ISSN 1392 – 2785 Inzinerine Ekonomika-Engineering Economics, **2010, 21(1)**
ECONOMICS OF ENGINEERING DECISIONS

Data Reduction Influence on the Accuracy of Credit Risk Estimation Models

Ricardas Mileris, Vytautas Boguslauskas

Kaunas University of Technology
K. Donelaicio str. 73, LT-44309, Kaunas, Lithuania
e-mail: ricardas.mileris@ktu.lt, vytautas.boguslauskas@ktu.lt

Credits in banks have risk of being defaulted. The main purpose of credit risk estimation in banks is the determination of company's ability to fulfil its financial obligations in future. It is very important to have a proper instrument for the estimation of credit risk in banks because it reduces potential loss due to crediting reliable clients. Banks develop internal credit risk estimation models and various data analysis methods can be applied for this purpose. Statistical predictive analytic techniques and artificial intelligence can be used to determine default risk levels. Banks must also have data about clients from the activity in the past. To understand risk levels of credits, banks usually collect information about borrowers. Financial ratios remain primary variables for predicting corporate financial distress. The principal financial ratios as variables for the analysis are indicators of company's financial structure, solvency, profitability and cash flow. Credit risk estimation models are based on the analysis of this data. Using these models it becomes possible to predict the default possibility of new clients. Credit risk estimation models in banks differ significantly in architecture and operating design. The main reason for these differences is that banks' models are assigned by bank personnel and are usually not revealed to outsiders.

The object of this research is credit risk estimation models. The purpose of research is to develop credit risk estimation models and to evaluate an influence of input data reduction on credit risk models accuracy. Two methods were applied in this research: the analysis of scientific publications about estimation of credit risk and the analysis of developed in this research credit risk estimation models performance.

Analyzing financial data of Lithuanaian companies three initial credit risk estimation models were developed wherein these data analysis methods were applied: discriminant analysis (DA), logistic regression (LR) and artificial neural networks (ANN) - multilayer perceptron. 60 financial rates of companies were analyzed. They were calculated from the financial reports of Lithuanian companies for 3 years. The variable selection for the DA model was accomplished applying the analysis of variance (ANOVA) and Kolmogorov-Smirnov test. The variable selection for the LR model was accomplished applying ANOVA. The actual variables for credit risk analysis in ANN model were selected by network according to their ranks of importance. The classification accuracy of models was evaluated by the correct classification rate (CCR). The highest classification accuracy was reached by LR model, which classified 97% of companies correctly. ANN model correctly classified 95.5%, DA model – 84% of companies.

Further situation was analyzed where 60 initial variables were reduced applying factor analysis and the changes in classification accuracy of models were estimated. The number of factors to retain were calculated by the Kaiser criterion ant the scree test. After the factor analysis the 6 new credit risk estimation models were developed applying the same data analysis methods: DA, LR and ANN. By every method 15 and 6 new variables obtained from the factor analysis were analyzed. The research has shown that the new 15 variables extract 89.37% of variance from initial variables. Analyzing these variables, the percent of correctly classified companies mostly decreased in ANN model (-14.6%). The classification accuracy of other models decreased from 2.0% to 7.1%. If an analyst includes into credit risk estimation only 6 new variables, which extract 63.92% of variance from initial variables, the highest decrease in classification accuracy will also be in ANN model (-15.7%).

Keywords: *artificial neural networks, credit risk, discriminant analysis, factor analysis, logistic regression.*

Introduction

Credit risk analysis has recently emerged as a necessity to help banks understand the importance of this risk and plan the appropriate countermeasures in advance. Usually such analysis is based on a number of indicators (parameters) that quantify the clients on which a bank designs credit risk measurement instruments.

Therefore there is no absolutely correct ways of estimating credit risk level for all situations. Different models allow to make the best decision in concrete cases. Under the conditions of objective existence of risk and connected with it financial and other losses, there is always a need for a certain mechanism which would allow to measure credit risk while making decisions.

Many researchers (Abdou, 2009; Fantazzini, Figini, 2009; Lieu, Lin, Yu, 2008; Liou, 2008; Yu, Wang, Wen, Lai, He, 2008; Ugurlu, Aksoy, 2006; and others) have developed various credit risk estimation models where different data analysis methods were applied: discriminant analysis, logistic regression, artificial neural networks, classification trees, support vector machines, etc. Often credit risk models are not able to operate with extremely large arrays of data about clients and the demand to reduce the amount of variables arises.

The object of this research is credit risk estimation models.

The purpose of research is to develop credit risk estimation models and to evaluate the influence of input data reduction on developed credit risk models accuracy.

The methods of the research:

1. Analysis of scientific publications.
2. Analysis of credit risk estimation models performance.

Credit scoring models and their classification techniques are under examination in this study. This study explores the performance of credit scoring models using traditional and artificial intelligence methods: discriminant analysis, logistic regression and artificial neural networks. Experimental studies using real financial data of Lithuanian companies have to demonstrate the influence of data reduction on models classification accuracy.

The purpose of credit scoring models

In recent years banking activity is increasing in many countries. Foreign banks entered the emerging markets to expand their business in various countries (Voinea, Mihaescu, 2008). Credits to private and public sectors increase money supply there, that influence investments, consumption and economic growth (Teresiene, Aarma, Dubauskas, 2008). Companies often need credits and their successful activity is one of the most important economics growth factors having the basic impact on the general development of the country's economy and social stability, creation of new work places (Tamosiunas, Lukosius, 2009).

The empirical findings of many studies suggest that bank's specific characteristics, in particular loans intensity, credit risk, and cost have positive and significant impacts on bank performance (Sufian, Habibullah, 2009). The goal of every bank is to achieve the highest profit at the lowest capital price. It determines the need of shareholders to operate using debt as much as possible. Long-term profitable activity of bank using high level of debt shows competence of management team (Gimzauskiene, Valanciene, 2009). The use of statistical analysis methods and information technologies in banks in pursuance of improving work processes is one of the most important opportunities of the application of these technologies in risk management, work and communication processes (Gatautis, 2008).

Bankruptcy predictions and solvency measurements have become important research topics for scientists and credit analysts. As the world's economy has been facing several challenges during the past decades, more and more companies are addressing the problems of insolvency. Within the current context of dynamic changes in business environment the theory and practice of financial management face a question: which method of evaluation of company's financial stability and credit risk to use (Koleda, Lace, 2009).

Categories of risk and return are becoming increasingly important in banks business process. The knowledge of risk indices allows to manage risk properly. The well provided risk management methods let make more reasonable and adequate managing decisions (Vlasenko, Kozlov, 2009). Bank cannot use primitive risk estimation models because proper risk management ensures the bank's profit (Jovarauskiene, Pilinkiene, 2009). The proper decisions lead to a higher level of efficiency. The quality of decisions may be perceived as a function of imperatives (requirements, regulations, orders, sophistication and knowledge) that is based on information needed for decision making (Gudonavicius, Bartoseviciene, Saparnis, 2009). Typical problems of risk estimation in banks are: how to decide which information is valuable and which one is useless, finally, how to assess the quality of the usable information (Ruzevicius, Gedminaite, 2007).

Credit scoring is a very important task for lenders to evaluate the loan applications they receive from clients. Credit scoring models are used to model the potential risk of loan applications, which have the advantage of being able to handle a large volume of credit applications quickly with minimal efforts, thus reducing operating costs, and they may be an effective substitute for the use of judgment among inexperienced loan officers, thus helping to control bad debt losses (Ince, Aktan, 2009). Information technology applications that support decision-making processes and problem solving activities have proliferated and evolved over the past few decades. These systems were further enhanced with components from artificial intelligence and statistics. Intelligent decision-support using advanced decision and optimization technologies are becoming increasingly important in banks and business management (Sakalauskas, Zavadskas, 2009). Banks use statistical methods such as discriminant analysis, linear probability models, probabilistic analysis, artificial intelligence methods, expert systems, artificial neural networks, genetic algorithms, etc., to identify credit risk, (Chen, Li, 2009).

The objective of credit scoring models is to decide whether or not to grant credit to an applicant. The majority of credit scoring models assign credit applicants to either a "good credit" group, which is likely to repay a financial obligation, or a "bad credit" group, with a high probability of defaulting on the financial obligation and hence their application should be denied. Therefore, credit scoring models basically belong to the field of classification problems (Mavri, Angelis, Ioannou, Gaki, Koufodontis, 2008). The accuracy of their estimations over a period of time is very important for financial results of a bank.

Credit scoring modeling has become a core component in risk management systems in banks and financial institutions. In fact, banks are prompt to develop or buy such models in order to make the whole procedure of evaluating credit applications faster, easier and more accurate.

Factor analysis as a data reduction method

The main applications of factor analytic techniques are:

- to reduce the number of variables;
- to detect structure in the relationships between variables, that is to classify variables.

The goal of a factor analysis is to explain the covariance relationships among the variables in terms of some unobservable and non measurable factors. A factor analysis describes groups of highly correlated variables by a single underlying factor that is responsible for the observed correlations (Li, Zhao, Ma, 2008). According to Oreski and Peharda (2008), the central aim of factor analysis is the orderly simplification of several interrelated measures using mathematical procedures. Traditionally, a factor analysis has been used to explore the possible underlying structure in a set of interrelated variables without imposing any preconceived structure on the outcome (Oreski, Peharda, 2008).

If the information on each variable X_i is decomposed to represent the linear combination of various information factors, then:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + ... + a_{ik}F_k + d_iU_i \quad (1)$$

where F_1, F_2, …, F_k – the common factors, which reflect certain information common in many variables;

$a_{i1}, a_{i2}, …, a_{ik}$ – the common factor loads;

U_i – a special factor, which is only related to the variable X_i, indicating a certain special character of this variable;

d_i – load of special factor.

The common factor is expressed by the linear combination of variables under investigation:

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jn}X_n \quad (2)$$

β_{ji} – factor score coefficients;

X_i – variables (Lu, Han, Gao, Cao, 2006).

Factor score coefficients are used to compute the factor scores. Factor scores can be estimated as new variables of individual cases. These factor scores are particularly useful when there is a need to perform further analysis involving the factors that were identified in the factor analysis. Factor scores in the next steps of analysis reduce the data and the research on the original problem can be continued (Chen, Li, 2009).

Each common factor is extracted according to the magnitude of variance contribution of each factor. The common factors extracted in this order are called the 1st principal factor, the 2nd principal factor, …, the *k*th principal factor respectively (Lu, Han, Gao, Cao, 2006).

The number of factors is an arbitrary decision. However, there are guidelines that seem to yield the best results. First, we can retain only factors with eigenvalues greater than 1. This is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. This criterion was proposed by Kaiser (1960). Also the graphical method can be applied - the scree test. It was proposed by Cattell (1966). We can plot the eigenvalues in a line plot. Cattell suggests to find the place where the smooth decrease of eigenvalues appears to level off to the right of the plot (Oreski, Peharda, 2008).

Developed credit risk estimation models

The essence of models was to classify companies into 2 groups:

- Group 1: reliable companies with low possibility of default.
- Group 2: not reliable companies with high possibility of default.

Data sample of this research consisted of 100 Lithuanian companies. In this sample 50 companies operated successfuly and 50 companies bankrupted. Every company was characterized by 20 financial ratios of 3 years:

1. Liquidity ratios: 1.1. Curent ratio; 1.2. Quick ratio; 1.3. Cash to current liabilities; 1.4. Working capital to total assets.

2. Profitability ratios: 2.1. Gross profitability; 2.2. Net profitability; 2.3. Net profit to total assets.

3. Leverage ratios: 3.1. Total liabilities to total assets; 3.2. Total debt to equity; 3.3. Long term debt to equity; 3.4. Equity to total assets.

4. Activity ratios: 4.1. Sales to total assets; 4.2. Sales to long term assets.

5. Other ratios: 5.1. Cash to total assets; 5.2. Current assets to total assets; 5.3. Unappropriate balance to total assets; 5.4. Working capital to sales; 5.5. Activity profit to total assets; 5.6. Activity profit to sales.

The financial reports of 3 years were analyzed, so overall 60 ratios are available for analysis.

Three initial credit risk estimation models were developed wherein these data analysis methods were applied:

1. Discriminant analysis (DA).
2. Logistic regression (LR).
3. Artificial neural networks (ANN).

1. Discriminant analysis (DA). The classification functions allow to determine which group each company most likely belongs to. There are 2 classification functions in the model as there are 2 groups of companies (reliable and not reliable). When analyzing data each function computes the classification score for the groups of reliable and not reliable clients:

$$f_j = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + ... + \beta_{jn}x_n \quad (3)$$

where j – number of the respective group;

α_j – the constant for the j group;

β_{ji} – the weight for the variables in the computation of the classification score for the j group;

x_i – the observed value of each variable for the respective company.

Result f_j is the classification score. Having computed the classification scores, we classify the bank client as belonging to the group for which it has the highest classification score f_j.

Variable selection was accomplished applying the analysis of variance (ANOVA) and Kolmogorov-Smirnov (K-S) test. The purpose of ANOVA was to test for significant differences between means. The variables were rejected which means in the groups of reliable and not reliable clients do not differ significantly. The Kolmogorov-Smirnov one sample test for normality is based on the maximum difference between the sample cumulative distribution and the hypothesized cumulative distribution. This test allowed to verify if the variable has the normal distribution. If the D statistic is significant, then the hypothesis that the respective distribution is normal should be rejected. These variables were not included in the credit risk analysis. So after variable selection 11 variables were analyzed by the discriminant analysis model (Figure 1).

2. Logistic regression (LR). Logistic regression is the method for modeling dichotomous dependent variables. LR modelling is widely used for the analysis of multivariate data involving the binary responses (Bensic, Sarlija, Zekic-Susac, 2005). This method is used for the prediction of the possibility of occurrence of an event by fitting data to a logistic curve. In credit risk estimation the event is company‘s default.

To model the relationship between creditworthiness of a client and its financial information, the bank can assume that the possibility of default (p) depends on the company‘s financial ratios (x_i) as follows:

$$p = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}} \quad (4)$$

where α is the intercept and $\beta_1, \beta_2, ..., \beta_n$, are the regression coefficients of variables $x_1, x_2, ..., x_n$ respectively. The above model (4) is called a logistic regression model (Xi, Lin, Chen, 2009).

All p values depend to range [0; 1]. They reflect enterprise‘s possibility of default from 0 to 100%. Because classification of clients was implemented into two groups the classification threshold was set to 0.5.

The variable selection for this model was accomplished applying ANOVA. After analysis of variance 25 variables were used in logistic regression model (Figure 1).

3. Artificial neural networks (ANN). The multilayer perceptron (MLP) network was developed to solve bank clients classification problem. MLP network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes. The input signal propagates through the network from one layer to another. The perceptron computes a single output from multiple inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. Neural network learns by appropriately changing the internal connection weights (Purvinis, Sukys, Virbickaite, 2005).

The actual variables for the credit risk analysis were selected by a network according to their ranks of importance. The range of ranks is from 0% (the variable is not important) to 100% (the variable is very important). 10 variables with ranks of 0% were rejected and 50 variables were used in this model (Figure 1).

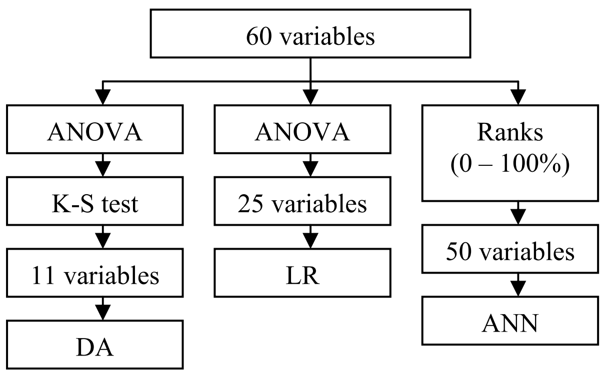

Figure 1. Variable selection for initial models

The classification accuracy of models was measured by the correct classification rate (CCR). It is the proportion of correctly classified companies:

$$CCR = (TP+TN)/N \quad (5)$$

where TP (True Positive) – correctly classified not reliable companies;

TN (True Negative) – correctly classified reliable companies;

N – number of companies analyzed.

The highest classification accuracy was reached by LR model, which classified 97% of companies correctly. ANN model correctly classified 95.5%, DA model – 84% of companies (Figure 2).

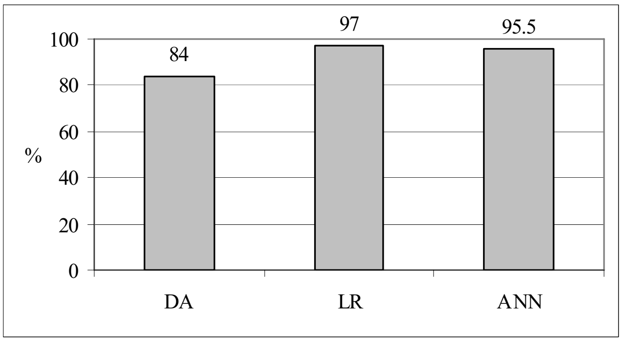

Figure 2. The correct classification rates (CCR) of developed models

Further situation was analyzed where 60 initial variables were reduced applying factor analysis. It was important to estimate the changes in classification accuracy of models.

Data reduction applying factor analysis

Factor analysis was applied as a data reduction method to reduce the quantity of data analyzed by credit risk estimation models.

Numbers of factors to retain were calculated by:

- The Kaiser criterion.
- The scree test.

According Kaiser criterion, we can retain only factors with eigenvalues greater than 1 (Table 1).

Table 1

Eigenvalues and extracted variance

Value	Eigen-value	Total variance, %	Cumulative eigenvalue	Cumulative variance, %
1	11.69	19.49	11.69	19.49
2	7.83	13.05	19.52	32.54
3	6.52	10.87	26.05	43.41
4	5.21	8.69	31.27	52.11
5	4.39	7.33	35.67	59.45
6	2.68	4.46	38.35	63.92
7	2.53	4.22	40.88	68.14
8	2.30	3.84	43.19	71.98
9	2.05	3.42	45.24	75.40
10	1.78	2.96	47.02	78.37
11	1.68	2.80	48.71	81.18
12	1.47	2.45	50.18	83.64
13	1.26	2.11	51.45	85.75
14	1.11	1.86	52.57	87.61
15	1.05	1.75	53.62	89.37

The variances extracted by the factors are called the eigenvalues. If a factor extracts less variance as the equivalent of one original variable, we reject it. Using this criterion, 15 factors (principal components) were retained.

Eigenvalues and extracted variance of factors after rotation *Varimax normalized* are shown in Table 1. In the second column of this table (Eigenvalue), we find the variance on the new factors that were successively extracted. The sum of the eigenvalues is equal to the number of variables (60). In the third column, these values are expressed as a percent of the total variance. As we can see, factor 1 accounts for 19.49 percent of the variance, factor 2 for 13.05 percent, and so on. The fourth column contains the cumulative variance extracted. The fifth column (Cumulative variance, %) indicates the cumulative percent of the total variance extracted by factors.

In order to accomplish the scree test, it is necessary to plot the eigenvalues in a simple line plot (Figure 3).

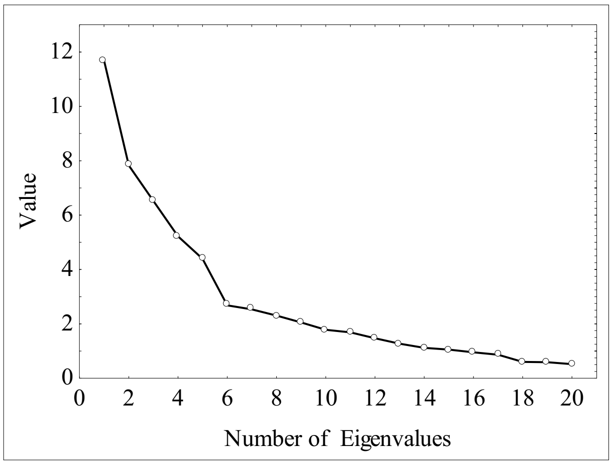

Figure 3. Plot of eigenvalues

We find the point where the smooth decrease of eigenvalues appears to the right of the plot. This point (number of eigenvalues) is equal to 6. According to this criterion, 6 factors were retained in the sample.

Classification accuracy of credit risk estimation models after data reduction applying factor analysis

Owing to problems of missing values in some variables, it was necessary to exclude some companies from the analysis, such that data collection resulted in a total of 89 companies.

The 6 new credit risk estimation models were developed applying the same data analysis methods: discriminant analysis, logistic regression and artificial neural networks. By every method 15 and 6 new variables obtained from factor analysis were analyzed.

Table 2

Correct classification rates of models, %

Model	**15 variables**		**6 variables**	
	CCR	**Change**	**CCR**	**Change**
DA	82.0	-2.0	73.0	-11.0
LR	89.9	-7.1	82.0	-15.0
ANN	80.9	-14.6	79.8	-15.7

The research has shown that the new 15 variables extract 89.37% of variance from initial variables. Analyzing these variables, the percent of correctly classified companies mostly decreased in ANN model (-14.6%). The classification accuracy of other models decreased from 2.0% to 7.1% (Table 2). If the analyst includes into credit risk estimation only 6 new variables, which extract 63.92% of variance from initial variables, the highest decrease in classification accuracy will be also in ANN model (-15.7%).

Conclusions

1. In this research credit risk estimation models were developed applying 3 data analysis methods: discriminant analysis, logistic regression and artificial neural networks.
2. The proposed models analyze the financial data of credit applicants and predict the future repayment behaviour for those who have been characterized as creditworthy and not creditworthy clients.
3. The influence of input data reduction by factor analysis on the classification accuracy of models was estimated. As the results of the factor analysis indicated, 15 variables extracted 89.37% of variance from initial variables. That reduced the classification accuracy of models from 2% to 14.6%. Also 6 new variables extracted 63.92% of variance from initial variables. That reduced the classification accuracy of models from 11% to 15.7%.

References

Abdou, H. A. (2009). An Evaluation of Alternative Scoring Models in Private Banking. *The Journal of Risk Finance*(1), 38-53.

Bensic, M., Sarlija, N., Zekic-Susac, M. (2005). Modelling Small-Business Credit Scoring By Using Logistic Regression, Neural Networks And Decision Trees. *Intelligent Systems In Accounting, Finance And Management*(13), 133-150.

Chen, W. D., Li, J. M. (2009). A Model Based On Factor Analysis And Support Vector Machine For Credit Risk Identification In Small-And-Medium Enterprises. *Machine Learning and Cybernetics*(2), 913-918.

Fantazzini, D., Figini, S. (2009). Random Survival Forests Models for SME Credit Risk Measurement. *Methodology and Computing in Applied Probability*(1), 29-45.

Gatautis, R. (2008). The Impact of ICT on Public and Private Sectors in Lithuania. *Inzinerine Ekonomika-Engineering Economics*(4), 18-28.

Gimzauskiene, E., Valanciene, L. (2009). Performance Measurement System In The Context Of Economics Changes. *Economics & Management*(14), 33-42.

Gudonavicius, L., Bartoseviciene, V., & Saparnis, G. (2009). Imperatives for Enterprise Strategists. *Inzinerine Ekonomika-Engineering Economics*(1), 75-82.

Ince, H., Aktan, B. (2009). A Comparison Of Data Mining Techniques For Credit Scoring In Banking: A Managerial Perspective. *Journal of Business Economics and Management*(10), 233-240.

Jovarauskiene, D., & Pilinkiene, V. (2009). E-Business or E-Technology? *Inzinerine Ekonomika-Engineering Economics*(1), 83-89.

Koleda, N., Lace, N. (2009). Development Of Comparative-Quantitative Measures Of Financial Stability For Latvian Enterprises. *Economics & Management*(14), 78-84.

Lieu, P. T., Lin, C. W., Yu, H. F. (2008). Financial Early-Warning Models on Cross-Holding Groups. *Industrial Management & Data Systems*(8), 1060-1080.

Liou, F. M. (2008). Fraudulent Financial Reporting Detection and Business Failure Prediction Models: a Comparison. *Managerial Auditing Journal*(7), 650-662.

Liviu, V., Flaviu, M. (2008). What Drives Foreign Banks to South-East Europe? *Transformations in Business & Economics*(15), 107-122.

Lu, Y. L., Han, J. Y., Gao, W. H., Cao, S. G. (2006). Using Factor Analysis For Indentifying Latent Contributing Factors Of Customer Loyalty Evaluation Of Mobile Communication Enterprise. *Machine Learning and Cybernetics*, 1913-1917.

Mavri, M., Angelis, V., Ioannou, G., Gaki, E., Koufodontis, I. (2008). A Two-Stage Dynamic Credit Scoring Model, Based On Customers' Profile And Time Horizon. *Journal of Financial Services Marketing*(13), 17-27.

Oreski, D., Peharda, P. (2008). Application of Factor Analysis in Course Evaluation. *Information Technology Interfaces*, 551-556.

Purvinis, O., Sukys, P., & Virbickaite, R. (2005). Research of Possibility of Bankruptcy Diagnostics Applying Neural Network. *Inzinerine Ekonomika-Engineering Economics*(1), 16-22.

Ruzevicius, J., & Gedminaite, A. (2007). Business Information Quality and its Assessment. *Inzinerine Ekonomika-Engineering Economics*(2), 18-25.

Sakalauskas, L., Zavadskas, E. K. (2009). Optimization And Intelligent Decisions. *Technological and Economic Development of Economy,* 15(2), 189-196.

Sufian, F., Habibullah, M. S. (2009). Determinants Of Bank Profitability In A Developing Economy: Empirical Evidence From Bangladesh. *Journal of Business Economics and Management*(10), 207-217.

Tamosiunas, T., & Lukosius, S. (2009). Possibilities for Business Enterprise Support. *Inzinerine Ekonomika-Engineering Economics*(1), 58-64.

Teresiene, D., Aarma, A., Dubauskas, G. (2008). Relationship Between Stock Market and Macroeconomic Volatility. *Transformations in Business & Economics*(14), 102-114.

Ugurlu, M., Aksoy, H. (2006). Prediction of Corporate Financial Distress in an Emerging Market: the Case of Turkey. *Cross Cultural Management: An International Journal*(4), 277-295.

Vlasenko, O., Kozlov, S. (2009). Choosing The Risk Curve Type. *Technological and Economic Development of Economy,* 15(2), 341-351.

Xi, R., Lin, N., Chen, Y. (2009). Compression and Aggregation for Logistic Regression Analysis in Data Cubes. *Knowledge And Data Engineering*(21), 479-492.

Yu, L., Wang, S., Wen, F, Lai, K. K., He, S. (2008). Designing a Hybrid Intelligent Mining System for Credit Risk Evaluation. *Journal of Systems Science and Complexity*(21), 527-539.

Ričardas Mileris, Vytautas Boguslauskas

Duomenų apimties mažinimo įtaka kredito rizikos vertinimo modelių tikslumui

Santrauka

Paskolų teikimas yra viena iš rizikingiausių bankų veiklos sričių, todėl labai svarbu turėti patikimą kredito rizikos vertinimo mechanizmą ir sugebėti tinkamai šią riziką valdyti. Kredito rizikos įvertinimas yra vienas iš kredito suteikimo etapų. Pagrindinis kredito rizikos vertinimo tikslas yra nustatyti, ar verta suteikti kreditą. Bankai stengiasi sumažinti kredito riziką, nesuteikdami kreditų klientams, kurie turi mažiausias galimybes įvykdyti prisiimtus finansinius įsipareigojimus.

Vienas iš kredito rizikos vertinimo būdų yra bankų vidaus kredito rizikos vertinimo modelių taikymas. Šiuos modelius bankai sudaro atsižvelgdami į savo poreikius ir turimus duomenis, pasirenka tinkamiausius duomenų analizės metodus. Modeliams sudaryti reikalingi duomenys apie klientus ir jų įsipareigojimų vykdymą. Skolininkai pagal rizikos lygį skirstomi į atskiras grupes.

Šio tyrimo objektas – kredito rizikos vertinimo modeliai.

Tyrimo tikslas – sudaryti kredito rizikos vertinimo modelius ir įvertinti duomenų mažinimo įtaką modelių klasifikavimo tikslumui.

Tyrimo metodai:

1. Mokslinių straipsnių analizė.
2. Kredito rizikos vertinimo modelių sudarymas ir jų klasifikavimo tikslumo analizė.

Buvo sudaryti 3 įmonių kredito rizikos vertinimo modeliai, kuriuose pritaikyti šie duomenų analizės metodai: diskriminantinė analizė, logistinė regresija, dirbtinių neuronų tinklai. Modeliais analizuojami 3 metų laikotarpio įmonių santykiniai finansiniai rodikliai. Iš viso turima 60 pradinių kintamųjų.

Diskriminantinės analizės modelio esmė tokia: tam, kad būtų galima įvertinti banko klientų kredito riziką, banko klientų grupėms (sėkmingai veikiančių įmonių ir bankrutavusių įmonių) buvo sudarytos klasifikavimo funkcijos. Analizuojamas klientas priskiriamas tai grupei, kurios klasifikavimo funkcija klientą atitinkančiam stebėjimui įgyja didesnę reikšmę. Prieš sudarant klasifikavimo funkcijas, buvo analizuojama, kurie kintamieji tinka tiriamų objektų diskriminavimui. Kintamieji, nepadedantys nustatyti grupių skirtumų, buvo pašalinti. Tam atlikta vienfaktorinė dispersinė analizė ir Kolmogorovo-Smirnovo testas. Atlikus vienfaktorinę dispersinę analizę buvo nustatyta, kurių kintamųjų vidurkiai skirtingų įmonių grupėse statistiškai reikšmingai nesiskiria. Kolmogorovo-Smirnovo testu nustatyti požymiai, kurių pasiskirstymas grupėse nėra normalusis. Į diskriminantinę analizę buvo įtraukta 11 kintamųjų.

Logistinės regresijos modelis savo esme yra glaudžiai susijęs su daugialype tiesine regresija. Logistinė regresinė analizė aprašo priklausomo kintamojo reikšmių priklausomybę nuo nepriklausomų kintamųjų reikšmių matematine formule. Taip pat ši analizė leidžia prognozuoti priklausomo

kintamojo reikšmes. Logistinė regresija dažniausiai taikoma tais atvejais, kai yra tam tikras nepriklausomų kintamųjų rinkinys, o priklausomas kintamasis gali įgyti tik dvi reikšmes (banko klientas įvykdys finansinius įsipareigojimus arba jų neįvykdys). Logistinės regresijos modelį galima sudaryti taip, kad būtų prognozuojamas ne dichotominis kintamasis, o tolydusis, kurio reikšmių intervalas yra [0; 1]. Taikant logistinės regresijos metodą, prognozuojamas ne priklausomas kintamasis Y, o galimybė $P(Y)$ nagrinėjamam objektui šią kintamojo reikšmę įgyti. Šiuo atveju modeliuojama galimybė, kad įmonė neįvykdys prisiimtų finansinių įsipareigojimų bankui. Apskaičiuotos $P(Y)$ reikšmės įvertina kliento finansinių įsipareigojimų neįvykdymo galimybę nuo 0 iki 100 proc. Į logistinės regresijos modelį įtraukti 25 kintamieji, kurie buvo atrinkti atlikus duomenų vienfaktorinę dispersinę analizę.

Dirbtinių neuronų tinklų (toliau – DNT) modeliais yra imituojami žmogaus smegenyse vykstantys procesai. Pagrindinis skirtumas, palyginti su kitais duomenų analizės metodais, yra tas, kad neuronų tinklai nereikalauja tikslaus duomenų analizės modelio, o sudaro jį patys pagal į tinklą įvestą informaciją. Įmonių kredito rizikai įvertinti buvo sudarytas daugiasluoksnis perceptronas. Šis DNT sudarytas iš įvesčių sluoksnio, vidinių sluoksnių ir išvesčių sluoksnio. Įvesčių sluoksnio paskirtis – gauti informaciją iš išorės ir perduoti ją tinklo vidiniams sluoksniams. Neuronų jungtys turi tam tikrus svorius ir perduoda informaciją tarp neuronų. Kiekvienai įvesties reikšmei yra skirtas vienas neuronas įvesčių sluoksnyje. Įvedamos į neuronų tinklą reikšmės šiame sluoksnyje standartizuojamos ir paverčiamos kintamaisiais, kurių intervalas yra [-1; 1]. Šios reikšmės perduodamos tolesnio vidinio sluoksnio neuronams. Kiekvieną neuroną apibūdina jo momentinė būsena: neuronas gali būti sužadintas arba nuslopintas. Jei įvesčių į neuroną svertinė suma viršija nustatytą ribinę reikšmę, neuronas yra sužadinamas. Priešingu atveju neuronas slopinamas. Neuronuose informacija apdorojama tam tikromis funkcijomis, gaunamos neuronų išvesčių reikšmės. Tinklo išvesčių sluoksnio reikšmė yra kredito rizikos analizės rezultatas. Analizei reikšmingų kintamųjų atranka buvo atlikta paties DNT. Kintamiesiems buvo priskirti rangai, kurių intervalas yra nuo 0 iki 100 proc. Į kredito rizikos vertinimą buvo įtraukta 50 kintamųjų, kurių rangai didesni už 0.

Modelių klasifikavimo tikslumui nustatyti buvo apskaičiuotas teisingo klasifikavimo rodiklis, kuris parodo, kokia dalis banko klientų buvo klasifikuota teisingai. Didžiausias tikslumas gautas logistinės regresijos modeliu – 97 proc. Dirbtinių neuronų tinklų modeliu buvo teisingai klasifikuota 95,5 proc. klientų, diskriminantinės analizės modeliu – 84 proc. klientų.

Atliekant tyrimą, buvo siekiama įvertinti duomenų mažinimo įtaką kredito rizikos vertinimo modelių tikslumui. Duomenų apimties mažinimas atliktas faktorinės analizės metodu.

Faktorinės analizės tikslas – pakeisti stebimą reiškinį apibūdinančių požymių rinkinį tam tikru skaičiumi faktorių. Atlikus duomenų faktorinę analizę, buvo siekiama sumažinti analizuojamų kintamųjų skaičių ir nustatyti duomenų kiekio sumažinimo įtaką kredito rizikos vertinimo modelių tikslumui. 60 pradinių kintamųjų suskirstyti į grupes atsižvelgiant į jų koreliacijas su tam tikrais tiesiogiai nestebimais (latentiniais) faktoriais. Faktorius – tai tiesiogiai nestebimas (latentinis) kintamasis F_j, kuris vienija tam tikrą susijusių kintamųjų X_i grupę. Faktorius yra tiesinė kintamųjų kombinacija. Bendrieji faktoriai buvo nustatyti taikant pagrindinių komponenčių analizę.

Vienas iš faktorinės analizės etapų yra faktorių sukimas. Tai faktorių svorių matricos transformavimas į lengviau interpretuojamą pavidalą. Atliekant sukimą „Varimax normalized", buvo siekiama supaprastinti faktorių svorių matricos struktūrą, t. y. tik kelių kintamųjų visų faktorių svoriai būtų nenuliniai. Buvo gauti tolesnei analizei tinkami faktorių reikšmių įverčiai. Faktorinė analizė leido didelį kiekį kintamųjų paaiškinti išskirtaisiais faktoriais. Taip sumažintas analizuojamų duomenų kiekis, o faktorinės analizės rezultatai įtraukti į atliktą analizę kitais daugiamatės statistikos metodais.

Faktorių skaičius buvo nustatytas remiantis Kaizerio kriterijumi ir faktorių tikrinių reikšmių grafiku. Remiantis Kaizerio kriterijumi, buvo išskirti tik tie faktoriai, kurių tikrinės reikšmės didesnės už 1. Kadangi faktoriai išskiriami taip, kad kiekvienas tolimesnis faktorius paaiškina vis mažesnę kintamųjų dispersijos dalį, tai faktorių tikrinės reikšmės yra išsidėsčiusios mažėjančiai. Jei faktorius nesuteikia daugiau informacijos nei vienas pradinis kintamasis, t. y. jo tikrinė reikšmė yra ≤ 1, tokie faktoriai buvo atmesti. Tokiu būdu buvo išskirta 15 faktorių, kurie paaiškina 89,37 proc. bendrosios pradinių kintamųjų dispersijos.

Faktorių skaičius taip pat buvo nustatytas ir remiantis faktorių tikrinių reikšmių grafiku. Grafiko x ašyje vaizduojamas pradinių kintamųjų skaičius (arba faktoriaus numeris), o y ašyje – tikrinės reikšmės. Tikslinga išskirti tiek faktorių, kur kreivės nuolydžio mažėjimo tempas sulėtėja. Taigi, remiantis tikrinių reikšmių grafiku, buvo išskirti 6 faktoriai, kurie paaiškina 63,92 proc. bendrosios pradinių kintamųjų dispersijos.

Sumažinus kintamųjų skaičių, buvo analizuojama duomenų mažinimo įtaka kredito rizikos vertinimo modelių tikslumui. Buvo sudaryti nauji kredito rizikos vertinimo modeliai, atskirai analizuojantys 15 ir 6 kintamuosius. Analizei pritaikyti šie metodai: diskriminantinė analizė, logistinė regresija ir dirbtinių neuronų tinklai. Nustatyta, kad 15 kintamųjų, kurie paaiškina 89,37 proc. bendrosios pradinių kintamųjų dispersijos, modeliuose teisingo klasifikavimo rodiklis sumažėjo nuo 2 iki 14,6 proc. 6 kintamųjų, paaiškinančių 63,92 proc. bendrosios pradinių kintamųjų dispersijos, modeliuose teisingo klasifikavimo rodiklio reikšmė sumažėjo nuo 11 iki 15,7 proc. Didžiausias klasifikavimo tikslumo sumažėjimas nustatytas dirbtinių neuronų tinklų modeliuose.

Raktažodžiai: *dirbtinių neuronų tinklai, diskriminantinė analizė, faktorinė analizė, kredito rizika, logistinė regresija.*

The article has been reviewed.

Received in December, 2009; accepted in February, 2010.