Supplementary Notebook (RTEP - Brazilian academic journal, ISSN 2316-1493)

INDIVIDUAL'S CREDITWORTHINESS ASSESSING MODELS CONSTRUCTION AND STUDY BASED ON BOOTSTRAPING METHOD

Ilyas Idrisovich Ismagilov [1],
Ajgul Ilshatovna Sabirova [2],
Dina Vladimirovna Kataseva [3] ,
Alexey Sergeevich Katasev[4]

[1]Doctor of Economics, Professor of the Department of Economic Theory and Econometrics of Management, Technical Sciences, Finance, Kazan Federal University; e-mail: iiismag@mail.ru.
[2]Ph.D. in Economics, Senior Lecturer at the Department of Accounting, Analysis and Audit of the Institute of Management, Economics and Finance, Kazan (Volga Region) Federal University; e-mail: aigylkinyes@mail.ru.
[3]Senior Lecturer, Department of Information Security Systems, Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI; e-mail: DVKataseva@kai.ru.
[4]Doctor of Technical Sciences, Professor of the Department of Information Security Systems of the Institute of Computer Technologies and Information Security, Kazan National Research Technical University named after A.N. Tupolev-KAI; e-mail: ASKatasev@kai.ru.

Abstract: This article solves the problem of constructing and studying models for individual's creditworthiness assessing. The relevance of solving this problem on the intelligent modeling technologies basis: neural networks, decision trees, logistic regression is noted. The initial data for the models constructing was a set of 35 columns and 149,000 rows. The model's construction and study were carried out in the Deductor Analytical Platform. Each model was tested on data set of 54827 records. For each model we constructed the corresponding classification matrices and calculated the 1st, 2nd kind errors, and the general error of the models. In terms of minimizing these errors, logistic regression showed the worst results, and the neural network showed the best. In addition, the constructed models' effectiveness was evaluated according to the «Income from loans» criterion. According to this criterion, the neural network model was also the best. Thus, the study results showed that to maximize profits and minimize classification errors, it is advisable to use a neural network model. This indicates its effectiveness and

practical use possibility in intelligent decision-making support systems for assessing the potential borrowers' creditworthiness.

Keywords: creditworthiness assessing, neural network, decision tree, logistic regression, data mining, decision-making support.

INTRODUCTION

In the information technology development era in conditions of processed information amount increase, the intellectual analysis of large data sets (data mining) becomes relevant (9; 10). Data mining allows us to reveal hidden patterns and use them in decision-making support systems (4; 13). This approach to modeling is advisable to use in the banking sector, particularly in the lending field (1; 3). Data analysis allows us to assess the creditworthiness of individuals and legal entities, considering various factors, to assess profit or loss of each credit approval, to assess credit repayment probability in case of late payments, etc. The individual's creditworthiness analysis is an integral part in the activities of each bank. In this process data mining is a necessary tool. It is relevant to use data mining due to the following main factors (12): 1) creditworthiness assessing model constructing does not require expensive equipment; 2) after setting the model parameters, the intervention of analysts is not required to decide in each specific case.

Currently, various methods are used in practice to construct models for individual's creditworthiness assessing (15; 23; 5; 6; 24.18): neural networks, decision trees, logistic regression, statistical methods, methods based on financial ratios, methods based on cash flows analysis, methods based on business risks analysis, etc. To increase the practical problems solving efficiency in individual's creditworthiness assessing, it is necessary to construct and study various models, and to select the best one according to the adequacy criterion. In this issue, to achieve this goal, data for the analysis are collected and prepared, a neural network model, a logistic regression and a decision tree for individuals creditworthiness assessing were constructed, and the study of the constructed model's effectiveness based on the bootstrapping method was conducted (22; 7).

METHODOLOGY

The modeling problems solution is often associated with the need to combine analytical processing methods. This is due to the fact that real data in a «raw» form is often unsuitable for analysis. Information collected from many sources must be cleaned, transformed, systematized, and only then the intellectual analysis methods can be applied to her. It is advisable to adhere to the standard stages of the *Knowledge Discovery in Databases* (2) technology (2), including data preparation, informative features selection, data cleaning, Data Mining methods application, data post-processing and the results interpretation. In this work, for data preparation and model's construction for individual's creditworthiness assessing we used the Deductor Studio analytical platform (19). This tool allows to prepare the initial data for analysis (to clear, to transform), to construct regression models, decision trees and to train neural networks.

For model's construction, it is necessary to create a file with the training set, in which examples of solving the individual's creditworthiness assessing problem («inputs» - «output») should be presented. The algorithm of data preparation for analysis is to create a set containing all the necessary data for creditworthiness assessing (data prepared by an expert containing all the necessary fields for further analysis). In this work we used a data set contained 149,000 lines. Each line contained the value of such fields as «Credit amount», «Credit cost», «Credit term», «Credit purpose», «Age», etc. The total set of input data consisted of 34 input fields and one output containing an expert decision to give or refuse to give a credit. The full name and fields types of the initial data are presented in table 1.

Table 1. The initial data set structure and field type

№	Field name	Field type
1	Credit amount	Integer
2	Credit cost	Integer
3	Credit term	Integer
4	Credit date	Date
5	Credit purpose	String
6	Amount	Integer
7	Age	Integer
8	Gender	String
9	Education	String
10	Private property	Logical
11	Flat	Logical
12	Apartment area	Integer
13	Property acquisition method	String
14	Location	String
15	Car	String
16	Vehicle life	Integer
17	Vacation home	Logical
18	Land plot	Logical
19	Residence in the area	Logical
20	Garage	Logical
21	Enterprise class	String
22	Enterprise lifetime	Integer
23	Enterprise branch	String
24	Specialization	String
25	Position	String
26	Duration of work at the enterprise	Integer
27	Duration of work in the specialty	Integer
28	Average monthly income	Integer
29	Average monthly consumption	Integer
30	The main area of expenditure	String
31	Number of dependents	Integer
32	Married or not	Logical
33	Employment of a spouse (wife)	Logical
34	Period of residence in the region	Integer
35	Credit decision	Logical

For intelligent models' construction and evaluation it is required to form the training and testing sets from the initial data set. There are various approaches for this (8): from direct initial set partitioning into two data sets in predetermined proportions to applying more complex procedures for the desired sets randomly generating.

In this work, the bootstrapping method was used (22; 7). It is multiple random sampling of *N* values from the initial set of volume *N*. Obtained training sets obviously have repeating elements. The corresponding testing sets consist of the original data set elements that are not used in training. On each training set, a model is built, and on the corresponding testing set, an assessment of the constructed model is performed.

Further, the obtained results are averaged. This method implies a better set dispersion; therefore, the obtained averaged results are more accurate than the results of the usual decomposition of the original data set into training and testing sets. On the generated training sets, the corresponding individual's creditworthiness assessing models were constructed: neural networks, decision trees, logistic regression models.

In constructed neural network models, the different number of hidden layers of neural networks, the number of neurons in each layer were set, the training method was selected, the input, output parameters and the maximum number of training eras were specified. After conducting many experiments, the following conclusion was made: 2-layer neural network models with 12 neurons in each hidden layer are optimal due to achieved accuracy. For the neural networks with selected architecture training, the *Back Propagation* method was used. The essence of this method is the propagation of error signals from the network outputs to its inputs in the direction, opposite to the direct propagation of signals in normal operation. During the training, the maximum number of eras was set, equal to 5000. Training also ended if the standard error of recognition reached a level of 0.05.

To construct the decision trees, we need to set input and output parameters, configure stop and cut off parameters, and select the construction method. Experimentally, we selected the following optimal training parameters for decision trees: minimum number of examples in a node is 2; build a tree with more reliable rules to the detriment of its compactness; cut off tree nodes at a confidence level of 20%. The first two parameters relate to early stop parameters. In this case, the next node will be divided into subnodes, if the number of unrecognized examples in the node is greater than the value of the «minimum number of examples in the node» parameter.

The third parameter refers to the cut-off parameters: the smaller level of trust, the more nodes will be cut off, and the smaller will be the tree. To set up the logistic regression, we need to set the input and output parameters, choose the variables selecting method, the estimate accuracy, the number of iterations, and the cut-off threshold. Full inclusion (*Enter*) was chosen as the selection method, the maximum number of iterations was set equal to 100, the accuracy of the estimation function was 10^{-7}, and the cut off threshold was 0.1.

RESULTS

To define the quality of the constructed individual's creditworthiness assessing models, in the Deductor we construct the «Give to everyone» model, which simulates a situation where all credit assessment decisions are positive. At the same time, we introduce new variables: *TP* (*true positive*) – the number of truly positive outcomes; *FP* (*false positive*) – the number of false positive outcomes; *FN* (*false negative*) – the number

of false negative outcomes; *TN* (*true negative*) – the number of true negative outcomes. These variables will later be used for 1st and 2nd kind model's errors evaluation. Let us compare the constructed model's effectiveness, including the «Give to everyone» model, according to the «Income from credits» criterion, calculated in conventional monetary units (see Table 2).

Table 2. Models comparison by the credit income criterion

Model	Credit income, c.u.
«Give to everyone»	190 000
Neural network	352 000
Decision tree	315 000
Logistic Regression	290 000

In this case, the «Give to everyone» model turned out to be the worst model. This is due to the large number of false positive outcomes in this model. Therefore, such model cannot be applied in practice. The best results were shown by the neural network model. It qualitatively analyzes the data and allows to get the maximum income from credits. For the constructed models practical use, it is required to determine their adequacy, that is, compliance with how accurately they solve the problem. The adequacy of each model can be determined by using bootstrap estimates. Obtaining these estimates is based on the sampling procedure, which allows us to select the same records from the initial data several times, forming training and testing sets. Consider a particular case of the bootstrap method, which is called «0.632-bootstrap» (16). A data set from n observations is selected with substitution to form another data set, also consisting of n cases. Since some elements in the second set will be repeated, and the initial and the resulting set contain the same number of examples, it turns out that some examples will not be selected to the second set. They will be used as test ones. The probability of choosing an observation is $1/n$. Accordingly, the probability that the observation will not be selected is $1-1/n$. Multiplying these probabilities by each other n times, we obtain $(1-1/n)n=e^{-1}=0{,}368$. This gives an estimate of the probability that a certain observation will not generally be selected. Thus, if the initial data set is large enough, the testing set will contain approximately 36.8% of the observations. On the created samples we tested the neural network model, the decision tree model and the logistic regression model (15; 21; 17). Testing results are presented in the corresponding contingency tables (24) (see tab. 3-5).

Table 3. The neural network model testing results

Actual Values	Classified by the model		
	0	1	Total
0	31839	2152	33991
1	819	20017	20836
Total	32658	22169	54827

Table 4. The decision tree model testing results

Actual Values	Classified by the model		
	0	1	Total
0	30776	3215	33991
1	1374	19462	20836
Total	32150	22677	54827

Table 5. The logistic regression model testing results

Actual Values	Classified by the model		
	0	1	Total
0	29711	4280	33991
1	1958	18878	20836
Total	31669	23158	54827

In tables 3-5 the number «1» in the column «Actual Values» means that the client is solvent, and he received a credit and the number «0» means the client is insolvent and he did not receive a credit. In addition, the number «1» in the «Classified by the model» column means that the client is recommended to receive a credit, and the number «0» - the client is recommended to refuse a credit. Based on the data presented in tables 3-5, the 1st, 2nd kind errors and the general model errors were calculated (see Table 6) (26).

Table 6. Models testing errors

Model	Testing results		
	1st kind errors,%	2st kind errors,%	Total error, %
Neural network	3,93	6,33	10,26
Decision tree	6,59	9,46	16,05
Logistic regression	9,4	12,59	21,99

As we can see from the table, from the point of view of the 1st, 2nd kind errors, and the general error of the models, logistic regression showed the worst results. At the same time, the neural network showed the best results in terms of minimizing these errors. In addition, the 1st kind error in the neural network model was less than 5%, which is considered as a high result in the intelligent model's construction. Other models, by the 1st kind errors, could not enter the 5% confidence interval.

SUMMARY

The study showed that to assess the individual's creditworthiness, it is possible and advisable to use modern methods of data mining for decision support models construction. Such models allow us to assess the maximum profit obtained as a result of lending. In addition, the intelligent model's application allows us to evaluate the possibility of potential borrowers to obtain cash automatically, without resorting to credit organizations. This saves the time of borrowers and the banking sector specialists.

CONCLUSION

Thus, the work solved the problem of intellectual models constructing and their effectiveness researching for the individual's creditworthiness assessing. The study results showed that in order to maximize profits from lending, it is advisable to use a model based on the multilayer neural network training. In addition, this model showed the greatest accuracy in terms of minimizing the 1st and 2nd kind errors. This indicates its effectiveness and practical use possibility in intelligent decision-making support systems for assessing the creditworthiness of borrowers (11; 20).

ACKNOWLEDGMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

1. Abedini, M., Ahmadzadeh, F., & Noorossana, R. (2016). Customer credit scoring using a hybrid data mining approach. *Kybernetes, 45*(10), 1576-1588.

2. Adhikari, A., Jain, L. C., & Prasad, B. (2017). A state-of-the-art review of knowledge discovery in multiple databases. *Journal of Intelligent Systems*, *26*(1), 23-34.

3. Alborzi, M., & Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. *International Journal of Business Information Systems, 23*(1), 1-22.

4. Alekseev, A.A., Katasev, A.S., Khassianov, A.F., Tutubalina, E.V., Zuev, D.S. (2018). Intellectual information decision support system in the field of economic justice. *CEUR Workshop Proceedings, 2260*, 17-27.

5. Ansori, M.F., Sidarto, K.A., & Sumarti, N. (2019). Logistic models of deposit and loan between two banks with saving and debt transfer factors. AIP Conference Proceedings. 2192,060002.

6. Asar, Y., & Wu, J. (2020). An improved and efficient biased estimation technique in logistic regression model. *Communications in Statistics - Theory and Methods, 49*(9), 2237-2252.

7. Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, *54*(12), 2976-2989.

8. Brunelli, A., & Rocco, G. (2006). Internal validation of risk models in lung resection surgery: bootstrap versus training-and-test sampling. *The Journal of thoracic and cardiovascular Surgery*, *131*(6), 1243-1247.

9. Chamikara, M.A.P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2020). Efficient privacy preservation of big data for accurate data mining. *Information Sciences, 527,* 420-443.

10. Dagaeva, M., Garaeva, A., Anikin, I., Makhmutova, A., & Minnikhanov, R. (2019). Big spatio-temporal data mining for emergency management information systems. *IET Intelligent Transport Systems, 13*(11), 1649-1657.

11. Dela Cruz Galapon, A. (2020). An assessment: Respiratory analysis using data mining method - A decision support system. *Test Engineering and Management, 83,* 4824-4829.

12. Gulsoy, N., & Kulluk, S. (2019). A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers. Wiley *Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), 1299.

13. Guo, Y., Wang, N., Xu, Z.-Y., & Wu, K. (2020). The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology. *Mechanical Systems and Signal Processing, 142*(106630).

14. Ismagilov, I. I., Khasanova, S. F., Katasev, A. S., & Kataseva, D. V. (2018). Neural network method of dynamic biometrics for detecting the substitution of computer. *Journal of Advanced Research in Dynamical and Control Systems*, *10*(10 Special Issue), 1723-1728.

15. Ismagilov, I.I., Molotov, L.A., Katasev, A.S., & Kataseva, D.V. (2019). Construction and efficiency analysis of neural network models for assessing the financial condition of enterprises. *Journal of Advanced Research in Dynamical and Control Systems, 11*(8),1842-1847.

16. Jiang, W., & Chen, B. E. (2013). Estimating prediction error in microarray classification: Modifications of the 0.632+ bootstrap when \bfn<\bfp. *Canadian Journal of Statistics*, *41*(1), 133-150.

17. Katasev, A. S., & Kataseva, D. V. (2016). Neural network diagnosis of anomalous network activity in telecommunication systems/Dynamics of Systems, Mechanisms and Machines (Dynamics). *Dynamics*, *7819020*.

18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).

19. Lomakin, N., Shokhnekh, A., Sazonov, S., Polianskaia, A., Lukyanov, G., & Gorbunova, A. (2019, October). Hadoop and Deductor Based Digital Ai System for Predicting Cost of Innovative Products in Conditions of Digitalization of Economy. In *Proceedings of the 2019 International SPBPU Scientific Conference on Innovations in Digital Economy* (pp. 1-8).

20. Mehdi, B., Hasna, C., & Tayeb, O. (2019). Intelligent credit scoring system using knowledge management. *IAES International Journal of Artificial Intelligence*, *8*(4), 391-398.

21. Mustafin, A. N., Katasev, A. S., Akhmetvaleev, A. M., & Petrosyants, D. G. (2018). Using Models of Collective Neural Networks for Classification of the Input Data Applying Simple Voting. *The Journal of Social Sciences Research*, 333-339.

22. Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical science*, 219-230.

23. Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy, 26*(2), 405-429.

24. Sulewski, P. (2019). Some contributions to practice of 2× 2 contingency tables. *Journal of Applied Statistics*, *46*(8), 1438-1455.

25. Swiderski, B., Kurek, J., Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. *Decision Support Systems, 52*(2), 539-547.

26. Zhang, Q., Xia, D., & Wang, G. (2017). Three-way decision model with two types of classification errors. *Information Sciences*, *420*, 431-453.