Financial innovation: Credit default hybrid model for SME lending

Kang Li a,1,*, Jyrki Niskanen a,2, Mikko Kolehmainen b,2, Mervi Niskanen a,2

a Department of Business, University of Eastern Finland, PO Box 1627, FIN-70211 Kuopio, Finland
b Environmental Informatics, Department of Environmental Science, University of Eastern Finland, PO Box 1627, FIN-70211 Kuopio, Finland

ARTICLE INFO

Article history:
Received 3 March 2015
Revised 16 May 2016
Accepted 18 May 2016
Available online 24 May 2016

JEL classification:
G32
G33

Keywords:
Credit default
Credit risk
Bankruptcy
Credit scoring hybrid model
Small and medium-sized firm (SME)

ABSTRACT

Credit risk evaluation is an integral part of any lending process. Therefore, an economic institution involved in lending to SMEs. The importance of credit scoring has increased recently because of the financial crisis and increased capital requirements for banks. There are, however, only few studies that develop credit scoring for SME lending. The objective of this study is to introduce a novel, more accurate credit risk estimation approach for SMEs business lending. Based on traditional statistical methods and recent artificial intelligence (AI) techniques, we propose a hybrid model which combines the logistic regression approach and artificial neural networks (ANN). In order to test the effectiveness and feasibility of the proposed hybrid model, we use the data of Finnish SMEs from the fiscal year 2004 to 2012. Our results suggest that the proposed ANN/logistic hybrid model is more accurate than either of the initial models. ANN or logistic regression. This improvement in the accuracy of the credit scoring model decreases evaluation errors and has thereby many potential practical implications. First of all, a more accurate credit scoring model can result in better performance of the whole SME loan portfolio. Second, it can also result in lower capital requirements from the banks perspective and lower interest rates from the individual firm’s perspective. Combined, these effects will enhance the banks competitiveness in the market for SME loans.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Most firms around the world are SMEs, and they are recognized worldwide as the engine of economic growth. It has also been suggested in the literature, that from a credit risk point of view, SMEs are different from large corporations for a number of reasons (Altman & Sabato, 2007). Jacobson, Lindé, and Roszbach (2005) and Dietsch and Petey (2004) show that bank loan portfolios of SME loans are usually riskier than corporate credit. Altman and Sabato (2007) further suggest that more accurate credit scoring models are needed in the market for SME loans have many potential benefits; First of all, if banks are able to improve the accuracy of their credit scoring models, their capital requirements may be lower. Second of all, if banks are able to reduce their capital requirements in SME lending, this could result in lower interest rates for their SME customers. Tsai and Wu (2008) further suggest that even a slight improvement in credit scoring accuracy might reduce credit risk and translate into significant future savings. It may therefore be plausible to assume, that more accurate credit risk evaluation can strengthen the competitive advantage of financial institutions and reduce the financing difficulties for SMEs.

Financial innovation has been described as the life blood of efficient and responsive capital markets, and one of the most important innovations since the 1980s is banks’ use of credit risk evaluation in SME business lending (Akhavein, Fram, & White, 2005). SME business lending based on credit risk evaluation is a relatively new technology that involves processing data about the firm and its owner by using statistical methods to predict applicants’ expected future loan performance (Hand & Henley, 1997). Previous studies indicate that many methods can be applied in credit risk evaluation. These methods can be classified into traditional statistical methods and recent artificial intelligence approaches.

Statistical methods are the first and most frequently used methods in credit scoring or credit risk evaluation. Many researchers used statistical methods to build a credit risk model (Altman, 1968; Altman & Sabato, 2007; Banasik, Crook, & Thomas, 2001; Boyes, Hoffman, & Low, 1989; Durand, 1941; Ewert, 1968; Makowski, 1985; Myers & Forgay, 1963; Orger, 1970; Šarlija, Benšič, & Bohaček, 2004; Steenackers & Goovaerts, 1989; Wiginton, 1980). With the development of information and computational technologies, recently more accurate credit risk models have been developed based on sophisticated intelligence approaches which are more

* Corresponding author.
E-mail addresses: kang.li@uef.fi (K. Li), jyrki.niskanen@uef.fi (J. Niskanen), mikko.kolehmainen@uef.fi (M. Kolehmainen), mervi.niskanen@uef.fi (M. Niskanen).
1 Researcher, University of Eastern Finland
2 Professor, University of Eastern Finland

http://dx.doi.org/10.1016/j.eswa.2016.05.029
0957-4174/© 2016 Elsevier Ltd. All rights reserved.
capable of modelling nonlinear or extremely complex functions. Credit risk modeling is one of the main areas in accounting and finance where artificial intelligent technologies have been applied into successfully (Angelini, di Tollo, & Roli, 2008; Armingero, Enache, & Bonne, 1997; Chatterjee & Barcun, 1970; Desai, Crook, & Overstreet, 1996; Piramuthu, 1999; Tsai & Wu, 2008; West, 2000). In recent years, hybrid prediction models, which combine traditional statistical methods and artificial intelligence technologies have been suggested to have better prediction ability than either of the two components. For instance, in environmental engineering, Schafer (2008) shows that the accuracy of prediction of a logistic regression model is lower than that of the ANN/logistic hybrid model. Lin (2009) obtained similar results when she investigated financial distressed banks. It has been suggested that neural network may not be as stable as standard statistical techniques, and combinations in certain applications might be more valuable (Paliwal & Kumar, 2009).

The bulk of previous studies in credit scoring models concentrate on either traditional statistical methods or artificial intelligence technologies and there is little evidence on hybrid credit risk models in the literature especially for SMEs business lending. The aim of this paper is to propose a novel hybrid credit risk model (ANN/logistic hybrid model) for SMEs lending integrating ANN with the logistic regression method. By using data Finnish SMEs from the fiscal years from 2004 to 2012, we are able to test whether the performance of credit risk models based on artificial intelligent technology can be improved by combining them with traditional statistical methods. To the best of our knowledge, our study is the first one to investigate the ANN/logistic hybrid model for data on SMEs.

Our main empirical results suggest that the ANN/logistic hybrid model is more accurate than either of the separate ANN or logistic regression approach. Additionally, we find that credit risky firms are less profitable than non-credit risky firms. Furthermore, they are more levered and have higher asset turnover than non-credit risky firms. Also, credit risky firms are more likely to be larger and younger than non-credit risky firms. This study extends the findings of previous studies on four aspects. First, our study is one of few that sheds light on the hybrid model for financial data, while most of the previous ones concentrate on statistical methods or computational (artificial) intelligence methods. Second, SMEs have unique accounting characteristics compared to large firms. In order to develop SME business lending, instead of using a credit risk model for corporations, a powerful and reliable credit risk estimation model just for SME is necessary. Third, we propose a credit risk model for Finnish firms, whereas most of existing studies have used data on the US and UK. Fourth, this study demonstrates that with the inclusion of traditional statistic methods (e.g. logistic regression), the accuracy of artificial intelligent technologies (e.g. ANN) in credit scoring models can be improved.

The remainder of this paper is structured as follows. Section two of the study discusses the relevant literature. Section three presents the methodology in this study. Section four presents the data and descriptive statistics on the variables. Section five presents the main empirical results. Section six is the discussions and conclusions.

2. Literature review

Previous studies have applied various methods to credit risk evaluation. These methods can be classified into traditional statistical methods and artificial intelligence approaches. Statistical methods are the first and the most frequently used methods in credit risk evaluation. These methods include linear regression, discriminate analysis, and logistic regression, etc. Artificial intelligence methods have been presented to credit risk evaluation much more recently. Even more recent and much less explored is the idea to combine traditional methods and artificial intelligence methods into so called hybrid models.

Myers and Forgy (1963) and Orgler (1970) were the first to apply linear regression into credit risk evaluation. Discriminant analysis, was first proposed by Fisher (1936). Following him, Durand (1941) applied discriminant regression method to evaluate credit risk. Later, Altman (1968) used discriminate regression model in his study and proposed one of the most famous credit scoring model, z-scores model. A number of studies have applied discriminant regression analysis to credit scoring since that (e.g. Abdou & Pointon, 2011; Hand, Oliver, & Lunn, 1998; and Mahmoudi & Duman, 2015). However, it has been suggested that discriminant regression is biased by extreme data points or satisfaction of equal group covariance assumption (Malhotra & Malhotra, 2003).

The logistic regression model for credit risk estimation was first introduced by Wiginton (1980). Logistic regression is much more relaxed and flexible in its assumptions than the discriminant analysis or linear regression, because logistic regression does not have the requirements of the independent variables to be normally distributed, linearly related, or equal variance within each group (Tabachnick & Fidell, 2001). Siddiqi (2005) indicated that credit default estimation can be viewed as a binary classification problem, and thus it is appropriate to apply logistic regression to credit risk estimation. This approach has been adopted by, e.g. Abdou, El-Masyr, & Pointon (2007), Baensens et al. (2003), Crook et al. (2007), Lee & Sung-Chang (2000). Evidence on traditional statistical methods applied to credit risk evaluation can also be found in Banasik et al. (2001), Boyes et al. (1989), Šarija et al. (2004) and Steenackers and Goovaerts (1989).

In recent years, artificial intelligence methods as the emerging technologies have been applied to credit risk evaluation, due to the development of information and computation technology. For example, Huang, Chen, and Wang (2007) find that artificial intelligent techniques perform better in credit scoring than traditional statistical methods. Genetic algorithms, nearest neighbors, support vector machine and artificial neural network (ANN) have been applied to credit scoring problems. Genetic algorithms were first applied in credit risk evaluation by Fogarty and Ireson (1993). During the following decades, genetic algorithms have been applied to evaluate credit risk by Etehadi, Rostamy, and Dekhordi (2009), Kozeny (2015), and Huang et al. (2007). Nearest neighbors is a non-parametric approach, which chooses a metric on the space of application data to measure how far apart applicants are. Chatterjee and Barcun (1970) first applied nearest neighbors into credit risk evaluation, Support vector machine (SVM) was first motivated in statistical learning theory by Cortes and Vapnik (1995). Due to the appealing classification performance and the theoretical foundation, it was soon introduced to credit scoring problems (e.g. Baensens et al., 2003; Harris, 2015; Huang et al., 2007, Marten et al., 2007; Yao, Crook, & Andreева, 2015).

ANN is an artificial intelligence algorithm that allows for some learning through experience to discern the relationship between borrower characteristics and the probability of credit default. Many previous studies suggest that ANN can be very effectively applied into credit risk evaluation (Angelini et al., 2008; Armingero et al., 1997; Desai et al., 1996; Khashman, 2011; Piramuthu, 1999; Soydaner & Kocadağlı, 2015; Tsai & Wu, 2008; West, 2000; Zhao et al., 2015).

More recently, some literature has emerged to suggest that hybrid credit models may outperform the individual models presented above. For instance, Chen and Li (2010) combined a SVM model with several other methods such as discriminate regression analysis. They found that the hybrid discriminate regression-SVM models outperforms the single classifiers. Liang, Tsai, and Wu (2015) find that if discriminate regression or genetic
algorithms, as pre feature selection methods, are combined with e.g. SVM, performance is improved. Also, other studies suggest that hybrid models with genetic algorithm and ANN perform better. (Arifovic & Gencay, 2001; Blanco et al., 2001; Hamadani, Shalbafzadeh, Rezvan, & Moghadam, 2013; Oreski & Oreski, 2014; Oreski, Oreski, & Oreski, 2012).

We build on the hybrid model literature by combining ANN with the most frequently used traditional credit scoring method, i.e., logistic regression analysis. We are aware of two previous studies that have applied a similar hybrid prediction model (ANN/logistic hybrid model) integrating ANN with a logistic regression model. In environmental engineering, Schafer (2008) used the ANN/logistic hybrid model to predict the failure possibility of broken rails. He found that the prediction accuracy of the logistic regression model was 66.3%; while the ANN/logistic regression model’s accuracy reached 67.9%. In the finance field, Lin (2009) first introduced the ANN/logistic hybrid model to predict the distress of banks using the data from 11 emerging countries. She found that accuracy rate of the ANN/logistic hybrid model was above 80.0% accuracy rate, while separate models have an accuracy between 71.2% and 75.3%. It has been suggested that neural network may not be as stable as standard statistical techniques, but using both neural network methods and statistical methods in certain applications together would be valuable (Paliwal & Kumar, 2009). Our study aims at estimating a hybrid model integrate ANN with traditional statistic method.

Previous studies suggest that linear regression models can be limited by the assumption of a linear relationship between input variables and the output variable (Süsterić, Mramor, & Zupan, 2009). Moreover, discriminant regression model may be biased by extreme data points or satisfaction of equal group covariance assumption (Malhotra & Malhotra, 2003). By contrast, logistic regression is much more relaxed and flexible in its assumptions, because logistic regression does not have the requirements of the independent variables to be normally distributed, linearly related, or equal variance within each group (Tabachnick & Fidell, 2001). Therefore, we choose logistic regression to be one component of our proposed hybrid model. Using Logistic regression method in evaluation has two advantages. First, as one of the traditional statistic approaches, logistic regression, known as parametric method, has the high interpretability to the parameters, since the model techniques like goodness of fit, coefficients and significance testing of input variables, etc. are available in logistic regression models. What’s more, compared to ANN, its cross-validation ability is stronger, which means the generalizability of logistic regression model is better. However, the disadvantage of logistic regression is that its prediction accuracy are not very high (Dreiseitl & Ohno-Machado, 2002).

Compared to other artificial intelligent technologies e.g. genetic algorithms, ANN has been found to share most similar roots in statistical pattern recognition with logistic regression, and considered as the generalization of the latter (Dreiseitl & Ohno-Machado, 2002). Due to the advantage of high learning ability of artificial intelligent system, ANN has a good forecasting capability. Duh, Walker, Pagano, and Kronlund (1998) found that network models always outperformance logistic regression in the training set. But as the non-parametric approach, one disadvantage of ANN is that it has a more complicated structure and is difficult to interpret the models based on coefficients and significance, etc. in ordinary language. Additionally, the generalizability of ANN is limited, which means that the prediction performance in the testing set is always found to be much lower than that in training set (Duh et al., 1998).

ANN/logistic hybrid model integrates ANN with logistic regression. The inclusion of logistic regression improves the interpretability and cross-validation ability of the model, while the use of ANN increases the accuracy of the proposed model. Thus, our proposed ANN/logistic hybrid model can overcome the shortcomings of both model and develop on the described advantages above. But our proposed model is not without limitations. It is also noteworthy that like ANN, the hybrid model usually requires a longer learning and training time with enough data in order to achieve the best accuracy.

3. Methodology

3.1. The principle of a feed-forward Neural Network

A BP neural network is one type of artificial intelligence algorithm in ANN. In this study, we chose a feed-forward neural network to build the model. The BP neural network is a two-tier or multi-layer feed forward neural network whose neurons transfer function is Sigmoid-function. The output of the network is continuous between 0 and 1 volume. The feed-forward neural network can realize the mapping from input to output of any non-linear function. A typical feed-forward neural network is an input layer, hidden layer, and output layer feed forward three-layer hierarchy network. The structure of a network can be illustrated like this

**Graph 1:**

Assume that the input pattern vector is \( A^K = (a^K_1, a^K_2, \ldots, a^n) \), the desired output vector is \( y^K = (y^K_1, y^K_2, \ldots, y^K_n) \); the middle layer unit output is \( B^K = (b^K_1, b^K_2, \ldots, b^K_n) \); output layer unit output is \( C^K = (c^K_1, c^K_2, \ldots, c^K_n) \); \( k=1, 2, \ldots, m \).

Input layer to middle layer connection weights is \( w_{ij} \), \( i=1, 2, \ldots, n; j=1, 2, \ldots, p; \) middle layer to output layer connection weights is \( v_{ij} \), \( j=1, 2, \ldots, q; \) output threshold of middle layer value of each unit is \( \{\theta_i\} \), \( j=1, 2, \ldots, p; \) output threshold of output layer value of each unit is \( \{\gamma_i\} \), \( t=1, 2, \ldots, q \).

Based on the above assumptions, the specific learning steps of a backpropagation (BP) neural network can be expressed as follows

**Graph 2:**

a. Initialization. Set all adjustable parameters. For instance, the input layer to middle layer connection weights \( w_{ij} \), the middle layer to output layer connection \( v_{ij} \), the output threshold of middle layer value of each unit \( \{\theta_i\} \), and the output threshold of output layer value of each unit \( \{\gamma_i\} \) is the random number between (-1, +1);

b. Calculate the input samples to get the output value of each layer. The formula is as follows:

\[
\text{Hidden layer: } b^j_t = f \left[ \sum w_{ij}a^i_t - \theta_j \right], j = 1, 2, \ldots, p
\]

\[
\text{Output layer: } c^t_j = f \left[ \sum v_{ij}b^j_t - \gamma_j \right], t = 1, 2, \ldots, q
\]

c. Calculate the error and the local gradient of the output layer.

d. Adjust the output layer weights, and \( \alpha \) is the learning coefficient.

\[
v_{ij}(N+1) = v_{ij}(N) + \alpha \delta^i_j b^j_t, \quad 0 < \alpha
\]

e. Adjust the hidden layer weights, and \( \beta \) is the learning coefficient.

\[
w_{ij}(N+1) = w_{ij}(N) + \beta \delta^i_j a^i_t, \quad \beta < 1
\]

f. The network continues the learning until it reaches the accuracy of the requirements.

\[
E(t) = \frac{1}{2} \sum_{k=1}^{m} \sum_{i=1}^{n} (y_i - c_i)^2
\]
Graph 1. The structure of feed-forward Neural Network.

After the above process, a BP learning process is completed. The backpropagation (BP) algorithm which is essentially a gradient steepest descent method, adjusts the weights in the steepest descent direction (negative of the gradient). However, it is found that although the negative of the gradient is fast decreased along, the most rapid convergence is not produced. Due to this, a lot of enhanced methods have been proposed (Zhang, Patuwo, & Hu, 1998). Levenberg-Marquardt (LM) algorithm (Hagan & Menhaj, 1994), is one of such robustness and faster convergence producing enhanced methods who has the ability to find good local minima. In practise, when the dataset is not too large, the Levenberg-Marquardt (trainlm in Matlab) converges probably fastest. In our study, we use trainlm in Matlab.

LM algorithm is the method based on numerical optimization to utilize the Jacobian matrix. Jacobian matrix contains the first order derivatives of the neural network errors in regard to the weights and biases. For LM algorithm, the most crucial procedure is to calculate the Jacobian matrix. The elements of the Jacobian matrix can be computed by the terms which can be calculated by using the standard backpropagation algorithm with one modification at the final layer. LM algorithm was designed to approximate the Hessian matrix in the following Newton-like update:

$$\Delta w = [J^T J + \mu I]^{-1} J^T e(t)$$  \hspace{1cm} (6)

where $\mu$ is multiplied by some factor $\beta$ when a step would rise the neural network performance index and divided by it when a step decline the performance index. When $\mu$ is large, the algorithm becomes steepest descent; while for small $\mu$, the LM becomes Gauss-Newton (Hagan & Menhaj, 1994).

3.2. The principle of logistic regression

Logistic regression uses the maximum likelihood method to estimate the model. A logistic regression model will have the
the following form:

\[ p_i = P(y_i = 1 | x_{i1}, x_{2i}, \ldots, x_{ki}) = \frac{\exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ki})}{1 + \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ki})} \]

\[ = \frac{e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}} \]  

(7)

Assuming “non-credit risky firm” is 0, then “non-credit risky firm’s” \( y_i = 0 \) conditional probability is: \( P(y_i = 0 | x_{i1}, x_{2i}, \ldots, x_{ki}) = 1 - p_i \). So we can get an observation of probability:

\[ P(y_i) = p_i y_i (1 - p_i)^{1 - y_i} \]  

(8)

Where \( y_i = 1 \) or \( y_i = 0 \). Because all observations are independent, so their joint probability is:

\[ L(\theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \]  

(9)

To simplify the calculation, the logarithmic likelihood function is as follows:

\[ \ln[L(\theta)] = \ln[\prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}] \]

\[ = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \]

\[ = \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{p_i}{1 - p_i} \right) + \ln(1 - p_i) \right] \]

\[ = \sum_{i=1}^{n} \left[ y_i (\alpha + \beta x_i) + \ln \left( 1 - \frac{e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}} \right) \right] \]

\[ = \sum_{i=1}^{n} \left[ y_i (\alpha + \beta x_i) - \ln \left( 1 + e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}} \right) \right] y \]  

(10)

These are the log-likelihood functions, and maximum likelihood estimation is derived parameters \( \alpha \) and \( \beta_j \) \((j = 1 \ldots k)\). As a result, we can get the following likelihood function:

\[ \frac{\partial \ln[L(\theta)]}{\partial \alpha} = \sum_{i=1}^{n} \left[ y_i - \frac{e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}} \right] x_l \]  

(11)

\[ \frac{\partial \ln[L(\theta)]}{\partial \beta_j} = \sum_{i=1}^{n} \left[ y_i - \frac{e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\alpha + \sum_{j=1}^{k} \beta_j x_{ij}}} \right] x_{lj} \]  

(12)

Calculate the solution of \( k + 1 \) equations to estimate \( \alpha \) and \( \beta_j \) \((j = 1 \ldots k)\) value. This process is the maximum likelihood estimation process.

3.3. ANN/Logistic hybrid model

The process of training the ANN/logistic hybrid model in our study is proposed as follows:

First step: Run principal components analysis with the independent variables \((X_1, X_2, \ldots, X_{22})\) in order to remove correlations and multicollinearity problems. From the unreported results of principle components analysis, it is suggested that the first six components (PC1, PC2, PC3, PC4, PC5, and PC6) are the final selected principle components which can represent 64.76% of the total variation of the original variables. Then run logistic regression with six components (PC1, PC2, PC3, PC4, PC5, and PC6) and the dependent variable (credit risk variable). The logistic regression model can provide the predicting probability of the logistic regression model. Define the predicting probability values into a new variable defined as Logistic_Results.

Second step: Train the basic neural network with the independent variables \((X_1, X_2, \ldots, X_{22})\) as the inputs, and with the dependent variable (credit risk variable) as output. The neural network can provide the predicting probability defined as ANN_Results.

Third step: Train the final neural network with six components (PC1, PC2, PC3, PC4, PC5, and PC6) and other Logistic_Results and ANN_Results as the inputs, and with the dependent variable (credit risk variable) as output. The process of our ANN/logistic credit risk hybrid model is described in Graph 3.

The structure of the final neural network is described as follows (Graph 4):

\[ a. \text{The first layer is the input layer. The number of nodes in first layer is six (the number of principle components) plus two (Logistic_Results and ANN_Results) plus one.} \]
\[ b. \text{The second layer is the hidden layer.} \]
\[ c. \text{The last layer is the output layer. The third layer is only one node, which means that there are only two types of output: “non-credit risk” (CREDIT_RISK=0) or “credit risky” (CREDIT_RISK=1).} \]

3.4. Validation indicators

In order to compare the performance among the different models, we define several validation indicators in this study.

3.4.1. Mean squared error (MSE)

Mean squared error (MSE) is one of the indicators which can describe the model performance. It can be calculated by the following equation:

\[ MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 \]  

(13)

where \( n \) is number of dataset; \( \hat{Y}_i \) is the prediction value of the model, \( Y_i \) is the observed value. The less the MSE is, the better the model performs.

3.4.2. R coefficient

The R coefficient is another indicator to describe the relationship between the outputs and targets. R is defined as the root of
the difference between one and normalized MSE. It is expressed as:

\[ R = \sqrt{1 - \text{NMSE}} \]  

(14)

where NMSE is the normalized MSE. The higher R value indicates the better model.

3.4.3. Accuracy rate

Accuracy rate is an indicator which expresses the model predicting capability. We define accuracy rate as the total number of the correct prediction in two different groups divided by the total number of the samples. It is obtained according to the following equation:

\[ \text{Accuracy} = \frac{p_1 + p_2}{z_1 + z_2} \times 100\% \]  

(15)

where \( p_1 \) is the No. of correct prediction in non-credit groups; \( p_2 \) is the No. of correct prediction in credit groups; \( z_1 \) is the No. of non-credit groups; and \( z_2 \) is the No. of credit groups.

4. Data and descriptive statistic

4.1. Sample selection

The data used in this study is a combination of credit related data and financial data of Finnish SMEs from the fiscal years from 2004 to 2012. The credit related data were collected through a VOITTO database from Suomen Asiakastieto Ltd, which is a credit rating and financial information company. The firms’ financial data were collected from Amadeus Database of Bureau van Dijk, which covers firms’ financial and business information all over the Europe. Finally, the total number of observations is 2681, with 1364 credit risky observations and 1317 non-credit risky observations. In this study, we use data on small and medium-sized firms, which have less than 250 employees and less than 50 million in revenue a year. This selection is based on the EU definition of a SME in 2003. Also, we randomly divided the data into two parts (training data 70% and testing data 30%) by the Bernoulli random function. The training data is to set for the training modeling, and the testing data is for testing the training models generalization capability.

4.2. Variables selection

The independent variable in this study is the credit risk variable. In the literature, there are several methods to proxy for credit risk. In this study, the proxy we chose for the firm’s credit risk variable is CREDIT_RISK, which is an indicator variable equal to 1 in case the firm is in a credit risky status (0 otherwise). We considered the firm to be credit risky if the firm is in any of the following statuses: payment default, insolvency proceedings, bankruptcy, and/or dissolution due to bankruptcy. Absent any of these statuses, the firm is considered to be in a non-credit risky status (Table 1). Many previous studies define credit risk in this manner
classifying industry into 10 types. The detailed definitions of all the variables are shown in Table 1.

4.3. Descriptive statistics and univariate analysis

Table 2 presents descriptive statistics for the variables. The reported statistics show that 50.90% of the sample firms are credit risky according to the credit risk definition in our study, while 49.10% are non-credit risk samples. The mean of ROA is 4.539 with the median of 3.149. LEVERAGE_A is on average 0.704. The median of TANGIBLE and CURRENT are 0.242 and 1.321, respectively. The maximum of SIZE_A is 10.620, while its minimum is 1.790. The maximum of AGE is 4.600, while its median is 2.639. WORKING_TASST is on average 0.241, ranging from -0.810 to 0.977.

Table 3 presents univariate comparisons with CREDIT_RISK. The results show that credit risky firms are more likely to have a lower profitability ratio (ROA, ROCE, ROE, PAT_ASST, EBIT_ASST) than non-credit risky firms, a higher capital structure ratio (such as LEVERAGE_A, LEVERAGE_E, LTTA, STTA, APR, ARR) than non-credit risky firms. Also, credit risky firms are more likely to be younger and a little bit larger (SIZE_A, SIZE_TURNOVER, SIZE_EMPLOYEE, AGE) than non-credit risky firms.

Like the linear regression, logistic regression is sensitive to multicollinearity. When there is not a serious multicollinearity, coefficient of logistic regression estimation is unbiased and effective; when the multicollinearity degree increases, a serious multicollinearity existing will cause coefficient of logistic regression estimation to be biased and ineffective (Berry & Feldman, 1985). Table 4 presents univariate Pearson correlations in the diagonal between 22 dependent variables. The result shows that some of the correlations are really high, such as the Pearson correlation between ROA and PAT_ASST (0.923). The correlation between variables does indicate serious multicollinearity problems. Principle components analysis is an efficient statistical method which can transform a set of correlated variables into a set of uncorrelated variables. The new generated uncorrelated variables are linear combinations of the original correlated variables. From the unreported results of principle components analysis, it is suggested that the first six components (PC1, PC2, PC3, PC4, PC5, and PC6) are the final selected principle components whose eigenvalues are above 1.0 according to Kaiser criterion and which can represent 64.76 % of the total variation of the original variables. These six components are not correlated with each other according to the rule of principle components analysis. Therefore, serious correlations and multicollinearity do not exist among the six components.

5. Empirical results

5.1. First step logistic regression in modeling samples

We put the selected six components as the inputs and CREDIT_RISK as the output to run a logistic regression model. In Table 5 we can see that the total number of the training samples for modeling is 1883 (70% of total sample 2681) and the pseudo R2 is 0.179. Based on the results, we define a new variable named Logistic_Results, which is the real prediction value from the first step logistic regression.

5.2. Second step basic neural network in modeling samples

We added 22 dependent variables as the inputs and CREDIT_RISK as the output to train the basic neural network. Based on the results, we define another new variable named ANN_Results, which is the real prediction value from the second step basic neural network.
Table 2
Descriptive statistics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Mean</th>
<th>std.dev.</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROA</td>
<td>2681</td>
<td>4.539</td>
<td>13.970</td>
<td>3.149</td>
<td>-100.00</td>
<td>80.220</td>
</tr>
<tr>
<td>ROCE</td>
<td>2681</td>
<td>12.653</td>
<td>46.686</td>
<td>11.481</td>
<td>-565.520</td>
<td>600.000</td>
</tr>
<tr>
<td>ROE</td>
<td>2681</td>
<td>0.297</td>
<td>121.977</td>
<td>14.512</td>
<td>-995.640</td>
<td>900.240</td>
</tr>
<tr>
<td>LEVERAGE_A</td>
<td>2681</td>
<td>0.704</td>
<td>0.222</td>
<td>0.754</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>LEVERAGE_E</td>
<td>2681</td>
<td>10.294</td>
<td>57.566</td>
<td>3.073</td>
<td>0.000</td>
<td>2719.720</td>
</tr>
<tr>
<td>LITTA</td>
<td>2681</td>
<td>0.260</td>
<td>0.222</td>
<td>0.218</td>
<td>0.000</td>
<td>0.975</td>
</tr>
<tr>
<td>STTA</td>
<td>2681</td>
<td>0.444</td>
<td>0.225</td>
<td>0.421</td>
<td>0.003</td>
<td>0.996</td>
</tr>
<tr>
<td>TANGIBLE</td>
<td>2681</td>
<td>0.321</td>
<td>0.271</td>
<td>0.242</td>
<td>0.000</td>
<td>0.999</td>
</tr>
<tr>
<td>CURRENT</td>
<td>2681</td>
<td>1.920</td>
<td>5.098</td>
<td>1.321</td>
<td>0.000</td>
<td>209.190</td>
</tr>
<tr>
<td>A_TURN</td>
<td>2681</td>
<td>7.263</td>
<td>18.406</td>
<td>3.562</td>
<td>0.010</td>
<td>390.510</td>
</tr>
<tr>
<td>WORKING_TASST</td>
<td>2681</td>
<td>0.241</td>
<td>0.253</td>
<td>0.205</td>
<td>-0.810</td>
<td>0.977</td>
</tr>
<tr>
<td>ARR</td>
<td>2681</td>
<td>0.176</td>
<td>0.153</td>
<td>0.141</td>
<td>0.000</td>
<td>0.916</td>
</tr>
<tr>
<td>APR</td>
<td>2681</td>
<td>0.137</td>
<td>0.133</td>
<td>0.098</td>
<td>0.000</td>
<td>0.810</td>
</tr>
<tr>
<td>SIZE_A</td>
<td>2681</td>
<td>6.570</td>
<td>1.494</td>
<td>6.492</td>
<td>1.790</td>
<td>10.620</td>
</tr>
<tr>
<td>SIZE_TURNOVER</td>
<td>2681</td>
<td>7.122</td>
<td>1.481</td>
<td>7.073</td>
<td>-1.390</td>
<td>10.740</td>
</tr>
<tr>
<td>SIZE_EMPLOYEE</td>
<td>2681</td>
<td>20.060</td>
<td>30.666</td>
<td>9.000</td>
<td>1.000</td>
<td>248.000</td>
</tr>
<tr>
<td>AGE</td>
<td>2681</td>
<td>2.456</td>
<td>0.896</td>
<td>2.639</td>
<td>0.000</td>
<td>4.600</td>
</tr>
<tr>
<td>PAT_ASS</td>
<td>2681</td>
<td>0.045</td>
<td>0.147</td>
<td>0.035</td>
<td>-1.000</td>
<td>1.040</td>
</tr>
<tr>
<td>EBIT_ASS</td>
<td>2681</td>
<td>0.086</td>
<td>0.167</td>
<td>0.068</td>
<td>0.030</td>
<td>1.050</td>
</tr>
<tr>
<td>SALE_A</td>
<td>2681</td>
<td>2.194</td>
<td>1.592</td>
<td>1.868</td>
<td>0.000</td>
<td>24.120</td>
</tr>
</tbody>
</table>

This table presents the descriptive statistics for the overall sample. N denotes the number of firm years. Data covers the years from 2004 to 2012. CREDIT_RISK is an indicator variable equal to 1 in case of the firm is in credit risk status (including the status of Active (default of payment), the status of Active (insolvency proceedings); the status of Bankruptcy, the status of Dissolved (bankruptcy)); 0 otherwise. ROA is a ratio of return on total assets. ROCE is a ratio of earnings before interest and tax to capital employed (capital employed—total assets-current liabilities). ROE is a ratio of return on equity. PAT_ASS is a ratio of profit after tax to total assets. EBIT_ASS is a ratio of EBIT to total assets. LEVERAGE_A is a ratio of total liabilities to total assets. LEVERAGE_E is a ratio of total liabilities to equity. LITTA is a ratio of noncurrent liabilities to total assets. STTA is a ratio of current liabilities to total assets. APR is a ratio of trade creditors to total assets. ARR is a ratio of operating revenue to net assets. SIZE_A is the natural logarithm of firms’ total assets. SIZE_TURNOVER is the natural logarithm of firms’ turnover. SIZE_EMPLOYEE is the number of the firms’ employees. AGE is the natural logarithm of 1 plus years since firm’s start to incorporate. YEAR is a year indicator variable defined year 2004 as 1; defined year 2005 as 2; defined year 2006 as 3; defined year 2007 as 4; defined year 2008 as 5; defined year 2009 as 6; defined year 2010 as 7; defined year 2011 as 8; defined year 2012 as 9. INDUSTRY is an industry indicator variable classifying industry into 10 types.

5.3. Third step final neural network in modeling samples

We put the selected six components with Logistic_Results and ANN_Results as the inputs and CREDIT_RISK as the output to train the final neural network. All the neural network related results were obtained with Matlab 2012b for windows. Graph 5 shows the training results of the ANN/logistic hybrid model compared to the ANN model. Obviously, the ANN/logistic hybrid model has the lower mean squared error (0.12752), but higher training R coefficient (0.71805) and test R coefficient (0.70426). We summarize the ANN/logistic hybrid model prediction results in Table 6. We also summarize the results of separate logistic regression and neural network as the comparison. Table 6A shows the comparison of predicting results between the ANN/logistic hybrid model and the separate models; Table 6B shows the comparison of the models’ predicting accuracy rate between the ANN/logistic hybrid model and separate models.

From Table 6A we find that, on the one hand, the logistic model predicted 609 observations correctly for non-credit risky firms, and left 320 observations in the wrong prediction results (the total non-credit risky firms were 929 in modeling samples). Neural network predicted 716 observations correctly for non-credit risky firms, and left 213 observations in the wrong prediction results. The ANN/logistic hybrid model predicted 767 observations correctly for non-credit risky firms, and left just 162 observations in the wrong prediction results. On the other hand, the logistic model predicted 713 observations correctly for credit risky firms, and left 241 observations in the wrong prediction results (the total number of credit risky firms was 954). Neural network predicted 775 observations correctly for credit risky firms, and left 179
observations in wrong prediction results. The ANN/logistic hybrid model predicted 798 samples correctly for credit risky firms, and left just 156 samples in the wrong prediction results. Generally speaking, in the modeling samples, the predicting rate of the logistic model was 70.21%, the predicting rate of neural network was 79.18%, and the predicting rate of the ANN/logistic hybrid model was 83.11% (Table 6B).

In the testing samples, the logistic model predicted 260 observations correctly for non-credit risky firms, and left 128 observations in the wrong prediction results (the total number of non-credit risky firms was 388 in the modeling sample). Neural network predicted 285 observations correctly for non-credit risky firms, and left 103 observations in the wrong prediction results. The ANN/logistic hybrid model predicted 334 observations correctly for non-credit risky firms, and left 54 observations in the wrong prediction results. On the other hand, the logistic model predicted 311 observations correctly for credit risky firms, and left 99 observations in the wrong prediction results (the total number of credit risky firms was 410 in the modeling sample). Neural network predicted 317 observations correctly for credit risky firms, and left 93 observations in the wrong prediction results. The ANN/logistic hybrid model predicted 341 observations correctly for credit risky firms, and left 69 observations in the wrong prediction results. Generally speaking, in the testing sample, the predicting rate of the logistic model was 71.55%, the predicting rate of neural network was 75.44%, and the predicting rate of the ANN/logistic hybrid model was 84.59%.

6. Conclusions and future research

The purpose of this study is to investigate the accuracy of a hybrid credit risk model (ANN/logistic hybrid model) for SME lending using data of Finnish SMEs for the period 2004 to 2012. To the best of our knowledge, our study is the first one to investigate the ANN/logistic hybrid model with data on SMEs.

Our main empirical results indicate that the ANN/logistic hybrid model is more accurate in evaluating credit risk in SMEs lending. Also, we find that, at least in our data on Finnish SMEs, credit risky firms are more likely to have a lower profitability ratio than non-credit risky firms, and a higher leverage and activity ratio than non-credit risky firms. Furthermore, credit risky firms are likely to be younger and larger than non-credit risky firms.

This study contributes to present literature on four aspects. First, our study is one of few that sheds light on the hybrid model, while most of previous ones concentrate on traditional statistical methods or computational (artificial) intelligence methods. Second, SMEs have unique accounting characteristics compared to big firms. In order to develop SME business lending, instead of a credit
| Table 4 |
| Correlations. |

<table>
<thead>
<tr>
<th>ROA</th>
<th>ROCE</th>
<th>ROE</th>
<th>LEVERAGE_A</th>
<th>LEVERAGE_E</th>
<th>LTTA</th>
<th>STTA</th>
<th>TANGIBLE</th>
<th>CURRENT</th>
<th>A_TURN</th>
<th>WORING_TASST</th>
<th>ARR</th>
<th>APR</th>
<th>SIZE_A</th>
<th>SIZE_TURN</th>
<th>SIZE_EMPLOYEE</th>
<th>AGE</th>
<th>YEAR</th>
<th>INDUSTRY</th>
<th>PAT_ASST</th>
<th>EBIT_ASST</th>
<th>SALE_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROA</td>
<td>1.00</td>
<td>.89**</td>
<td>.83**</td>
<td>-.44**</td>
<td>-.44**</td>
<td>-.25**</td>
<td>-.13**</td>
<td>-.09**</td>
<td>.25**</td>
<td>.94**</td>
<td>.06**</td>
<td>.11**</td>
<td>-.10**</td>
<td>-.14**</td>
<td>-.04**</td>
<td>-.10**</td>
<td>-.07**</td>
<td>.94**</td>
<td>.91**</td>
<td>.91**</td>
<td></td>
</tr>
<tr>
<td>ROCE</td>
<td>.89**</td>
<td>1.00</td>
<td>.80**</td>
<td>-.19**</td>
<td>-.19**</td>
<td>-.32**</td>
<td>-.16**</td>
<td>-.15**</td>
<td>.06**</td>
<td>.28**</td>
<td>.08**</td>
<td>.26**</td>
<td>.05**</td>
<td>.05**</td>
<td>-.15**</td>
<td>-.02**</td>
<td>-.13**</td>
<td>-.07**</td>
<td>.75**</td>
<td>.85**</td>
<td>.92**</td>
</tr>
<tr>
<td>ROE</td>
<td>.83**</td>
<td>.80**</td>
<td>1.00</td>
<td>-.13**</td>
<td>-.13**</td>
<td>-.12**</td>
<td>-.02**</td>
<td>-.07**</td>
<td>.08**</td>
<td>.12**</td>
<td>.05**</td>
<td>.12**</td>
<td>-.02**</td>
<td>-.15**</td>
<td>-.05**</td>
<td>-.11**</td>
<td>-.14**</td>
<td>-.08**</td>
<td>.07**</td>
<td>.83**</td>
<td>.93**</td>
</tr>
<tr>
<td>LEVERAGE_A</td>
<td>-.44**</td>
<td>-.19**</td>
<td>-.13**</td>
<td>1.00</td>
<td>.00</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.07**</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.03**</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.02**</td>
<td>-.13**</td>
<td>-.06**</td>
<td>-.11**</td>
<td>-.14**</td>
<td>-.08**</td>
<td>-.03**</td>
<td>-.05**</td>
</tr>
<tr>
<td>LEVERAGE_E</td>
<td>-.44**</td>
<td>-.19**</td>
<td>-.13**</td>
<td>1.00</td>
<td>.00</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.07**</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.03**</td>
<td>-.02**</td>
<td>-.10**</td>
<td>-.02**</td>
<td>-.13**</td>
<td>-.06**</td>
<td>-.11**</td>
<td>-.14**</td>
<td>-.08**</td>
<td>-.03**</td>
<td>-.05**</td>
</tr>
<tr>
<td>LTTA</td>
<td>-.14**</td>
<td>.00</td>
<td>-.02**</td>
<td>-.10**</td>
<td>.00</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>STTA</td>
<td>-.05**</td>
<td>-.03</td>
<td>.10**</td>
<td>-.11**</td>
<td>-.11**</td>
<td>.00</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>TANGIBLE</td>
<td>-.97**</td>
<td>-.06</td>
<td>-.03</td>
<td>.10**</td>
<td>-.11**</td>
<td>-.11**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>CURRENT</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
<td>-.25**</td>
<td>-.02**</td>
<td>-.02**</td>
<td>-.02**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>A_TURN</td>
<td>0.07**</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>WORING_TASST</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>ARR</td>
<td>0.95**</td>
<td>.91**</td>
<td>.95**</td>
<td>.68**</td>
<td>.68**</td>
<td>.68**</td>
<td>.68**</td>
<td>.68**</td>
<td>.68**</td>
<td>.68**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>APR</td>
<td>0.11</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>SIZE_A</td>
<td>-.11**</td>
<td>-.03**</td>
<td>-.07**</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>-.007</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>SIZE_TURN</td>
<td>-.01</td>
<td>-.00</td>
<td>-.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>SIZE_EMPLOYEE</td>
<td>-.01</td>
<td>-.01</td>
<td>-.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>YEAR</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.06</td>
<td>-0.08**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>-0.04**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>PAT_ASST</td>
<td>-.05</td>
<td>.02</td>
<td>.04</td>
<td>-.08**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>EBIT_ASST</td>
<td>0.05</td>
<td>.04</td>
<td>.06</td>
<td>-.08**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>-.04**</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>SALE_A</td>
<td>0.09</td>
<td>.08</td>
<td>.06</td>
<td>-.03</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>-.02</td>
<td>1.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

This table presents Pearson (Spearman) correlations below (above) the diagonal. CREDIT_RISK is an indicator variable equals to 1 in case of the firm is in credit risk status (including the status of Active (default of payment), the status of Active (insolvency proceedings), the status of Bankruptcy, the status of Dissolved (bankruptcy)); 0 otherwise. ROA is a ratio of return on total assets. ROCE is a ratio of earnings before interest and tax to capital employed (capital employed=total assets-current liabilities). ROE is a ratio of return on equity. PAT_ASST is a ratio of profit after tax to total assets. EBIT_ASST is a ratio of EBIT to total assets. LEVERAGE_A is a ratio of total liabilities to total assets. LEVERAGE_E is a ratio of total liabilities to equity. LTTA is a ratio of noncurrent liabilities to total assets. STTA is a ratio of current liabilities to total assets. APR is a ratio of trade creditors to total assets. ARR is a ratio of trade receivables to total assets. TANGIBLE is a ratio of tangible assets to total assets. WORING_TASST is a ratio of working capital to current assets. CURRENT is a ratio of current assets to current liabilities. A_TURN is a ratio of operating revenue to net assets. SALE_A is a ratio of operating revenue to total assets. SIZE_A is the natural logarithm of firms’ total assets. SIZE_TURN is the natural logarithm of firms’ turnover. SIZE_EMPLOYEE is the number of the firms’ employees. AGE is the natural logarithm of 1 plus years since firm’s start to incorporate. YEAR is a year indicator variable defined year 2004 as 1; defined year 2005 as 2; defined year 2006 as 3; defined year 2007 as 4; defined year 2008 as 5; defined year 2009 as 6; defined year 2010 as 7; defined year 2011 as 8; defined year 2012 as 9. INDUSTRY is an industry indicator variable classifying industry into 10 types. Data covers the years 2004 to 2012 and contains 2681 observations.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).
risk model for overall firms, a powerful and reliable credit risk estimation model just for SMEs is necessary. Our study is one of few that just focuses on a credit risk model only for SMEs. Third, our database can enable us to propose the credit risk model in Finnish cases, whereas most of the existing studies have used data on the US and UK. Fourth, this study demonstrates that with the inclusion of traditional statistic methods (e.g. logistic regression), the accuracy of artificial intelligent technologies (e.g. ANN) in credit scoring models can be improved.

The proposed ANN/logistic hybrid model has practical implications. Credit scoring is one of the main areas in accounting and finance which expert and intelligent technologies have been applied into. With the recognition of SMEs economic importance, increasing capital requirements for banks and technological
**Acknowledgements**

We would also like to acknowledge the financial support from OP-Pohjola Group Research Foundation, Liikesivistysrahasto, CSC and Jenny ja Antti Wihurin Rahasto.

**References**


This table presents the logistic regression models in modeling sample. N denotes the number of the firm years. Data covers the years from 2004 to 2011. Coefficient significant at 10% level or better are reported in bold. CREDIT_RISK is an indicator variable equals to 1 in case the firm is in credit risk status (including the status of Active (default of payment), the status of Active (insolvency proceedings), the status of Bankruptcy, the status of Dissolved (bankruptcy)); 0 otherwise. ROA is a ratio of return on total assets. ROCE is a ratio of earnings before interest and tax to capital employed (capital employed = total assets-current liabilities). ROE is a ratio of return on equity. PAT_AST is a ratio of profit after tax to total assets. EBIT_AST is a ratio of EBIT to total assets. LEVERAGE_A is a ratio of total liabilities to total assets. LEVERAGE_E is a ratio of total liabilities to equity. LTTA is a ratio of noncurrent liabilities to total assets. STTA is a ratio of current liabilities to total assets. APR is a ratio of trade creditors to total assets. ARR is a ratio of trade receivables to total assets. TANGABLE is a ratio of tangible assets to total assets. WOERING_TASS is a ratio of working capital to total assets. CURRENT is a ratio of current assets to current liabilities. A_TURN is a ratio of operating revenue to net assets. SALE_A is a ratio of operating revenue to total assets. SIZE_A is the natural logarithm of firms’ total assets. SIZE_TURNOVER is the natural logarithm of firms’ turnover. SIZE_EMPLOYEE is the number of the firms’ employees. AGE is the natural logarithm of 1 plus years since firm’s start to incorporate. YEAR is a year indicator variable defined year 2004 as 1; defined year 2005 as 2; defined year 2006 as 3; defined year 2007 as 4; defined year 2008 as 5; defined year 2009 as 6; defined year 2010 as 7; defined year 2011 as 8; defined year 2012 as 9. INDUSTRY is an industry indicator variable classifying industry into 10 types.

This is a comparison of predicting results. Logistic regression means to use the logistic regression model to investigate the credit risk model; Neural network means to use the neural network only to investigate the credit risk model; the ANN/logistic hybrid model means to use the ANN/logistic hybrid model which we proposed to investigate the credit risk model.

### Table 5

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Sample Restriction</th>
<th>CREDIT_RISK</th>
<th>Training model sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>-0.029</td>
<td>0.580</td>
</tr>
<tr>
<td>PC1</td>
<td></td>
<td>-0.279</td>
<td>0.000</td>
</tr>
<tr>
<td>PC2</td>
<td></td>
<td>0.298</td>
<td>0.000</td>
</tr>
<tr>
<td>PC3</td>
<td></td>
<td>-0.117</td>
<td>0.000</td>
</tr>
<tr>
<td>PC4</td>
<td></td>
<td>0.460</td>
<td>0.000</td>
</tr>
<tr>
<td>PC5</td>
<td></td>
<td>-0.515</td>
<td>0.000</td>
</tr>
<tr>
<td>PC6</td>
<td></td>
<td>-0.441</td>
<td>0.000</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>1883</td>
<td>0.179</td>
</tr>
</tbody>
</table>

This is a comparison of models’ predicting accuracy rate. This is a comparison of predicting results. We define accuracy rate as the total number of the correct prediction in two different groups divided by the total number of the samples [Accuracy rate = (No. of correct prediction in non-credit groups + No. of correct prediction in credit groups) / (No. of non-credit groups + No. of credit groups)]. Logistic regression means to use the logistic regression model to investigate the credit risk model; Neural network means to use the neural network only to investigate the credit risk model; the ANN/logistic hybrid model means to use the ANN/logistic hybrid model which we proposed to investigate the credit risk model.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Observed</th>
<th>Predicted</th>
<th>Modeling(1883)</th>
<th>Testing(798)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>CREDIT_RISK</td>
<td>0</td>
<td>609</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>241</td>
<td>713</td>
</tr>
<tr>
<td>Neural Network</td>
<td>CREDIT_RISK</td>
<td>0</td>
<td>716</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>179</td>
<td>775</td>
</tr>
<tr>
<td>ANN/logistic Hybrid Model</td>
<td>CREDIT_RISK</td>
<td>0</td>
<td>767</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>156</td>
<td>798</td>
</tr>
</tbody>
</table>

This is a comparison of predicting results. Logistic regression means to use the logistic regression model to investigate the credit risk model; Neutral network means to use the neural network only to investigate the credit risk model; the ANN/logistic hybrid model means to use the ANN/logistic hybrid model which we proposed to investigate the credit risk model.