SATBAYEV UNIVERSITY

Computing & Engineering

Volume 1 (2023), Issue 4, 6-11

https://doi.org/10.51301/ce.2023.i4.02

Research of existing machine learning methods for borrower credit scoring

L. Maralbayeva*

International Information Technology University, Almaty, Kazakhstan

**Corresponding author: maralbaevalaura@gmail.com*

Abstract. This research thoroughly explores how machine learning methods are used to evaluate the creditworthiness of borrowers, specifically those associated with second-tier banks. The main focus of the article is on using current scientific literature to explain the latest trends in credit scoring. The study gives a detailed overview of the credit process, pointing out important stages and factors that affect decision-making. The authors take a deep dive into various data sources used in scoring, explaining how they help make credit assessments more accurate and fairer. They analyze the strengths and weaknesses of different machine learning methods, figuring out how effective they are and if they suit second-tier banking. The article provides a detailed comparison of various machine learning methods, explaining where they work well and where they might have limitations. The research's importance lies in giving a broad view of machine learning methods, including recent updates and comparisons, acting as a starting point for future studies in this area. From a practical perspective, the article is useful for professionals in the banking sector by offering insights for the more effective use of modern machine learning methods in evaluating borrowers' creditworthiness. The study contributes significantly to understanding and applying contemporary approaches to credit scoring, offering valuable recommendations and practical strategies for those interested in this field.

Keywords: *credit scoring, scoring models, machine learning algorithms, logistic regression, decision trees, random forest, support vector machine, extreme gradient boosting.*

1. Introduction

In the contemporary financial landscape, the process of credit scoring plays a pivotal role in shaping the lending practices of banks, particularly second-tier institutions. As technological advancements continue to redefine the financial sector, the integration of machine learning methods into credit scoring has emerged as a critical area of research and application. This article seeks to provide a comprehensive overview of the existing machine learning methods in credit scoring, highlighting their significance, challenges, and potential for advancement.

The overarching problem addressed by this research is the dynamic nature of credit risk management, especially for second-tier banks. The evolving economic landscape and intricate borrower behaviors pose unprecedented challenges in accurately assessing creditworthiness. In an era marked by increasing financial complexities, unresolved issues persist in the formulation of robust credit scoring models that effectively navigate the intricate web of risks associated with lending.

The importance of addressing these challenges is underscored by the integral role second-tier banks play in fostering economic growth [1]. As crucial pillars of financial inclusion, these banks often grapple with unique credit risks that demand sophisticated methodologies for risk mitigation [2]. The need for precise credit scoring methods is paramount in enabling these institutions to strike a delicate balance between risk management and profitability [3].

Against this backdrop, the goal of this article is to conduct a meticulous analysis of the existing machine learning methods employed in credit scoring. By providing a comparative examination of these methods, the article aims to lay the groundwork for future research initiatives in this domain. The primary task is to offer insights into the strengths and weaknesses of different methods, facilitating a nuanced understanding that can inform subsequent studies and advancements in credit scoring methodologies.

Practically, the significance of this research extends to the specific context of credit scoring in Kazakhstan. As the country navigates its path towards economic development, second-tier banks stand to benefit significantly from the application of suitable machine learning methods tailored to the unique data and demands they encounter. A judicious choice of methods can enhance the precision of credit scoring models, mitigating risks, attracting more borrowers, and contributing to the overall economic prosperity of the nation.

In conclusion, this article serves as a foundational exploration into the realm of machine learning methods in credit scoring, with a distinct focus on the context of second-tier banks in Kazakhstan. Through a detailed analysis, it aspires to drive future research endeavors, offering practical insights that align with the evolving dynamics of the financial sector and the broader economic landscape.

1.1. Description of the credit scoring process using various data sources

The credit scoring process involves a systematic assessment of a borrower's creditworthiness using various financial and personal data. Financial institutions, such as banks or credit organizations, conduct this process to make informed decisions about loan approvals.

© 2023. L. Maralbayeva
https://ce.journal.satbayev.university/. Published by Satbayev University
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Here is a more detailed description of the key stages of the credit scoring process:

Information Gathering:

1. Personal Data: The borrower provides personal information, such as name, address, date of birth, and other identifying details.

2. Financial Information: This includes details about income, current obligations, savings, and other financial parameters.

Credit History:

1. Credit Reports: Financial institutions access credit reports from credit bureaus, including payment history, credit limits, and outstanding debts.

Credit Score Calculation:

1. Algorithms and Models: Using various algorithms and models, the bank calculates the borrower's credit score. This score assesses the likelihood of loan repayment and serves as a primary criterion for decision-making.

Decision Making:

1. Loan Terms: Based on the credit score and other borrower parameters, the bank determines the terms of the loan, including interest rates, duration, and loan amount.

2. Decision Making: The bank, relying on the obtained information and credit score, decides whether to approve the loan and, if so, under what conditions.

Communicating the Decision to the Borrower:

1. Decision Notification: The borrower is informed about the bank's decision. In the case of an approval, loan terms are provided, and in the case of a denial, the reasons are explained.

This process is based on a systematic analysis of a multitude of data and serves as a key tool for financial institutions to make reasoned decisions about loan approvals while minimizing risks. The approval process is schematically presented in Figure 1 below.

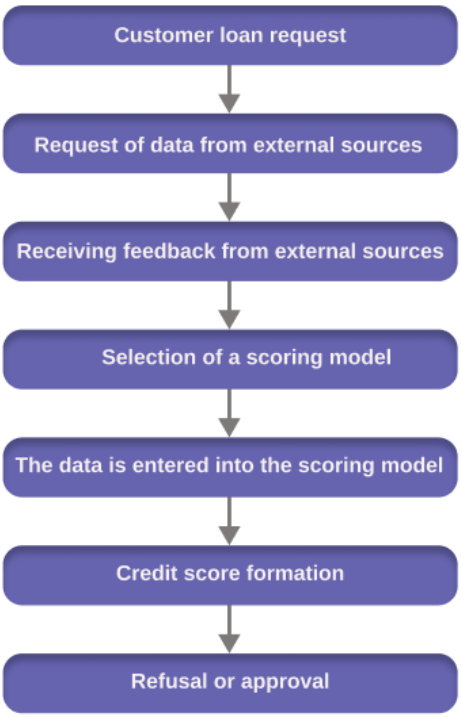

Figure 1. The process of calculating a client's personal credit score using data from external sources

2. Materials and methods

2.1. Most common machine learning algorithms in credit scoring process

2.1.1. Logistic regression

Logistic regression stands out as one of the most prevalent and widely utilized machine learning methods in credit scoring models. Commonly applied for predicting binary outcomes, especially in scenarios where the outcome variable is dichotomous, such as in the case of credit default [4]. The logical regression equation is presented as follows:

$$y = \frac{1}{1+e^{-\beta x}}$$

where x – input value, y – predicted value, β – coefficient for input(x), e – Euler's number

In credit scoring, a common practice involves utilizing a binary variable, representing the likelihood of a borrower defaulting on a loan (coded as 1 or 0), as the dependent variable. Numerous independent factors, such as the borrower's salary, credit history, employment status, and personal details, are considered. The logistic regression model calculates coefficients to illustrate the impact of each independent variable on the likelihood of default. These coefficients are subsequently employed in assessing loan applications, generating a credit score—a numerical representation indicating the probability of default. Figure 2 illustrates the application of logistic regression in credit scoring.

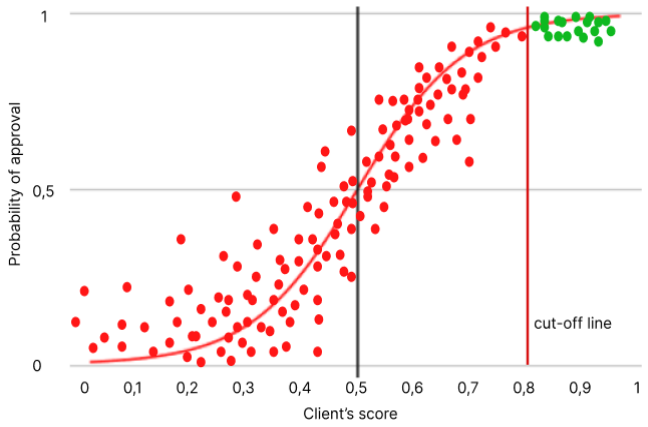

Figure 2. Credit scoring model based on logistic regression. The cut-off value divides clients into rejected and approved

Its computational efficiency is crucial for processing large datasets common in credit scoring. The model's assumption of linearity aligns well with the often-linear associations in credit-related datasets, and its reduced risk of overfitting enhances generalization to new data.

Logistic regression models in credit scoring may struggle to capture complex nonlinear relationships present in credit data, potentially leading to reduced predictive accuracy. Additionally, logistic regression assumes linearity between features and outcomes, which may not always hold true in real-world credit scenarios, limiting its effectiveness.

Despite the availability of more complex models, logistic regression remains a foundational and practical tool in credit scoring, offering a balance between simplicity and predictive power.

2.2 Decision trees

Decision trees, a visual and versatile machine learning method, are widely applied in credit scoring. Offering interpretability and adaptability, decision trees assess borrower attributes, creating a structured framework for credit decisions. They effectively capture non-linear relationships and reveal the importance of different variables in determining creditworthiness [5]. Decision trees have become frequently employed for data fitting and default prediction. The algorithms within decision trees adopt a top-down methodology, selecting the variable that optimally splits the dataset at each step [6]. Decision tree models offer simplicity and interpretability, making them valuable for understanding credit decisions, and they can handle both numerical and categorical data effectively, allowing for flexible modeling of credit risk factors.

Decision tree models may suffer from overfitting [7], particularly with complex data or a large number of features, which can lead to reduced generalization performance on unseen data. Additionally, decision trees may struggle to capture complex nonlinear relationships between variables, potentially limiting their predictive accuracy in credit scoring tasks. The operating principle of the decision tree algorithm in credit scoring is presented in Figure 3 below.

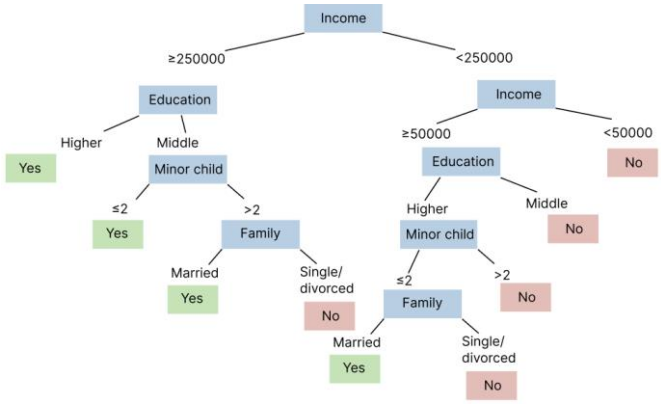

Figure 3. An example of an algorithm for using decision trees in a credit scoring model

2.3 SVM

The Support Vector Machine (SVM) algorithm is extensively utilized in credit scoring models for its adeptness in managing high-dimensional data and capturing intricate relationships between variables. Within the context of credit scoring, SVM functions by identifying the optimal hyperplane that effectively segregates creditworthy applicants from non-creditworthy ones within the feature space. This process entails maximizing the margin between the hyperplane and the nearest data points of each class while minimizing classification errors [8].

In implementing SVM for credit scoring, the dataset undergoes preprocessing to address missing values and scale features appropriately. Subsequently, the algorithm is trained on the preprocessed data, optimizing critical hyperparameters such as the selection of kernel function and regularization parameter via methodologies like cross-validation. Lastly, the trained SVM model is rigorously evaluated using a validation dataset to gauge its accuracy and generalization capabilities prior to its deployment in real-world credit scoring applications.

Support Vector Machine (SVM) in credit scoring provides accurate risk assessments by effectively handling high-dimensional data and identifying creditworthy applicants through optimal hyperplane separation. However, SVM models may demand significant computational resources and intricate hyperparameter tuning, while their interpretability compared to other algorithms might be limited [9]. An example graph of customer classification using the SVM method is presented below in Figure 4.

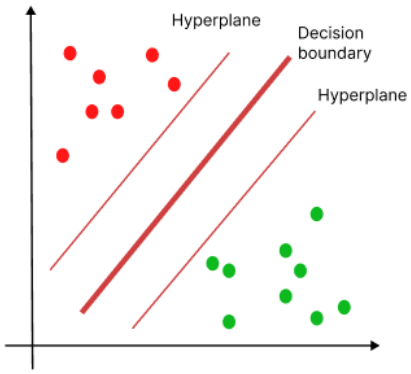

Figure 4. An example of client classification using SVM

2.4 Random Forest

Random Forest is a powerful ensemble learning algorithm widely used in credit scoring models due to its ability to provide accurate predictions and handle complex data structures [10]. In a credit scoring context, Random Forest operates by constructing multiple decision trees during the training phase [11]. Each decision tree is trained on a random subset of the data and a random subset of features, ensuring diversity among the trees. During prediction, the algorithm aggregates the predictions of individual trees to produce a final prediction. This ensemble approach helps to mitigate overfitting and improve the overall predictive performance of the model.

Implementing Random Forest in a credit scoring model involves several steps. Firstly, the dataset containing relevant credit features such as credit history, income, and debt-to-income ratio is preprocessed to handle missing values and encode categorical variables if necessary. Then, the Random Forest algorithm is trained on the preprocessed data using techniques like bootstrapping and feature bagging [12]. Hyperparameters such as the number of trees and maximum tree depth are tuned to optimize model performance, typically using techniques like cross-validation. Finally, the trained Random Forest model is evaluated on a separate validation dataset to assess its accuracy and generalization performance before deploying it for credit scoring purposes.

In credit scoring, the Random Forest machine learning method offers notable advantages, including its capability to effectively manage large and intricate datasets, facilitating accurate predictions while accommodating complex relationships among credit variables. Moreover, its ensemble approach mitigates overfitting, enhancing the model's ability to generalize to unseen data. However, the computational demands of Random Forest, especially with extensive datasets, present a notable challenge, as does its interpretability compared to simpler models. Furthermore, optimizing parameters for Random Forest models can be intricate [13], necessitating meticulous tuning to achieve optimal performance, thereby adding

complexity to the modeling process in credit scoring research. An example of a random forest algorithm with input data and an ensemble of decision trees is presented in Figure 5.

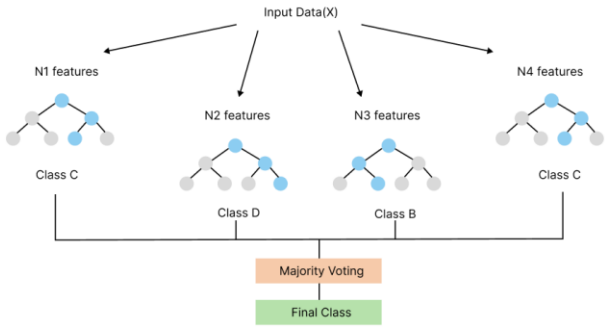

Figure 5. Schematic model of the random forest algorithm

2.5 XGBoost

XGBoost, an Extreme Gradient Boosting algorithm, is extensively applied in credit scoring models for its exceptional predictive performance and adaptability [14]. In credit scoring, XGBoost iteratively refines decision trees to minimize a predefined loss function, thereby enhancing model accuracy. Implementation involves preprocessing the credit dataset to address missing values and categorical variables, followed by training the XGBoost algorithm with optimized hyperparameters. Model evaluation on a separate dataset assesses its performance before deployment in credit scoring applications, ensuring accurate prediction of creditworthiness and effective risk management.

XGBoost offers several advantages in credit scoring, including superior predictive accuracy, adaptability to handle complex data structures, and the ability to capture intricate relationships between credit features. Its ensemble approach mitigates overfitting, enhancing generalization performance on unseen data, while its feature importance analysis provides valuable insights into credit risk factors. However, XGBoost models may be computationally intensive, particularly with large datasets, and tuning hyperparameters can be complex and time-consuming [15]. Additionally, XGBoost's black-box nature may limit interpretability, making it challenging to explain credit decisions to stakeholders. The operating algorithm of the XGBoost method is presented schematically in Figure 6 below.

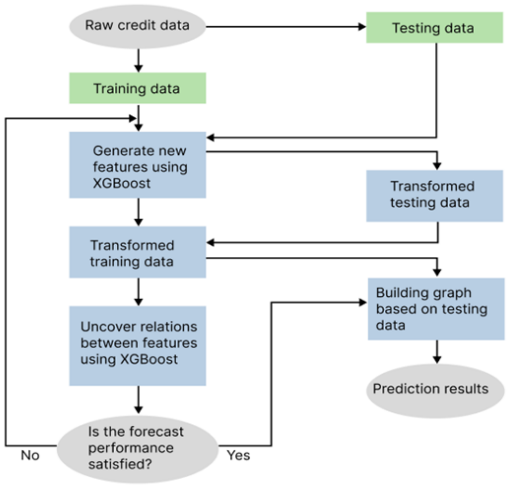

Figure 6. An example of a scoring model algorithm using the XGBoost machine learning method

3. Results and discussion

Table 1. Advantages and disadvantages of the methods

ML method	Advantages	Disadvantages
Logistic regression	Clear result, easy to implement.	Reduced accuracy when dealing with non-linear relationships.
DTs	The outcome is readily understandable, and it can manage relationships that aren't linear.	Propensity for overfitting.
SVM	Superior accuracy, adeptness in managing non-linear relationships, and minimal data requirements.	Requires significant effort to implement, and has a higher likelihood of overfitting.
Random Forest	High level of accuracy, capability to handle an extensive range of features.	Result that is not easily understood or interpreted.
XGBoost	High accuracy, adeptness in managing numerous features, and improved alignment between predictions and actual values.	Susceptible to overfitting, presenting greater implementation complexity.

4. Conclusions

In conclusion, this article has provided a comprehensive overview of common machine learning algorithms used in credit scoring models, including logistic regression, decision trees, random forest, XGBoost, and SVM. Each algorithm offers distinct advantages and disadvantages in handling credit data, from simplicity and interpretability to robustness and predictive accuracy. Logistic regression, for instance, provides straightforward interpretations but may struggle with capturing nonlinear relationships, while ensemble methods like random forest and XGBoost excel in handling complex data structures but may require more computational resources.

Understanding the strengths and limitations of these algorithms is crucial for future research in credit scoring. By exploring novel techniques for feature engineering, model interpretation, and hyperparameter tuning, researchers can enhance the predictive performance and interpretability of machine learning models in credit scoring. Moreover, investigating the impact of emerging data sources such as alternative credit data and incorporating ethical considerations into model development are important avenues for future exploration. Ultimately, this article serves as a foundation for researchers to advance the field of credit scoring by developing more accurate, transparent, and fair machine learning models that empower financial institutions to make informed lending decisions while effectively managing credit risk.

References

[1] Jittima Tongurai, Chaiporn Vithessonthi. (2018). The impact of the banking sector on economic structure and growth. *International Review of Financial Analysis, (56), 193-207*

[2] Tokaev, N., Gokoev, A. (2023). Commercial bank credit risk management. *Russian Journal of Management, 11(2), 82-90*

[3] Saeed, M., Zahid N. (2016). The Impact of Credit Risk on Profitability of the Commercial Banks. *Journal of Business & Financial Affairs, 5(2)*

[4] Hosmer, Jr, D.W., Lemeshow, S. & Sturdivant, R.X. (2013). Applied logistic regression. *John Wiley & Sons*

[5] Lee, S. (2022). Decision Trees: A Non-Parametric Supervised Learning Approach for Classification and Regression Tasks. *Machine Learning Journal, 12(3), 45-53*

[6] Kim, J. (2022). Visualizing Decision Trees: The Tree Flow View. *Journal of Visualization, 15(1), 67-72*

[7] Nadar, J.B. (2023). Overfitting in Decision Tree Models: Understanding and Overcoming the Pitfalls. *Retrieved from*: https://joelnadarai.medium.com/overfitting-in-decision-tree-models-understanding-and-overcoming-the-pitfalls-880cf7af7d8b#:~:text=Overfitting%20occurs%20when%20a%20decision,recursively%20partitioning%20the%20feature%20space
[8] Putri, N.H., Fatekurohman, M. & Tirta, I.M. (2021). Credit risk analysis using support vector machines algorithm. *Journal of Physics: Conference Series, 1836*
[9] Karamizadeh, S., Abdullah, S.M., Asl, M.H. & Shayan, J. (2014). Advantage and Drawback of Support Vector Machine Functionality. *Conference: Computer, Communications, and Control Technology (I4CT)*
[10] Donges, N. (2024). Random Forest: A Complete Guide for Machine Learning. *Retrieved from*: https://builtin.com/data-science/random-forest-algorithm
[11] Xingzhi Zhang, Yan Yang & Zhurong Zhou. (2018). A novel credit scoring model based on optimized random forest. *IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*
[12] Ullah, A. & Wang, R. (2020). Bootstrap Aggregating and Random Forest. *Macroeconomic Forecasting in the Era of Big Data, 389-429*
[13] Mohapatra, N., Shreya, K. & Chinmay, A. (2020). Optimization of the Random Forest Algorithm. *Advances in Data Science and Management, 201-208*
[14] Kui Wang, Meixuan Li, Jingyi Cheng, Xiaomeng Zhou & Gang Li. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science, (199), 1128-1135*
[15] Zeravan Arif Ali, Ziyad H. Abduljabbar. (2023). Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. *Academic Journal of Nawroz University, 12(2)*

Несие алушылар үшін машиналық оқу әдістерін зерттеу

Л. Маралбаева*

Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан

**Корреспонденция үшін автор: maralbaevalaura@gmail.com*

Аңдатпа. Бұл зерттеу қарыз алушылардың, әсіресе екінші деңгейлі банктермен байланысқандардың несиелік қабілеттілігін бағалау үшін машиналық оқыту әдістерінің қалай қолданылатынын егжей-тегжейлі қарастырады. Мақаланың негізгі бағыты кредиттік скорингтің соңғы тенденцияларын түсіндіру үшін заманауи ғылыми әдебиеттерді пайдалану болып табылады. Зерттеу маңызды қадамдар мен шешім қабылдауға әсер ететін факторларды бөліп көрсете отырып, несиелеу үдерісіне егжей-тегжейлі шолу жасайды. Авторлар скорингте қолданылатын әртүрлі деректер көздеріне терең еніп, несиелік скорингті дәлірек және әділ етуге қалай көмектесетінін түсіндіреді. Олар машиналық оқытудың әртүрлі әдістерінің күшті және әлсіз жақтарын талдайды, олардың қаншалықты тиімді екенін және екінші деңгейлі банктерге жарамдылығын анықтайды. Мақалада машинада оқытудың әртүрлі әдістерін егжей-тегжейлі салыстыру, олардың қай жерде жақсы жұмыс істейтінін және қай жерде шектеулер болуы мүмкін екенін түсіндіреді. Зерттеудің маңыздылығы машиналық оқыту әдістеріне, соның ішінде соңғы жаңартулар мен салыстыруларға кең шолу жасау болып табылады, бұл осы саладағы болашақ зерттеулер үшін бастапқы нүкте болады. Практикалық тұрғыдан алғанда, мақала қарыз алушылардың несиелік қабілетін бағалауда заманауи машиналық оқыту әдістерін тиімдірек пайдалану идеяларын ұсынатын банк секторының мамандары үшін пайдалы. Зерттеу осы салаға қызығушылық танытқандар үшін құнды нұсқаулар мен практикалық стратегияларды ұсына отырып, несиелік скорингтің заманауи тәсілдерін түсінуге және қолдануға елеулі үлес қосады.

Негізгі сөздер: *несиелік скоринг, баллдық үлгілер, машиналық оқыту алгоритмдері, логистикалық регрессия, шешім ағаштары, кездейсоқ орман, қолдау векторлық машинасы, экстремалды градиентті арттыру*.

Исследование существующих методов машинного обучения для кредитного скоринга заемщика

Л. Маралбаева*

Международный университет информационных технологий, Алматы, Казахстан

**Автор для корреспонденции: maralbaevalaura@gmail.com*

Аннотация. В этом исследовании подробно изучается, как методы машинного обучения используются для оценки кредитоспособности заемщиков, особенно тех, кто связан с банками второго уровня. Основное внимание в статье уделяется использованию современной научной литературы для объяснения последних тенденций в кредитном скоринге. Исследование дает подробный обзор кредитного процесса, указывая на важные этапы и факторы, влияющие на принятие решений. Авторы глубоко погружаются в различные источники данных, используемые при скоринге, и объясняют, как они помогают сделать кредитную оценку более точной и справедливой. Они анализируют сильные и слабые стороны различных методов машинного обучения, выясняя, насколько они эффективны и подходят ли они банкам второго уровня. В статье представлено подробное сравнение различных методов машинного обучения, объяс-

нено, где они работают хорошо, а где могут иметь ограничения. Важность исследования заключается в том, чтобы дать широкий обзор методов машинного обучения, включая недавние обновления и сравнения, что послужит отправной точкой для будущих исследований в этой области. С практической точки зрения статья полезна для специалистов банковского сектора, предлагая идеи для более эффективного использования современных методов машинного обучения при оценке кредитоспособности заемщиков. Исследование вносит значительный вклад в понимание и применение современных подходов к кредитному скорингу, предлагая ценные рекомендации и практические стратегии для тех, кто интересуется этой областью.

Ключевые слова: *кредитный скоринг, скоринговые модели, алгоритмы машинного обучения, логистическая регрессия, деревья решений, случайный лес, машина опорных векторов, экстремальное повышение градиента*.

Received: 13 September 2023
Accepted: 15 December 2023
Available online: 31 December 2023