from the population, for evaluating the adjusted both investigated models.

After simulating for each *K*, we obtained vectors of length 1000, i.e. vectors with 1000 records for each of the studied performance measures. Then we could obtain 95% confidence intervals for each measure, by recording the 2.5% and 97.5% percentiles of the ordered vectors with components in ascending order.

We calculated the optimal cutoff point for every adjusted model. Finally, we also compared the original probabilities, estimated from the naive logistic regression models, with the adjusted probabilities, predicted from the logistic regression with state-dependent sample selection models, that are the original probabilities corrected by considering the principles of sample selection. Such comparison was done as follows. After performing the simulation study, we obtained 1000 vectors of estimated default probabilities, for each *K* (*K* = 1, 3, 9) and for each of the two techniques studied. Then we sorted, in ascending order, each of the 1000 vectors (columns). Thus, the first row of the resulting spreadsheet contained the lowest estimated probabilities in each of the 1000 simulations, while the last line comprised the highest estimated probabilities. We got 90% confidence bands for the distributions of estimated probabilities, original or adjusted, by recording the 5% and 95% percentiles. The simulations were performed using SAS version 9.0. Interested readers can obtain the codes by email the authors.

4.2. Simulation results

In this Section we present the main results obtained from the simulation study performed with balanced and unbalanced samples, regarding the model performance measures, i.e. the models' predictive capacities. We also present the main results obtained from the simulation study, relating to the distributions of default probabilities estimated from the two studied techniques. The naive logistic regression models give us the original probabilities while the logistic regression with state-dependent sample selection models, the adjusted ones.

4.2.1. Performance measures

This Section presents the main results regarding the performance, i.e. the predictive capacity of models adjusted using the two studied techniques. Thus, Tables 2 and 4 show the 95% confidence intervals for the performance measures: naive logistic regression results are presented in Table 2 and logistic regression with state-dependent sample selection results are presented in Table 4, while Table 3 shows the validation results of the naive logistic regression in a balanced test sample.

The empirical results presented in Table 2 reveal that the naive logistic regression technique produces good results, with high values of sensitivity and specificity among others, only when the sample used for the development of the model is balanced, i.e. when $K = 1$. As the degree of imbalance increases ($K = 3$ and 9) sensitivity decreases considerably, assuming values less than 0.5 when $K = 9$, whereas specificity increases, reaching values near 1 for $K = 9$. Note that the values of *MCC*, *IC* and *AC* are also decreasing as *K* increases.

We observe similar results regardless of the value of *K*, i.e. regardless the fact that the sample used for building the model is balanced or not, when an independent balanced sample is considered for investigating the model performance. Moreover, these results are indicative of good predictive capacity as it is shown in Table 3.

Discussions about the results of the logistic regression with state-dependent sample selection technique (Table 4 summarizes these results) are analogous to those made previously, when we use the naive logistic regression technique.

4.2.2. Estimated probabilities

Fig. 1 presents the 90% confidence band curves for both original and fitted model. It can be observed that regardless of the *K* value, the estimated probabilities without the adjustment to the constant term of the equation are lower than those where the adjustment was made. So the naive logistic regression model underestimates the probability of default. Note also that the distance between the curves decreases as the degree of sample imbalance increases. The curves are closer when the degree of imbalance is close to the real one present in the full sample, that is, approximately 91% of good payers and 9% of bad payers. For instance, the distance between the curves is largest for $K = 1$ (Fig. 1 upper panel), i.e. for balanced samples, while the curves are very close to each other for $K = 9$ (Fig. 1 bottom panel), i.e. for 90,000 good payers and 10,000 bad payers in the selected samples.

The results presented in Tables 2 and 4 are very close, without statistically significant difference between the majority of corresponding empirical intervals (i.e. almost every corresponding empirical confidence intervals present intersection).

5. Brazilian bank credit scoring data

In order to illustrate the measures of a model performance discussed so far, let us examine a data set taken from the overall clients from a Brazilian bank portfolio. The data classifies 4504 clients into good or bad payers, according to their credit histories. The variables considered are: client type, gender, age, marital status and length of residence. The dataset used for model-fitting, i.e. the training sample, has 3153 clients (70% of the original sample) and the validation one, i.e. the test sample, has 1351 clients (the remaining 30% of the original sample), both with roughly the same proportion of bad payers (8%). Here, we considered the training sample as the full sample. Then the selected samples were obtained by maintaining all bad payers of the full sample, plus $258 * K$ good payers selected randomly from the full sample. As well as in the simulation study, we only examined the cases where $K = 1$, 3 and 9, which correspond to 258, 774 and 2322 good payers in the selected sample, respectively. Similarly, three new test samples were obtained by retaining all bad payers of the original test sample, plus $102 * K$ ($K = 1$, 3 and 9) good payers chosen randomly from the original test sample. Thus, a naive logistic regression model (Hosmer & Lemeshow, 1989) and a logistic regression with state-dependent sample selection model (Cramer, 2004) were applied to each of the three selected samples and then the three test samples were used to analyze the appropriateness of the adjusted models.

Fig. 2 shows the ROC curves (panels on the left) together with the cumulative probabilities curves (panels on the right) for each fitting. As can be seen, the ROC curves for the naive logistic regression model and the logistic regression with state-dependent sample selection model are very close, regardless of the value of *K*. With the help of such curves, we selected cutoff points equal to 0.53, 0.26 and 0.09 (for $K = 1$, 3 and 9, respectively) for the naive logistic regression models and cutoff points equal to 0.92, 0.57 and 0.13 (for $K = 1$, 3 and 9, respectively) for the logistic regression with state-dependent sample selection models. In the Fig. 2 right panels we observe the distributions of estimated default probabilities in the balanced test sample from the two studied techniques. The results for the unbalanced test samples are very similar to the ones of the balanced test sample, so they don't need to be presented here. Note that the naive logistic regression model, which provides us the original probability of a client being a bad payer, always underestimates the probability of default, especially in the case where the sample used to build the model is balanced, i.e. when $K = 1$ (Fig. 2 upper right panel). However, the distance between the curves, i.e. the difference between original and adjusted probabilities decreases as the degree of imbalance of the selected sample increases. The measures relating to the predictive capacity of the adjusted models are showed in Tables 5–7 for the balanced, unbalanced with 75% good payers and 25% bad payers and unbalanced with 90% good payers and 10% bad payers test samples, respectively. Note that, for each table, the measures of both models are very close and are indicative of good predictive capacity, which is confirmed by the ROC curves (Fig. 2 left panels). We also observe that the performance does not change or just changes a little as the degree of imbalance of the selected samples increases, i.e. the predictive capacity is almost the same regardless of the sample used to build the models is balanced or not. As the degree of imbalance of the test samples increases (Tables 6 and 7) *PPV* decreases considerably, assuming values less than 0.5 in the case where the test sample is unbalanced with 90% good payers and 10% bad payers (Table 7). Except for the *NPV*, the values of the others performance measures are also decreasing as the degree of imbalance of the test samples increases.

6. Final comments

Credit scoring techniques have become one of the most important tools currently used by financial institutions such as banks, in the measurement and evaluation of the loans credit risk. Furthermore, credit scoring is regarded as one of the basic applications of misclassification problems that have attracted most attention during the past decades.

This paper compares, via simulation, the statistical technique widely used in modeling credit scoring data: the naive logistic regression (Hosmer & Lemeshow, 1989) with the logistic regression with state-dependent sample selection (Cramer, 2004).

A real case study on a Brazilian bank data was also performed in order to illustrate the presented procedures on a real data set.

Based on the simulation results, we can conclude that there is difference between the cumulative distributions of the default estimated probabilities by the use of the two studied techniques. Particularly, the naive logistic regression models underestimate such probabilities. However, there is no difference concerning to the performance of adjusted models from such techniques when we use measures like sensitivity, specificity and accuracy among others, in the evaluation of such models. The simulation study also showed that regardless of which of these two statistical modeling techniques is used, there is a need for working with balanced samples, which ensure models with good measures of sensitivity and specificity and high accuracy rate. This finding confirms the subjective results reported by Thomas et al. (2002) who suggest the use of balanced samples in building credit scoring models.

Thus, the ideal is always to work with balanced and consider samples using the logistic regression with state-dependent sample selection, since in this situation the logistic regression with state-dependent sample selection lead us to the true probability of default and has predictive capacity similar to the naive logistic regression model, while the naive logistic regression model underestimates the probability of default.

Acknowledgments

The authors are grateful to the Editorial Boarding for suggestions and comments. The research is partially founded by CNPq.

Appendix A. Appendix

Sketch of the proof of Eq. (3). If the event $I_i = 1$ means that element i is included in the sample, from the Bayesian Theorem (Baron, 1994), it is possible to see that,

$$\tilde{p}_i = P(Y_i = 1 | I_i = 1) = \frac{P(Y_i = 1, I_i = 1)}{P(Y_i = 1, I_i = 1) + P(Y_i = 0, I_i = 1)}$$
$$= \frac{\alpha p_i}{\alpha p_i + \gamma \alpha (1 - p_i)} = \frac{p_i}{p_i + \gamma (1 - p_i)},$$

completing the proof.

References

Abreu, H. J. (2004). *Aplicação de análise de sobrevivência em um problema de credit scoring e comparação com a regressão logística*. Dissertação de Mestrado, DEs-UFSCar.

Baesens, B., Gestel, T. V., 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.

Baldi, P., & Brunak, S. (1998). *Bioinformatics: The machine learning approach*. Cambridge, MA: MIT Press.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics Review, 16*(5), 412–424.

Banasik, J., Crook, J., & Thomas, L. (2001). Scoring by usage. *Journal of the Operational Research Society, 52*(9), 997–1006.

Baron, Jonathan (1994). *Thinking and deciding* (2 ed.). London: Oxford University Press.

Boyes, W. J., Hoffman, D. L., & Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics, 40*(1), 3–14.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics, 26*, 801–849.

Burset, M., & Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics, 34*, 353–367.

Caouette, J. B., Altman, E. I., & Narayanan, P. (1998). *Managing credit risk: The next great financial challenge*. New York: John Wiley & Sons Inc..

Cramer, J. S. (2004). Scoring bank loans that may go wrong: A case study. *Statistica Neerlandica, 58*(3), 365–380.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37.

Dunn, G., & Everitt, B. S. (1995). *Clinical biostatistics: An introduction to evidence based medicine*. London: Edward Arnold.

Fisher, R. A. (1936). The use multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188.

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy, 10*(3), 299–316.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science, 21*(1), 1–14.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

Hosmer, W. D., & Lemeshow, S. (1989). *Applied logistic regression*. Wiley.

Kang, S., & Shin, K. (2000). *Customer credit scoring model using analytic hierarchy process*. Informs & Korms, Seoul, Korea, pp. 2197–2204.

Kullback, S. (1959). *Information theory and statistics*. New York: Dover Publications.

Kullback, S., & Leibler, R. A. (1986). On information and sufficiency. *Annals of Mathematical and Statistical, 22*, 79.

Lee, T., & Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743–752.

Lee, T., Chiu, C., Lu, C., & Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245–254.

Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications, 32*(2), 427–431.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta, 405*, 442–451.

Orgler, Y. E. (1971). Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research, 2*(1), 31–37.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology, 232*, 584–599.

Rost, B., Sander, C., & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology, 235*, 13–26.

Sarlija, N., Bensic, M., & Bohacek, Z. (2004). *Multinomial model in consumer credit scoring*. 10th International Conference on Operational Research. Trogir: Croatia.

Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics, 8*(8), 31–34.

Thomas, L. C. (2000). *A survey of credit and behavioural scoring; forecasting financial risk of lending to consumers*. Edinburgh, UK: University of Edinburgh.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: SIAM.

Wang, Z. X. (1994). Assessing the accuracy of protein secondary structure. *Nature Structural Biology, 1*, 145–146.