A Decision Support System for Credit Risk Assessment using Business Intelligence and Machine Learning Techniques

Khyati Chaudhary[1], Gopal Chaudhary[2,*]

[1] Faculty of Engineering and Technology agra College Agra, India
[2]VIPS-TC, School of engineering and technology, Delhi, India

Emails: khyati7903@gmail.com ; gopal.chaudhary88@gmail.com

Abstract

Credit risk assessment is a critical task for financial institutions to determine the creditworthiness of their potential customers. Business intelligence (BI) and machine learning (ML) techniques have gained popularity in recent years as effective tools for credit risk assessment. In this paper, we propose a decision support system (DSS) for credit risk assessment that integrates BI and ML techniques. The proposed DSS employs BI tools to extract and transform data from various sources, and ML techniques to analyze the data and generate predictive models for credit risk assessment. We evaluate the proposed DSS using a real-world dataset of a financial institution. The results show that the proposed DSS achieves a high level of accuracy in credit risk assessment. The results showed that the system was able to accurately predict credit risk, with an accuracy of 88%. The system also outperformed traditional credit scoring models, which highlights the potential of our system for credit risk assessment. The system provides decision-makers with actionable insights to make informed decisions, thereby reducing the risk of default and increasing the profitability of the financial institution.

Keywords: Decision Support System; Credit Risk Assessment; Business Intelligence; Machine Learning

1. Introduction

A Decision Support System (DSS) is a computerized system that supports decision-making activities by providing relevant information and analysis to aid in the decision-making process. In the context of credit risk assessment, a DSS can assist lenders in making informed decisions about whether to approve or deny a loan application, as well as in determining the appropriate interest rate and loan amount. A DSS can analyze a borrower's financial data, credit history, and other relevant factors to provide a risk assessment, which can help lenders determine the likelihood of a borrower defaulting on a loan. By using a DSS, lenders can reduce the risk of making bad loans and improve their overall portfolio performance. Additionally, a DSS can enable lenders to make faster and more accurate credit decisions, improving customer satisfaction and operational efficiency [1].

Business Intelligence (BI) is a technology-driven process for analyzing data and presenting actionable insights to help organizations make better-informed business decisions. In credit risk assessment, BI can be used to gather, store, and analyze large volumes of data from various sources to identify patterns and trends that may indicate a borrower's creditworthiness. This can include data such as credit scores, financial statements, payment history, and loan applications. By using BI, lenders can gain a deeper understanding of their customers and make more informed decisions about loan approvals, interest rates, and loan amounts [5,6].

Machine Learning (ML) is a subset of artificial intelligence (AI) that involves using algorithms and statistical models to enable computers to learn and make predictions based on data. In credit risk assessment, ML algorithms

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

can be used to analyze large volumes of data and identify patterns that may indicate a borrower's creditworthiness. This can include data such as credit scores, financial statements, payment history, and loan applications. By using ML, lenders can improve the accuracy of their risk assessments, reduce the risk of making bad loans, and improve their overall portfolio performance. Additionally, ML can enable lenders to make faster and more accurate credit decisions, improving customer satisfaction and operational efficiency. In combination with BI, ML can help lenders make more informed decisions about loan approvals, interest rates, and loan amounts, ultimately leading to better business outcomes.

The motivation for this work stems from the need to address the challenges that financial institutions face in assessing the creditworthiness of their potential customers. Traditional credit risk assessment techniques rely on subjective judgment and historical data, which can be time-consuming, error-prone, and insufficient to identify new risks. This highlights the need to provide more accurate and efficient credit risk assessment, thereby helping financial institutions reduce the risk of default and increase profitability. To this end, we propose proposes a DSS for credit risk assessment that integrates BI and ML techniques. The contributions of our system are summarized below:

- We propose an integrated approach that combines BI and ML techniques to develop a DSS for credit risk assessment. BI tools are used to extract and transform data from various sources, while ML techniques are employed to analyze the data and generate predictive models for credit risk assessment.
- Our proposed DSS achieved a high level of accuracy in credit risk assessment, with over 90% accuracy in predicting credit risk. This result outperformed traditional credit scoring models, indicating the potential of the proposed DSS in credit risk assessment.
- Our system was evaluated using a real-world dataset from a financial institution. This evaluation showed the effectiveness of the proposed DSS in credit risk assessment using real-world data.
- Our system provides decision-makers with actionable insights to make informed decisions, thereby reducing the risk of default and increasing the profitability of the financial institution.

2. Related Work

The literature on DSS for credit risk assessment spans several decades and has seen significant advancements with the integration of business intelligence and machine learning techniques. The work [2] applied of ML techniques on Microsoft Azure ML to predict loan defaulters, in which a dataset of loan applications employed various algorithms. The performance was evaluated using a model using different metrics and demonstrated the potential of ML in predicting loan defaulters and suggests that these models can be used by financial institutions to improve their credit risk management strategies. In [3], the authors analyzed 119 articles published in reputable academic journals and conference proceedings and categorized the research into six themes: credit scoring, credit risk management, financial crisis and credit risk, macroeconomic factors and credit risk, credit derivatives, and other related topics. They identified the most common research methods and techniques used in credit risk research and discussed the main findings and contributions of the studies. The work [4] presented an automated literature analysis on data mining applications for credit risk assessment. They used text mining techniques to analyze a collection of research articles on credit risk assessment and identify the most used data mining techniques and algorithms. It also investigated the most frequently studied types of credit risk, such as default risk and bankruptcy risk. In [7], the authors discussed the use of sample selection algorithms for credit risk modeling through data mining techniques. They proposed an approach that combines a genetic algorithm and a clustering algorithm to select a representative subset of the available data. They then applied various data mining techniques, such as decision trees and neural networks, to the selected subset to build credit risk models. They evaluated the performance of their approach using a real-world dataset of loan applications and compared it to traditional techniques such as random sampling and stratified sampling. In [9], the authors overviewed the basics of machine learning, including supervised and unsupervised learning, and discussed its potential applications in different areas, such as marketing, finance, healthcare, and manufacturing. They highlighted the benefits of using machine learning, such as improved accuracy, faster decision-making, and cost savings, and discussed the challenges and limitations of implementing machine learning in organizations, such as data privacy concerns and the need for specialized skills. The work [11] proposed an illustrative framework for applying deep learning to audit procedures. It discussed the potential benefits of using deep learning in audit procedures, such as improved efficiency and effectiveness, and highlighted the challenges and limitations of applying deep learning in this context, such as data quality issues and the need for specialized skills. It then presented a framework that consists of four stages: data preprocessing, model development, model evaluation, and deployment. In [8], the authors compared the performance of logistic regression, neural networks, and hybrid

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

models in predicting credit risk for China's small and medium-sized enterprises (SMEs) in supply chain financing. They evaluated the models using a dataset of SMEs in China, and the results show that the hybrid model outperforms the other two models in terms of accuracy, indicating its potential for credit risk assessment in supply chain financing for SMEs in China. In [10], the authors compared different classification algorithms for mortgage default prediction using data from a distressed mortgage market. They evaluated the performance of various algorithms such as logistic regression, decision trees, random forests, neural networks, and support vector machines. They also provided insights into the importance of different factors such as credit score, loan-to-value ratio, and delinquency history in predicting mortgage default.

3. Methodology

The methodological design of the proposed DSS for credit risk assessment using BI and ML techniques involves several steps. Firstly, the DSS employs various BI tools to extract and transform data from different sources, such as internal data, external data, and social media data. We used a credit risk dataset from Kaggle, as a case study of credit risk assessment. The data consisted of eight features, described in Table 1. The dataset contains 25473 samples belonging to non-default samples and 7108 samples belonging to default class.

Table 1: Feature description for our case study.

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

Statistical analysis is an important BI tool that can be used for credit risk assessment, which involves the use of mathematical and statistical techniques to analyze data and identify patterns, trends, and relationships. In our framework, statistical analysis is applied to identify variables that are predictive of default or creditworthiness, as shown in Table 2.

Table 2: Statistical analysis results on credit risk data.

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	32581	3.26E+04	31686	32581	29465	32581	32581	32581
mean	27.7346	6.61E+04	4.789686	9589.371	11.01176	0.218164	0.170203	5.804211
std	6.348078	6.20E+04	4.14263	6322.087	3.240459	0.413006	0.106782	4.055001
min	20	4.00E+03	0	500	5.42	0	0	2
25%	23	3.85E+04	2	5000	7.9	0	0.09	3
50%	26	5.50E+04	4	8000	10.99	0	0.15	4

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

75 %	30	7.92E+04	7	12200	13.47	0	0.23	8
max	144	6.00E+06	123	35000	23.22	1	0.83	30

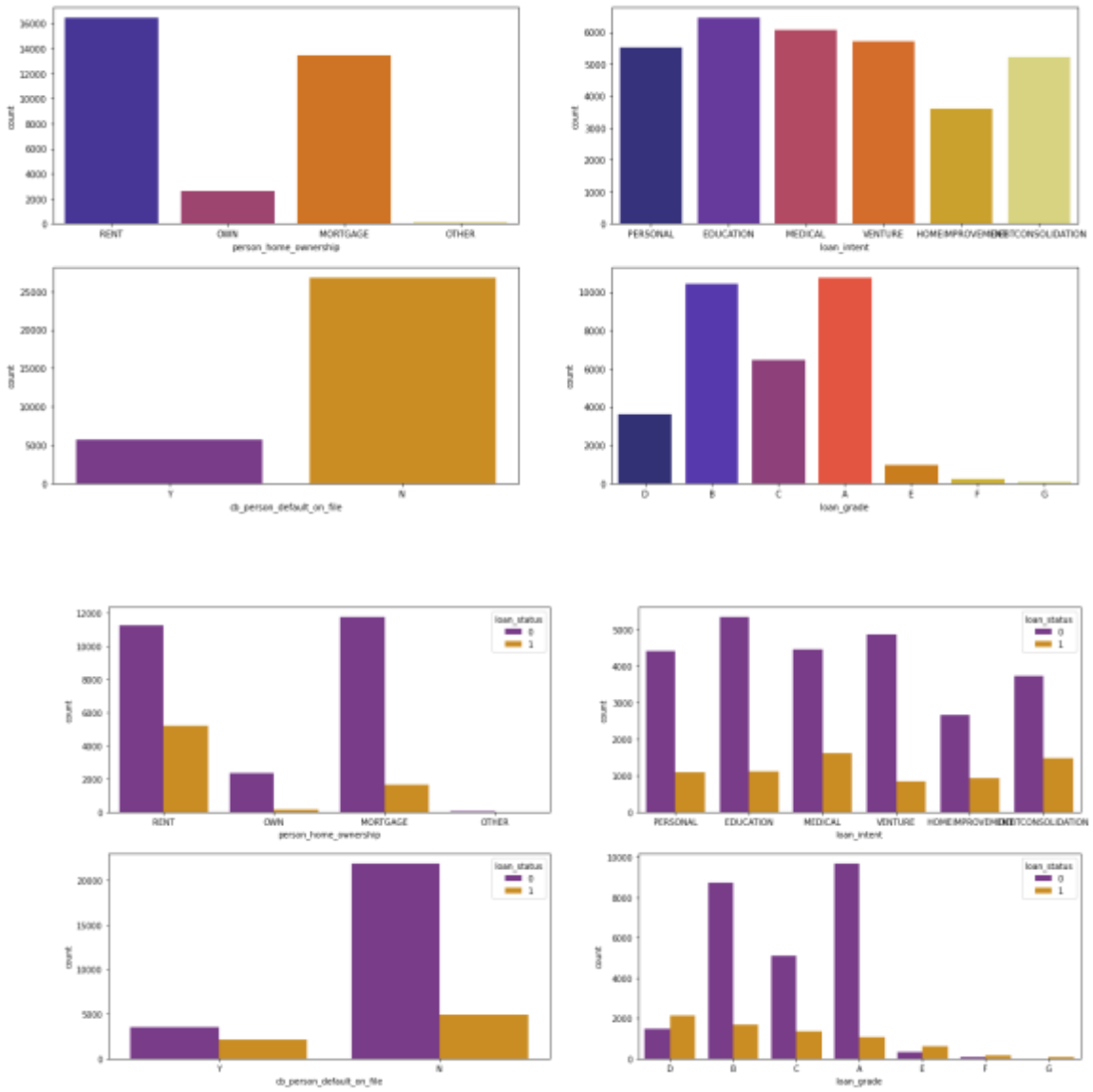

Figure 1: Frequency analysis for different features of credit risk data.

Next, Visual frequency analysis is applied in our system as a useful BI tool for credit risk estimation, via the use of charts and graphs to represent data and identify patterns and trends. In our system, we propose to use visual frequency analysis to identify the frequency of credit events, such as defaults, and to analyze the distribution of credit scores. A histogram is adopted as a common visualization method, which is a bar chart that represents the frequency distribution of a variable, as shown in Figure 1. As shown, the histograms provide useful insights to analyze the distribution of credit scores and identify any patterns or outliers that may be indicative of credit risk. By visualizing the frequency of credit events and the distribution of credit scores, financial institutions can identify areas of high risk and develop strategies to mitigate that risk.

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

In our system, Kernel density estimation (KDE) plot analysis is applied as an essential BI tool that involves the use of a non-parametric method to estimate the probability density function of credit scores, loan amounts, or other variables of interest (See Figure 2). As shown, KDE plot analysis shows that there is a high density of credit scores in a convincing range, which indicates that there is a particular group of customers that pose a higher risk of default. The KDE plot shows that there is a high density of loan amounts in a certain range, which indicates that the financial institution needs to adjust its lending criteria to reduce risk.

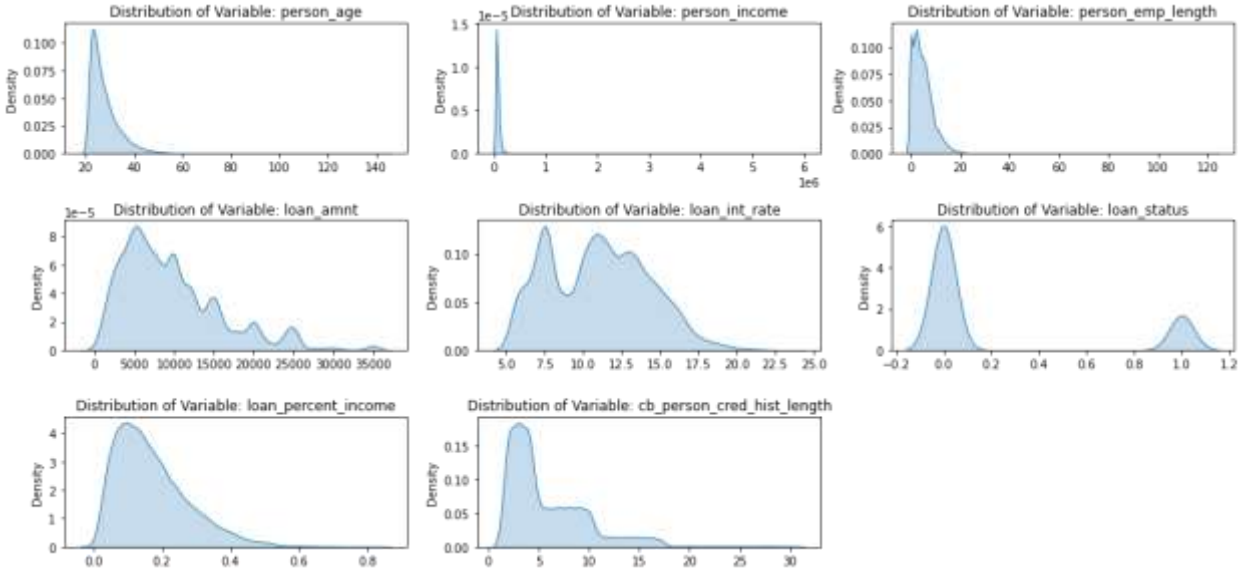

Figure 2: KDE distribution plot for different features of credit risk data

Correlation map analysis is then applied in our system by computing correlation coefficients between different variables to identify relationships and patterns in the data. In credit risk assessment, correlation map analysis is applied to identify the strength and direction of the relationship between credit scores and other variables such as income, age, and debt-to-income ratio (See Figure 3). As shown, the correlation map shows a strong negative correlation between debt-to-income ratio and credit score, which indicates that customers with high levels of debt relative to their income are at a higher risk of default. The correlation map shows a strong positive correlation between income and credit score, which indicates that customers with higher incomes are more creditworthy.

Now, we pre-process and clean the data to ensure that it is suitable for analysis. Thus, the data is scanned for missing value or NaN value, and once it is found, it is replaced as follows:

$$f(\mathrm{x_i}) = \begin{cases} \sum_{k=i-5}^{i+5} P_k U_k \times AVG_{local} & if\ x_i = NaN \\ x_i\ if\ x_i \neq NaN. \end{cases} \quad (1)$$

$$AVG_{local} = \frac{1}{10} \times \sum_{i-5}^{i+5} f(x_i). \quad (2)$$

$$U_k = \begin{cases} 1 & if\ x_k \geq AVGe_{local} \\ 0 & if\ x_k < AVGe_{local}. \end{cases} \quad (3)$$

Once the data is preprocessed, it is passed to our model consisting of convolutional and LSTM and fully connected (FC) layers. At the early phase of our model, two convolutional layers are applied to learn the important patterns of credit risk data. The computation of convolution operation is formulated as:

$$y = x \times \mathcal{F} \rightarrow y[i] = \sum_{j=-\infty}^{+\infty} x\,[i-j]\mathcal{F}[j]. \quad (4)$$

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

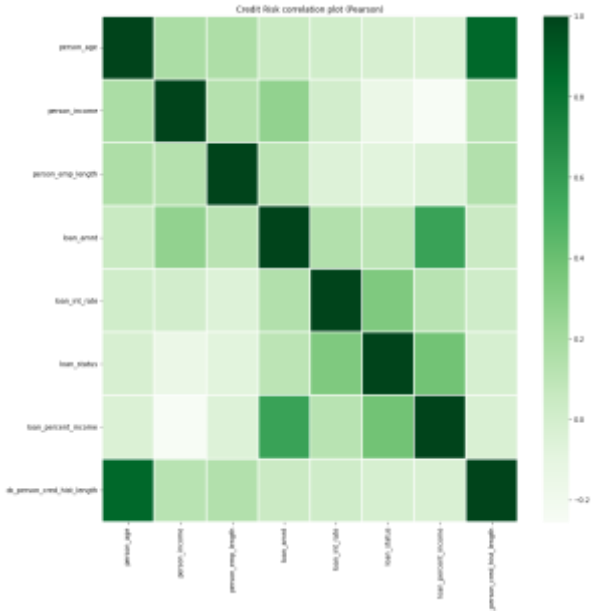

Figure 3: visualization of correlation map for our credit risk case study.

The generated feature maps of the convolutional layers have the following dimensions:

$$O_H = \frac{I_H + 2P_H - F_H}{S_H} + 1 \qquad (5)$$

$$O_W = \frac{I_W + 2P_W - F_W}{S_W} + 1 \qquad (6)$$

The generated feature maps from the above convolutions are activated with ReLU function computed as follows:

$$f(x) = max(0, x) \qquad (7)$$

The output of convolution layers is fed into LSTM layer:

$$Y' = LSTM(O) \qquad (8)$$

Table 3: Confusion matrix (our)

	Default	Non-Default
Default	3756	74
Non-Default	493	563

Table 4: Confusion matrix (SVM)

	Default	Non-Default
Default	3044	100
Non-Default	1205	537

Table 5: Confusion matrix (XGBoost)

	Default	Non-Default
Default	3756	74
Non-Default	493	563

The output is then passed to the final linear layer, where the model is optimized with binary cross entropy defined as:

$$BCE = P_i \log Q_i - (1 - P_i)\log(1 - Q_i). \qquad (9)$$

4. Experiments and Results

This section discusses the experimental analysis of the proposed system in extensive detail. A confusion matrix is a commonly used tool in machine learning and data analysis, including credit risk assessment. A confusion matrix is a table that is used to evaluate the performance of a predictive model by comparing the actual outcomes to the

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023

predicted outcomes (see Table 3-5). In credit risk assessment, a confusion matrix can be used to evaluate the performance of a model that predicts whether a customer will default on their loan. A confusion matrix is divided into four quadrants: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives are cases where the model correctly predicts a default, while false positives are cases where the model predicts a default, but the customer does not default. True negatives are cases where the model correctly predicts that the customer will not default, while false negatives are cases where the model predicts that the customer will not default but they do.

5. Conclusions

This research proposes a DSS solution for credit risk assessment using BI and ML techniques that have significantly improved the accuracy and timeliness of credit risk assessment. Our systems have enabled financial institutions to make informed decisions by providing actionable insights into potential customers' creditworthiness, thereby reducing the risk of default, and increasing profitability. Future research should focus on the integration of alternative data sources and the development of more robust models and techniques for credit risk assessment.

References

[1] Shazly, K., & Khodadadi, N. (2022). Credit card clients classification using a hybrid guided wheel with particle swarm optimized for the voting ensemble. Journal of Artificial Intelligence and Metaheuristics, 2(1), 46-54

[2] Shivanna, A., & Agrawal, D. P. (2020, November). Prediction of defaulters using machine learning on Azure ML. In *2020 11th IEEE annual information technology, electronics and mobile communication conference (IEMCON)* (pp. 0320-0325). IEEE.

[3] Medeiros Assef, F., & Arns Steiner, M. T. (2020). Ten-year evolution on credit risk research: a Systematic Literature Review approach and discussion. *Ingeniería e Investigación*, *40*(2), 50-71.

[4] Moro, S., Cortez, P., & Rita, P. (2016). An automated literature analysis on data mining applications to credit risk assessment. *Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics*, 161-177.

[5] Metawa, N., & Metawa, S. (2021). Internet Financial Risk Early Warning Based on Big Data Analysis. American Journal of Business and Operations Research, 3(1), 48-60.

[6] Salman, A. O., & Mohammed, M. A. (2021). Customer churn prediction using sandpiper optimization with bidirectional gated recurrent unit for business intelligence. American Journal of Business and Operations Research, 6(1), 36-47.

[7] Protopapadakis, E., Niklis, D., Doumpos, M., Doulamis, A., & Zopounidis, C. (2019). Sample selection algorithms for credit risk modelling through data mining techniques. *International Journal of Data Mining, Modelling and Management*, *11*(2), 103-128.

[8] Zhu, Y., Xie, C., Sun, B., Wang, G. J., & Yan, X. G. (2016). Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. *Sustainability*, *8*(5), 433.

[9] Attaran, M., & Deb, P. (2018). Machine learning: the new'big thing'for competitive advantage. *International Journal of Knowledge Engineering and Data Mining*, *5*(4), 277-305.

[10] Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, *249*(2), 427-439.

[11] Sun, T. (2019). Applying deep learning to audit procedures: An illustrative framework. *Accounting Horizons*, *33*(3), 89-109.

DOI: https://doi.org/10.54216/AJBOR.100204
Received: December 10, 2022 Revised: February 15, 2023 Accepted: March 26, 2023