International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 5, October 2024, pp. 5534~5542
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i5.pp5534-5542

Forecasting creditworthiness in credit scoring using machine learning methods

Ayagoz Mukhanova[1], Madiyar Baitemirov[1], Azamat Amirov[2], Bolat Tassuov[3], Valentina Makhatova[4], Assemgul Kaipova[5], Ulzhan Makhazhanova[1], Tleugaisha Ospanova[1]
[1]Department of Information Systems, L. N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
[2]Digitalization Department, Abylkas Saginov Karaganda Technical University, Karaganda, Republic of Kazakhstan
[3]Faculty of Natural Sciences, Non-profit Limited Liability Company, M.H. Dulaty Taraz State University, Taraz, Republic of Kazakhstan
[4]Department of Software Engineering, Atyrau State University Kh. Dosmukhamedova, Atyrau, Republic of Kazakhstan
[5]Department of Biostatistics, Bioinformatics and Information Technologies, Astana Medical University, Astana, Republic of Kazakhstan

Article Info

Article history:

Received Feb 27, 2024
Revised Jun 20, 2024
Accepted Jul 1, 2024

Keywords:

Creditworthiness
Decision tree classifier
Gradient boosting classifier
Linear discriminant analysis
Logistic regression
Machine learning

ABSTRACT

This article provides an overview of modern machine learning methods in the context of their active use in credit scoring, with particular attention to the following algorithms: light gradient boosting machine (LGBM) classifier, logistic regression (LR), linear discriminant analysis (LDA), decision tree (DT) classifier, gradient boosting classifier and extreme gradient boosting (XGB) classifier. Each of the methods mentioned is subject to careful analysis to evaluate their applicability and effectiveness in predicting credit risk. The article examines the advantages and limitations of each method, identifying their impact on the accuracy and reliability of borrower creditworthiness assessments. Current trends in machine learning and credit scoring are also covered, warning of challenges and discussing prospects. The analysis highlights the significant contributions of methods such as LGBM classifier, LR, LDA, DT classifier, gradient boosting classifier and XGB classifier to the development of modern credit scoring practices, highlighting their potential for improving the accuracy and reliability of borrower creditworthiness forecasts in the financial services industry. Additionally, the article discusses the importance of careful selection of machine learning models and the need to continually update methodology in light of the rapidly changing nature of the financial market.

Corresponding Author:

Madiyar Baitemirov
Department of Information Systems, L. N. Gumilyov Eurasian National University
010000 Astana, Republic of Kazakhstan
Email: madiyar.baytemirov@inbox.ru

1. INTRODUCTION

In the modern world of finance, where competition in the lending market [1]–[3] is constantly growing, the relevance of developing effective methods for predicting creditworthiness is high. The accuracy and reliability of such methods are key factors for financial institutions seeking to minimize risk and ensure the sustainability of their loan portfolios. Light gradient boosting machine (LGBM) classifier [4]–[6], logistic regression [7]–[9], linear discriminant analysis [10]–[13], decision tree classifier [13], [14], gradient boosting classifier [15]–[17] and extreme gradient boosting (XGB) classifier [18], [19] are a variety of machine training, each with its own unique characteristics and applications. Their use in credit scoring [20]–[22] opens up new opportunities for improving the accuracy of forecasts, especially when working with large volumes of data and complex credit models. In this study, we will also address issues of model interpretability, computational complexity, and potential limitations. These aspects play an important role in the implementation of research results in real financial practices. Research into creditworthiness using machine learning methods is not only of academic interest, but also has direct application value for financial institutions, insurance companies and other market players. It is expected that the results of this study can serve as a basis for optimizing decision-making processes in the lending industry and provide more effective risk management.

Koc *et al.* [23] explore the role of credit ratings in assessing financial stability and the criteria for issuing a loan. They review eight machine learning methods, including support vector machines (SVM), Gaussian naive Bayes, decision trees (DT), random forest (RF), XGB, k-nearest neighbors (KNN), multi-layer perceptron (MLP), and logistic regression (LR). The main objective of the study is to demonstrate the beneficial application of these methods for predicting loan default risk and identifying influencing factors. The paper provides an extensive comparison evaluating which machine learning models perform better with and without their own feature selection method. Jiang *et al.* [24] explore the problem of credit scoring with a focus on identifying anomalies and maintaining order in financial transactions. They highlight the class imbalance problem that arises from the limited number of default records in financial data. To address this problem, the authors analyze various classical approaches to learning from imbalanced data, including resampling methods, cost-of-error strategies, and the use of generative adversarial networks (GANs) as a tool for learning from imbalanced data.

Abdoli *et al.* [25] examine the importance of automated credit scoring as a risk management tool for banks and financial institutions, noting its attractiveness in recent decades. They highlight that the unbalanced nature of credit scoring datasets, as well as feature heterogeneity, pose challenges to developing efficient models that can generalize to previously unseen data. The paper proposes the bagging supervised autoencoder classifier (BSAC), a model that combines the benefits of supervised learning for autoencoders and a bagging mechanism to handle heterogeneities in feature space, and the results of extensive experiments confirm the superiority and robustness of the proposed method in predicting the outcome of loan applications.

The relevance of the topic of predicting the creditworthiness of credit scoring using machine learning methods cannot be overestimated in the light of modern challenges in the financial industry. With the increasing volume of data and the variety of factors affecting the financial situation of borrowers, standard methods of assessing creditworthiness are not effective enough. The use of advanced machine learning methods provides the opportunity not only for more accurate forecasting, but also for deeper data analysis, which in turn helps to identify early signs of financial risks. Solutions based on LGBM classifier, LR, LDA, DT classifier, gradient boosting classifier and XGB classifier promise improved credit scoring results, which are critical to ensuring the sustainability of financial institutions and reducing the likelihood of financial crises.

2. METHOD

The purpose of this study is a comparative analysis of machine learning methods for predicting creditworthiness in credit scoring. We set ourselves the task of determining the optimal method among LGBM classifier, LR, LDA, DT classifier, gradient boosting classifier and XGB classifier, as well as evaluating their effectiveness based on standard classification metrics. As a basis for our research, we used a large and diverse data set that included information about borrowers' financial situation, credit history, social factors and other relevant variables. This dataset provides us with the opportunity to more comprehensively analyze and evaluate the proposed methods. Before applying machine learning methods, careful data preprocessing was carried out, including handling missing values, encoding categorical features, normalizing numerical data, and processing outliers. This stage allows you to ensure the correctness and stability of the models.

LGBM is a gradient amplification method optimized for efficient work with large volumes of data. This method draws attention to taking into account unbalanced classes, which is an important aspect in credit scoring problems. Logistic regression is a classic binary classification method based on a logistic function. We use it in the context of credit scoring to assess the likelihood of a borrower's creditworthiness and make a decision based on that likelihood. LDA is a linear discriminant analysis method designed to maximize differences between classes. In credit scoring, this method can be effective for highlighting key features that affect creditworthiness. Decision tree classifiers provide a visual representation of decision making and are capable of capturing complex relationships in data. We use this method to identify the structure of criteria that influence the forecasting of creditworthiness. The gradient boosting classifier allows the construction of ensembles of decision trees, which can improve the predictive power of the model. We'll explore its use in credit scoring and evaluate how it handles complex data structures. XGB is a gradient boosting implementation that provides additional optimizations and regularizations. We will look at its impact on the accuracy of credit forecasts.

To objectively compare machine learning methods, we used standard metrics such as accuracy, recall, precision and F1-measure. These metrics evaluate both the overall performance of the models and their ability to correctly identify borrowers with problematic credit histories. We conducted a series of experiments, training each model on the training set and testing on the test set. The results are analyzed using evaluation metrics to identify the best methods that can effectively solve the problem of creditworthiness forecasting.

3. RESULTS AND DISCUSSION

To build the models, we used the home credit dataset from *kaggle.com*, containing 65 columns. The data set includes a column called TARGET, which represents the target variable (1 - customer with payment difficulties: he was late in payment, 0 - all other cases). Column names and descriptions are given in Table 1 (see in Appendix).

In experiments with credit prediction in a binary classification task, we applied a common thresholding method to convert predicted probabilities (y_{pred}) into binary class labels. This process is carried out using a threshold set at 0.5: probabilities equal to or greater than 0.5 are rounded to 1 (positive class), while probabilities below 0.5 are rounded to 0 (negative class). This approach allows us to obtain clear categories of object membership in classes, which simplifies the interpretation of classification results and prevents ambiguities associated with threshold values. In the considered methods (LGBM classifier, LR, LDA, DT classifier, gradient boosting classifier and XGB classifier), the key metrics for each of them were evaluated after applying a set threshold, as shown in Figure 1(a) to (f), where the results are present: ROC_AUC, accuracy, precision, recall, specificity and F1-score. These results reflect the performance of each method after applying a standard threshold of 0.5. This probability rounding technique plays an important role in the construction of binary classification models, ensuring their interpretability and applicability in various subject areas, including credit scoring.

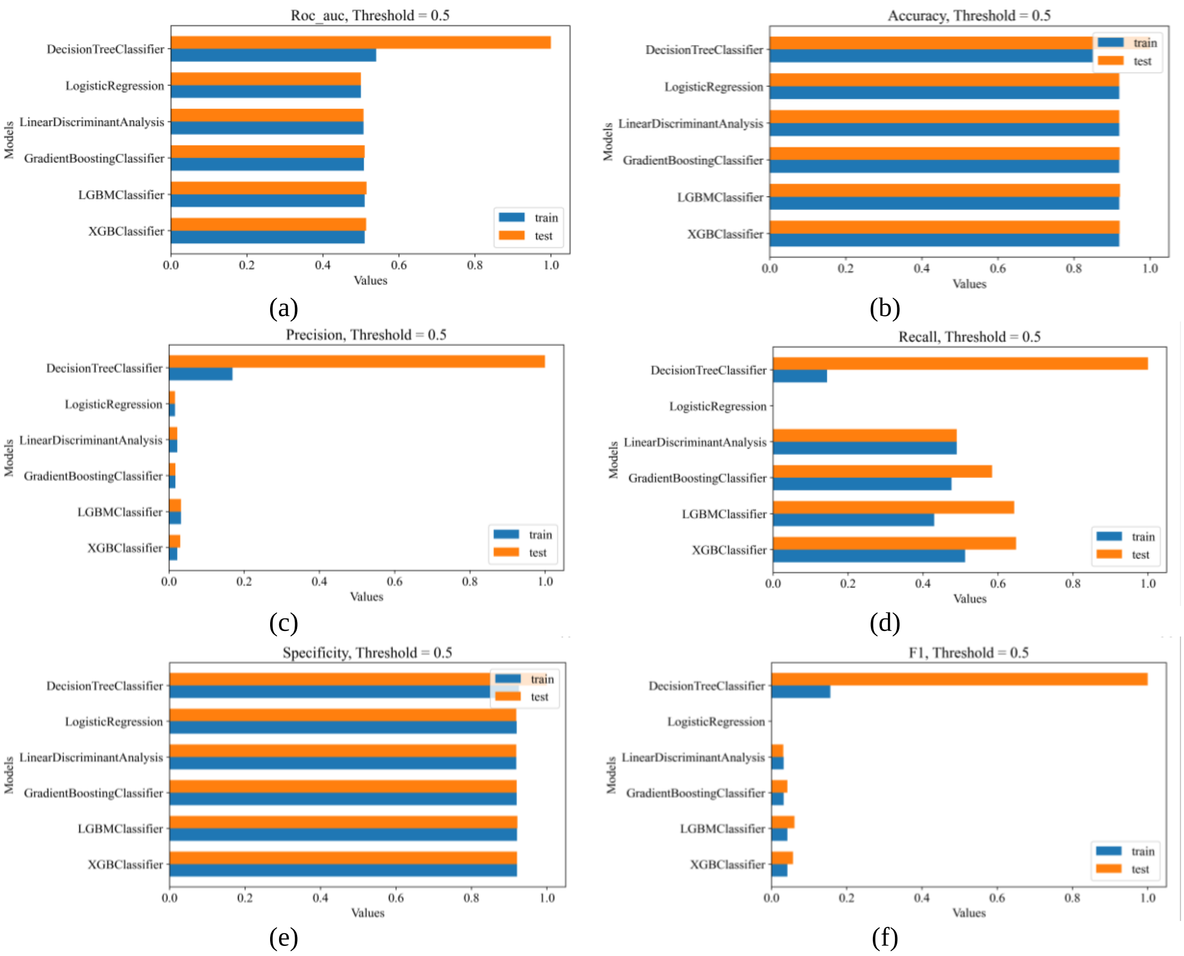

Figure 1. Metric results after adjustment by methods: (a) ROC_AUC, (b) accuracy, (c) precision, (d) recall, (e) specificity, and (f) F1-score

Based on the analysis of the graphs in Figure 1, we can conclude that negative classes were correctly predicted in most cases, which is confirmed by high specificity values. However, the need to adjust the decision threshold is an integral step in optimizing models, especially when balancing between false positives and false negatives. In this context, it was decided to conduct experiments with different threshold values and evaluate their impact on key metrics such as precision (Precision), recall (Recall) and F1-measure. Using different thresholds for classification allows you to tune the sensitivity of the model to specific classes in accordance with the requirements of the application domain. This is especially important in the context of credit scoring, where the weight of various types of errors can be critical. By finding the optimal threshold, a balance can be achieved between minimizing false positives and false negatives, which in turn will improve the quality of the model's predictions, as shown in Figure 2(a) to (f), which presents metric results after adjustment by methods: ROC_AUC, accuracy, precision, recall, specificity, and F1-score.

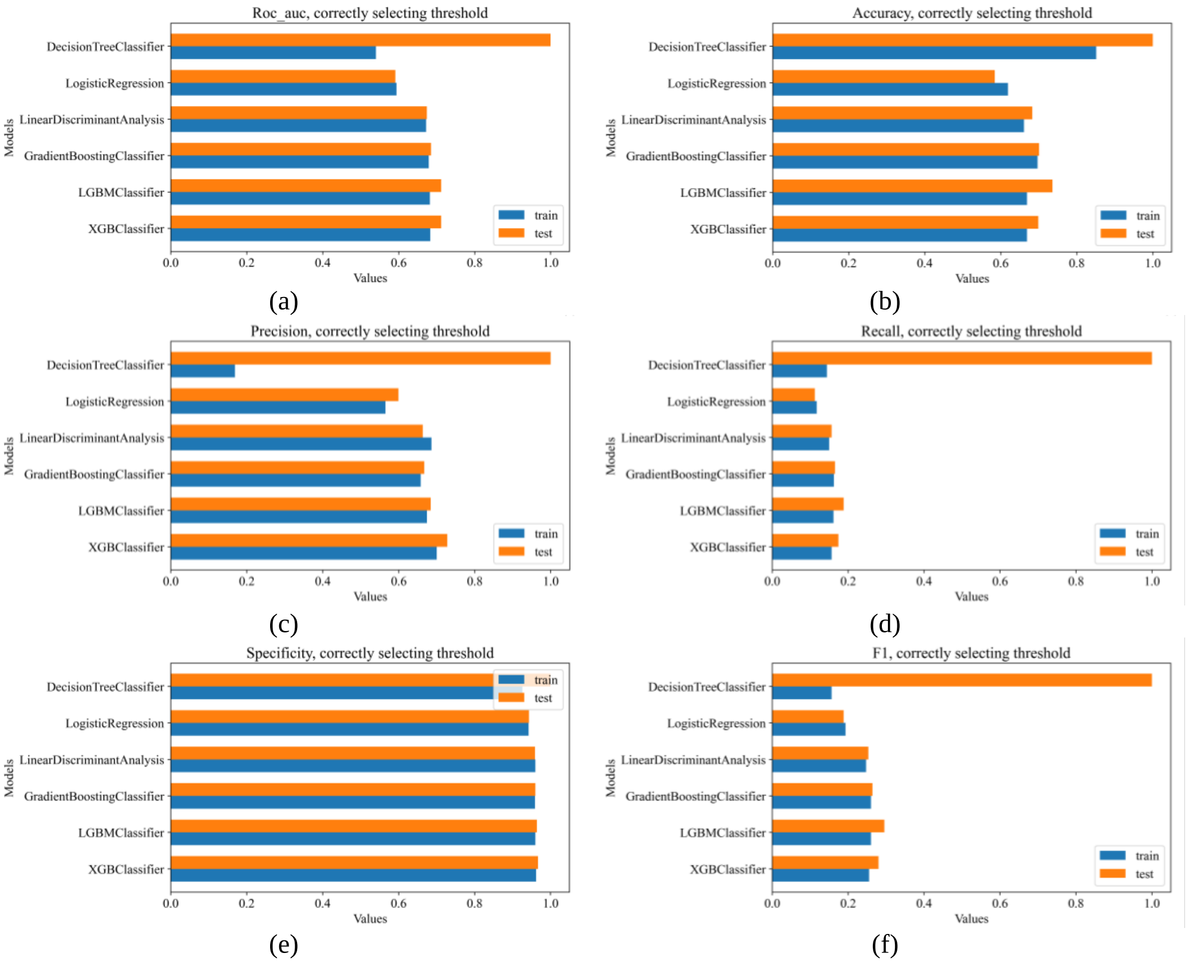

Figure 2. Metric results after adjustment by methods: (a) ROC_AUC, (b) accuracy, (c) precision, (d) recall, (e) specificity, and (f) F1-score

Experiments with different decision thresholds provide additional exploratory evidence, expanding our understanding of the model's sensitivity to different levels of decision confidence. Analysis of the confusion matrix in Figure 3, or error matrix, based on different thresholds allows you to look in more detail at the impact of changing the threshold on the quality of classification. This approach is important for refining the model settings in accordance with the specific requirements and preferences of the business. The results obtained can serve as the basis for more accurate and flexible adjustment of the model within the framework of credit scoring requirements.

In the figures, the metrics of the decision tree classifier model for the training and test data sets remained unchanged. This is because the tree returns predictions not as probabilities, but as integers. The accuracy of the models is high in the original tables with a threshold of 0.5, since in the data under study the number of one class significantly exceeds the number of another. Models are good at predicting bad customer data, but bad at predicting good ones. Therefore, it is advisable to rely on other indicators. After adjusting the threshold, the main metrics increased significantly for all models except the decision tree, indicating the positive impact of choosing the right thresholds.

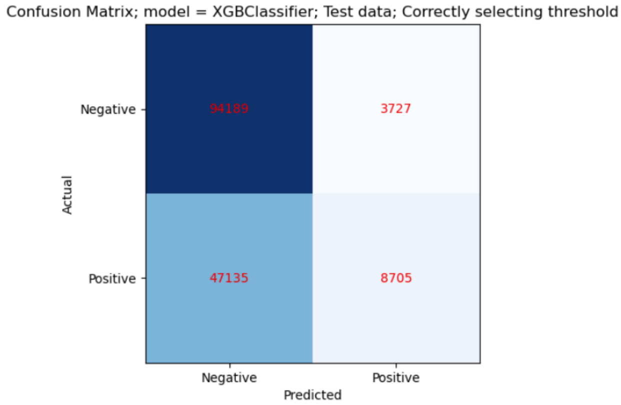

Figure 3. Confusion matrix of XGBoost method

4. CONCLUSION

In this study, analyzing models in the context of credit scoring, six different classification algorithms were examined. Each of these models has demonstrated its unique characteristics and performance, enriching our understanding of their applicability in credit forecasting tasks. The results highlight the outstanding performance of XGB classifier, which stands out among other models in all metrics reviewed, including ROC_AUC, accuracy, F1-score, and specificity. This demonstrates the high performance of XGB classifier in the context of credit scoring and its ability to predict creditworthiness with a high degree of accuracy. Additional research into the variation of decision thresholds when rounding class membership probabilities revealed a significant improvement in the predictive ability of the models. Analyzing metrics such as ROC_AUC, precision, recall, specificity, and F1-score at different thresholds highlights the importance of finding a trade-off between false positives and false negatives. Overall, the results of our study not only enrich the understanding of the performance of various models in credit scoring, but also highlight the importance of careful calibration and selection of optimal thresholds to improve forecasting performance. These findings provide valuable guidance for decision making in the field of credit scoring and in the context of financial risk management.

APPENDIX

Table 1. Data set with descriptions

Column name	Describe
EXT_SOURCE_3	Normalized score from external data source
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
DAYS_BIRTH	Client's age in days at the time of application
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
OWN_CAR_AGE	Age of client's car
DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_REGISTRATION	How many days before the application did client change his registration
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
AMT_INCOME_TOTAL	Income of the client
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
ENTRANCES_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

Table 1. Data set with descriptions (*continue*)

Column name	Describe
AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)
YEARS_BUILD_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix)*, modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
LIVINGAREA_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix)*, modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
OBS_30_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings with observable 30 DPD (days past due) default
FLOORSMAX_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
CNT_FAM_MEMBERS	How many family members does client have
YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
FLOORSMIN_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
OBS_60_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings with observable 60 DPD (days past due) default
BASEMENTAREA_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

Table 1. Data set with descriptions (*continue*)

Column name	Describe
YEARS_BUILD_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
DEF_60_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings defaulted on 60 (days past due) DPD
AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
TOTALAREA_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
DEF_30_CNT_SOCIAL_CIRCLE	How many observations of client's social surroundings defaulted on 30 DPD (days past due)
YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
FLAG_DOCUMENT_18	Did client provide document 18
FLAG_DOCUMENT_16	Did client provide document 16
FLAG_DOCUMENT_8	Did client provide document 8
FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLAG_DOCUMENT_3	Did client provide document 3
FLAG_DOCUMENT_6	Did client provide document 6
FLAG_DOCUMENT_14	Did client provide document 14
APARTMENTS_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MEDI	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MODE	Normalized information about building where the client lives, what is average (*_AVG suffix*), modus (*_MODE suffix*), median (*_MEDI suffix*) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)

ACKNOWLEDGEMENTS

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19677451).

REFERENCES

[1] Z. He, J. Huang, and J. Zhou, "Open banking: credit market competition when borrowers own the data," *Journal of Financial Economics*, vol. 147, no. 2, pp. 449–474, Feb. 2023, doi: 10.1016/j.jfineco.2022.12.003.
[2] U. T. Makhazhanova, F. A. Murzin, A. A. Mukhanova, and E. P. Abramov, "Fuzzy logic of Zadeh and decision-making in the field of loan," *Journal of theoretical and applied Information Technology*, vol. 98, no. 06, pp. 1076–1086, 2020.
[3] U. Makhazhanova *et al.*, "The evaluation of creditworthiness of trade and enterprises of service using the method based on fuzzy logic," *Applied Sciences*, vol. 12, no. 22, Nov. 2022, doi: 10.3390/app122211515.
[4] Y. Wang, Y. Liu, J. Zhao, and Q. Zhang, "Low-complexity fast CU classification decision method based on LGBM classifier," *Electronics*, vol. 12, no. 11, May 2023, doi: 10.3390/electronics12112488.
[5] İ. F. Kilinçer and O. Katar, "A new intrusion detection system for secured IoT/IIoT networks based on LGBM," *Gazi University Journal of Science Part C: Design and Technology*, 2023.
[6] T. Liu, X. Zhang, R. Chen, X. Deng, and B. Fu, "Development, comparison, and validation of four intelligent, practical machine learning models for patients with prostate-specific antigen in the gray zone," *Frontiers in Oncology*, vol. 13, Jun. 2023, doi: 10.3389/fonc.2023.1157384.
[7] J. Tussupov *et al.*, "Analysis of formal concepts for verification of pests and diseases of crops using machine learning methods," *IEEE Access*, vol. 12, pp. 19902–19910, 2024, doi: 10.1109/ACCESS.2024.3361046.
[8] W.-Y. Loh, "Logistic regression tree analysis," in *Springer Handbook of Engineering Statistics*, 2023, pp. 593–604.
[9] G. Troiano *et al.*, "Development and international validation of logistic regression and machine-learning models for the prediction of 10-year molar loss," *Journal of Clinical Periodontology*, vol. 50, no. 3, pp. 348–357, Mar. 2023, doi: 10.1111/jcpe.13739.
[10] R. Graf, M. Zeldovich, and S. Friedrich, "Comparing linear discriminant analysis and supervised learning algorithms for binary classification—a method comparison study," *Biometrical Journal*, vol. 66, no. 1, Jan. 2024, doi: 10.1002/bimj.202200098.
[11] T. Suesse, A. Brenning, and V. Grupp, "Spatial linear discriminant analysis approaches for remote-sensing classification," *Spatial Statistics*, vol. 57, Oct. 2023, doi: 10.1016/j.spasta.2023.100775.
[12] G. Singh, Y. Pal, and A. K. Dahiya, "Classification of power quality disturbances using linear discriminant analysis," *Applied Soft Computing*, vol. 138, May 2023, doi: 10.1016/j.asoc.2023.110181.
[13] G. Devisetty and N. S. Kumar, "Prediction of bradycardia using decision tree algorithm and comparing the accuracy with support vector machine," *E3S Web of Conferences*, vol. 399, Jul. 2023, doi: 10.1051/e3sconf/202339909004.
[14] H. Chen, G. Zhang, X. Pan, and R. Jia, "Using dual evolutionary search to construct decision tree based ensemble classifier," *Complex & Intelligent Systems*, vol. 9, no. 2, pp. 1327–1345, Apr. 2023, doi: 10.1007/s40747-022-00855-x.
[15] Abdullah-All-Tanvir, I. A. Khandokar, A. K. M. M. Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15163.
[16] R. Suhendra *et al.*, "Evaluation of gradient boosted classifier in atopic dermatitis severity score classification," *Heca Journal of Applied Sciences*, vol. 1, no. 2, pp. 54–61, Sep. 2023, doi: 10.60084/hjas.v1i2.85.
[17] H. Nhat-Duc and T. Van-Duc, "Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification," *Automation in Construction*, vol. 148, Apr. 2023, doi: 10.1016/j.autcon.2023.104767.
[18] V. Jain and M. Agrawal, "Heart failure prediction using XGB classifier, logistic regression and support vector classifier," in *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, May 2023, pp. 1–5, doi: 10.1109/InCACCT57535.2023.10141752.
[19] K. Konar, S. Das, and S. Das, "Employee attrition prediction for imbalanced data using genetic algorithm-based parameter optimization of XGB classifier," in *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, Jan. 2023, pp. 1–6, doi: 10.1109/ICCECE51049.2023.10085402.
[20] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: a systematic literature survey," *Applied Soft Computing*, vol. 91, Jun. 2020, doi: 10.1016/j.asoc.2020.106263.
[21] G. Teles, J. J. P. C. Rodrigues, K. Saleem, S. Kozlov, and R. A. L. Rabêlo, "Machine learning and decision support system on credit scoring," *Neural Computing and Applications*, vol. 32, no. 14, pp. 9809–9826, Jul. 2020, doi: 10.1007/s00521-019-04537-7.
[22] E. S. Kamimura, A. R. F. Pinto, and M. S. Nagano, "A recent review on optimisation methods applied to credit scoring models," *Journal of Economics, Finance and Administrative Science*, vol. 28, no. 56, pp. 352–371, 2023, doi: 10.1108/JEFAS-09-2021-0193.
[23] O. Koc, O. Ugur, and A. S. Kestel, "The impact of feature selection and transformation on machine learning methods in determining the credit scoring," *arXiv preprint arXiv:2303.05427*, 2023.
[24] C. Jiang, W. Lu, Z. Wang, and Y. Ding, "Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring," *Expert Systems with Applications*, vol. 213, Mar. 2023, doi: 10.1016/j.eswa.2022.118878.
[25] M. Abdoli, M. Akbari, and J. Shahrabi, "Bagging supervised autoencoder classifier for credit scoring," *Expert Systems with Applications*, vol. 213, Mar. 2023, doi: 10.1016/j.eswa.2022.118991.

BIOGRAPHIES OF AUTHORS

Ayagoz Mukhanova received her PhD in 2015 in information systems from L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, she is an associate professor of the Department of Information Systems at the same university. Her research interests include artificial intelligence and decision making. She can be contacted at email: ayagoz198302@mail.ru.

Madiyar Baitemirov in 2023, he graduated from the Eurasian National University named after L.N. Gumilyov with a degree in information systems. Currently, he is a master's student at the Eurasian National University. Scientific interests - machine learning, data science, data mining, and artificial intelligence. He can be contacted via email at: madiyar.baytemirov@inbox.ru.

Azamat Amirov in informatics, computing, and control systems from L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, he is the acting associate professor and director of the Digitalization Department at Abylkas Saginov Karaganda Technical University. His research interests include information technology, smart cities, big data, and the internet of things. He can be contacted at email: a.amirov@kstu.kz.

Bolat Tassuov in 1997 he graduated from Zhambyl University with a degree in mathematics and computer science. In 2006, he defended his dissertation and received the degree of Candidate of Technical Sciences. He started his career in 1998 as a software engineer at M.H. Dulaty Taraz State University. Currently, he is the dean of the Faculty of "Natural Sciences" of the Non-profit limited Liability Company "M.H. Dulaty Taraz State University". He is the author of more than 60 scientific papers, including 1 monograph, 3 textbooks and 2 articles in the Scopus database. Research interests – information security, computer modeling. He can be contacted at email: bolat_ktn@mail.ru.

Valentina Makhatova candidate of Technical Sciences, currently professor of the Department of Software Engineering at Atyrau State University Kh. Dosmukhamedova, Atyrau, Kazakhstan. She has more than 140 scientific papers, including 8 papers in Web of Science and Scopus rating publications, 3 monographs, 8 textbooks and 5 copyright certificates of intellectual property. H-index – 5. She was the executor of the project of search and initiative research work on the topic fundamental patterns of rheological properties of nanocomposite materials". Grant funding from the Ministry of Education and Science of the Republic of Kazakhstan 2018-2020. She can be contacted at email: mahve@mail.ru.

Assemgul Kaipova Master of Technical Sciences. She is currently a senior lecturer in the Department of Biostatistics, Bioinformatics and Information Technologies. She is the author of more than 10 scientific papers, 1 article in the Scopus database. She can be contacted by email: kaipova.a@amu.kz.

Ulzhan Makhazhanova in 2008 she graduated from Eurasian National University with a degree in computer science. In 2011 he received a master's degree in computer science. In 2021, she graduated from Eurasian National University doctoral studies with a specialty 6D070300 – information systems. From 2020 to the present, he is a senior lecturer at the Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana. She is the author of more than 22 works. Research interests: data analysis, big data, machine learning, fuzzy logic. She can be contacted at email: makhazhan.ut@gmail.com.

Tleugaisha Ospanova 1981 graduated from the Kazakh State University named after. S.M. Kirova, Faculty of Mechanics and Applied Mathematics, specialty applied mathematics. Candidate of Technical Sciences, specialty 13/05/2018 - mathematical modeling, numerical methods and software packages. Since 2011, she has been working at ENU named after. L.N. Gumileva, Department of Information Systems. Her research interest's information and computer technologies and systems, artificial intelligence. She can be contacted at email: Tleu2009@mail.ru.