Modeling Small Business Credit Scoring by Using Logistic Regression, Neural Networks, and Decision Trees

Authors (in alphabetical order):
Mirta Bensic
J.J. Strossmayer University of Osijek
Department of Mathematics
Gajev trg 6, 31000 Osijek, Croatia
e-mail: mirta@mathos.hr
Tel.: +385 31 224 800
Fax: +385 31 224 801

Natasa Sarlija
e-mail: natasa@efos.hr
J.J. Strossmayer University of Osijek
Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
Tel.: +385 31 224 442
Fax: +385 31 211 604

Marijana Zekic-Susac
e-mail: marijana@efos.hr
J.J. Strossmayer University of Osijek
Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
Tel.: +385 31 224 442
Fax: +385 31 211 604

ABSTRACT:

Previous research on credit scoring that used statistical and intelligent methods was mostly focused on commercial and consumer lending. The main purpose of this paper is to extract important features for credit scoring in small business lending on a dataset with specific transitional economic conditions using a relatively small dataset. To do this, we compare the accuracy of best models extracted by different methodologies, such as logistic regression, neural networks, and CART decision trees. Four different neural network algorithms are tested, including backpropagation, radial basis function network, probabilistic and learning vector quantization, by using the forward nonlinear variable selection strategy. Although the test of differences in proportion and McNemar's test do not show a statistically significant difference in the tested models, the probabilistic NN model produces the highest hit rate and the lowest type I error. According to the measures of association, the best NN model also shows the highest degree of association with the data, and it yields the lowest total relative cost of misclassification for all examined scenarios. The best model extracts a set of important features for small business credit scoring for the observed sample, emphasizing credit program characteristics, as well as entrepreneur's personal and business characteristics as the most important ones.

KEYWORDS: **CART decision trees, credit scoring modeling, logistic regression, neural networks, small business loans**

INTRODUCTION

For few decades credit scoring models have been used in commercial and consumer lending, and only recently in small business lending. Variables that are found important in small business credit differ from those that effect company loans (Feldman, 1997). Specific economic conditions of transitional countries also influence credit scoring modeling. Belonging to the group of postcommunist transitional countries, Croatia shares all the typical transitional characteristics with other countries of that type, such as economic changes focused on the market economy, then political, institutional, and social changes. Economic changes were marked with a rapid fall of economic activity, although the falling trend has been stopped in all countries, and even changed to the growth in some of them. Institutional changes include the development of institutions needed for enabling market economy, then liberalization of prices and foreign trade, and finally restruction and privatization of businesses (Mervar, 2002). Croatian economy is additionally featured by some specific war-caused characteristics (Kasapovic, 1996). Due to a negative influence of the war to economic development (Selowsky and Martin, 1997, in Mervar, 2000), it can be expected that the specific conditions caused by the war will also influence credit scoring modeling.

Small business development in all transitional countries is characterized by difficult access to the turnover capital, legal and law limitations, undeveloped infrastructure, high transactional costs, as well as high loan interest rates (Skare, 2000). Those characteristics are important for creating credit conditions in a country, and are incorporated in our model by introducing the following input variables: the way of interest repayment, the grace period in credit payment, the way of the principal payment, the length in months of the repayment period, and the interest rate.

Therefore, the paper aims to extract important features in modeling small business credit scoring in such environment. For this purpose, the predictive power of logistic regression (LR) in comparison to neural networks (NNs) and CART decision trees (CART) is investigated.

The paper consists of a brief overview of previous research results, data and methodology part, description of experiments, comparison of results, and conclusion which discusses the best LR, NN, and CART models as well as the features that are most important for credit scoring for the observed dataset.

OVERVIEW OF PREVIOUS RESEARCH RESULTS

Most of the credit scoring systems vary regarding the type and quantity of the data needed for decision making. Personal and business activities have both been found relevant in a small business credit scoring system (Friedland, 1996; Feldman, 1997; Arriaza, 1999; Frame *et al.* 2001). Willingness and ability of the business owner to repay personal borrowings could be assumed to correlate with the ability and willingness of the firm managed by the owner to repay their loans (Feldman, 1997).

Developers of scoring systems, especially Fair, Isacc and Co. Inc., found that the same variables determining the owner's probability of loan repayment also determine a large part of the credit score for small firms. They also found ratios from financial statements not crucial in determining repayment prospects of the small firm. Arriaza (1999) reports on Analytical Research & Development team at Experian who undertook a study using three risk scoring models: (i) The consumer risk scoring model (consumer credit performance data), (ii) The commercial risk scoring model (business credit performance data), (iii) The blended small business risk scoring model (combining both business and owner consumer credit performance data). The blended model was significantly more effective in identifying and appropriately scoring applicants defined as bad. Friedland (1996) reports that the following variables are important in deciding whether to grant a loan to a small business or not: financial reports, credit bureau report of the business owner and credit bureau report of the firm. Credit analysts have found that the personal credit history of the business owner is highly predictive in determining the repayment prospects of the small firm (Frame *et al.,* 2001).

One of the first investigations of NNs in credit scoring was done by Altman et al. (1994) who compared linear discriminant analysis (LDA) and LR with NNs in distress classification and prediction. Rates of recognition in the holdout sample showed that LDA was best in performance by recognizing 92.8% healthy and 96.5% unsound firms. Desai *et al.* (1996) obtained different results. Tested on a credit union dataset, the multilayer perceptron correctly classified the highest percentage of the total and the bad loans (83.19% and 49.72% for total and bad, respectively) and was significantly better than the LDA, whereas the difference was not significant when compared to LR. A more recent research by Desai *et al.* (1997) on credit scoring shows that LR outperforms the multilayer perceptron at the 0.05 significance level. Yobas *et al.* (2000) compared the predictive performance of LDA, NN (multilayer perceptron), genetic algorithms and decision trees by distinguishing between good and slow payers of bank credit card accounts. In their experiments the best neural network was able to classify only 64.2% payers and the mean proportion over ten decision trees was 62.3%. LDA was superior to genetic algorithms and NNs, although NNs showed almost identical results to LDA in predicting the slow payers. Galindo and Tamayo (2000) made a comparative analysis of CART decision-tree models, neural networks, the k-nearest neighbor and probit algorithms on a mortgage loan dataset. The results showed that CART decision-tree models provide the best estimation for default with an average 8.31% error rate. Much effort in testing various NN algorithms for credit scoring was done by West (2000), who compared five NN algorithms to five more traditional methods. McNemar’s test showed that mixture-of-experts, radial basis function NN, multi layer perceptron, and LR produced superior models, while learning vector quantization and fuzzy adaptive resonance, as well as CART decision trees performed as inferior.

Following suggestions of the previous research, the universe of variables in our research is characterized by small business as well as entrepreneurs themselves. Concerning methodology, the previous research showed that the best methodology for credit scoring modeling has not been extracted yet since it depends on the dataset characteristics. Since most authors, except West (2000), test a single NN algorithm, mostly a multi-layer perceptron such as backpropagation, we were challenged to compare the efficiency of more of them on a specific Croatian dataset. Results of NN models are compared to logistic regression and decision trees.

DATA

Data were collected randomly in a Croatian savings and loan association specialized for financing small and medium enterprises, mostly start-ups. The sample size consisted of 160 applicants. The reasons for such small dataset were the following: (1) relatively low entrepreneurial activity in the country (TEA index for 2003 = 2.56; TEA index for 2004 = 3.73)[1] which means a low number of people actually looking for a credit in order to start or grow a business, and (2) a certain proportion of start-ups applying for a credit was found too risky and rejected by the savings and loan association. Therefore, gathering a larger dataset would be possible after few years, while the models are needed even before that time.

Variable selection for credit scoring

Input variables describe the owner's profile, small business activities, and financial data. Due to problems of missing values and insufficient reliability in some variables, it was necessary to exclude some of them from the model, such that data collection resulted in the total of 31 variables. In order to extract the set of variables that add some information to the model, an information value was calculated (Hand and Henley, 1997), which extracted seven groups of variables shown in Table 1. Descriptive statistics of the variables used, separately for good (G) and bad (B) applicants, is also presented, where variables marked with "*" were found significant in the best model.

Table 1. Input variables and their statistical distribution

Variable code	Variable description	Descriptive statistics
Group 1	Small business characteristics	

V6	Main activity of the small business *	Textile production and sale (G: 11.54% B: 15.15%);
		Cars sale (G: 7.69% B: 9.09%);
		Food production (G: 20.51% B: 13.64%);
		Medical, intellectual services (G: 19.23% B: 4.54%);
		Agriculture (G: 29.49% B: 39.39%);
		Building (G: 6.41% B:10.61 %);
		Tourism (G: 5.13% B: 7.58%)

V14	Starting a new business undertaking	Yes (G: 19.23% B: 25.76%); No G: (80.77% B: 74.24%)
V16	Equipment necessary for a business	Yes (G: 75.64%, B: 65.15%); No (G: 24.36%, B: 34.85%)
V19	Number of employees in a small business firm being granted a credit	Mean G: 2.29 (σ=3.08), Median=1; Mean B:1.64 (σ=1.34), Median=1
Group 2	**Personal characteristics of entrepreneurs**	

V3	Entrepreneur's occupation *	Farmers (G: 48.72% B: 43.94%);
		Retailers (G: 6.41% B: 15.15%);
		Construction (G: 8.97% B:10.61 %);

[1] TEA = the ratio of the number of people per each 100 adults (aged between 18 and 64) who are trying to start their own businesses or are owners/managers in an active enterprise not older than 42 months (Global entrepreneurship monitor 2003, 2004)

		Elect. engineering, medicine (G: 21.79% B: 21.21%); Chemists (G: 14.10% B: 9.09%)
V7	Entrepreneur's age	Mean G: 43.36 (σ=10.2), Mean B: 40.21 (σ=8.34)
V13	Business location	Region 1 - G: 37.18% B: 46.97%; Region 2 - G: 26.92% B: 16.67%; Region 3 - G: 10.26% B: 12.12%; Region 4 - G: 25.64% B: 24.24%
Group 3	**Relationship characteristics with the financial institution**	
V12	This is the first time an entrepreneur is granted a credit by this bank	For the first time (G: 84.62% B: 15.38%); For the second or third time (G: 92.42% B: 7.58%)
Group 4	**Credit program characteristics**	

V5	Way of interest repayment *	Monthly (G: 78.21% B: 71.21%);
		Quarterly (G: 17.95%, B: 12.12%);
		Semi-annually (G: 3.85% B: 16.67%)

V9	Grace period in credit repayment *	Yes (G: 58.97% B: 69.7%); No G: (41.03% B: 30.3%)
V10	Way of the principal repayment *	Monthly (G: 78.21% B: 68.18%); Annually (G: 21.79% B: 31.82%)
V17	Length in months of the credit repayment period *	Mean G: 18.36 (σ=6.77) Mean , Median G: 24; B: 19.76 (σ=6.45), Median B: 24
V18	Interest rate *	Mean G: 13.5 (σ=1,98); Mean B: 13.14 (σ=1.81)
V20	Amount of credit (in HRK)	Mean G: 48,389 kunas (σ=32,750); Mean B: 49,044.27 kunas (σ=32,961.72)
Group 5	**Growth plan**	
V2	Planned value of the reinvested profit (percentage) *	G: 50-70% B: 30-50%
Group 6	**Entrepreneurial idea**	

V1	Clear vision of the business *	No (G: 1.28% B: 18.18%);
		Yes (G: 17.95% B: 7.58%);
		Existing business (G: 80.77% B: 74.24%)

V11	Main characteristics of entrepreneur's goods/services comparing to others	Quality (G: 37.18% B: 34.85%);
		Production (G: 7.69% B: 9.09%);
		Service, price (G: 17.95% B: 6.06%);
		Reputation (G: 17.95% B: 16.67%);
		No answer (G: 19.23% B: 33.33%)

V15	Sale of goods/services	Local level (G: 55.13% B: 46.97%);
		Defined customers (G: 15.38: B: 28.79%);
		One region (G: 10.26% B: 7.58%);
		Whole country (G: 15.38% B: 13.64%);
		No answer G: 3.85% B: 3.03%)

Group 7	**Marketing plan**	

V4	Advertising goods/services	No adds (G: 15.39% B: 9.09%);
		All media (G: 21.79% B: 34.85%);
		Personal sale (G: 19.23% B: 1.52%);
		Internet (G: 5.13% B: 4.55%);
		No answer (G: 38.46% B: 50%)

V8	Awareness of competition *	No competition (G: 12.82% B: 16.67%);
		Broad answer (G: 57.69% B: 46.97%);
		Defined competition (G: 16.67% B: 9.09%);
		No answer (G: 12.82% B: 27.27%)

As the output, a binary variable with one category representing good applicants and the other one representing bad applicants was used. An applicant is classified as good if there have never been any payments overdue for 45 days or more, and bad if the payment has at least once been overdue for 46 days or more. In the initial sample of accepted applicants, 66 applicants (or 45.83%) were good, while 78 of them (or 54.17%) were bad.

METHODOLOGY FOR CLASSIFYING CREDIT APPLICANTS

Logistic regression classifier

Traditionally, different parametric models are used for classifying input vectors into one of two groups, which is the main objective of statistical inference on the credit scoring problem. Due to specific characteristics of small business data, most of the variables in our research are categorical. Thus, logistic regression is chosen as the most appropriate for such type of data. Logistic regression modeling is widely used for the analysis of multivariate data involving binary responses we deal with in our experiments. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses. Since the likelihood function of mutually independent variables $Y_1,\ldots,Y_n$ with outcomes measured on a binary scale is a member of the exponential family with $\left(\log\left(\frac{\pi_1}{1-\pi_1}\right),\ldots,\log\left(\frac{\pi_n}{1-\pi_n}\right)\right)$ as a canonical parameter (π_j is a probability that Y_j becomes 1), the assumption of the logistic regression model is a linear relationship between a canonical parameter and the vector of explanatory variables $\mathbf{x}_j$ (dummy variables for factor levels and measured values of covariates):

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \mathbf{x}_j^\tau \boldsymbol{\beta} \qquad (1)$$

This linear relationship between the logarithm of odds and the vector of explanatory variables results in a nonlinear relationship between the probability of Y_j equals 1 and the vector of explanatory variables:

$$\pi_j = \exp(\mathbf{x}_j^\tau \boldsymbol{\beta}) / (1 + \exp(\mathbf{x}_j^\tau \boldsymbol{\beta})) \qquad (2)$$

As we had a small dataset along with a large number of independent variables, in order to avoid overestimation we included only the main effects in the analysis. In order to extract important variables, we used the forward selection procedure available in SAS software, with standard overall fit measures. Since the major cause of unreliable models lies in overfitting the data (Harrel, 2001), especially in datasets with a relatively large number of variables as candidate predictors (mostly categorical) and a relatively small dataset such as the case in this experiment, we cannot expect to improve our model due to addition of new parameters. That was the reason to investigate if some non-parametric methods, such as neural networks and decision trees can give better results on the same dataset.

Neural network classifiers

Although many research results show that NNs can solve almost all problems more efficiently than traditional modeling and statistical methods, there are some opposite research results showing that statistical methods, in particular data samples, outperform NNs. A variety of results is sometimes due to non-systematic use of neural networks, such as testing only one or two NN algorithms and not using all the possibilities of optimization techniques that will lead to the best network structure, training time and learning parameters. The lack of standardized paradigms that can determine the efficiency of certain NN algorithms and architectures, particularly problem domains, is emphasized by many authors (Li, 1994).

Therefore, we test four different NN classifiers by NeuralWorks Professional II/Plus software: backpropagation with SoftMax activation function, radial basis function network with SoftMax, probabilistic, and learning vector quantization. The first two algorithms were tested using both sigmoid and tangent hyperbolic functions in the hidden layer, and the SoftMax activation function in the output layer. Learning is improved by the Extended Delta-Bar-Delta (EDBD) rule. The learning rate and the momentum for the EDBD learning rule were set as follows: 0.3 for the first hidden layer, 0.15 for the output layer, whereas the initial momentum term was set to 0.4. and exponentially decreased during the learning process. Saturation of weights is prevented by adding a small F-offset value to the derivative of the sigmoid transfer function. It is experimentally proved that value 0.1 is adequate for the sigmoid transfer function (Fahlmann in NeuralWare, 1998). Overtraining is avoided by the “save best” cross-validation procedure which alternatively trains and tests the network until the performance of the network on the test sample improves for *n* number of iterations.

The initial training sample (approximately 75% of the total sample) was divided into two subsamples: approximately 85% and 15% for training and testing, respectively. After training and testing the network for the maximum of 100,000 iterations, all the NN algorithms were validated on the out-of-sample data (approximately 25% of the total sample) in order to determine its generalization ability.

The probabilistic neural network (PNN) algorithm was chosen due to its fast learning and efficiency. It is a stochastic-based network developed by Specht (Masters, 1995). In order to determine the width of Parzen windows (σ parameter), we follow a cross-validation procedure for optimizing σ proposed by Masters (Masters, 1995). LVQ was used as a supervised version of the Kohonen algorithm with an unsupervised kernel in the hidden layer. Improved versions called LVQ1 and LVQ2 were applied in our experiments (NeuralWare, 1998).

The topology of networks consisted of an input layer, a hidden or a pattern layer (or Kohonen layer in LVQ), and an output layer. The number of neurons in the input layer varied due to the forward modeling strategy while the number of hidden neurons was optimized by a pruning procedure. The maximum number of hidden units was initially set to 50 in the backpropagation and radial basis function networks. The number of hidden units in the probabilistic network was set to the size of the training sample, while the LVQ networks consisted of 20 hidden nodes. The output layer in all network architectures consisted of two neurons representing classes of bad and good credit applicants.

As in LR models, variable selection in NN models is also performed by forward modeling starting from one input variable and gradually adding another one which improves the model most. The best overall model is selected on the basis of the best total hit rate. The advantage of such nonlinear forward strategy allows NN to discover nonlinear relationships among variables that are not detectable by linear regression.

CART decision tree classifier

In order to compare the ability of NNs not only to a parametric method such as LR, but also to another non-parametric method, we tested CART decision trees, because of their suitability for classification problems. Benchmarking LR to NNs and decision trees is also present in the previous research (West, 2000).

CART algorithm is used as one of the most popular methods for building a decision tree. The approach was pioneered in 1984 by Breiman *et al.* (in Witten and Frank, 2000), and it builds a binary tree by splitting the records at each node according to a function of a single input field. The evaluation function used for splitting in CART is the Gini index, which can be defined in the following way (Apte, 1997):

$$Gini(t) = 1 - \sum_i p_i^2 \qquad (3)$$

where t is a current node and p_i is the probability of class i in t. The CART algorithm considers all possible splits in order to find the best one. The algorithm can deal with continuous as well as with categorical variables. All possible splits are considered in the sequence of values for continuous valued attributes [*(n-1)* splits for *n* values]. Concerning categorical attributes, if *n* is small, [2^{n-1}*-1* splits for *n* distinct values] are considered, whereas [*n* splits for *n* distinct values] are considered if *n* is large (Apte, 1997). CART determines the best split for each attribute at each node and selects the winner by using the Gini index. The decision tree is growing until new splits are found that improve the ability of the tree to separate the records into classes. Since each following split has a less representative population to work with, it is necessary to prune the tree to get a more accurate prediction. The aim is to identify the branches that provide the least additional predictive power per leaf. In order to accomplish that, the complexity/error trade off of the original CART algorithm is used (Breiman *et al.* 1984 in Galindo, Tamayo, 2000). The winning subtree is selected on the basis of its overall error rate when applied to the task of classifying the records in the test set (Berry and Linoff, 1997). Therefore, we used pruning on the misclassification error as the procedure for selecting subtrees, where the sets of branches were pruned from the complete classification tree, similarly to elimination of predictors in a discriminant analysis. The right-sized classification tree is then selected using the specified standard error rule. The trees were created on the basis of 15 categorical variables, 5 continuous predictor variables, while the parameters were the following: (i) minimal number of cases that controls when the split selection stops and pruning begins was 5, (ii) equal prior probabilities were used, (iii) stopping rule was the misclassification error, (iv) standard error rule was 1. After the 7-fold cross-validation, the best classification tree was selected and validated on the same out-sample data as LR and NN models.

INCLUSION OF REJECTED APPLICANTS INTO THE MODEL

According to Lewis (1991), the credit scoring system is intended to be applied to the entire population that enters a bank, not only to those approved by the previous bank system. If a scoring system is constructed using only accepted applicants, it will contain the effect of the previously used methods that are to be replaced. In order to avoid that, Lewis suggests to make an inference of the future performance of the rejected applicants – a process called augmentation – which enables addition of the inferred goods and bads to the sample of the known goods and bads. Among different approaches to the way of estimating the performance of rejected applicants, we follow the one suggested by Meester (Crook and Banasik, 2002) who proposes to build a model on the accepted and rejected cases, estimate the performance of the rejects with this model and appropriate cut-off and then build a new model.

First-level experiments

The first level of experiments consisted of classifying applicants that were accepted by the bank in the past, with their known real credit scoring history (sample 1). We will use the term "first-level" LR and NN models that were used separately to classify the accepted credit applicants. The best extracted LR and NN models were then applied to applicants that were rejected by the bank in the past in order to obtain credit scoring probabilities for those applicants.

Datasets on both levels were divided into the in-sample data (approximately 75% of data used to estimate models), and the out-of-sample data (approximately 25% of data used for final validation of the best models of all three methodologies). Because of the nature of their objective function, NNs require an equal number of good and bad applicants in their training sample. Subsamples were created using a random selection of cases into the train, the test and the validation sample, while keeping the equal distribution of good and bad applicants in the train and test sample. The best LR, NN and CART models were validated on the same validation data in order to enable comparison. ~~D~~istribution of goods and bads in subsamples on the first level of experiments is given in Table 2.:

Table 2. Distribution of good and bad credit applicants in sample 1 (accepted credit applicants)

Subsample	No. of cases	Good applicants		Bad applicants	
		Actual no.	%	Actual no.	%
Train	92	46	50.00	46	50.00
Test	14	7	50.00	7	50.00
Validation	38	13	34.21	25	65.79
Total	144	66	45.83	78	54.17

NN models were trained on the train sample, optimized by using the test sample for cross-validation, and finally validated on the validation sample. LR models on the first level of experiments were estimated using 106 cases (train + test subsamples). Due to the equal distribution of goods and bads in the estimation sample, a cut-off of 0.5 is used in both LR and NN models on the first level of experiments.

Second-level experiments

The applicants rejected by the bank entered the model in the second-level experiments, with their output vector containing values estimated by the best NN and by LR and an appropriate cut-off. Therefore, two independent datasets were created:

- sample 2.a: all applicants entered the bank using NN values for rejected applicants
- sample 2.b: all applicants entered the bank using LR values for rejected applicants

Both of the above mentioned samples contain accepted and rejected applicants divided into in-sample and out-of-sample data as described earlier. NNs, LR, and CART decision trees were applied on both samples using the same in-sample and out-of-sample data. Distribution of good and bad applicants on the second level of experiments is presented in Table 3. and Table 4.

Table 3. Distribution of good and bad credit applicants in sample 2a (NN probabilities for rejects)

Subsample	No. of cases	Good applicants		Bad applicants	
		Actual no.	%	Actual no.	%
Train	106	54	50.94	52	49.06
Test	16	7	43.75	9	56.25
Validation	38	13	34.21	25	65.79
Total	160	74	46.25	86	53.75

Table 4. Distribution of good and bad credit applicants in sample 2b (LR probabilities for rejects)

Subsample	No. of cases	Good applicants		Bad applicants	
		Actual no.	%	Actual no.	%
Train	106	50	47.17	56	52.83
Test	16	8	50.00	8	50.00
Validation	38	13	34.21	25	65.79
Total	160	71	44.37	89	55.63

According to the distribution of good and bad applicants in Tables 3. and 4., the cut-off in LR models was estimated on the basis of distribution of goods and bads in the estimation sample.

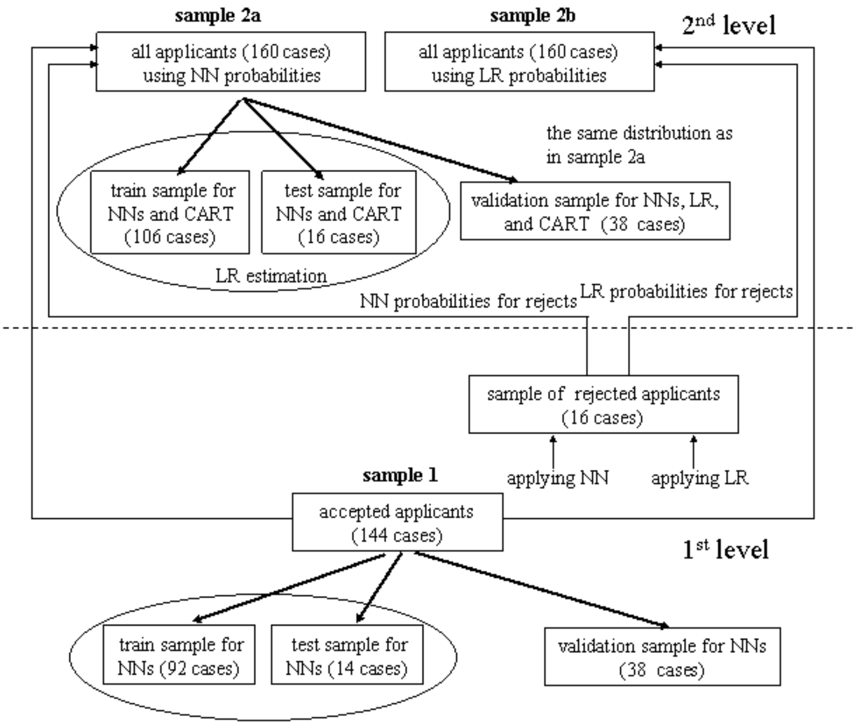
sample 2a
sample 2b
2nd level
all applicants (160 cases) using NN probabilities
all applicants (160 cases) using LR probabilities
the same distribution as in sample 2a
train sample for NNs and CART (106 cases)
test sample for NNs and CART (16 cases)
validation sample for NNs, LR, and CART (38 cases)
LR estimation
NN probabilities for rejects
LR probabilities for rejects
sample of rejected applicants (16 cases)
sample 1
applying NN
applying LR
accepted applicants (144 cases)
1st level
train sample for NNs (92 cases)
test sample for NNs (14 cases)
validation sample for NNs (38 cases)

Figure 1. Sampling procedure for two levels of experiments

RESULTS OF CREDIT SCORING MODELS

Estimating probabilities for rejected credit applicants

The first scoring model is used for the purpose of classifying the rejected credit applicants. NN results on the first level show that the best total hit rate of 76.3% is obtained by the backpropagation algorithm which also had the best hit rate for bad applicants (84.6%). Hit rates of the first scoring model for LR estimation gave the total hit rate of 83.08%, the good hit rate of 89.92% and the bad hit rate of 69.69%. Fitting measures such as Wald=45.6581 (p=0.0033); Score=77.3657 (p<0.0001) show that the model fits well.

The best estimated LR and NN models extracted on the first-level tests were applied separately to the subsample of rejected applicants. Table 5 shows the obtained results.

Table 5. Classification of rejected credit applicants

Model	% of goods*	% of bads*
LR model	30.00	70.00
Best NN (Backprop, 4-50-2)	50.00	50.00

* computed from the obtained probabilities using the threshold 0.5

As can be seen from Table 5., the estimated LR model classified 30% applicants into good and 70% into bad ones. Since the two methods produced different results on rejected applicants, we were challenged not to rely on LR or NN individually but to include both NN and LR probabilities of rejected applicants into further credit scoring assessments. Furthermore, we aimed to examine the efficiency of LR and NN models when using their own probabilities, and also to examine the efficiency of integrated methods by using NN probabilities in LR estimation and vise versa.

Classification of the overall set of applicants

a) Classification of the overall set of applicants with NN probabilities for rejected applicants

Table 6. shows the results of the LR, NN, and CART models obtained on sample 2a (when rejected applicants were included in the sample using NN probabilities). Since four different NN algorithms were tested, their results are given separately with the topology for each NN architecture.

Table 6. LR, NN and CART results on the hold-out validation sample using NN probabilities for rejected applicants

Model	Total hit rate (%)	Hit rate of bads (%)	Hit rate of goods (%)	Extracted input variables
Logistic – estimated on the same in-sample as NN	76.32	61.54	84.00	V1, V4, V5
Backprop NN, 6-50-2	73.68	46.15	88.00	V1, V2, V3, V7, V15, V20
RBFN, 5-50-2	71.05	69.23	72.00	V1, V2, V3, V5, V6
Probabilistic NN, 10-106-2, σ=0.9	84.21	84.62	84.00	V1,V2, V3, V5, V6, V8, V9, V10, V17, V18
LVQ NN, 2-20-2	65.79	15.39	92.00	V1, V2
CART - estimated on the same in-sample as NN	65.79	61.54	68.00	V1, V2, V4, V20

The LR model produced the total hit rate of 76.32%, and it classified good applicants more accurately than the bad ones (the hit rate of good applicants was 84%, while the hit rate of bad ones was 61.54%). Concerning the variable extraction, it can be seen that (V1) vision of the business, (V4) advertising goods/services and (V5) way of the interest repayment have been recognized as important ones in deciding whether to grant a loan or not.

As presented in Table 6. the best NN result is also the overall best result according to all three methodologies, obtained by probabilistic NN (the total hit rate of 84.21%) using 10 input units and 106 pattern units with optimized σ parameter 0.9. The 95% confidence interval for this total hit rate is (0.726 – 0.958). The best NN model also classified bad credit applicants more accurately than all other models (the hit rate of bad applicants was 84.62%). Among other NN algorithms, LVQ was the worst at performance, unable to classify more than 65.79% of the total number of applicants, while backpropagation was the second best, followed by RBFN. The best NN model extracted most of the input variables as important (10 of them), while V1 (clear vision of the business), and V2 (planned value of the reinvested profit) were the variables extracted by all models. In addition to those two, the most important variables extracted by the best NN model were V3 (entrepreneur's occupation), V5 (way of interest repayment), V6 (main activity of the small business), V8 (awareness of competition), V9 (grace period in credit payment), V10 (way of the principal payment), V17 (length in months of the repayment period), and V18 (interest rate).

Similarly to the LR model, the CART model classifies good applicants better than the bad ones (68% of bads and 61.54% of goods). It also selected more variables than LR, but less than NNs. One variable that was extracted by CART and was not found in the best NN and LR models is V20 (amount of credit).

b) Classification of the overall set of applicants with LR probabilities for rejected applicants

Table 7. LR, NN and CART results on the hold-out validation sample using LR probabilities for rejected applicants

Model	Total hit rate (%)	Hit rate of bads (%)	Hit rate of goods (%)	Extracted input variables
Logistic – estimated on the same in-sample as NN	76.32	53.58	88.00	V1, V4, V5
Backprop NN, 4-50-2	71.05	23.08	96.00	V1, V2, V3, V5
RBFN, 8-50-2	71.05	76.92	68.00	V1, V2, V3, V6, V8, V13, V14, V18
Probabilistic NN, 7-106-2, σ=0.7	78.95	84.62	76.00	V1, V2, V3, V5, V6, V8, V9
LVQ NN, 2-20-2	65.79	15.39	92.00	V1, V2
CART - estimated on the same in-sample as LR	50.00	100.00	24.00	V4

Table 7. shows the results of LR, NN, and CART models when LR probabilities were used for rejected applicants. It can be seen that LR yields the same total hit rate (76.32%) regarding NN or LR probabilities used in the model. Once more, better accuracy of the LR model is obtained for good applicants (88%) than for the bad ones (53.58%). Concerning variable selection, the same 3 variables are isolated here as in estimation using NN probabilities for rejects. These are: (V1) vision of the business, (V4) advertising goods/services and (V5) way of the interest repayment.

When NNs use probabilities for rejected applicants obtained by LR, they generally perform worse in the sense of the total hit rate. The best NN model is once more obtained by the probabilistic algorithm, yielding the total hit rate of 78.95%. Its hit rate for bad applicants is the same as with NN probabilities (84.62%). In both samples 2a. and 2b. NNs classify bad applicants more accurately than the good ones. Backpropagation and RBFN attained equal total hit rates, although RBFN was again better at classifying bad applicants. LVQ was the last in performance among NN algorithms. The best NN model extracted 7 input variables as important: V1 (clear vision of the business), V2 (planned value of the reinvested profit), V3 (entrepreneur's occupation), V5 (way of the interest payment), V6 (main activity of the small business), V8 (awareness of competition), and V9 (grace period in credit repayment). It is interesting that all of them were also included in the best NN model that used NN probabilities, indicating that logistic probabilities did not add new information in the sense of model dimensionality.

The CART model gives the worst result among the three methodologies with LR probabilities for rejected applicants (the total hit rate of 50%). However, it is interesting that it correctly classifies 100% of bad applicants, and only 24% of the good ones. By extracting only one variable as important (advertising goods/services), it drastically discards much of the model complexity.

If we compare results obtained by LR, NN, and CART estimation, it can be noticed that the best NN model is more accurate when using NN probabilities. The LR model gives the same accuracy regarding probabilities for rejects used in the sample, although the bad hit rate of LR is higher if NN probabilities were used. The LR model selected the same set of variables in both samples: (V1) vision of the business, (V4) advertising goods/services and (V5) way of the interest repayment. Although the CART total hit rates are higher if NN probabilities were used, it classifies bad applicants better if LR probabilities were used.

Best model extraction

In order to compare the classification accuracy of the best LR, NN, and CART models, we use standard statistical tests: the test of differences in proportions (z-test), McNemar's test, and association measures. Table 8. shows results of the z-test.

Table 8. Two-way comparison of the best LR, NN, and CART model total hit rates

Hypothesis	Model	Total hit rates	Test results

H_0: $p_{LR} = p_{NN}$	LR	76.32	z=-0.8684
			p=0.3903

	NN	84.21	

H_0: $p_{LR} = p_{CART}$	LR	76.32	z=1.0189
			p=0.3148

	CART	65.79	

H_0: $p_{CNN} = p_{CART}$	NN	84.21	z=-1.8977
			p=0.0677

	CART	65.79	

P-values from Table 8 show that no significant difference is found among NN, LR and CART best models at the 0.05 level. Based on this test, it cannot be concluded that the NN model outperforms LR and CART in prediction accuracy, although its total hit rate is higher than the LR and CART total hit rates.

McNemar’s test is aimed to evaluate an experiment in which a sample of *n* subjects is evaluated on a dichotomous dependent variable and assumes that each of the *n* subjects contributes two scores on the dependent variable (Sheskin, 1997). Since Dietterich (in West 2000) suggests that McNemar's test is more appropriate for the supervised learning models than the test of difference in proportions, we performed a pairwise testing of the best models based on the same hold-out validation sample. The hypotheses: H_0: $p_b=p_c$, and H_1: $p_b \neq p_c$ were tested using the McNemar's test:

$$\chi^2 = \frac{(b-c)^2}{b+c} \qquad (6)$$

where b is the number of cases in which model 1 results in a hit, while model 2 results in a non-hit, and c is the number of cases in which model 1 produces a non-hit, while model 2 produces a hit. If one model outperforms another, there should be a significant difference in probabilities in the distribution table in positions of score b and score c. If the difference is not shown, it should be concluded that both models produce their hits with the same probability.

Table 9. McNemar's distributions for the comparison of NN, LR and CART best models

Model	LR		
NN		hit	non-hit
	hit	25	7
	non-hit	4	2

a) NN vs. LR model (χ^2=0.82)

Model	CART		
NN		hit	non-hit
	hit	20	12
	non-hit	5	1

b) NN vs. CART model (χ^2=2.88)

Model	CART		
LR		hit	non-hit
	hit	21	8
	non-hit	4	5

c) LR vs. CART model (χ^2=1.33)

The results of the McNemar's test show that the difference between NN and LR models is not significant (χ^2=0.82, df=1) at 0.05 level, as well as the difference between NN and CART (χ^2=2.88, df=1), and LR and CART (χ^2=1.0, df=1).

In order to examine the difference between the three models we computed different measures of association (Sheskin, 1997) such as phi coefficient, contingency coefficient, Kendal tau-c, and Spearman rank R between experimental and estimated values for good and bad applicants in the validation sample. The results are given in Table 10.

Table 10. Measures of association between real and estimated values for LR, NN and CART models

Association measure	LR	NN	CART
phi coefficient	0.45	0.67	0.28
contingency coefficient	0.41	0.55	0.27
Kendal tau-c	0.37	0.61	0.26
Spearman Rank R	0.45	0.67	0.28

According to all computed association measures, the NN model is the best in performance, while CART is the worst. Considering the fact that the value of phi coefficient for the NN model is very high (0.67), it can be stated that the NN model shows a high degree of association with experimental data, while the association measures for the CART model are so low that they can be considered inapplicable.

The most accurate prediction is operationally defined as the prediction with minimum costs. Since the cost of accepting a bad applicant (type I error) is greater for the bank than the cost of rejecting a good applicant (type II error), according to Koh (Swicegood and Clark, 2001), it is desirable to see the accuracy of each model separately for good and bad applicants.

Table 11. Best model accuracy in predicting good and bad credit applicants

Model	Actual	Predicted		Total
		Bad	Good	
Logistic Regression	Bad	53.85%	46.15%*	100.00%
	Good	12.00%**	88.00%	100.00%
Neural Network	Bad	84.62%	15.38%*	100.00%
	Good	16.00%**	84.00%	100.00%

CART Decision Tree	Bad	61.54%	38.46%*	100.00%
	Good	32.00%**	68.00%	100.00%

*Type I error: Predicting bad credit applicants as good.
**Type II error: Predicting good credit applicants as bad.

Table 11. shows that LR correctly identifies 53.85% of bad credit applicants, resulting in type I error rate of 46.15%. When classifying good applicants, LR showed a much higher hit rate (88%) producing type II error rate of 12%. NN correctly classified 84.62% and 84% of bad and good applicants, respectively, yielding type I error rate of 15.38%, while type II error rate for the NN model was 16%. The CART decision tree model classified good applicants more accurately than the bad ones, with type I error of 38.46%, and type II error of 32%. It is evident that the lowest type I error was obtained by the NN model, while the LR model produced the lowest type II error. The LR model had the highest type I error (46.15%).

It is also valuable to compute the total relative cost of misclassification (RC) according to Swicegood and Clark (2001):

$$RC = \alpha(P_I C_I) + (1-\alpha)(P_{II} C_{II}) \tag{7}$$

where α is the probability of being a bad applicant, P_I is the probability of Type I error, C_I is the relative cost of Type I error, P_{II} is the probability of Type II error, and C_{II} is the relative cost of Type I error. RC of each model is computed for eight scenarios, while the best model for each scenario is the model with the lowest RC value. Since the real costs of misclassification are unknown and relative to a bank, we follow the proposal by Hopwood *et al.* and Ethridge and Srinam (in Swicegood and Clark, 2001) to use the range of cost ratios as different scenarios. The cost ratio is represented by the cost of Type I error divided by the cost of Type II error, and we compute RC for eight scenarios (1:1, 2:1, 3:1, 10:1, 20:1, 30:1, 40:1, and 50:1). The RC values for each scenario and each model are presented in Table 12.

Table 12. Estimated relative costs of misclassifications

Cost ratio (CI:CII)	LR P_I=0.4615 P_{II}=0.12	NN P_I=0.1538 P_{II}=0.16	CART P_I=0.3846 P_{II}=0.32
1:1	0.171056	0.132451	0.22634
2:1	0.284877	0.192058	0.339643
3:1	0.398699	0.251665	0.448043
10:1	1.195446	0.668911	1.206848
20:1	2.333658	1.264978	2.290856
30:1	3.471869	1.861045	3.374863
40:1	4.61008	2.457111	4.458871
50:1	5.748292	3.053178	5.542878

It is evident from Table 12 that the NN model yielded the lowest RC for all given scenarios.

Concerning all above mentioned results, the best NN model can be suggested as the most successful model for credit scoring on our specific dataset.

Pertaining to variable selection, the best NN model extracted 10 input variables as important. C*lear vision of the business* belongs to the group of features that describe entrepreneurial idea, whereas *planned value of the reinvested profit* features the growth plan. Among personal characteristics of entrepreneurs, *entrepreneur's occupation* was found to be the most important one. Small business characteristics enter the model in the form of the *main activity of the small business*. The group describing a marketing plan also entered the model with the variable *awareness of competition*. All other extracted variables belong to the sixth group of variables describing credit program characteristics (*way of interest repayment, grace period in credit payment, way of the principal payment, length in months of the repayment period, and interest rate*). It was surprising that most of the extracted variables belong to the group of credit program characteristics, emphasizing that specific credit conditions in Croatian banks influence credit repayment. The fact that at least one variable from other variable groups was found relevant, implies that both entrepreneur's personal and business characteristics influence credit scoring on the observed dataset.

CONCLUSION

So far credit scoring has been investigated by using traditional statistics and intelligent methods, but without focusing on small business in transitional countries. The paper aimed to identify important features for the credit scoring model for small business, as well as to compare the performance of LR, NN and CART decision tree models on the Croatian dataset. The sample included not only previously accepted credit applicants, but also the rejected ones who were estimated by LR and NN and entered the model on the second level of experiments.

The results show that the highest total hit rate, and the lowest type I error are obtained by the probabilistic NN. The best model extracted 10 variables as important belonging to entrepreneur's personal and business characteristics, as well to credit program characteristics.

Regarding variable selection, this research partially confirms some previous findings of other authors recognizing the following features as important for small business credit scoring: clear vision of the business, personal characteristics of entrepreneurs, as well as business characteristics. In addition to that, the best model in this research extracted all variables influenced by macroeconomic conditions, such as the *way of interest repayment*, *grace period*, *way of the principal payment*, *length in months of the repayment period*, and *interest rate*, thus confirming the initial assumption that specific economic transitional conditions also influence variable selection.

The research also confirmed that previously described limitations of LR methodology (while dealing with small datasets and a large number of variables) prevent it from being successful with this type of problems. The nonparametric CART methodology has shown the ability to extract more features but it produced the lowest accuracy. A nonlinear nature and a multi-layer structure of NN methodology have shown more success in reaching higher hit rates, although the difference was not statistically significant according to the test of differences in proportion and McNemar's test. However, the association measures show the highest degree of association between real and estimated values in the NN model, and the NN model yielded the lowest cost of misclassification in all tested scenarios. Due to the small dataset tested, the other methodologies should also be taken into consideration in order to find the best one that suits the specific dataset.

In accordance with the above given reasons, the results can be applied on the model basis, providing some new information about credit scoring modeling in a transitional country.

As guidelines for further research we suggest adding credit bureau data to the existing dataset as well as adding more datasets from transitional countries in order to obtain more generalized results. A comparative methodology analysis can also be extended by adding more NN algorithms such as unsupervised classifiers. Other artificial intelligence techniques such as genetic algorithms, expert systems and others are also worth of exploring in credit scoring modeling.

References

Altman EI, Marco G, Varetto F. 1994. Corporate distress diagnosis: Comparison using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance* **18**: 505-529.

Apte C, Weiss S. 1997. Data Mining with Decision Trees and Decision Rules. *Future Generation Computer Systems* **13**: 197-210.

Arriaza BA. 1999. Doing business with small business. *Business Credit* Nov/Dec: **101**, Issue 10: 33-36.

Berry MJA., Linoff G. 1997. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley & Sons, Inc.: New York, Toronto.

Crook J, Banasik J. 2002. Does Reject Inference Really Improve the Performance of Application Scoring Models? *Working Paper Series No.02/3.* The Credit Research Centre, The School of Management, The University of Edinburgh: 1-27.

Desai VS, Crook JN, Overstreet GA. 1996. A comparison of neural network and linear scoring models in credit union environment. *European Journal of Operational Research* **95**: 24-35.

Desai VS, Conway DG, Crook JN, Overstreet GA. 1997. Credit scoring models in credit union environment using neural network and generic algorithms. *IMA Journal of Mathematics Applied in Business & Industry* **8**: 323-346.

Feldman R. 1997. Small business loans, small banks and a big change in technology called credit scoring. *Region* **11**, Issue 3: 18-24.

Frame WS, Srinivasan A, Woosley L. 2001. The effect of credit scoring on small business lending. *Journal of Money, Credit and Banking* **33**., No.3, August: 813-825.

Friedland M. 1996. Credit scoring digs deeper into data. *Credit World*, May/June **84**, Issue 5: 19-24.

Galindo J, Tamayo P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modelling Applications. *Computational Economics* **15**: 107-143.

Hand, DJ, Jacka SD(eds). 1998. *Statistics in Finance.* Arnold Application in Statistics, John Wiley & Sons Inc.: New York

Hand DJ, Henley WE. 1997. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society A* **160**: 523-541.

Harrel FE Jr. 2001. Regression modelling strategies with applications to linear models, logistic regression and survival analysis. Springer: Berlin.

Kartalopoulos, SV. 1996. *Understanding Neural Networks and Fuzzy Logic, Basic Concepts and Application*. IEEE Press.

Kasapovic M. 1996. Democratic transition and political institutions in Croatia; *Politicka misao (Political mind)*, **33**: No. 2-3: 84-99.

Lewis EM .1992. *An Introduction to Credit Scoring*, Fair Isaac and Co., Inc: San Rafael.

Li EY. 1994. Artificial Neural Networks and Their Business Applications. *Information & Management*. **27**: 303-313.

Masters T. 1993. *Practical Neural Network Recipes in C++*. Academic Press.

Masters T. 1995. *Advanced Algorithms for Neural Networks, A C++ Sourcebook.* John Wiley & Sons.

Mervar A. 2002. Ekonomski rast i zemlje u tranziciji (Economic development and transitional countries). *Privredna kretanja i ekonomska politika (Economic movements and economic policy).* **12**: 53-87.

NeuralWare. 1998. *Neural Computing, A Technology Handbook for NeuralWorks Professional II/Plus and NeuralWorks Explorer, NeuralWare*. Aspen Technology, Inc.

Patterson DW. 1995. *Artificial Neural Networks*, Prentice Hall.

Sheskin DJ. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press: Washington D.C.

Refenes AN, Zapranis A, Francis G. 1997. Stock Performance Modeling Using Neural Networks: A Comparative Study with Regression Models. *Neural Networks* **7**, No. 2: 375-388.

Skare M. 2000. Ogranicenja razvoja malih i srednjih poduzeca u zemljama tranzicije (Limitations of small and medium enterprises development). *Hrvatska gospodarska revija (Croatian economic review)* **49,** No. 7: 1-9.

Swicegood P, Clark JA. 2001. Off-site Monitoring Systems for Predicting Bank Underperformance: A Comparison of Neural Networks, Discriminant Analysis, and Professional Human Judgement. *International Journal of Intelligent Systems in Accounting, Finance and Management.* **10**: No. 3: 169-186.

Wasserman PD. 1993. *Neural Computing: Theory & Practice*. Van Nostrand Reinhold: New York.

West D. 2000. Neural Network Credit Scoring Models. *Computers & Operations Research*. **27**: 1131-1152.

Westphal C, Blaxton T. 1998. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley & Sons, Inc: New York.

Witten IH, Frank E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufman Publishers: San Francisco.

Wong BK, Bodnovich TA, Selvi Y. 1997. Neural Network Applications in Business: A Review and Analysis of the Literature (1988-95). *Decision Support Systems* **19**: 301-320.

Yobas MB, Crook JN, Ross P. 2000. Credit scoring using and evolutionary techniques. *IMA Journal of Mathematics Applied in Business & Industry* **11**: 111-125.

Zahedi F. 1996. A Meta-Analysis of Financial Applications of Neural Networks. *International Journal of Computational Intelligence and Organizations* **1**, No. 3: 164-178.