Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, June 12-14, 2006 (pp164-169)

Modeling customer revolving credit scoring using logistic regression, survival analysis and neural networks

NATASA SARLIJA [a], MIRTA BENSIC [b], MARIJANA ZEKIC-SUSAC [c]

[a] Faculty of Economics, J.J.Strossmayer University of Osijek, Gajev trg 7, Osijek, CROATIA
http://www.efos.hr/nastavnici/nsarlija

[b] Department of Mathematics, J.J.Strossmayer University of Osijek, Gajev trg 7, Osijek, CROATIA
http://www.mathos.hr/~mirta

[c] Faculty of Economics, J.J.Strossmayer University of Osijek, Gajev trg 7, Osijek, CROATIA
http://www.efos.hr/nastavnici/mzekic

Abstract: The aim of the paper is to discuss credit scoring modeling of a customer revolving credit depending on customer application data and transaction behavior data. Logistic regression, survival analysis, and neural network credit scoring models were developed in order to assess relative importance of different variables in predicting the default of a customer. Three neural network algorithms were tested: multilayer perceptron, radial basis and probabilistic. The radial basis function network model produced the highest average hit rate. The overall results show that the best NN model outperforms the LR model and the survival model. All three models extracted similar sets of variables as important. Working status and client's delinquency history are the most important features for customer revolving credit scoring on the observed dataset.

Key-Words: credit scoring modeling, logistic regression, revolving credit, survival analysis, neural networks

1 Introduction

According to Hand and Henley [9] credit scoring is a process of determining how likely applicants are to default on their repayments. It includes various decision models and tools that assist credit managers in making credit decisions. The aim is to assess the risk of default associated with a credit product/decision. Although credit scoring has been used in commercial and consumer lending for a few decades, exploring new models that will produce more accurate predictions is still of research interest. Scoring models can be divided into two types [12]: (i) credit scoring in targeting and application, which deals with new applicants, and (ii) credit scoring in managing existing accounts called behavior scoring.

This paper is focused on combining the characteristics from the behavior data and socio-economic data, in order to build a successful model for making on-going decisions about open accounts. Selection of characteristics used in developing a credit scoring model for a consumer credit depends on the type of the credit, the amount of the credit and the purpose of the credit. The paper deals with personal open-end accounts where clients can make different kinds of payments, withdrawals and revolving credits. There is a lack of published works that reveal composition of the models in that area, and the previous research is mostly concentrated on credit scoring modeling in developed countries. In case of an undeveloped country, or a transitional country with no credit bureau and no previous research about credit scoring models, it is challenging to see what key features will be extracted. The experiments are conducted using Croatian dataset which is influenced by specific economic conditions of a transitional country such as a fall of economic activity and high unemployment rate. Three methods were used in modeling: logistic regression (LR), Cox survival analysis, and neural networks (NN).

2 Methodology

Traditionally, different parametric methods are used for classifying input vectors into one of the two groups, which is the main objective of statistical inference on the credit scoring problem. Here we used

LR as a classical one, survival analysis in order to observe the client's behavior in time, and NN as a nonparametric intelligent method which was frequently used in data mining to discover nonlinear relationships among variables.

2.1 Logistic regression

The reason for using LR is to find determinants of default in revolving credit in personal open-end accounts, as well as to provide the probability that a client will default in the next 6 months. LR modeling is widely used for analyzing multivariate data involving binary responses that we deal with in credit scoring modeling. 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

$$\left(\log\left(\frac{\pi_1}{1-\pi_1}\right), \ldots, \log\left(\frac{\pi_n}{1-\pi_n}\right)\right) \quad (1)$$

as a canonical parameter (π_j is a probability that Y_j becomes 1), the assumption of the LR model is a linear relationship between canonical parameter and the vector of explanatory variables $\mathbf{x}_j$ (dummy variables for factor levels and measured values of covariates):

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \mathbf{x}_j^{\tau}\boldsymbol{\beta} \quad (2)$$

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:

$$\pi_j = \exp\left(\mathbf{x}_j^{\tau}\boldsymbol{\beta}\right) / \left(1 + \exp\left(\mathbf{x}_j^{\tau}\boldsymbol{\beta}\right)\right) \quad (3)$$

In order to extract important variables, the forward selection procedure, available in SAS software, is used with standard overall fit measures.

2.2 Survival analysis

Survival-based model is created using the length of time before a loan defaults (let us denote it by T). The main interest in survival modeling is in describing the probability that T is larger then a given time t (the survivor function $S(t)$). The most interesting object for this is a hazard function, $h(t)$, which can be described by the equation:

$$h(t)\Delta t = \text{Prob}\{t \le T \le t+\Delta t \mid T \ge t\} \quad (4)$$

which denotes the probability that a loan defaults in the next instant period conditioned by the fact that it survives at least to the time t [7]. In the proportional hazard models, it is a hazard function that is modeled using the explanatory variable. The main assumption of a proportional hazard model is that the explanatory variables have a multiplier effect on the hazard rate which does not depend on time t:

$$h(t) = \exp\left(\mathbf{x}^{\tau}\boldsymbol{\beta}\right) h_0(t) \quad (5)$$

Cox's regression is used in order to find determinants of default in personal open-end accounts, including time to default and to provide the likelihood of default in the period of next 6 months. The Cox procedure is used because it enables estimation of the regression parameters without knowing the type of baseline hazard function $h_0(t)$. As we are not, at this moment, interested in the form of the hazard function, but only in the significant variables, it was sufficient for this purpose [7]. The forward selection procedure, available in SAS software, is used for the variable extraction.

2.3 Neural network classifiers

Although many research results show that NNs can solve almost all problems more efficiently than traditional modelling and statistical methods, there are some opposite research results showing that statistical methods in particular data samples outperform NNs. The lack of standardized paradigms that can determine the efficiency of certain NN algorithms and architectures in a particular problem domain is emphasized by many authors [11]. Therefore, we test three different NN classifiers: backpropagation, radial basis function network, and probabilistic network. First two algorithms were tested using tangent hyperbolic functions in the hidden layer, and the softmax activation function in the output layer in order to obtain probabilities. The learning rate ranged from 0.01 to 0.08 during the training phase. The momentum was set to 0.3. Overtraining was avoided by a cross-validation procedure which alternatively trains and tests the network until the performance of the network on the test sample does not improve for n number of iterations. After training and cross-validating the network on maximum 10000 iterations, all the NN algorithms were tested on the out-of-sample data 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 [11], which uses nonparametric estimation methods for classification.

The topology of networks consisted of an input layer, a hidden or a pattern layer, and an output layer. 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 RBFN. The number of hidden units in the probabilistic NN was set to the size of the training sample.

The inputs with small fan-out weights were pruned during the learning phase in order to select important variables. Sensitivity analysis is performed on the test sample in order to determine the significance of selected variables to the output.

3 Review of previous research results

Previous research has shown that common issues in behavior scoring models are delinquency history, payment history, and usage history [12]. Hamilton and Khan [8] found that important discriminating variables between those who defaulted and those who did not are behavior characteristics such as cash advances, minimum payment due, and interest paid in the previous period. Avery et al. [2] emphasize that a failure to consider situational circumstances in credit scoring may influence the accuracy of the model. Their results show that unemployment rate is positively associated with estimated likelihood of default, than there is a higher probability of default with lower income individuals. Marital status also affects probability of default. Andreeva et al. [1] analyzed a revolving store card and found out that a combination of application, purchase and behavior characteristics is the most important in predicting time of default. The research of Noh et al. [13] shows that the default increases with a higher utilization rate, but decreases with a higher contracted overdraft, a higher number of sales transactions made over the account, a client's age and her/his relationship with the bank. Dey and Mumy [6] found that the higher clients' income, age, longer usage of the account, the lower their likelihood to default. Also, less riskier clients are those who are employed and haven't previously experienced bankruptcy.

Concerning the methodology used in credit scoring modeling, the most frequent method is the LR [9]. Survival analysis is also often used in behavior scoring modeling. Banasik et al. [3] compare logistic regression to survival analysis in analysing a personal loan data set. Andreeva et al. [1] showed that survival analysis is competitive with LR and there is a little difference in classification accuracy between parametric models, non-parametric Cox PH model and LR. Baesens et al. [4] showed that LR and Cox have the same accuracy in predicting default in the first 12 months, while NNs were more accurate, but not significantly. In predicting loan default between 12 and 24 months LR had a hit rate of 78.24%, Cox model 77.50% and NNs 78.58%, where NNs were significantly better than the Cox model. Malhotra and Malhotra [10] used NNs and LDA to classify consumer loans and obtained the overall mean accuracy of 71.98% by NNs and 69.32% by LDA, where the NNs significantly outperformed LDA.

4 Variables and data

Data was collected randomly in a Croatian bank, covering the period of 12 months in 2004. An observation point is settled in the middle of the period, on June 30. A period preceding this point is the performance period and the characteristics of the performance in this period were used for developing scoring models. On the basis of its performance, in the period of 6 months after the observation point, a client is defined as "good" or "bad" [14]. A client is "bad" if she/he exceeds a contracted overdraft for more than 35 days during the period of 6 months. Otherwise, a client is considered to be "good". The total number of 35 input variables that deal with personal open-end accounts is used. The variables can be divided into three main groups: (i) demographic data; (ii) socio-economic data; (iii) behavior data – repayment and usage (average values for the period of 6 months). Due to the lack of credit bureau in the country it was not possible to examine the influence of some other variables that create credit bureau score.

5 Sampling procedure

The total sample consisted of 44087 cases. Majority of applicants in the sample were good (96.80%), and 3.20% were bad. Using the appropriate cut-off [5], the LR model was estimated on the in-sample data (34879 or approximate 80% of cases), and validated on the out-of-sample data (9208 or approximate 20% of cases). Because of the nature of its objective function, NNs require equal number of good and bad applicants in the training sample. Therefore, the surplus of good applicants was removed from the in-sample data and NN models were trained on the equal number of good and bad applicants, cross-validated in order to optimize the training time and the network topology, and finally tested on the out-of-sample test data. The subsamples were created using a random selection of cases into the train, cross-validation and test sample, while keeping the equal distribution of good and bad applicants in the train and cross-validation sample. The proportion of good and bad applicants in the validation sample corresponds to the proportion of good and bad applicants in the whole sample. In order to enable the comparison, the same out-of-sample validation data was used to test LR, Cox and NN models. The distribution of applicants in the subsamples separately for LR, Cox, and NN models is given in Tables 1 and 2.

Table 1. Distribution of applicants in the sample - LR and Cox models

Subsample	No. of cases	Good applicants		Bad applicants	
		No.	%	No.	%
Estimation	34879	33745	96.75	1134	3.25
Test	9208	8930	96.98	278	3.02
Total	44087	42675	96.80	1412	3.20

Table 2. Distribution of applicants in the sample - NN models

Subsample	No. of cases	Good applicants		Bad applicants	
		No.	%	No.	%
Train	1800	900	50.00	900	50.00
Cross-validation	450	225	50.00	225	50.00
Test	9208	8930	96.98	278	3.02
Total	11458	10055	87.76	1403	12.24

6 Results of credit scoring models

6.1 Logistic regression scoring model

The final LR scoring model, using forward selection, ended with 18 variables, and has the following standard overall fit measures: Score= 8309.4855, (p <.0001), Wald= 2919.9757 (p <.0001)].

In order to measure score performance ROC Curve is used as a plot of the cumulative score distribution of the bad accounts versus good accounts (Fig. 1), where the cut-off maximizes expected profit [5]. LR results of the out-of-sample data showed the average hit rate of 77,58%, good hit rate of 76,01% and bad hit rate of 79,14%. The percentage of the good clients estimated as bad ones is 23,99%, and the percentage of bad clients estimated as good is 20,86%.

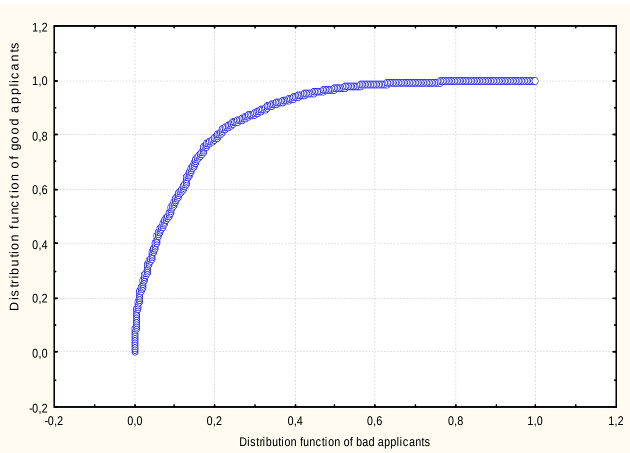

Fig. 1 ROC curve of the LR model

On the basis of the LR model results it can be seen that default is increased: (i) with higher withdrawals by cheques; (ii) with higher number of days client's overdraft exceeds contracted overdraft; (iv) with higher number of continuous months with overdraft exceeded; (v) if the client doesn't have a permanent job; (vi) if the client doesn't have a second job. The opposite influence on probability of default has been found in the following variables: (i) higher amount of contracted overdraft; (ii) higher salaries; (iii) higher amount of cash paid from the account; (iv) higher amount of payments to the account from other account; (v) higher balance; (vi) higher number of days from the last time the client exceeded contracted overdraft in the previous period; (vii) older age of the client and higher number of years a client owns the account.

6.2 Survival analysis model results

The final Cox's regression scoring model, using forward selection, ended with 10 input variables. The Cox's regression model has the following standard overall fit measures: Score=8777.4851 (p<.0001), Wald=3407.2117 (p<.0001). The ROC curve (Fig. 2) has also been used as a measure of performance in the same way as in the LR model. Test results of the model obtained on the holdout sample showed the average hit rate of 77,7%, good hit rate of 75,91% and bad hit rate of 79,50%.

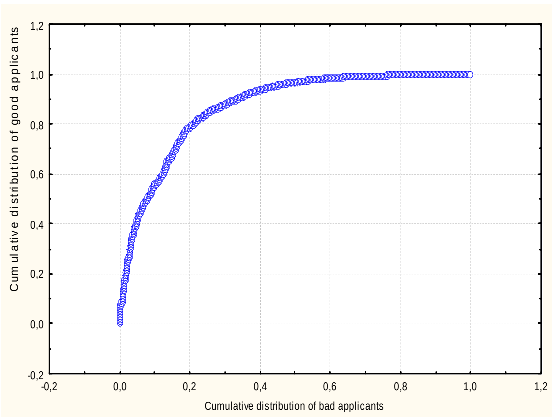

Fig. 2 ROC curve of the Cox survival model

The Cox model indicates that as the amount of contracted overdraft increases, the rate of default decreases. The same applies for the number of days since the last time when a client exceeded contracted overdraft. Also, if the client is older, as well as his account, it will decrease probability of default. On the contrary, as the number of days the client's overdraft exceeded contracted overdraft increases, the rate of default also increases. Female clients are more likely to have an account in default compared to male clients. If the client is without a permanent job it is more likely that the account will be in default compared to employed clients. Also, if the client doesn't have a second job it is more likely that he/she will be in default compared to the client with an extra job.

Both, LR and Cox ROC curves, are very similar, which has also been shown by statistical measures of associations between predicted output values of the two models (Kendall's tau b = 0,93 and Spearman Rank R = 0,93).

6.3 Neural network model results

The results of the three tested NN architectures are presented in Table 3. After training and cross-validation, the backpropagation model ended with 30 input variables and 30 hidden units. The RBFN yielded best results with 15 input variables and 28 hidden units, while the probabilistic NN found all 35 input variables as important. The highest average hit rate of 78.05% is produced by the RBFN; its bad hit rate was 82.71%, while the good hit rate was 73.28%.

Table 3. Hit rates of the NN models

Model	Ave. hit rate	Bad hit rate (%)	Good hit rate (%)
Backprop 30-30-1	77.86	83.45	72.27
RBFN 15-28-1	**78.05**	82.71	73.38
Probabilistic 35-1798-2	77.28	82.98	77.28

The five highest ranked variables extracted by the best NN algorithm are concerned with a client's delinquency history (see Table 4). The next important set of features, according to the RBFN model, describes the payment history and the account characteristics, followed by the client's demographic characteristics and other variables.

The variables extracted by LR, Cox, and NN models are presented in Table 5. P values are given for the LR and Cox model, while the rank of a variable importance according to the sensitivity analysis is given for the best NN model.

Table 5. Variables extracted by the three models

Variable	LR p value	Cox p value	NN rank
Working status - Retired	**<.0001**	**<.0001**	11
Working status – Employed	**0.0162**	**<.0001**	10
Working status - No answer	**0.0106**	**<.0001**	-
Does the client have a second job	**0.0398**	**0.0525**	14
Contracted overdraft	**<.0001**	**0.0001**	5
Payments to the account from other accounts	**0.0212**	>.05	-
Regular payments (salaries)	**<.0001**	>.05	7
Interests paid to the client	**<.0001**	>.05	-
Withdrawals by cheques	**<.0001**	>.05	-
Cash withdrawals from the account	**<.0001**	>.05	-
Interests paid from the client	**<.0001**	>.05	6
Total payments on the account minus total withdrawals from the account	**<.0001**	>.05	-
Balance	**<.0001**	>.05	-
Total payments to the account/ total withdrawals from the account	**0.0019**	>.05	-
Amount of money that exceeded contracted overdraft	**0.0284**	**<.0001**	-
Number of days the client's overdraft exceeded contracted overdraft	**<.0001**	**<.0001**	4
Number of continues months with overdraft exceeded	**0.0003**	**<.0001**	3
Number of days since the last time when the client exceeded her/his contracted overdraft	**<.0001**	**<.0001**	1
Number of years a client owns the account	**<.0001**	**<.0001**	8
Client's age	**<.0001**	**0.032**	9
Gender	>.05	**0.0097**	15
Does the client live in a town	>.05	>.05	13
Does the client have savings account	>.05	>.05	12
How many times client's overdraft exceeded contracted overdraft	>.05	>.05	2

The LR model extracted the largest number of variables, and all the variables that were extracted by the Cox were also found important in the LR model. It is evident that the working status, client's age, contracted overdraft, and second job are found as important client's characteristics in all three models. The another set of variables describing payment history that was also extracted by all three models consists of the following variables: number of days the client's overdraft exceeded contracted overdraft, number of continues months with overdraft exceeded, number of days since the last time when the client exceeded her/his contracted overdraft, and the number of years a client owns the account. It is evident from the above that payment history is the most important for credit scoring modeling of open ended accounts.

6 Conclusion and discussion

The paper aimed to identify important features for customer revolving credit scoring modeling for open-end accounts on a Croatian dataset. A standard LR is used in addition to Cox survival analysis and three NN algorithms in order to classify clients into good or bad ones due to the probability of their default. The most successful NN algorithm was the radial basis function network producing higher average hit rate than the LR and Cox models.

The above findings confirm some previous results in consumer behavior in open-end accounts, in the sense that delinquency, payment, and usage history were also found important by other authors. However, it also shows that the working status (having a permanent and a second job) is a specific feature extracted on the examined dataset. It is not surprising due to the specific economics conditions present in the observed dataset, such as high unemployment rate. In order to develop models that will be more sensitive to macroeconomic factors, additional variables such as unemployment rate, interest rate, stock market, housing market, etc., should be included into the data. Also, more significant results in testing the impact of economy would be accomplished if time seria approach is taken with more than one data sample. In order to achieve generalization of the results, more datasets from different transitional countries should be included. The additional NN models that deal with time series and survival analysis could also be examined.

References:

[1] G. Andreeva, J. Ansell, J.N. Crook, Modelling the Purchase Propensity: Analysis of a Revolving Store Card, Journal of the Operational Research Society, Vol. 56, 2005, pp. 1041-1050.

[2] R.B. Avery, P.S. Calem, G.B. Canner, Consumer Credit Scoring: Do Situational Circumstances Matter?, Journal of Banking and Finance, Vol. 28, 2004, pp. 835-856.

[3] J. Banasik, J.N. Crook, L.C. Thomas, Not If But When Will Borrower Default, Journal of the Operational Research Society, Vol. 50, 1999, pp. 1185-1190.

[4] B. Baesens, T. Gastel, M. Stepanova, J. Vanthienen, Neural Network Survival Analysis for Personal Loan Data, Journal of the Operational Research Society, Vol. 56, 2005, pp.1089-1098.

[5] P. Beling, Z. Covaliu, R.M. Oliver, Optimal Scoring Cutoff Policies and Efficient Frontiers, Journal of the Operational Research Society Vol. 56, 200, pp. 1016-1029.

[6] S. Dey, G. Mumy, Determinants of Borrowing Limits on Credit Cards, Bank of Canada Working Paper 2005-7, 2005.

[7] E. Frank, J.R. Harrell, Regression Modeling Strategies, Springer-Verlag, New York, 2001.

[8] R. Hamilton, M. Khan, Revolving Credit Card Holders: Who Are They and How Can They Be Identified?, The Service Industries Journal, Vol. 21, No. 3, 2001, pp. 37-48.

[9] D.J. Hand, W.E. Henley, Statistical Classification Methods in Consumer Credit Scoring: A Review, Journal Royal Statistical Society A, Vol. 160, 1997, pp. 523-541.

[10] R. Malhotra, D.K. Malhotra, Evaluating Consumer Loans Using Neural Networks, Omega, Vol. 31, 2003, pp. 83-96.

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

[12] H. McNab, A. Wynn, Principles and Practice of Consumer Credit Risk Management, CIB Publishing, Canterbury, 2000.

[13] H.J. Noh, T.H. Roh, I. Han, Progonostic Personal Credit Risk Model Considering Censored Information, Expert Systems with Applications, Vol. 28, 2005, pp. 753-762.

[14] L.C. Thomas, J. Ho, W.T. Scherer, Time Will Tell: Behavioural Scoring And The Dynamics of Consumer Credit Assessment, IMA Journal of Management Mathematics 12, 2001, pp. 89-103.