SHORT PAPERS

UDC: 336.77
JEL Classification: B41, C14, D81, G21, P43
Keywords: credit scoring; discrimination analysis; banking sector; pattern recognition; retail loans

Credit Scoring Methods

Martin VOJTEK – Evžen KOČENDA*

1. Introduction

Despite the proliferation of banking services, lending to industry and the public still constitutes the core of the income of commercial banks and other lending institutions in developed as well as post-transition countries. From the technical perspective, the lending process in general is a relatively straightforward series of actions involving two principal parties. These activities range from the initial loan application to the successful or unsuccessful repayment of the loan. Although retail lending belongs among the most profitable investments in lenders' asset portfolios (at least in developed countries), increases in the amounts of loans also bring increases in the number of defaulted loans, i.e. loans that either are not repaid at all or cases in which the borrower has problems with paying debts. Thus, the primary problem of any lender is to differentiate between "good" and "bad" debtors prior to granting credit. Such differentiation is possible by using a credit-scoring method. The goal of this paper is to review credit-scoring methods and elaborate on their efficiency based on the examples from the applied research. Emphasis is placed on credit scoring related to retail loans.

We survey the methods which are suitable for credit scoring in the retail segment. We focus on retail loans as sharp increase in the amounts of loans for this clientele has been recorded in the last few years and another increase can be expected. This dynamic is highly relevant for post-transition countries. In the last few years, banks in the Czech and Slovak Republics have allocated a significant part of their lending to retail clientele. In 2004 alone, Czech and Slovak banks recorded 33.8% and 36.7% increases in retail loans, respectively. Hilbers et al. (2005) review trends in bank lending to the private sector, with a particular focus on Central and Eastern European countries, and find that rapid growth of private sector credit continues to be a key challenge for most of these countries. In the Czech and Slovak Republics the financial liabilities of households formed 11 % and 9 %

* CERGE-EI, joint-place of Charles University and Academy of Sciences, Prague
(martin.vojtek@cerge-ei.cz), (evzen.kocenda@cerge-ei.cz)

The authors thank Petr Franěk, Vilém Semerák, an anonymous referee and participants of the Algoritmy 2005 (Podbanské) and MME 2005 (Hradec Králové) conferences for their helpful comments. The support of the Grant Agency of the Czech Republic under Grant No. 402/05/0931 is gratefully acknowledged. The usual disclaimer applies.

of the GDP, respectively, in 2004.[1] Since the average ratio of financial liabilities to GDP in the older 15 members of the European Union is about five times higher, and taking into account the recent trend of decreasing interest rates, it is expected that the amount of loans to retail clientele will continue to increase.

Empirical studies on credit scoring with respect to retail loans are infrequent in the relevant literature on developed markets and, to our best knowledge, no such empirical study exists with respect to retail loans in post-transition countries. We conjecture that this is due to the sensitivity of information associated with privacy laws that results in an aversion of banks to provide this data. Thus, most of the credit-scoring literature deals with industry loans, i.e. loans received by firms. Industry-loans credit scoring (or, in general, rating assignment) is different from that of commercial loans in several instances. Primarily, the amounts lent are much lower in the case of retail lending. Most importantly, there are different decision variables used in the decision process regarding industry loans: various ratios of financial indicators, for example. To assess the efficiency of various scoring methods we primarily draw on the empirical studies that deal with retail loans. When appropriate, we also use those that deal with industry loans since they often provide a focused illustration of a particular scoring method's abilities.[2] In order to properly distinguish the different aspects of retail credit scoring, we also provide a comprehensive discussion on the indicators that are typically important in the credit-scoring models and are employed in the literature.

The rest of the paper is organized as follows: in Section 2 we introduce the concept of incurred costs that is essential to credit scoring; then, in the following sub-sections, we review the most widely used credit-scoring methods and provide representative results from the empirical literature. Section 3 concludes with a brief assessment of the reviewed methods. The appendix contains indicators typical to retail credit-scoring models as known from the literature and practice.

2. Techniques Used in Credit Scoring

The process of credit scoring is very important for banks as they need to segregate "good borrowers" from "bad borrowers" in terms of their creditworthiness. This is a classic example of asymmetric information, where a bank has to reveal hidden data about its client. Credit scoring in its automated form is often the only way to assess creditworthiness, as banks do not have enough resources to treat each small exposure individually. The methods generally used for credit scoring are based on statistical pattern-recognition techniques.[3] These sophisticated methods defy the perception that often regards credit scoring as being a simple data-mining process. Statistical pattern recognition is a growing research area. Renault and

[1] These numbers cover only the banking sector and not various financing companies, etc. Hence there is a large space for expansion in the financial liabilities in both countries.

[2] Altman and Narayanan (1997) provide a broad review of the corporate failure models and their classification.

de Servigny (2004) claim that “[...] this means that most current statistical models applied to credit risk are lagging far behind the state-of-the-art methodologies. As a consequence, we anticipate that in coming years banks will be catching up with the integration of nonparametric techniques and machine learning models.”

In this paper we review the best developed and most frequently applied methods of credit scoring employed by banks when evaluating applications for loans based on the (private) information provided by the borrower.[4] We will discuss problems connected with implementing these approaches as well.

An essential concept relevant to credit scoring is that of incurred costs associated with the probabilities of repayment of loans. For simplicity let us assume that the population of loans consists of two groups or classes G and B that denote loans that (after being granted) will turn good and bad in the future, respectively. Good loans are repaid in full and on time. Bad loans are subject to various degrees of default.[5]

Usually the class sizes are very different, so that for the probability that a randomly chosen candidate (customer) belongs to group G, denoted as p_G, we have $p_G > p_B$. Let x be a vector of independent variables[6] (called measurement vector) used in the process of deciding whether an applicant belongs to group G or B. Let the probability that an applicant with measurement vector x belongs to group G is $p(G|x)$, and that of B is $p(B|x)$. Let the probability $p(x|G)$ indicate that a good applicant has measurement vector x. Similarly, for a bad applicant the probability is $p(x|B)$. The task is to estimate probabilities $p(\cdot|x)$ from the set of given data about applicants which turn out to be good or bad and to find a rule for how to partition the space X of all measurement vectors into the two groups A_G and A_B based on these probabilities so that in A_G would be the measurement vectors of applicants who turn out to be good and vice versa. However, it is usually not possible to find perfect classification as it may happen that the same vector is given by two applicants where one is good and the other is bad. Therefore it is necessary to find a rule that will minimize the costs of the agency providing credit connected with the misclassification of applicants. Let us denote c_G as the costs connected with misclassifying a good applicant as bad and c_B as the costs connected with classifying a bad applicant as good. Usually $c_B > c_G$, because costs incurred due to misclassify-

[3] The first major use of credit scoring dates back to the 1960s, when credit card business grew up and the automatized decision process was a must. Credit scoring was fully recognized in the USA by the 1975 Equal Opportunity Act, which stated that any discrimination can be based only on statistical assessments.

[4] Discussing the issue that applicants are subject to various external influences such as external social or financial changes, which can change their probability of default, is beyond the scope of this paper, however. This can be incorporated into the behavioral scoring model. There is a practice to offer loans to clients based mainly on behavioral credit scoring (average balance on their checking accounts, etc.). This practice is common with banks trying to aggressively increase their credit retail portfolio.

[5] For a precise definition of default according to the new Basel II framework, see (BIS, 2004).

[6] In practice the variables are often correlated. However, some methods can deal with such a data problem as multicollinearity. We comment on this issue later on when describing various techniques. Assumption of independence is maintained here for ease of exposition.

ing a bad customer are financially more damaging than costs associated with the former kind of error. If applicants with x are assigned to class G the expected costs are $c_B\, p(B\,|\,x)$ and the expected loss for the whole sample is $c_B \sum_{x \in A_G} p(B\,|\,x)p(x) + c_G \sum_{x \in A_B} p(G\,|\,x)p(x)$, where $p(x)$ is a probability that the measurement vector is equal to x. This is minimized when into group G such applicants are assigned who have their group of measurement vectors

$$A_G = \{x\,|\,c_B p(B\,|\,x) \leq c_G p(G\,|\,x)\}$$

which is equivalent to

$$A_G = \left\{ x \,\middle|\, p(G\,|\,x) \geq \frac{c_B}{c_B + c_G} \right\} \tag{1}$$

Without loss of generality we can normalize misclassification costs to $c_B + c_G = 1$. In that case the rule for classification is to assign an applicant with x to class G if $p(G\,|\,x) > c_B$ and otherwise to class B.

Of course, an important task is to specify the cost of lending errors for retail loans and, in doing so, more accurately specify the optimal cutoff-score approach to credit scoring. Namely the bank has to choose the optimal trade-off between profitability and risk. Loan policies that are too restrictive may ensure minimal costs in terms of defaulted loans, but the opportunity costs of rejected loans may exceed potential bad debt costs and thus profit is not maximized. Conversely, policies that are too liberal may result in high losses from bad debts.

2.1 Linear Discriminant Analysis

The aim of Linear Discriminant Analysis (hereinafter "LDA") is to classify a heterogeneous population into homogeneous subsets and further the decision process on these subsets. We can assume that for each applicant there is a specific number of explanatory variables available. The idea is to look for such a linear combination of explanatory variables, which separates most subsets from each other. In a simple case of two subsets, the goal is to find the linear combination of explanatory variables, which leaves the maximum distance between means of the two subsets.

In a general case we consider the distributions $p(x\,|\,G)$ and $p(x\,|\,B)$ which are multivariate normal distributions with the common variance. Then equation (1) reduces to

$$A_G = \left\{ x \,\middle|\, \sum w_i x_i > c \right\} \tag{1'}$$

as follows from the econometrics theory. Here x_i are explanatory variables, w_i are associated coefficients (weights) in the linear combination of explanatory variables. If one takes $s\ (x) = \Sigma w_i x_i$ then it is possible to discriminate according to this *score* and thus to reduce the problem to only one dimension.

The discriminant analysis was introduced by Fisher (1936) who searched for the best way to separate two groups using linear combination of variables.[7] Eisenbeis (1977) criticized this method by stating that the rule is optimal only for a small class of distributions. However, Hand and Henley (1997) claim that "if the variables follow a multivariate ellipsoidal distribution (of which the normal distribution is a special case), then the linear discriminant rule is optimal".[8]

Other critiques state that there is a selection bias due to the fact that a learning sample for a credit-scoring system is made of applicants to whom credit has been granted. That means that results are biased when applied to the whole population. Eisenbeis (1977) also saw problems in the definition of bad and good groups in the case when no clear boundary is between them and under the assumption that the covariance matrices of the two distributions are equal. In this case the use of quadratic discriminant analysis instead of the linear case is appropriate. Problems also arise when one wants to test for the significance of individual variables as one does not have the assumption of normality and therefore cannot perform statistical inference.[9]

Altman (1968), who was the first to apply discriminant analysis, constructed the so-called *z*-score, which is a linear combination of several explanatory variables for the case of the corporate credit granting problem.[10] He found the model to be extremely accurate in correctly predicting bankruptcy.

As we have mentioned, the advantages of the LDA method are that it is simple, it can be very easily estimated and it actually works very well; it is often used by banks for credit-scoring purposes. The disadvantage is that LDA requires normally distributed data but the credit data are often non-normal (and categorized).

2.2 Logit Analysis

As stated earlier, distribution of the credit information data is usually non-normal, and this fact may theoretically pose a problem when conducting an LDA.[11] One way to overcome the problems with non-normality of data is to use an extension of the LDA model that allows for some parametric distribution. In this case a suitable extension is a generalized linear model known as logit model. Given a vector of application characteristics x, the probability of default p is related to vector x by the relationship:

$$\log\left(\frac{p}{1-p}\right) = w_0 + \sum_i w_i \log x_i \tag{2}$$

[7] Fisher (1936) suggested (under assumption of common sample variance) looking for the linear combination of explanatory variables which leaves the maximum distance between means of the two classes.

[8] The proof of this claim can be found in (Webb, 2002).

[9] Many of these issues are addressed in the review by Rosenberg and Gleit (1994).

[10] The variables used are Sales/Total assets (TA), Working capital/TA, Retained Earnings/TA, Earnings before Interest and Taxation/TA, Market Value of Equity/Book Value of Total Debt.

[11] On the other hand, Reichert, Cho and Wagner (1983) argue that the non-normality of credit information is not a limitation for the use of LDA from the empirical point of view.

One of the advantages over linear discriminant analysis is the use of the maximum likelihood method for the estimation of parameters w_i. Another advantage is that one can provide the probabilities of being in some concrete class. The logit method can also deal with categorized data; the solution is to take dummy variables for each category of data. Several studies found that that logit model outperforms discriminant analysis.[12]

More recently, logit analysis became one of the main approaches of classification in credit scoring in the practices of banks. The coefficients obtained have the same values as in the studies that employed the LDA decision rule. Nevertheless, they are obtained under much weaker assumptions. Actual classification results are similar for both types of regression and because both are sensitive to high correlation among explanatory variables, one should ensure that there are no such variables left in the training set.[13] Another disadvantage of this method is the sensitivity to missing values (all observations with missing values have to be deleted). Similarly, as with linear discriminant analysis, the logit approach is limited by a parametric form of model.

Lawrence and Arshadi (1995) used the logit model for the analysis of the management of problem loans and of determinants of resolution choices using a series of borrower and bank variables. In the area of mortgage lending, Campbell and Dietrich (1983) utilized a logit model to show that the age of a mortgage, the loan-to-value ratio, interest rates, and unemployment rates are significant in explaining mortgage prepayments, delinquencies and defaults. Gardner and Mills (1989), recognizing that delinquent borrowers do not necessarily end up in default, employ a logit regression model to estimate the probability of default for currently delinquent loans. They recommend that bankers use this method to identify the severity of the problem and thereby formulate an appropriate response to the delinquency. Recently, Charitou, Neophytou and Charalambous (2004) found that the logit method is superior to other methods in predicting defaults.

2.3 *k*-nearest Neighbor Classifier

The k-nearest neighbor classifier serves as an example of the non-parametric statistical approach. This technique assesses the similarities between the pattern identified in the training set and the input pattern. One chooses a metric on the space of applicants and takes k-nearest neighbor (hereinafter "k-NN") of the input pattern that is nearest in some metric sense. A new applicant will be classified in the class to which the majority of the neighbors belong (in the case when the costs of misclassification are equal) or according to the rule expressed by equation (1). This means that this method estimates the $p(G \mid x)$ (or $p(B \mid x)$) probability by the proportion of G (or B) class points among the k-nearest neighbors to the point x to be classified.

[12] See for example (Wiginton, 1980).

[13] The training set is the data sample with known properties that serves to calibrate and verify the model performance before the model is applied.

The first to use this method were Chatterjee and Barcun (1970). Identifying the advantages of this method, Henley and Hand (1996) stated that the non-parametric nature of the method enables modeling of irregularities in the risk function over the feature space. The k-NN method has been found to perform better than other non-parametric methods such as kernel methods when the data are multidimensional. It is a fairly intuitive procedure and as such it could be easily explained to business managers who would need to approve its implementation. It can also be used dynamically by adding applicants when their class becomes known and deleting old applicants to overcome problems with changes in population over time.[14]

When performing the k-NN methodology, a very important step is the choice of the metric used. Henley and Hand (1996) describe the choice of the metric and the choice of the number of nearest neighbors to consider. A commonly used metric is the standard Euclidean norm given by

$$\rho_1(x, y) = [(x - y)^T (x - y)]^{1/2} \tag{3}$$

where x and y are measurement vectors.

However, when the variables are in different units or categorized[15], it is necessary to use some appropriate standardization of variables as well as to select some data-dependent version of the Euclidean metric such as:

$$\rho_2(x, y) = [(x - y)^T A(x - y)]^{1/2} \tag{4}$$

where A is a $n \times n$ matrix with n number of variables. As matrix A can depend on x we can define two types of metrics according to how A is selected: local metrics are those where A depends on x; global metrics are those where A is independent of x.[16]

The choice of the number of nearest neighbors chosen (k) determines the bias/variance trade-off in the estimator. The k has to be much smaller than the smallest class. A simulation study by Enas and Choi (1986) suggested that $k \approx n^{2/8}$ or $n^{3/8}$ is reasonable.[17]

Recently, Hand and Vinciotti (2003) observed that in problems where there are two unbalanced classes, the fact that k is finite (and thus asymptotic

[14] The method is often superior over linear discriminant analysis; this is shown for example by Ripley (1994), who compares data linear discriminant, neural networks, k-NN and various other classification methodologies.

[15] As mentioned earlier, many variables in measurement vector x are categorized and measured on different scales. One way to overcome this problem is to introduce convenient dummy variables. Another strategy is to use the so-called weights of evidence (w_{ij}) where the j-th attribute of the i-th characteristic is given by $w_{ij} = (p_{ij}/q_{ij})$, where p_{ij} is the number of those classified in G class in attribute j of characteristic i divided by the total number of good risks and similarly q_{ij} is a proportion of bad risks in attribute j of characteristic i.

[16] Henley and Hand (1996) proposed a global metric given by the parametrization $A = I + Dww^T$, where w is the direction orthogonal to equiprobability contours for $p(G\,|\,x)$, and D is a distance parameter. The direction w can be calculated using linear regression weights.

[17] On the other hand a choice of k via cross-validation on the misclassification rate is often adopted in empirical literature. The cross-validation consists of dividing the training sample on m subsets and then using $m - 1$ subsets to train; the last set is used as the test set, and this is repeated for each subset.

properties do not hold) results in a non-monotonic relationship between the k and the proportion of each class correctly classified. That means, in general, that a larger k may not yield better performance than a smaller k. For example if the number of points from the smaller class is less than $(1 - c_B)^{-1}$, then the best classification rule for predicting class G membership is to use $k = 1$.

Holmes and Adams (2002) realized that there is a lack of a formal framework for choosing the k and that the method can only make discrete predictions by reporting the relative frequencies which have no probabilistic interpretation. They tried to overcome these difficulties by proposing the Bayesian approach, which integrates over the choice of k. Such approach leads to the conclusion that marginal predictions are given as proper probabilities.

2.4 Classification and Regression Trees (CART)

Classification and Regression Trees (CART) is a nonparametric method that is due to Breiman, Friedman, Olshen and Stone (1984). It is a flexible and potent technique; however, it is used in banking practice chiefly only as a supporting tool to accompany the parametric estimation methods described earlier. It serves, for example, in the process to select regressors or characteristics with the highest explanatory power. The CART method employs binary trees and classifies a dataset into a finite number of classes. It was originally developed as an instrument for dealing with binary responses and as such it is suitable for use in credit scoring where the default and non-default responses are contained in data. The CART method was later refined in subsequent editions of Breiman, Friedman, Olshen and Stone (1984).

Similarly as with the methodologies reviewed earlier, we make the assumption of having a training set of measurement vectors $X_T = \{x^j\}$ along with information whether an individual j defaulted, or not (hence, y^j is coded 1 or 0 respectively). The CART tree consists of several layers of nodes: the first layer consists of a root node; the last layer consists of leaf nodes. Because it is a binary tree, each node (except the leaves) is connected to two nodes in the next layer. The root node contains the entire training set; the other nodes contain subsets of this set. At each node, this subset is divided into 2 disjoint groups based on one specific characteristic x_i from the measurement vector. If x_i is ordinal, the split results from the fact, related to a particular individual, as to whether $x_i > c$, for some constant c. If the previous statement is true, an individual j is classified into the right node; if not, an individual is classified into the left node. A similar rule applies, if x_i is a categorized variable.

The characteristic x_i is chosen among all possible characteristics and the constant c is chosen so that the resulting subsamples are as homogeneous in y as possible. In other words: x_i and c are chosen to minimize the diversity of resulting subsamples (diversity in this context will be defined presently). The classification process is a recursive procedure that starts at the root node and at each further node (with exception of the leaves) one single characteristic and a splitting rule (or constant c) are selected. First, the best split is found for each characteristic. Then, among these characteristics the one with the best split is chosen.

The result of this procedure is that subsamples are more homogeneous than the parental sample. The procedure ends when the node contains only individuals with the same y^j or it is not possible to decrease diversity further. For illustration purpose, let $p(\cdot \mid t)$ be the proportion of G and B groups present at node t. As an example of the diversity functions can be taken a *Gini index* function, which is defined as $d(t) = p(G \mid t)p(B \mid t)$. The value of constant c that is instrumental to splitting between nodes is defined as to minimize the weighted diversify in the daughter nodes (i.e. nodes to which node t is parental). Formally, the aim is to choose c that minimizes $p_L d\ (t_L) + p_R d(t_R)$,where p_L and p_R are the proportions of individuals going into nodes t_L and t_R respectively. The completed tree is usually very large but algorithms exist for pruning it into a simpler final tree.[18] The advantages of the CART method in credit scoring are that it is very intuitive, easy to explain to management, and it is able to deal with missing observations. The major disadvantage is the computational burden in case of large datasets since at each node every characteristic has to be examined. Very often the resulting tree is quite large so that the process of model--learning becomes too time-consuming. Some empirical studies also note that often the trees are not stable since small change in a training set may considerably alter the structure of the whole tree.[19] A significant problem is also the fact that CART optimizes only locally on a single variable at a time and thus it may not minimize the overall costs of misclassification.

The first to use the CART method in the credit scoring area were Frydman, Altman and Kao (1985) who found it to outperform LDA.[20] Relevant to retail lending is the study by Devaney (1994), who used logit and CART methods to choose which financial ratios are the best predictors of households default and found that both methods differ substantially in selecting the ratios: the ones chosen by the CART method were not so important according to logit regression. A very recent addition to the empirical literature dealing with the CART method is (Feldman – Gross, 2005). The authors use this method for mortgage default data and discuss the pros and cons of CART in relation to traditional methods.

Extensions of the CART method cover the Bayesian approach to tree construction that uses Bayesian techniques for node splitting, pruning and averaging of multiple trees (Denison – Mallick – Smith, 1988). A comparative study of pruning methods for CART is provided in (Esposito – Malerba – Semeraro, 1997). An acceleration procedure in splitting the tree is discussed in (Mola – Siciliano, 1997).

[18] The choice of the pruning algorithms can be found, for example, in (Breiman – Friedman – Olshen – Stone, 1984). The most common and efficient ones are based on the fact that if one tries to select a subtree of the maximal tree that minimizes the estimated misclassification costs, a large number of trees yield approximately the same estimated misclassification costs. Therefore it is reasonable to stop the search for the best pruned tree once a subtree with similar misclassification costs to the maximal tree is found.

[19] See for example (Hastie – Tibshirani – Friedman, 2001).

[20] Chandy and Duett (1990) compared trees with logit and LDA and found that these methods are comparable in results to a sample of commercial papers from Moody's and S&P's. Zhang (1998) generalized CART to multiple binary responses and used it for medical data.

2.5 Neural Networks

The last reviewed method is so-called neural networks. A neural network (NNW) is a mathematical representation inspired by the human brain and its ability to adapt on the basis of the inflow of new information. Mathematically, NNW is a non-linear optimization tool. Many various types of NNW have been specified in the literature.[21]

The NNW design called multilayer perceptron (MLP) is especially suitable for classification and is widely used in practice. The network consists of one input layer, one or more hidden layers and one output layer, each consisting of several neurons. Each neuron processes its inputs and generates one output value that is transmitted to the neurons in the subsequent layer. Each neuron in the input layer (indexed $i = 1,\dots,n$) delivers the value of one predictor (or the characteristics) from vector x. When considering default/non-default discrimination, one output neuron is satisfactory. In each layer, the signal propagation is accomplished as follows. First, a weighted sum of inputs is calculated at each neuron: the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron. A transfer function $g(x)$ is then applied to this weighted sum to determine the neuron's output value. So, each neuron in the hidden layer (indexed $j = 1,\dots, q$) produces the so-called activation:

$$a_j = g\left(\sum_i w_{ij} x_i\right) \tag{5}$$

The neurons in the output layer (indexed $k = 1,\dots, m$) behave in a manner similar to the neurons of the hidden layer to produce the output of the network:

$$y_k = f\left(\sum_j w'_{ik} a_j\right) = f\left[\sum_j w'_{jk} g\left(\sum_i w_{ij} x_i\right)\right] \tag{6}$$

where w_{ij} and w_{jk}' are weights.[22]

There are two stages of optimization. First, weights have to be initialized, and second, a nonlinear optimization scheme is implemented. In the first stage, the weights are usually initialized with some small random number. The second stage is called learning or training of NNW. The most popular algorithm for training multilayer perceptrons is the back-propagation algorithm. As the name suggests, the error computed from the output layer is back-propagated through the network, and the weights are modified according to their contribution to the error function. Essentially, back-propagation performs a local gradient search, and hence its implementation; although not computationally demanding, it does not guarantee reaching

[21] For a thorough exposition see (Bishop, 1995).

[22] The Sigmoid (or logistic) function $f(x) = 1/(1 + \exp(x))$ or hyperbolic tangent function $f(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$ is usually employed in the above network output for functions f and g. The logistic function is appropriate in the output layer if we have a binary classification problem, as in credit scoring, so that the output can be considered as default probability. According to the theory (Bishop, 1995), the NNW structure with a single hidden layer is able to approximate any continuous bounded integrable function arbitrarily accurately.

a global minimum. For each individual, weights are modified in such a way that the error computed from the output layer is minimized.

NNWs were described in the 1960s but their first use in the credit-scoring-related literature appears only at the beginning of 1990s. Altman (1994) employed both LDA and a neural network to diagnose corporate financial distress for 1,000 Italian firms and concluded that neural networks are not a clearly dominant mathematical technique compared to traditional statistical techniques such as discriminant analysis, and that LDA compares rather well to the neural network model in decision accuracy. On the other hand, with different results, Tam and Kiang (1992) studied the application of the neural network model to Texas bank-failure prediction for the period 1985–1987; they compared the NNW prediction accuracy with that of LDA, logistic regression, *k*-NN, and a decision tree model. Their results suggest that the NNW is the most accurate, followed by linear discriminant analysis, logistic regression, decision trees, and *k*-NN.

In direct relation to retail credit scoring, Desay, Crook and Overstreet (1996) investigate a multilayer perceptron neural network, a mixture of an expert's neural network, linear discriminant analysis, and logistic regression for scoring credit applicants in the credit union industry. Their results indicate that customized neural networks offer a very promising avenue if the measure of performance is the percentage of bad loans correctly classified. However, if the measure of performance is the percentage of good and bad loans correctly classified, logistic regression models are comparable to the neural networks approach.

West (2000) investigated the credit-scoring accuracy of five various neural network models. The neural network credit-scoring models were tested using 10-fold cross-validation with two real-world data sets (with both German and Australian credit data). Results were benchmarked against more traditional methods reviewed in this article. Their research suggested that logistic regression is a good alternative to the neural models. Logistic regression was slightly more accurate than the neural network models for the average case, which includes some inferior neural network training iterations.

The major drawback of NNWs is their lack of explanation capability. While they can achieve a high prediction accuracy rate, the reasoning behind why and how the decision was reached is not available. For example, in a case of a denied loan it is not possible to determine which characteristic(s) was exactly the key one(s) to prompt rejection of the application. Consequently, it is very difficult to explain the decision results to managers.[23]

3. Concluding Remarks

In this paper we have identified the most common methods used in the process of credit scoring of applicants for retail loans. Our review concentrates on the most relevant methods, which correspond to their use in the prac-

[23] Baesens, Setiono, Mues and Vanthienen (2003) presented the results from analyzing three real-life datasets using NNW rule extraction techniques, i.e. how to clarify NNW decisions by explanatory rules that capture the learned knowledge embedded in networks.

tice of banks. According to the personal experience of the authors, most local banks in the Czech and Slovak Republics use models based on the logit method, which is an extension of the linear discriminant analysis and is very tractable and convenient. Other methods such as CART or neural networks are used mainly as support tools, either in the process of selecting variables or in the process of the model-quality evaluation. The lesser known *k*-NN method is not used at all or is used very rarely. These facts are very surprising, as the alternative (nonparametric) methods have excellent potential in pattern recognition and they are very competitive with logit regression. It seems that this potential is unrecognized by the local banks and this reality is not far from the claim by Renault and de Servigny (2004) mentioned in Section 2.

Answering the question of *which method to choose* is not straightforward and depends mainly on the bank's preferences, data availability, its characteristics, etc. As follows from our short survey, the various methods are often very comparable in results. This fact can be partly explained by the mathematical relationships between these models: for example, the NNW can be seen as a generalization of the logit method. Often, there is no superior method for diverse data sets. However, the logit method is the most favored method in practice, mainly due to (almost) no assumptions imposed on variables, with the exception of missing values and multicollinearity among variables. Contrary to this, non-parametric methods can deal with missing values and multicollinearity (or correlations) among variables, but often are computationally demanding. The rules that are constructed on the basis of some of these methods can be hard to explain to a manager as well as to a client, however. For example, despite the fact that neural networks have the potential to produce excellent results, the inability to explain *why* can present a serious drawback to applying this method in practice.

In any event, as the amounts of retail loans increase, the quality and methodology improvements in the credit-scoring processes become imperatives for commercial banks. This is especially important during periods of sharp increase in optimism about future earnings, which often prompts households to borrow and spend. If real performance falls below these expectations, the severity of incurred losses may be very high for commercial banks. The post-transition developments on the retail-loans market in the Czech and Slovak Republics can serve as an example of the necessity to advance credit methods in order to protect banks as well as their customers.

APPENDIX

Indicators that Are Typically Important in Retail Credit Scoring Models

Variables or indicators that are typical to the retail segment of credit scoring are given in this appendix. To answer the question of what the main determinants of default are, it is necessary to design a model specification containing the right variables. The following brief outline concentrates on those variables that most frequently come from the relevant literature; the variables are presented in *Table 1*. They have two common features: first is their soundness in helping to estimate the probability of default of an applicant; second is their explanatory power when a credit-scoring method is employed to analyze a loan application. The variables can be divided into four main categories as Table 1 indicates. In the following text we will briefly discuss the importance of each of these categories.

TABLE 1 Indicators that Are Typically Important in Retail Credit Scoring Models

Demographic Indicators	**Financial Indicators**	**Employment Indicators**	**Behavioral Indicators**
1. Age of borrower	1.Total assets of borrower	1. Type of employment	1. Checking account (CA)
2. Sex of borrower	2. Gross income of borrower	2. Length of current employment	2. Average balance on CA
3. Marital status of borrower	3. Gross income of household	3. Number of employments over the last x years	3. Loans outstanding
4. Number of dependants	4. Monthly costs of household		4. Loans defaulted or delinquent
5. Home status			5. Number of payments per year
6. District of address			6. Collateral/guarantee

The first category contains *demographic indicators*. These variables typically do not have the highest importance, but they are useful in capturing various regional, gender, and other relevant differences. For example, it is often found that older women are less risky than young men. In general, the risk of default decreases with age and is also lower for married applicants with dependants. Home owners also represent a less risky category due to a house as collateral (more on collateral in the fourth category). Relations like this can help to better discriminate between good/bad applicants.

The second category contains data on a *financial situation*. When considering a loan application, a bank needs to know what other available resources a household has, what its incomes and costs are, and consequently from these items of information, what the realistic or potential maximum possible monthly payment is. The importance of these variables is evident.

The source of income and *employment status* constitute the third set of variables. Typically, in developed countries, a large proportion of people are self-employed and this category frequently receives a lower score in the assessment of loan applications than employed people. This is due to the fact that stability of employment may provide a sign of stability of payments. The character and length of employment are also decisive factors: frequent change of low-skilled jobs invites a low score.

The *behavioral characteristics* of the fourth category are the first-rate information that can be used for credit scoring. This type of information significantly lowers the problem of asymmetric information between a bank and a client. If a client has some history with a bank, then that bank can easily verify, for example, the history of average balances in a checking account(s), the inflow and outflow of money from that checking account(s), etc. The bank knows if the client has already had a loan, whether this loan was successfully repaid, or whether it involved some problems. Banks often share this type of information, since previous default/delinquency is a serious determinant of future problems with repaying debts. The existence, type, and value of collateral are also a part of the category. Collateral is often a key, and in the case of certain loans the dominant, factor in determining a bank's lending decision. Collateral is also a forceful factor in a client's decision to repay the debt. From this point of view, real estate serve as the best collateral. The threat of losing one's house in the event of default is a critical factor for a client.

REFERENCES

ALTMAN, E. (1968): Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, vol. 23, 1968, pp. 589–609.

ALTMAN, E. (1994): Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience). *Journal of Banking and Finance*, vol. 18, 1994, pp. 505–529.

ALTMAN, E. – NARAYANAN, P. (1997): An International Survey of Business Failure Classification Models. *Financial Markets, Institutions and Instruments*, vol. 6, 1997, no. 2, pp. 1–57.

BAESENS, B. – SETIONO, R. – MUES, CH. – VANTHIENEN, J. (2003): Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. *Management Science*, vol. 49, 2003, pp. 312–329.

BIS (2004): *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel Committee on Banking Supervision, Bank for International Settlements, June 2004.

BISHOP, C. M. (1995): *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

BREIMAN, L. – FRIEDMAN, J. H. – OLSHEN, R. A. – STONE, C. J. (1984): *Classification and Regression Trees*. Pacific Grove, CA, Wadsworth, 1984.

CAMPBELL, T. S. – DIETRICH, J. K. (1983): The Determinants of Default on Insured Conventional Residential Mortgage Loans. *Journal of Finance*, vol. 38, 1983, pp. 1569–1581.

CHANDY, P. R. – DUETT, E. H. (1990): Commercial Paper Rating Models. *Quarterly Journal of Business and Economics*, vol. 29, 1990, pp. 79–101.

CHARITOU, A. – NEOPHYTOU, E. – CHARALAMBOUS, C. (2004): Predicting Corporate Failure: Empirical Evidence for the UK. *European Accounting Review*, vol. 13, 2004, pp. 465–497.

CHATTERJEE, S. – BARCUN, S. (1970): A Nonparametric Approach to Credit Screening. *Journal of American Statistical Association*, vol. 65, 1970, pp. 150–154.

DENISON, D. G. T. – MALLICK, B. K. – SMITH, A. F. M. (1988): A Bayesian CART Algorithm. *Biometrica*, vol. 85, 1988, pp. 363–377.

DESAY, V. – 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*, vol. 95, 1996, pp. 24–37.

DEVANEY, S. (1994): The Usefulness of Financial Ratios as Predictors of Household Insolvency: Two Perspectives. *Financial Counseling and Planing*, vol. 5, 1994, pp. 15–24.

EISENBEIS, R. A. (1977): Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics. *Journal of Finance*, vol. 32, 1977, pp. 875–900.

ENAS, G. G. – CHOI, S. C. (1986): Choice of the Smoothing Parameter and Efficiency of *k*-nearest Neighbor Classification. *Computers and Mathematics with Applications*, vol. 12, 1986, pp. 235–244.

ESPOSITO, F. – MALERBA, D. – SEMERARO, G. (1997): A Comparative Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, 1997, pp. 476–491.

FELDMAN, D. – GROSS, S. (2005): Mortgage Default: Classification Tree Analysis. *Journal of Real Estate Finance and Economics*, vol. 30, 2005, pp. 369–396.

FISHER, R. A. (1936): The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenic*, vol. 7, 1936, pp. 179–188.

FRYDMAN, H. – ALTMAN, E. – KAO, D.-L. (1985): Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. The *Journal of Finance*, vol. 40, 1985, pp. 269–291.

GARDNER, M. J. – MILLS, D. L. (1989): Evaluating the Likelihood of Default on Delinquency Loans. *Financial Management*, vol. 18, 1989, pp. 55–63.

HAND, D. J. – HENLEY, W. E. (1997): Statistical Classification Methods in Consumer Credit Scoring. *Journal of the Royal Statistical Society*, *Series A (Statistics in Society)*, vol. 160, 1997, pp. 523–541.

HAND, D. J. – VINCIOTTI, V. (2003): Choosing *k* for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes. *Pattern Recognition Letters*, vol. 24, 2003, pp. 1555–1562.

HASTIE, T. – TIBSHIRANI, R. – FRIEDMAN, J. H. (2001): The *Elements of Statistical Learning: Data Mining, Inference and Prediction*. (Springer Series in Statistics) New York, Springer Verlag, 2001.

HENLEY, W. E. – HAND, D. J. (1996): A *k*-nearest Neighbour Classifier for Assessing Consumer Credit Risk. *Statistician*, vol. 45, 1996, pp. 77–95.

HILBERS, P. – OTKER-ROBE, I. – PAZARBASIOGLU, C. – JOHNSEN, G. (2005): Assessing and Managing Rapid Credit Growth and the Role of Supervisory and Prudential Policies. *International Monetary Fund Working Paper*, no. 05/151.

HOLMES, C. C. – ADAMS, N. M. (2002): A Probabilistic Nearest Neighbour Method for Statistical Pattern Recognition. *Journal of Royal Statistical Society. Series B*, vol. 64, 2002, pp. 295–306.

LAWRENCE, E. – ARSHADI, N. (1995): A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking. *Journal of Money, Credit and Banking*, vol. 27, 1995, pp. 202–216.

MOLA, F. – SICILIANO, R. (1997): A Fast Splitting Procedure for Classification Trees. *Statistics and Computing*, vol. 7, 1997, pp. 209–216.

REICHERT, A. K. – CHO, C.-C. – WAGNER, G. M. (1983): An Examination of Conceptual Issues Involved in Developing Credit-scoring Models. *Journal of Business and Economic Statistics*, vol. 1, 1983, pp. 101–114.

RENAULT, O. – SERVIGNY, A. de (2004): The *Standard & Poor's Guide to Measuring and Managing Credit Risk*. 1st ed. McGraw-Hill, 2004.

RIPLEY, B. D. (1994): Neural Networks and Related Method for Classification. *Journal of the Royal Statistical Society. Series B*, vol. 56, 1994, pp. 409–456.

ROSENBERG, E. – GLEIT, A. (1994): Quantitative Methods in Credit Management: A Survey. *Operations Research*, vol. 42, 1994, pp. 589–613.

TAM, K. Y. – KIANG, M. Y. (1992): Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science*, vol. 38, 1992, pp. 926–947.

WEBB, A. R. (2002): *Statistical Pattern Recognition*. 2nd edition. John Willey & Sons, 2002.

WEST, D. (2000): Neural Network Credit Scoring Models. *Computers & Operations Research*, vol. 27, 2000, pp. 1131–1152.

WIGINTON, J. C. (1980): A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, vol. 15, 1980, pp. 757–770.

ZHANG, H. (1998): Classification Trees for Multiple Binary Responses. *Journal of the American Statistical Association*, vol. 93, 1998, pp. 180–193.

SUMMARY

JEL Classification: B41, C14, D81, G21, P43
Keywords: banking sector; credit scoring; discrimination analysis; pattern recognition; retail loans

Credit-Scoring Methods

Martin VOJTEK – Evžen KOČENDA: CERGE-EI, Prague (martin.vojtek@cerge-ei.cz)
(evzen.kocenda@cerge-ei.cz)

The paper reviews the best-developed and most frequently applied methods of credit scoring employed by commercial banks when evaluating loan applications. The authors concentrate on retail loans – applied research in this segment is limited, though there has been a sharp increase in the volume of loans to retail clients in recent years. Logit analysis is identified as the most frequent credit-scoring method used by banks. However, other nonparametric methods are widespread in terms of pattern recognition. The methods reviewed have potential for application in post-transition countries.