УДК 330.4
DOI 10.33111/nfmte.2020.156

RESEARCH OF THE DEPENDENCE OF THE EFFICIENCY OF MODELING THE CREDITWORTHINESS OF BORROWERS ON THE METHOD OF FORMING A CONTROL SUBSET

Viacheslav Pyrohov

Barclays Bank PLC
1683/127 Na Pankráci, Praha, 14000, Czech Republic
ORCID: 0000-0002-0948-5230, E-mail: viacheslav.pyrohov@gmail.com

Stanislav Turchenko

Sberbank Europe AG
3 Schwarzenbergplatz, Vienna, 1010, Austria
E-mail: stanislav.turchenko@sberbank.at

In the article has been conducted a research aiming increase of classification result stability of commercial bank's debtor creditworthiness with usage of boosted decision trees and neural network algorithms due to the use of stratified sampling. It is proposed to improve the classical procedure of stratified sampling by taking into account not only the target variable, but also the most significant predictors of the model when forming the control subset.

Experimental calculations to test the proposed hypotheses were carried out using the program packages LGBM and H2O on the data of international consumer finance provider Home Credit. In the article checked and confirmed that the use of stratified sampling in the process of forming a control subset during training of machine learning models makes possible to increase their stability and accuracy of forecasts on new data sets.

As per the achieved results, the authors' approach of stratified sampling during forming a control dataset by target variable and the most significant characteristics of a model demonstrates a higher average accuracy for boosted decision trees on the test subset compared to the standard stratified sampling algorithm and random selection.

Keywords: *decision tree, gradient boosting, neural network, stratified sampling*

ДОСЛІДЖЕННЯ ЗАЛЕЖНОСТІ ЕФЕКТИВНОСТІ МОДЕЛЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ ВІД СПОСОБУ ФОРМУВАННЯ КОНТРОЛЬНОЇ ВИБІРКИ

В. І. Пирогов

ТОВ «Барклайс Банк»
вул. На Панкраці, 1683/127, м. Прага, 14000, Чехія
ORCID: 0000-0002-0948-5230, E-mail: viacheslav.pyrohov@gmail.com

С. В. Турченко

АТ «Сбербанк Європа»
пл. Шварценберга, 3, м. Відень, 1010, Австрія
E-mail: stanislav.turchenko@sberbank.at

У статті проведено дослідження з підвищення стійкості результатів класифікації кредитоспроможності боржників комерційного банку з використанням бустингових дерев рішень та нейромережевих алгоритмів за рахунок застосування стратифікованого семплінгу. Запропоновано удосконалення класичної процедури стратифікованого семплінгу шляхом врахування при формуванні контрольної вибірки не тільки цільової змінної, але й найбільш значущих предикторів моделі.

Експериментальні розрахунки для перевірки висунутих гіпотез проведено з використанням програмних пакетів LGBM і H2O на даних міжнародного провайдера споживчого кредитування Home Credit. У статті перевірено та підтверджено, що використання стратифікованого семплінгу в процесі формування контрольної вибірки під час навчання моделей машинного навчання дозволяє підвищити їх стабільність і точність прогнозів на нових наборах даних.

Відповідно до отриманих результатів, авторський підхід до стратифікованого семплінгу при формуванні контрольного набору даних за цільовою змінною та найбільш значущими характеристиками моделі демонструє вищу середню точність для бустингових дерев рішень на тестовій вибірці в порівнянні зі стандартним стратифікованим алгоритмом семплінгу та випадковим відбором.

Ключові слова: *дерево рішень, градієнтний бустинг, нейронна мережа, стратифікований семплінг*

JEL Classification: C38, C45, C51, C52, C63

Introduction

With the formation of the modern information society on the edge of XX–XXI centuries the economy faced new challenges and opportunities. The updated economic system generates huge flows of information that can be used to obtain additional economic effect through the correct interpretation of data using modern mathematical methods.

According to recent research [1], just in 2016-2017, humanity has generated more information than in the previous 5,000 years of its development. Despite the large amount of information generated, only a small percentage of it is used to make operational decisions – 0.5% [1].

One very important task for the sustainability of the economy, which has recently been affected by the explosive increase in data, is the assessment of the creditworthiness of borrowers. Not so long ago, a borrower was characterized by several dozen indicators (primarily from a questionnaire for a loan and data from credit bureaus), which were very successfully processed by statistical methods (for example, logistic regression, discriminant analysis, etc.).

But recently, new sources of information about customers related to big data have appeared, such as a digital footprint, social networks, photos and videos of a potential borrower when applying for a loan, etc. Processing these arrays of information makes it possible to generate thousands of features that can be used in predicting the credit behavior of customers. As a result, the use of statistical methods for processing such information becomes irrelevant. Therefore, researchers and practicing data scientists are increasingly turning to the use of artificial intelligence and machine learning methods to assess borrowers' credit risks.

The authors of the paper [2] analyze 258 academic papers since 1976 to detect trends and changes in the credit scoring literature and to reveal the challenges and opportunities big data bring to credit scoring. This paper presents study on how big data challenges traditional credit scoring models and addresses the need to develop new credit models that identify new and secure data sources, and convert them to useful insights that are in compliance with regulations.

To solve the problems of assessing credit risks, artificial intelligence and machine learning methods began to be used at the end of the last century [3-6].

In the last decade, artificial intelligence methods, such as neural networks [7-11], fuzzy logic [12-15], genetic algorithms [16, 17], as well as machine learning techniques [18-20], have been more actively developed and applied in practice.

However, when training artificial intelligence and machine learning models with a large number of adjustable parameters, the negative effect of overfitting appears (when it is strongly adjusted to the training set and begins to poorly describe patterns on test data). The problem of overfitting and the ways of model validation during the training process are covered by various authors [21-23]. The article [24] investigates the dependence of the adequacy of credit scoring models based on logistic regression and perceptron-type neural networks of different configurations on the size of the training sample. The causes of the overfitting effect and ways to prevent it are identified.

The article [25] studies the issue of finding optimal architectures of neural network of multilayer perceptron and RBF types for the problem of assessing the creditworthiness of individual borrowers. Experimental studies have confirmed that the combination of several models in an ensemble makes it possible to compensate for the possible errors of each of them, caused, among other reasons, by the effect of overfitting.

Authors of the paper [26] propose to use the boosting procedure to prevent overfitting. Additionally, it reduces the training time of models (which is important for big data) and gives a bit higher performance for the problem being solved.

Also, a typical problem in modeling complex processes, in particular credit risks, is that the event under study is rare (default occurs much less often than the loan is repaid), and this class imbalance strongly affects the performance of traditional classifiers. The paper [27] discusses the performance of standard boosting procedures to deal with unbalanced classes and proposes a new boosting-based sampling algorithm.

As can be seen, there are many scientific papers devoted to the study of the dependence of the accuracy of credit risk modeling on the choice of mathematical tools and approach to the formation of a training sample. At the same time, the little-studied issue of analyzing the performance of scoring models depending on the choice of control sample is of scientific and practical interest.

The main purpose of the article is to study the influence of the control sample formation procedure on improving the quality of binary classification in the problem of credit risk modeling.

Methodological basis

Stratified sampling method

As justified above, during the creation of machine learning models, an important task of the researcher is the formation of a control sample. The use of control sample during the preparation of the model makes possible to prevent overfitting of the predicative model – to avoid cases when the model is able to successfully recognize only a specific set of training data.

The process of creation of a control sample includes the selection of observations from the general set so that the control sample was as close as possible to the general totality in its properties. To preserve the properties of the complete data set in the control set is used the method of stratified sampling [28].

Stratified sampling is a method of random selection that involves dividing the general totality into smaller subgroups (strata) and combining random sampling from strata. The strata are formed based on the homogeneous characteristics of observations, which makes it possible to reproduce the heterogeneity of the general totality in the sample. A classic work describing the use of stratified sampling in statistics is the paper of J. Neyman "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection" [29]. The basic idea of stratified sampling:

1) splitting the heterogeneous sample into smaller groups (or strata), such that the selection groups are:

- homogeneous with respect to the target characteristics within the strata;
- heterogeneous in terms of target characteristics between strata;

2) random selection of observations from each stratum in accordance with the distribution of its target characteristics in the initial data.

The general approach to stratified selection is shown in Fig. 1.

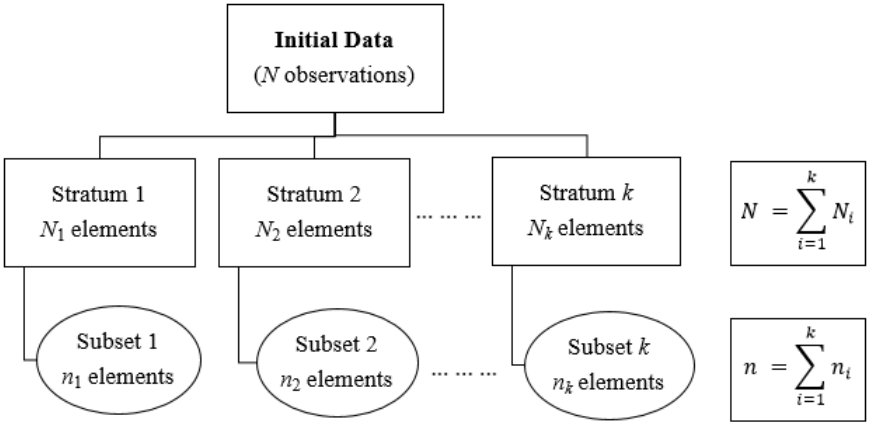

Fig. 1. Description of stratified sampling method

Note that the stratified sampling method is one of the ways to form a control sample. However, it is possible to examine its effectiveness in comparison with other approaches (for example, random selection) only on the basis of mathematical models built on these data sets. We will conduct this study using the boosted decision trees and neural networks.

The method of boosted decision trees

One of the most popular and effective algorithms used in classification problems is the gradient boosted trees method. A classic work that laid the theoretical foundation for the creation of boosting

decision trees is the work of J. Friedman "Greedy approximation of functions: a gradient boosting machine" [30].

Friedman's work is based on the idea that the basic predicative model itself is "weak" and can be strengthened by constructing ensembles of models whose characteristics will be redefined using optimization algorithms (such as a gradient descent algorithm). Once the result of the final ensemble of models is aggregated, the original model is considered "strong" by reducing the variance of the original result and optimizing the parameters. The general representation of the original model will look like:

$$F(\mathbf{x}; \{b_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} b_m h(\mathbf{x}; \mathbf{a}_m), \quad (1)$$

where $\mathbf{x}$ is the set of random "input" or "explanatory" variables ($\mathbf{x} = \{x_1, x_2, \dots, x_n\}$), $h(\mathbf{x}; \mathbf{a}_m)$ is a parametric function (predictive model) with vectors of input variables $\mathbf{x}$ and parameters $\mathbf{a}_m = \{a_1^m, a_2^m, \dots, a_{n_m}^m\}$ (note that in the general case, each model $m = \overline{1, M}$ can have its own number of parameters n_m), and b_m is a the weight of the corresponding model.

Let's consider the case where each basic model is a decision tree. For this method, the parameters $\mathbf{a}_m$ are the splitting variables, split locations and the end node means of the individual trees. In this case, each decision tree has an additive form:

$$h\left(\mathbf{x}; \{a_j, R_j\}_1^J\right) = \sum_{j=1}^{J} a_j 1(\mathbf{x} \in R_j), \quad (2)$$

where $\{R_j\}_1^J$ is the space of the J end nodes of the decision tree, which completely covers the range of values of the independent variables $\mathbf{x}$, $1(\cdot)$ is an indicator function that has the value 1 if its argument is true and 0 otherwise, $\{a_j\}_1^J$ are the parameters of the model which define the boundaries of spaces $\{R_j\}_1^J$, which in its turn represent the distributions of non-end nodes of the tree.

For the decision tree, the definition of the boosting algorithm have the form:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + p_m \sum_{j=1}^{J} a_{jm} 1(\mathbf{x} \in R_j), \quad (3)$$

where $\{R_{jm}\}_1^J$ – spaces which are defined by the end nodes $j = \overline{1,J}$ of the decision tree during the iteration m (for new decision tree model), p_m is the scaling factor.

Formula (3) can be reduced to

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} 1(x \in R_j), \quad (4)$$

where $\gamma_{jm} = p_m a_{jm}$.

The boosting algorithm is described in [30]. Gradient boosting of decision trees creates competitive, reliable, interpretable models for solving classification problems, and good results are achieved even in conditions of low quality of input data.

For program implementation of gradient boosting of decision trees we will use the program packages XGBoost and LGBM.

XGBoost is an open source software library which supports the gradient boosting algorithm for C++, Java, Python, R, Scala, and Julia programming languages, created in 2014. The ability to use different programming languages has expanded the circle of developers and brought XGBoost popularity among the Kaggle community. The work on XGBoost was published by the library's authors T. Chen and C. Guestrin [26].

LightGBM (LGBM) is a framework for gradient boosting that uses training algorithms for decision trees. LightGBM is a more modern optimized software implementation of the algorithm for greedy approximation of functions using decision trees, which characterized by greater speed of model learning and their higher accuracy [31].

Both XGBoost and LGBM packages were used for creation of the optimal architecture of the model. As both packages implement similar

algorithms, accuracy of the achieved result is also the same. The main difference between the packages is technical implementation – as LGBM is optimized in C++, its training speed is significantly higher in comparison with XGBoost. As a result, for estimation of the efficiency of the experiment with a control subset was used only LGBM package.

The method of neural networks

In addition to boosted decision trees, this study used the method of neural networks for modelling the debtor creditworthiness. The architecture of the neural network was chosen as a result of an experimental study. Thus, a feed-forward neural net with hyperbolic tangent activation function with dropout was selected.

The neural network architecture included 2 hidden neuron layers, 100 neuron each. Hidden layer dropout ratios were set at 5% for each layer for better model generalization. Neural net training has run for 1000 iterations, with enabled early stopping rounds option, set to 20 iterations based on the AUC metric improvement. For program implementation of this neural network was used H2O artificial intelligence package for R programming language [32].

Collecting data for the study

To solve the stated problem, devoted to the study of influence of the control sample formation procedure on improving the quality of binary classification, we will use the open data of Home Credit international consumer finance provider, uploaded for the Home Credit Default Risk competition [33] to the Kaggle platform. This platform contains the largest database of contests for analytics and predictive modeling, in which statisticians and data mining specialists compete to create the best models for predicting and describing data offered by companies or users.

Within the framework of the Home Credit Default Risk competition, Home Credit provided data on loan applications for 2 retail loan products: consumer loans and credit cards. The specifics of the

incoming sample was the selection of a population of unbankable customers whose loan applications would be denied under a one of the existing credit rules, but were financed with a credit by Home Credit to improve existing decision-making models and expand the coverage of potential borrowers.

Given that Home Credit selected a sample of customers with low credit ratings to ensure sufficient predictive power of analytical models, additional data sources were provided for the competition, such as:

1) detailed behavioral information on the balance of existing and previous loans of the client and his payments according to 3 external credit bureaus (EXT_SRC) and internal data of the Home Credit;

2) information from real estate registers on the condition and average values of factors that characterize real estate owned by the client;

3) assessment of the client's region of residence.

Description of the database structure is shown in Fig. 2.

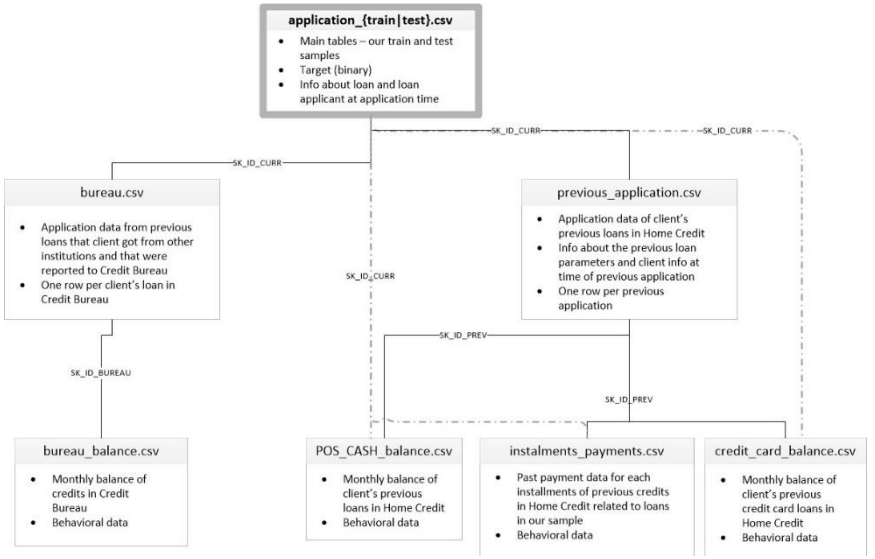

Fig. 2. The structure of the input data of the Home Credit Default Risk competition [33]

The training sample provided by Home Credit finance provider for the construction of the scoring models included 307 511 observations (see Table 1).

Table 1

THE STRUCTURE OF THE TRAINING SAMPLE OF HOME CREDIT COMPETITION

Type of credit	Number of observations	% observations
Credit repaid	282 686	91.93%
Default	24 825	8.07%
Grand Total	**307 511**	**100.00%**

As can be seen from Table 1, in the database for the Home Credit competition, the above-mentioned typical problem of imbalance in the classes of borrowers is observed (the "Default" class is much less common than the "Loan repaid" class). Thus, this dataset is appropriate for comparative analysis of different sampling procedures to improve the quality of binary classification in the problem of credit risk modeling.

The test sample provided by Home Credit to validate the constructed models included 48 744 observations.

The first stage in the preparation of data set for building predicative models was the initial data processing and the development of predictors based on them. The total number of predictors that were included in the final data set is 591, among which were both initial indicators:

- Income per Person – the borrower's income per 1 member of his family;
- Children ratio – the ratio of the number of children in the borrower's family to the total number of members of his family;
- Credit to Goods ratio – the ratio of the loan amount to the value of goods purchased in credit;
- DPD – days past due;
- DBD – days before due;

- Loan to Income ratio – the ratio of the loan amount to the borrower's income, etc.,

and auxiliary predictors calculated on their basis: ratios and other combinations of initial factors, average value, minimum, maximum, amount, volatility for different periods, number of unique records and others.

LGBM and Neural Network modeling using different sampling methods

During the research, an experiment was conducted on the use various approaches to form a control sample to improve the efficiency of economic and mathematical models on the basis of tools of boosted decision trees and neural networks.

For the experiment, 3 methods of forming a control sample were used:

1) random selection;

2) stratified selection by dependent variable;

3) stratified selection by dependent variable and the most significant variables of the model from the group EXT_SRC.

Note that the 3rd method is proposed by the authors of the paper. It represents a departure from conventional method by linking sampling not only with the classes of the dependent variable, but also with the characteristics of the clusters within the training dataset itself. Since the dependent variable is binary (it only takes the value 0 or 1), it makes sense to add additional parameters to the stratification. It is suggested to choose the most significant independent variables of the model as the extension of standard method. As a result, more specific sampling groups are formed, providing higher level of accuracy by being able to capture the inherent structure of the data in a more sophisticated way.

A common models' architecture and input predictors were used for all methods. For every selection method, 10% of a training sample (Table 1) was used in a control subset. As a result of 5 iterations for each of the types of selection, the following results of testing LGBM and neural network models on the control sample were obtained (Tables 2-4).

Table 2

RANDOM SELECTION

Model type	№ of model, AUC					Variance ∂^2
	1	2	3	4	5	
LGBM	0.7963	0.7902	0.7860	0.7873	0.7751	$6.02 * 10^{-5}$
NN	0.6683	0.6880	0.6672	0.6629	0.6595	$12.34 * 10^{-5}$

Table 3

STRATIFIED SELECTION BY DEPENDENT VARIABLE

Model type	№ of model, AUC					Variance ∂^2
	1	2	3	4	5	
LGBM	0.7952	0.7904	0.7896	0.7850	0.7908	$1.31 * 10^{-5}$
NN	0.6747	0.6695	0.6789	0.6757	0.6807	$1.85 * 10^{-5}$

Table 4

STRATIFIED SELECTION BY DEPENDENT VARIABLE AND VARIABLES EXT_SRC

Model type	№ of model, AUC					Variance ∂^2
	1	2	3	4	5	
LGBM	0.7934	0.7911	0.7871	0.7895	0.7877	$0.66 * 10^{-5}$
NN	0.6740	0.6721	0.6771	0.6819	0.6732	$1.55 * 10^{-5}$

As can be seen from Tables 2-4, neural nets showed systematically less accurate results, than decision trees. Such behavior is linked to the fact, that during solving the binary classification tasks based on neural network approach all variables (quantitative, and qualitative as well) needed to be transformed using a binning approach. This significantly increases its efficiency [34].

Note that decision trees effectively work based on as-is input information, without its additional transformation. Taking into account that the main purpose of the paper is not to obtain the most accurate creditworthiness model, but a research of an impact of a stratified sampling procedure on an increase of a classification quality, binning hasn't been applied. Accordingly, all conclusions in the context

of the current task for neural networks remain valid, the same as for decision trees.

Based on the obtained results, is possible to conclude that the use of stratified selection by indicators that have the greatest impact on the model, reduces the variance of model errors for the control sample. This result makes it possible to increase the stability of the model calculation, which is useful when choosing an architecture.

The next stage of the experiment is to compare the predicative power of models, built using different control sample selection methods, on independent data set. To assess the predictive ability, a test sample was used, which, according to the rules of the competition, was not available to researchers. Such an assessment of the prediction accuracy of the target variable was carried out on the side of the Kaggle system. The result of the constructed models' application for the test sample is shown in Table 5. Here, the AUC values for both the control and test subsets are calculated as an average of the accuracy of the models presented in Tables 2-4.

Table 5

EFFICIENCY OF MODELS WITH DIFFERENT CONTROL SAMPLE SELECTION METHODS ON THE TEST SAMPLE

Type of control set selection	Result (AUC)			
	Boosted decision tree		Neural net	
	Control subset	Test subset	Control subset	Test subset
Random selection	0.7870	0.7907	0.6692	0.6139
Stratified selection by dependent variable	0.7902	0.7916	0.6759	0.6182
Stratified selection by dependent variable and variables EXT_SRC	0.7898	0.7935	0.6757	0.6165

As can be seen from Table 5, on both control and test subsets, the predictive power of both types of models increases if the general totality was stratified when creating the control sample. In addition, the standard approach to stratified selection only on the dependent variable and the method proposed by the authors of selection on the dependent variable

and the most significant predictors demonstrate approximately equal accuracy rates. Although for boosted decision trees on the test subset, the authors' method demonstrated even higher average accuracy.

Conclusion

In solving the problem of assessing the creditworthiness of borrowers (which belong to the class of binary classification tasks), artificial intelligence and machine learning methods are widely used, in particular, Boosted decision trees and neural networks.

In the course of constructing the best model for a given classification task, there is a need to ensure a stable result of its work, and accordingly, it becomes necessary to eliminate fluctuations in the distribution of features in the control sample compared to the main data set. To solve the described task, it is proposed to improve the classical procedure of stratified sampling by taking into account not only the target variable, but also the most significant predictors of the model when forming the control subset.

According to the results of the study, it can be concluded that:

1) additional stratification during the selection of the control sample positively affects the predicative power of both boosted decision tree and neural net models by maintaining the heterogeneity of the overall data set in the control sample;

2) in addition to the positive effect on predicative power, the use of stratified selection by the most significant indicators of the models, led to a decrease in the variance of the results of calculations of different models on the control sample (which indicates an increase in their stability).

Thus, the use of stratified sampling in the process of forming a control sample during the training both boosted decision tree models and neural networks improves the stability of the model, which increases the efficiency of the process of choosing its architecture to improve the accuracy of forecasts on new data sets. The obtained findings are valid not only for creditworthiness estimation tasks – it's expected that a positive effect from the use of sampling will be achieved if it is used for any classification task.

References

1. Harris, R. (2016, December 23). More data will be created in 2017 than the previous 5,000 years of humanity. *App Developer Magazine*. https://appdevelopermagazine.com/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/

2. Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance, 26*(3), 382-405. https://doi.org/10.1108/JFRC-06-2017-0054

3. Desai, V.S., Crook, J.N., & Overstreet, G.A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37. https://doi.org/10.1016/0377-2217(95)00246-4

4. Desai, V. S., Conway, D. G., Crook, J. N., & Overstreet, G. A. (1997). Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics, 8*(4), 323-346. https://doi.org/10.1093/imaman/8.4.323

5. Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics, 15*(1), 107–143. https://doi.org/10.1023/A:1008699112516

6. West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11-12), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

7. Chuang, Ch.-L., & Huang, S.-T. (2011). A hybrid neural network approach for credit scoring. *Expert Systems, 28*(2), 185-196. https://doi.org/10.1111/j.1468-0394.2010.00565.x

8. Hryhorovych, O. (2019). Application of multilayer perceptrons to legal entities borrowers classification. *Neiro-Nechitki Tekhnolohii Modelyuvannya v Ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics), 8*, 48-64. http://doi.org/10.33111/nfmte.2019.048 [in Ukrainian]

9. Munkhdalai, L., Lee, J.Y., & Ryu, K.H. (2020). A Hybrid Credit Scoring Model Using Neural Networks and Logistic Regression. In J.S. Pan, J. Li, P.W. Tsai, & L. Jain (Eds.), *Smart Innovation, Systems and Technologies: Vol. 156. Advances in Intelligent Information Hiding and Multimedia Signal Processing* (pp. 251–258). Springer. https://doi.org/10.1007/978-981-13-9714-1_27

10. Kocadağlı, O., & Soydaner, D. (2015). Artificial Neural Networks with Gradient Learning Algorithm for Credit Scoring. *Istanbul University*

Journal of the School of Business, 44(2), 3-12. https://dergipark.org.tr/en/pub/iuisletme/issue/9259/115847

11. Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access*, *7*, 2161-2168. https://doi.org/10.1109/ACCESS.2018.2887138

12. Akkoç, S. (2019). Exploring the nature of credit scoring: a neuro fuzzy approach. *Fuzzy Economic Review*, *24*(1), 3–24. http://doi.org/10.25102/fer.2019.01.01

13. Doskocil, R. (2017). Evaluating the Creditworthiness of a Client in the Insurance Industry Using Adaptive Neuro-Fuzzy Inference System. *Engineering Economics, 28*(1), 15-24. https://doi.org/10.5755/j01.ee.28.1.14194

14. Mehdiyev, N. (2020). Application of Fuzzy TOPSIS for Credit Scoring. In C. Kahraman, S. Cebi, S. Cevik Onar, B. Oztaysi, A. Tolga, & I. Sari. (Eds.), *Advances in Intelligent Systems and Computing: Vol. 1029. Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making* (pp. 779–786). Springer. https://doi.org/10.1007/978-3-030-23756-1_93

15. Sohn, S., Kim, D.-H., & Yoon, J. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing, 43*, 150-158. https://doi.org/10.1016/j.asoc.2016.02.025

16. Eghbali, A., Razavi Hajiagha, S. H., & Amoozad, H. (2017). Performance Comparison of Genetic Algorithm Fitness Function in Customer Credit Scoring. *Industrial Management Journal*, *9*(2), 245-264. http://doi.org/10.22059/imj.2017.226860.1007191

17. Kozeny, V. (2015). Genetic algorithms for credit scoring: alternative fitness function performance comparison. *Expert Systems with Applications, 42*(6), 2998-3004. https://doi.org/10.1016/j.eswa.2014.11.028

18. Frydman, H., & Matuszyk, A. (2020). Random survival forest for competing credit risks. *Journal of the Operational Research Society, 73*(1), 15-25. http://doi.org/10.1080/01605682.2020.1759385

19. Rudra Kumar, M., & Kumar Gunjan, V. (2020). Review of Machine Learning models for Credit Scoring Analysis. *Ingeniería Solidaria*, *16*(1), Article 11. https://doi.org/10.16925/2357-6014.2020.01.11

20. Veeramanikandan, V., & Jeyakarthic, M. (2019). An ensemble model of outlier detection with random tree data classification for financial credit scoring prediction system. *International Journal of Recent Technology and Engineering, 8*(3), 7108-7114. http://doi.org/10.35940/ijrte.C5850.098319

21. Bramer, M. (2020). Avoiding Overfitting of Decision Trees. In *Principles of Data Mining* (4th ed., pp. 121–136). Springer. https://doi.org/10.1007/978-1-4471-7493-6_9

22. Caruana, R., Lawrence, S., & Giles, L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In T.K. Leen, T.G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 402-408), MIT Press.

23. Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series, 1168*(2), Article 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

24. Velykoivanenko, H., Korchynskyi, V., & Chernyshova, V. (2016). Study of the neural networks overfitting effect on the example of the problem of application scoring. *Neiro-Nechitki Tekhnolohii Modelyuvannya v Ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics), 5*, 3-23. https://doi.org/10.33111/nfmte.2016.003 [in Ukrainian]

25. Savina, S., & Ben, V. (2016). Selection of neural network architecture for solving problem of borrowers-individuals trustability classification. *Neiro-Nechitki Tekhnolohii Modelyuvannya v Ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics), 5*, 123–151. https://doi.org/10.33111/nfmte.2016.123 [in Ukrainian]

26. Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.48550/arXiv.1603.02754

27. Menardi, G., Tedeschi, F., & Torelli, N. (2011). On the Use of Boosting Procedures to Predict the Risk of Default. In B. Fichet, D. Piccolo, R. Verde, & M. Vichi (Eds.), *Classification and Multivariate Analysis for Complex Data Structures* (pp. 211–218). Springer. https://doi.org/10.1007/978-3-642-13312-1_21

28. Thompson, S. (2012). Stratified Sampling. In *Sampling* (3rd ed., pp. 139-156). John Wiley & Sons. https://doi.org/10.1002/9781118162934.ch11

29. Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558-625. https://www.jstor.org/stable/2342192

30. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

31. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3149-3157). Neural Information Processing Systems Foundation. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

32. H2O. (2016). Welcome to H2O 3. R Users. *UpToDate*. Retrieved February 20, 2020, from https://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html#r-users

33. Kaggle. (2020). *Home Credit Default Risk. Dataset Description* [Data set]. Retrieved Fabruary 10, 2020, from https://www.kaggle.com/c/home-credit-default-risk/data

34. Kleban, Y. (2019). Studying the methods of data transformation in the context of increasing the effectiveness of credit scoring models. *Neiro-Nechitki Tekhnolohii Modelyuvannya v Ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics)*, *8*, 94—123. https://doi.org/10.33111/nfmte.2019.094 [in Ukrainian]

Стаття надійшла до редакції 01.07.2020