COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 1, Jan-Jun 2023, pp. 51-60
DOI: 10.17492/computology.v3i1.2305
https://www.journalpressindia.com/cjacsit

Credit Default Prediction System using Machine Learning

Hassan J. Bature, Daniel D. Wisdom**, Tolulope T. Dufuwa*** and Isaac O. Ayetuoma*****

ABSTRACT

The use of innovative technologies and services has allowed financial institutions to offer microcredit to low-income earners. Digitization has allowed lenders to automate loan application components, including underwriting and e-signatures; resulting in more efficient loan delivery while maintaining traditional underwriting and compliance practices. Traditional credit score models have limitations in applying big data technology to build risk models, and machine learning based credit risk models have emerged as a more effective way to predict defaults. This paper proposed a Credit Default Prediction System Using Machine Learning. The scheme successfully designed a classification model to predict loan default before the loan is approved.

Keywords: Machine Learning, Credit, default, Loan-prediction, classification model

1.0 Introduction

Digital lending, also referred to as social or online lending, is a rapidly growing phenomenon in emerging markets. This type of lending involves credit products that are delivered solely via digital channels, such as mobile apps and the web. These loans are typically unsecured cash loans that are obtained via digital channels without the involvement of a salesperson [1]. Digital lending leverages digital data to make lending decisions via automated processes. This has allowed millions of low income consumers to borrow money from the convenience of their location. However, digital lending carries a higher risk compared to traditional banking due to insufficient credit checking, inadequate intermediation, lack of transparency, and the financial status of online borrowers [2]. Credit risk prediction and management become vitally important in this domain.

Digital lending has disrupted the traditional consumer banking sector, offering more effective loan processing, and the loan application decision is made automatically with electronic data-driven algorithms. This has allowed online lenders to offer small loans with short-term maturities, making it easier for borrowers who are excluded from traditional banking systems to access credit [17]. However, credit risk is a major concern among commercial organizations, which may result in a dire condition known as credit default. To minimize credit risk, lenders thoroughly evaluate and verify the ability of a borrower to deliver on their obligation of repaying the loan [2].

**Federal University Oye Ekiti,Ekiti State, Nigeria (E-mail: joshua.hassan@fuoye.edu.ng)*
***Corresponding author; Chrisland University Abeokuta, Ogun State, Nigeria (E-mail: danieldaudawisdom1@gmail.com)*
****Federal University Oye Ekiti,Ekiti State, Nigeria (E-mail: tolulope.odufuwa@fuoye.edu.ng)*
*****Chrisland University Abeokuta, Ogun State, Nigeria (E-mail: iayetuoma@chrislanduniversity.edu.ng)*

Digital lending platforms use predictive analytics to arrive at creditworthiness score and limit, which is based on factors such as past business records, call logs, text messages, contact lists, age, education, and income [10]. Recently, researchers and lending institutions have turned to One could utilize machine learning algorithms and neural networks to predict an individual's credit score by analyzing their past data. By doing so, the system could also identify potential credit defaulters and reject their loan application. Despite the risks, loan lending is still considered an essential part of the financial organization, and lenders strive to minimize credit risk to ensure smooth functioning of their businesses.

2.0 Traditional Credit Risk Assessment

The conventional credit scoring method relies on historical data, comprising bank transactional data like previous credit transactions, records of delayed payments, credit bureau inquiries, and commercial data like financial statements and duration of credit history (World Bank, 2019).

2.1 Discriminant analysis

Discriminant analysis is a statistical technique to classify two groups, and it is still commonly used to classify customers as good or bad credit. This approach has been used for credit scoring and can classify group variables into different categories.

2.2 Judgment-based models

Various methods, including the analytic hierarchy process (AHP), are employed to develop judgment-based models. The AHP is a structured approach that breaks down complex decisions into easier-to-understand sub problems for analysis; human judgments are utilized alongside the data to complete the evaluations in the AHP.

2.3 Overview of machine learning

Arthur Samuel introduced the term "Machine Learning" in 1959, defining it as the study of enabling computers to learn without explicit programming. Machine Learning is a subfield of AI that aims to understand data structure and fit it into models that can be utilized by humans [16]. Machine learning implementations are divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves learning a function that maps input to output based on labeled data, with the algorithm modifying the model based on errors. Unsupervised learning involves inferring patterns from unlabeled input data and discovering hidden patterns within a dataset [13]. Reinforcement learning involves maximizing rewards by getting an agent to act in the world.

2.4 Machine learning approaches in credit scoring

The supervised machine learning algorithms implemented in this paper focus on the Decision Tree, Random Forest, and Extremely Randomized Tree classifiers, with more detailed explanations on their theory provided in the subsequent sections.

2.5 Decision tree

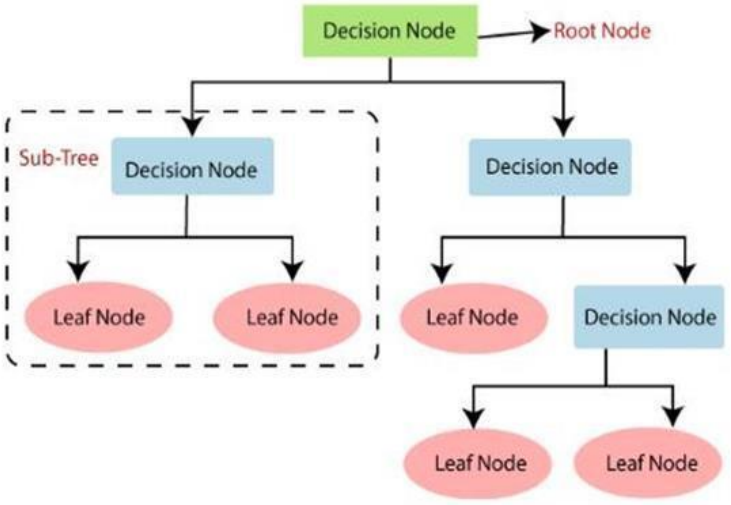
The Decision Tree is a supervised learning technique utilized for classification and regression

tasks, where classification is more frequently used. This classifier comprises a tree structure, where the internal nodes correspond to the dataset's features, decision rules represented by branches, and outcomes by leaf nodes. Two primary nodes, known as the Decision Tree, comprises two essential components: Decision Nodes and Leaf Nodes. Decision nodes, characterized by multiple branches, make determinations by assessing dataset features, whereas leaf nodes furnish outputs without any additional branches [9]. This tree-shaped structure represents all conceivable solutions to a problem or decision based on specified conditions, commencing with the root node and expanding into branches. The CART algorithm constructs this tree, and the decision tree functions by presenting a query and partitioning the tree into subtrees according to the response. The structure of a decision tree is illustrated in Figure 1 as depicted below.

2.5.1 Decision tree terminologies

1. The root node initiates the decision tree, representing the entire dataset and dividing it into subsets.
2. Leaf nodes are the final output nodes that cannot be divided further.
3. Splitting involves dividing a decision or root node into smaller nodes based on specific conditions.
4. A branch or sub-tree is created by splitting a node.
5. Pruning refers to the removal of unnecessary branches from the tree.
6. The root node is the parent node, while other nodes are considered child nodes.

Figure 1: Decision Tree



2.5.2 How the decision tree algorithm works

i. When employing a decision tree for classifying a particular dataset, the process commences by positioning itself at the tree's root node. It then proceeds to evaluate the attribute value at the root in relation to the dataset record's attribute, and based on this evaluation, it selects a branch and advances to the subsequent node.

ii. As the algorithm advances to the subsequent node, it iterates the attribute comparison within the sub-nodes, continuing this procedure until it eventually reaches the final leaf node of the tree.

2.5.3 Advantages

i. Decision Trees have the ability to imitate human reasoning when making a decision, which makes them easy to comprehend.
ii. The decision tree's rationale can be easily comprehended due to its tree-like structure.

2.6 Random Forest

The Random Forest is a supervised machine learning method that employs ensemble learning to tackle complex issues and enhance model performance. Its methodology entails creating numerous decision trees using different subsets of the dataset and then combining their outcomes to improve the predictive accuracy of the dataset [9]. This algorithm combines forecasts from each individual tree and reaches its final prediction by counting the most frequently occurring prediction among these forecasts. Increasing the quantity of trees in the forest enhances accuracy and acts as a safeguard against overfitting issues, as demonstrated in Figure 2 depicting the Random Forest.

2.6.1 Advantages and Disadvantages

i. In comparison to other algorithms, the training time required is shorter.
ii. Even when handling large datasets, it can efficiently predict outputs with high accuracy.
iii. It is capable of maintaining accuracy even when a significant amount of data is missing.
iv. Random Forests are resilient and can perform well in both regression and classification tasks.
v. RF algorithms are effective with large datasets and various data types, including numerical, binary, and categorical.
vi. However, when the number of trees is high, the complexity and computational time are relatively increased, leading to a longer training time. Additionally, the sampling of subsets may introduce bias.

2.6.2 Applications

Random Forests can be utilized in virtually any classification or regression scenario, although they are frequently employed in areas such as Remote Sensing, predicting stock market trends, identifying fraudulent activity, analyzing sentiment, and recommending products.

2.7 Extra Trees Classifier

The Extra Trees Classifier, akin to the Random Forest Classifier, is an ensemble machine learning technique, but it distinguishes itself through its approach to constructing decision trees. It generates several unpruned decision trees from the training dataset, where each tree employs a random subset of k features at every testing node to divide the data (Gupta, 2020). Subsequently, the algorithm combines the predictions from these trees to produce the ultimate classification outcome, relying on majority voting for classification tasks or an arithmetic average for regression tasks.

2.7.1 Extra Trees has two main advantages:

- It reduces bias by sampling from the entire dataset during the construction of the trees, which prevents the introduction of different biases that may arise when different subsets of the data are used.
- It reduces variance by randomizing the splitting of nodes within the decision trees, thereby preventing the algorithm from being heavily influenced by specific features or patterns in the dataset.

2.7.2 Applications

Extra Trees, much like Random Forests, find application in both classification and regression tasks. Furthermore, in specific scenarios, Extra Trees are employed for the purpose of feature selection as well.

2.7.3 Extra Trees versus Random Forest

The two ensembles, Random Forest and Extra Trees, share many similarities. They both consist of a significant number of decision trees, and finally the overall decision is made according to the predictions of every tree. In classification problems, this is achieved through majority voting, while in regression problems, it is done through the arithmetic mean.

2.8 The main differences between Extra Trees and Random Forest are the following

1. In Random Forest, bootstrap replicas are employed, indicating that it subsamples the input data with replacement, whereas Extra Trees utilizes the entire original dataset.
2. These two algorithms diverge in their methods for selecting cut points to split nodes. Random Forest opts for the optimal split, whereas Extra Trees opts for random selection. Nevertheless, both algorithms choose the best split point from subsets of features. Consequently, Extra Trees introduces randomness to mitigate both bias and variance while still optimizing the outcome.
3. Using the complete original sample, as opposed to a bootstrap replica, diminishes bias, while the random selection of split points for each node diminishes variance.
4. When it comes to computational efficiency, the Extra Trees algorithm is quicker because it randomly selects split points instead of computing the optimal ones, leading to time savings during the process.

3.0 Related Research Literatures

3.1 Evaluating Machine Learning Prediction Techniques for Peer-to-Peer Lending Credit Default

The objective of this research is to develop a tree-based classification approach capable of forecasting the likelihood of loan default in peer-to-peer lending before granting approval. The study employed a binary PSO with SVM (BPSOSVM) for feature selection and utilized Extremely Randomized Tree (ERT) and Randomized Forest (RF) as classification models [14] . The findings of the study indicated that BPSOSVM can effectively identify a subset of features without affecting performance, and ERT demonstrates superior performance compared to RF across various metrics.

3.2 Utilizing Machine Learning and Artificial Neural Networks to Develop a Credit Scoring Model for P2P Lending: A Case Study Using Lending Club Data

The study introduced a customized credit-scoring model designed for P2P lending institutions, incorporating a variety of machine learning techniques and artificial neural networks (ANNs). To achieve this, researchers employed publicly available P2P loan data from Lending Club and conducted feature engineering to identify critical factors associated with loan defaults. They used XGBoost for data preprocessing and feature importance assessment [3]. The primary discovery of the research was that gradient-boosting decision tree methods, particularly XGBoost and LightGBM, outperformed traditional credit-scoring approaches such as ANN and LR, with XGBoost demonstrating the most remarkable results**.**

3.3 Research on Loan Default Prediction Utilizing the Random Forest Algorithm

This study introduces a model for predicting loan defaults by employing the random forest algorithm with Lending Club user loan data. The researchers highlight the susceptibility of P2P online lending platforms to the risk of user loan defaults, which can impact their long-term sustainability. To tackle the issue of class imbalance within the dataset, the SMOTE method was applied, alongside data cleaning and dimensionality reduction operations [19]. The findings indicated that the RF algorithm outperformed alternative machine learning techniques like logistic regression and decision trees when it came to predicting default cases.

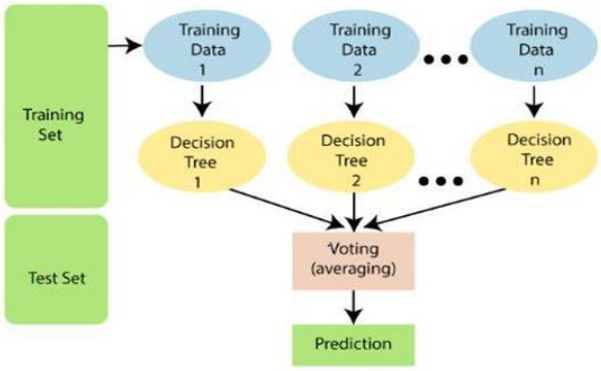
3.4 An Empirical Investigation into Models for Predicting Loan Default

In this research, the study performed a comprehensive review of the literature concerning the application of machine learning algorithms in assessing lending-related risks. They observed that lending plays a crucial role in the financial sector and is associated with substantial risks, specifically referred to as credit risk. In the assessment of creditworthiness, industry professionals and researchers utilize numerical ratings referred to as credit scores [2]. The study revealed that machine learning algorithms have gained growing popularity in quantifying and forecasting credit risk by analyzing an individual's historical data.

3.5 Comparative Analysis of Loan Default Prediction Using Decision Trees and Random Forest

The primary objective of this study was to create machine learning models capable of forecasting whether an individual should be approved for a loan, streamlining the loan selection process for banking authorities. The research involved a performance evaluation of the Random Forest and Decision Trees algorithms using the same dataset, and the results indicated that Random Forest exhibited greater accuracy in predicting loan eligibility [12]. Figure 2 illustrates a typical loan default prediction. The researchers underscored the escalating rate of loan defaults in the banking sector, which presents challenges in evaluating loan applications and mitigating the risk of loan default.

Figure 2: Loan Default Prediction



3.6 Machine Learning-Driven Systems for Credit Risk Prediction: A Comprehensive Review

This paper conducted an extensive examination of 76 research papers published over the last eight years, focusing on credit risk assessment through statistical, machine learning, and deep learning techniques. The authors introduced a novel approach for categorizing credit risk algorithms employing

machine learning and a methodology for ranking their effectiveness using publicly accessible data [15]. The study revealed that, in general, deep learning models outperform traditional machine learning and statistical methods when it comes to estimating credit risk, and ensemble techniques prove to be more accurate than individual models. However, the researchers also identified various challenges, including imbalanced data, inconsistent datasets, model transparency issues, and the underutilization of deep learning models.

3.7 Study on Efficiency in Credit Risk Prediction Using Logistic-SBM Model

Proposed a study developing a new method to predict credit risk in online loans using risk efficiency analysis. The researchers introduced a new approach of borrower risk efficiency, established risk efficiency characteristics, and conducted feature selection using a combination of logistic regression and slack-based measure framework [11]. The study discovered that the logistic-SBM model is more appropriate for credit risk prediction than the commonly use logistic approach, thus achieving efficient credit risk prediction based on the logistic SBM model.

3.8 Investigating Credit Scoring Through Integration of Social Media Details in Online Peer-to-Peer Lending

In their research, the scholars introduced a credit assessment framework that integrates information from social media through the utilization of decision tree analysis. They conducted their analysis using a dataset obtained from a prominent Chinese P2P lending platform to study loan default cases. Their findings demonstrated the model's remarkable proficiency in accurately categorizing default cases, as reported by [18]. The study emphasized that loan data, social media information, and credit history emerged as the most crucial variables in forecasting defaults.

3.9 Enhancing the Transparency of Machine Learning Credit Scoring Models in Peer-to-Peer Lending

This study underscores the importance of developing effective and transparent credit risk models in the context of P2P lending. While machine learning algorithms excel in prediction accuracy, they often fall short in providing explanations for their predictions. To address this limitation, the study recommends the utilization of interpretability tools such as SHAP values. A comparative analysis revealed that machine learning algorithms not only outperform other models in terms of classification accuracy but also offer superior transparency, allowing for the capture of dispersion, nonlinearity, and structural shifts in data. Consequently, the study's findings indicate that machine learning-based credit scoring models are both more precise and comprehensible, thereby instilling the trust required by P2P lending industries, regulatory bodies, and end-users.

3.10 Determinants of Default in P2P Lending

This study investigates the factors contributing to loan defaults in the context of peer-to-peer (P2P) lending, emphasizing the significance of this issue, especially because individual investors bear the credit risk in P2P lending, unlike traditional financial institutions. P2P lending platforms aim to reduce information disparities by providing lenders with borrower information and assigning grades to each loan. The research utilized data from Lending Club spanning the period from 2008 to 2014. It pinpointed loan purpose, annual income, current housing status, credit history, and debt levels as crucial elements in explaining loan defaults. The analysis was carried out using univariate statistical tests and survival analysis techniques, as reported by [4].

3.11 A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction

This research conducts a comparative analysis of statistical and machine learning techniques for bankruptcy prediction, highlighting the importance of a bankruptcy prediction model (BPM) in determining creditworthiness and preventing socio economic effects caused by inaccurate predictions. Statistical techniques are effective for large data sets, while machine learning techniques provide greater prediction accuracy for smaller data sets. Optimization techniques like GA and PSO can improve prediction accuracy for large data sets when integrated with machine learning [5]. A credit default prediction model was designed using historical loan data and categorical variables of loan status from the Prosper Lending Place dataset. The Random Forest algorithm was found to be superior in times of performance to the Decision Tree and Extra Tree Classifier algorithms, achieving a precision of 0.93% and an accuracy of 0.81%. The model successfully predicted the likelihood of default on new loan applications by recognizing borrower behavior patterns.

4.0 Proposed Work

This paper proposed a Credit default prediction model show in Figure 3 in the proposed model a unique confusion matrix, decision Tree confusion Matrix, Random forest confusion matrix as well as Extra Tree Confusion Matrix were used to depict the results in Table 1, 2., 3 as follows:

Credit default prediction system is proposed. In the proposed scheme, a confusion matrix displays the ratio of accurately classified and inaccurately classified items in each category, presenting a comprehensive view of the test results. The results are rounded to the nearest whole number, the proposed model is shown in Figure 3.

Figure 3: The Proposed Model

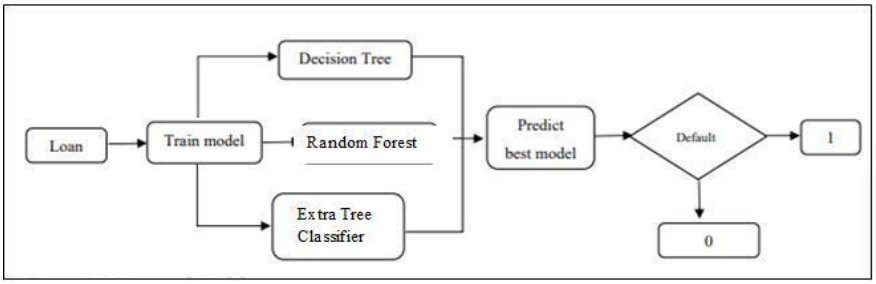


Table 1: Decision Tree Confusion matrix

```
# Printing the report
from sklearn import metrics
print(metrics.classification_report(expected, predicted))
```

	precision	recall	f1-score	support
0	0.28	0.30	0.29	5701
1	0.86	0.84	0.85	28481
accuracy			0.75	34182
macro avg	0.57	0.57	0.57	34182
weighted avg	0.76	0.75	0.76	34182

Table 2: Random Forest Confusion matrix

```
# Printing the report
from sklearn import metrics
print(metrics.classification_report(expected, predicted))
```

	precision	recall	f1-score	support
0	0.39	0.24	0.30	5701
1	0.86	0.92	0.89	28481
accuracy			0.81	34182
macro avg	0.63	0.58	0.60	34182
weighted avg	0.78	0.81	0.79	34182

Table 3: Extra Tree Classifier Confusion Matrix

```
# Printing the report
from sklearn import metrics
print(metrics.classification_report(expected, predicted))
```

	precision	recall	f1-score	support
0	0.28	0.29	0.29	5701
1	0.86	0.85	0.86	28481
accuracy			0.76	34182
macro avg	0.57	0.57	0.57	34182
weighted avg	0.76	0.76	0.76	34182

5.0 Conclusion

In the digital age, there is a growing interest in harnessing machine learning techniques to optimize profits within the financial sector, with a specific emphasis on areas like risk assessment, credit scoring, and bankruptcy prediction. This particular research project has set out to identify the specific characteristics of financial data that contribute to the likelihood of loans defaulting. By conducting thorough data analysis, the study explores the connections between various features and loan defaults to determine the most suitable attributes for training a predictive model. Subsequently, three distinct machine learning algorithms are put to the test, using both training and testing datasets, and essential performance metrics are employed to assess which algorithm yields the most accurate predictions for loan defaults.

References

1. Akitunde, A. (2020). Digital lending in Nigeria. Retrieved from https://www.linkedin.com/pulse/digital-lending-nigeria akitobi-akitunde-sfc-ssyb acib-2020.
2. Uzair, A., Tariq, A. H. I., Asim, S. & Nowshath, B. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience. 16*, 3483-3488. 10.1166/jctn.2019.8312.
3. Chang, A. H., Yang, L. K., Tsaih, R. H., & Lin, S. K. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model*:* A case using lending club data. *Quantitative Finance and Economics*, *6*(2), 303-325.
4. Serrano-Cinca, C., Gutiérrez-Nieto, B. & López-Palacios, L. (2015). Determinants of default in P2P lending. PLOS ONE, 10. e0139427. Retrieved from 10.1371/journal.pone.0139427.

5. Devi, S. & Yalavarthi, R. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing, 8*, 133-139. Retrieved from 10.18178/ijmlc.2018.8.2.676.
6. Ariza, M., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE access. Retrieved from 10.1109/ACCESS.2020.2984412.
7. Gupta, A. (2020). ML/ extra tree classifier for feature selection. Retrieved from https://www.digitalocean.com/community/tutorials/an introduction-to-machine learning.
8. IBM. (2021). CRISP-DM help overview. Retrieved from https://www.ibm.com/docs/en/spss modeler/saas?topic=dm crisp-help-overview.
9. JavaPoint. (2021). Decision tree classification algorithm. Retrieved from https://www.javatpoint.com/machine-learning-decisiontree-classification algorithm.
10. Kisutsa, G. T. (2021). Loan default prediction using machine learning: A case of mobile lending. University of Nairobi Digital Repository. Retrieved from http://erepository.uonbi.ac.ke/ handle/11295/155863
11. Li, D. & Li, L. (2022). Research on efficiency in credit risk prediction using logistic-SBM model. *Wireless Communications and Mobile Computing*. Retrieved from doi:10.1155/2022/5986295.
12. Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering, 1022*, 012042. Retrieved from doi:10.1088/1757-899X/1022/1/012042.
13. Alhasan, S., Ajayi, E. A., & Wisdom, D. D. (2020). A comparative performance study of machine learning algorithms, for efficient data mining management of intrusion detection systems. *International Journal of Engineering Applied Sciences and Technology, 5*(6), 85-110.
14. Setiawana, N., Suharjitoa, & Dianab. (2019). A comparison of prediction methods for credit default on peer to peer lending using machine learning. *4th International Conference on Computer Science and Computational Intelligence*, *157*, 38-45.
15. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk prediction systems: A systematic review. *Neural Computing and Applications (34)*, 14327- 14339.
16. Tagliaferri, L. (2017). An introduction to machine learning. Retrieved from https://www.digitalocean.com /community/tutorials/anintroduction-to-machine learning.
17. Yu, X. (2017). Machine learning application in online lending risk prediction. Retrieved from https://www.researchgate.net/publication/318488108_Machine_learning_application_in_online_lending_risk_prediction.
18. Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2019). Research on credit scoring by fusing social media information in online peer to peer lending. *Information Technology and Quantitative Management, 91*, 168-174.
19. Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *7th International Conference on Information Technology and Quantitative Management, 162*, 503-513.