Proceedings of the Second International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS-2024)
IEEE Xplore Part Number: CFP24DN7-ART; ISBN: 979-8-3503-6841-3

2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-6841-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSAS64001.2024.10760746

Leveraging Machine Learning and Forecasting Techniques to Enhance Credit Risk Analysis and Prediction

Satheeshkumar S
Assistant Professor & Department of Artificial intelligence and data science,
Bannari Amman Institute of Technology, Tamilnadu, India
sathishsoundarc@gmail.com

Dakshana M
Assistant Professor & Department of Physics, Sathyabama Institute of Science and Technology, Tamilnadu,
India drmdakshana@gmail.com

Gunalan K
Assistant Professor & Department of Computer Technology,
Bannari Amman Institute of Technology, Tamilnadu, India
info.kvguna@gmail.com

Anandan P
Assistant Professor & Department of Computer Technology,
Bannari Amman Institute of Technology, Tamilnadu, India
anandan@bitsathy.ac.in

Saveetha R
Assistant Professor & Department of Computer Technology,
Bannari Amman Institute of Technology, Tamilnadu, India
saveethar@bitsathy.ac.in

Nithya M
Assistant Professor & Department of Computer Technology,
Bannari Amman Institute of Technology, Tamilnadu, India
nithyam@bitsathy.ac.in

***Abstract*- This research presents a novel data-driven framework to improve credit risk assessment and minimize loan defaults. The framework leverages a combination of machine learning techniques, including Random Forest and XGBoost, to analyze a comprehensive dataset from Axis Bank spanning 2007-2015. The study focuses on key factors such as loan amount, debt-to-income ratio, interest rate, and debt burden to predict the likelihood of default. By employing oversampling techniques to address data imbalance and integrating ensemble methods, the framework enhances prediction accuracy. The results demonstrate the superior performance of the XGBoost model in accurately classifying loans as default or non-default. This research contributes to the advancement of credit risk management by providing a powerful tool for financial institutions to make informed lending decisions and mitigate financial risks.**

Keywords: Credit risk analysis, loan default prediction, machine learning (Random Forest, XGBoost, LSTM), forecasting, bank viability, informed lending decisions, risk management, loan amount, debt-to-income ratio, interest rate, and debt burden.

I. INTRODUCTION

This research involves an in-depth analysis of a loan dataset, focusing on loan details, borrower information, and payment records. The study includes data preprocessing, exploratory data analysis, and feature engineering. Techniques like SMOTE for oversampling and ensemble models such as Random Forest and XGBoost are applied for predictive modeling. This research aims to assess credit risk, enhance the accuracy of credit risk assessments, and support informed decision-making. The research significance lies in its contribution to risk assessment, operational efficiency, handling imbalanced datasets, and fostering innovation in financial data analysis.

II. LITERATURE REVIEW

Recent research highlights the superiority of deep learning in credit risk evaluation compared to traditional methods. This study addresses the gap by exploring deep learning ensemble algorithms for credit evaluation and tackls imbalanced credit data challenges. The proposed model combines an enhanced synthetic minority oversampling technique (SMOTE) with a deep learning ensemble classification method using long short-term memory (LSTM) and adaptive boosting (AdaBoost). Performance comparisons with widely used credit scoring models show the competitive edge of the proposed deep learning ensemble model in addressing imbalanced credit risk evaluation challenges [1].

Another study proposes an expert system called ESCRPSE, integrating the Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbor (ENN) to handle class imbalance. The ensemble bagging technique Extra Trees (ET) enhances predictive accuracy. Comparative analysis against single-classifier and ensemble models in the literature demonstrates the superior performance of ESCRPSE in achieving higher accuracy and f1-score [2].

A study focusing on China's Peer-to-Peer (P2P) lending industry analyzes 126,090 loan deals from RenRen Dai. In response to recent industry challenges, the research employs a stacking ensemble machine-learning model for predicting credit default probabilities in P2P lending. Feature selection using Max-Relevance and Min-Redundancy (MRMR) enhances model accuracy, while k-means clustering eliminates irrelevant features. The stacking ensemble model demonstrates high performance, excelling in prediction accuracy,

precision, and recall compared to single classifiers. The results highlight the effectiveness of the proposed model, confirming its efficiency through the area under the ROC curve [3].

To address challenges in credit risk assessment, a novel integrated learning approach tackles high- dimensional features and data imbalance. The method involves hybrid feature selection using Relief, maximum information coefficient, and Random Forest. An adaptive Borderline-SMOTE method is applied to balance minority samples, employing a new interpolation technique for improved sample distribution. Focal Loss enhances LightGBM's loss function, prioritizing minority and challenging samples for better classification accuracy. The modified algorithm is the base classifier, integrated using AdaBoost and random subspace methods to establish an effective credit risk prediction model. Comparative experiments demonstrate improved - mean and AUC values; showcasing enhanced default prediction performance [4].

Credit scores are pivotal for determining loan eligibility in the banking sector. Automating the identification of potential loan defaulters is crucial to mitigate credit risks. Employing machine learning techniques such as gradient boosting, random forest, feature selection, and decision trees aids in accurately classifying customers. The model effectively distinguishes between good and bad loan applicants, enhancing accuracy in customer data analysis for improved decision-making [5].

Credit risk, or default risk, involves a customer or counterparty's potential inability or unwillingness to fulfill commitments related to financial transactions. It encompasses transaction and portfolio risks (intrinsic and concentration risks). External factors like economic conditions and internal factors such as loan policies, administration, and risk pricing influence it. This research addresses credit risk assessment, mitigation, and monitoring [6].

III. PROPOSED METHODOLOGY

Research methodology is crucial for conducting and analyzing research systematically. It provides a framework for formulating research questions, collecting data, and interpreting findings. The chosen methodology outlines methods, procedures, and techniques to ensure the reliability and validity of outcomes. It involves decisions on research design, sampling, data collection, and analysis, enhancing the credibility and replicability of the study. Methodologies can vary from qualitative to quantitative or mixed-methods, depending on the study's nature, aiming to understand the research problem comprehensively. The methodology section is essential, guiding scholars and readers to understand the rigor and appropriateness of the methods used in knowledge pursuit.

A. PROBLEM STATEMENT

The main challenge for businesses is managing credit risk from loan applicants. Uncertainty about whether borrowers will repay on time can lead to financial losses. Accurately predicting creditworthiness is crucial to avoid adverse impacts on revenue and financial health. Developing and implementing predictive modeling techniques is vital to identifying and mitigating potential default risks and ensuring operational and economic sustainability.

B. STRATEGIES FOR IMPROVING CREDIT RISK ANALYSIS AND PREDICTION PERFORMANCE

To improve the performance of machine learning algorithms, optimization techniques such as hyperparameter tuning, cross-validation, and the use of ensemble methods like Random Forest and Gradient Boosting can be employed to effectively handle non-linear data. Achieving high accuracy can be facilitated by utilizing advanced models such as Deep Neural Networks (DNNs), XGBoost, or Long Short-Term Memory (LSTM) networks for time-series data, along with feature selection methods to eliminate noise and enhance model accuracy. Enhancing precision involves fine-tuning models based on precision-recall trade-offs, balancing datasets, and applying techniques like Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalances. For effective credit risk analysis, a combination of supervised learning for default prediction and unsupervised learning to uncover hidden patterns in borrower behavior can be utilized, integrating macroeconomic and behavioral data for a comprehensive assessment. Improving the prediction process can be achieved through the use of ensemble models and the integration of real-time data, allowing for dynamic updates and more accurate forecasts.

IV. EXPLORATORY DATA ANALYSIS

Data analysis and interpretation are core to this research, focusing on extracting insights from the dataset for credit risk assessment. This involves using statistical methods, machine learning, and forecasting to uncover patterns and trends. The goal is to transform raw data into actionable knowledge for informed decision-making. Interpreting the data involves understanding and contextualizing findings within the credit risk management framework. Rigorous techniques are used to ensure reliability and validity. Exploratory Data Analysis (EDA) is pivotal, offering a preliminary understanding of the data's characteristics and guiding further analyses.

A. DATA PREPROCESSING

Before conducting the exploratory data analysis (EDA), it is essential to perform several data preprocessing steps to ensure the accuracy and integrity of the analysis.

Step – 1: Identify and fill/remove missing values to prevent biased analysis.

Step – 2: Transform the data as required for analysis, such as normalization or log transformation.

Step – 3: Reduce the dimensionality of the data if necessary to improve efficiency and reduce complexity.

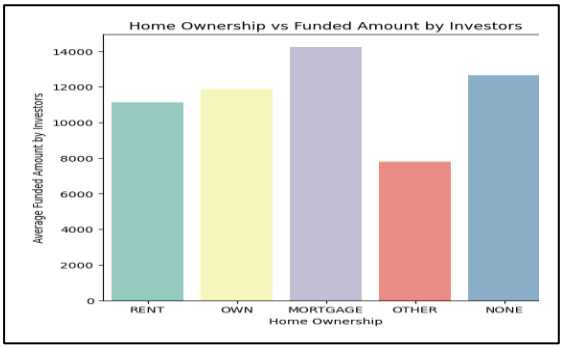
Step – 4: Split the dataset into training and test sets to accurately evaluate the model's performance.

Step – 5: Combine variables if they enhance the interpretability or predictive power of the model.

Step – 6: Convert categorical variables into numerical values for analysis.

Step – 7: Convert date columns into a standardized data format for consistency.

Step – 8: Ensure all columns have a consistent data type for proper analysis.



Loan Status	Sum of the funded amount
Default	Rs.29,90,40,575
Fully Paid	Rs.2,76,41,40,500
Grand Total	Rs.3,06,31,81,075

To increase transparency in model decision-making, Explainable AI tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be applied to make the processes more understandable for stakeholders. Additionally, optimizing preprocessing steps is crucial for ensuring clean and optimal data, which includes data normalization, missing value imputation, outlier detection, and dimensionality reduction. Lastly, achieving reliability involves building robust models through cross-validation, stress testing in diverse market conditions, and regularly retraining on new data to maintain reliability over time.

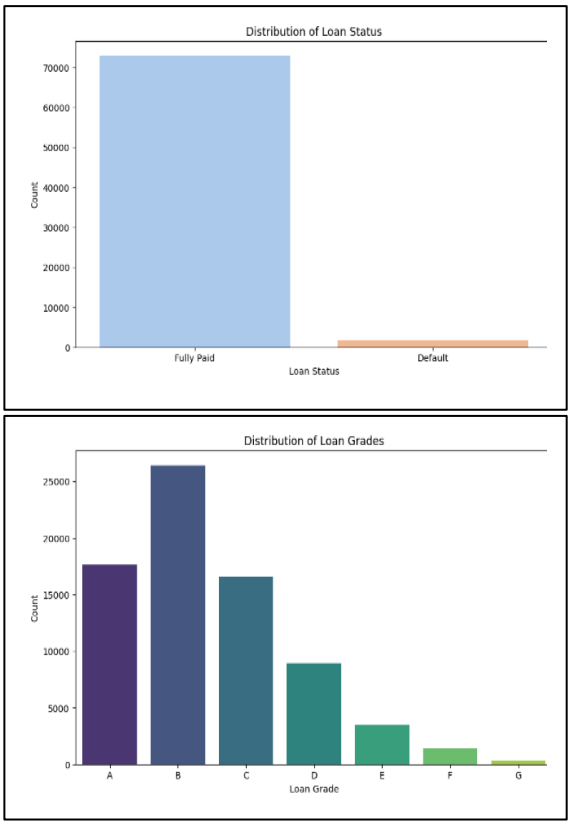


Figure -1 -Distribution loan status

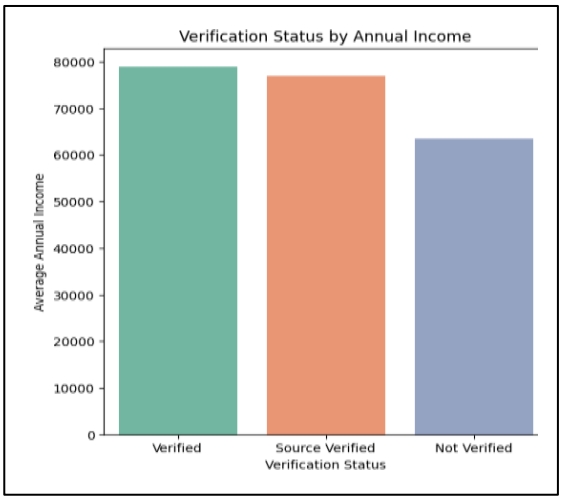


Figure 2- Verification status

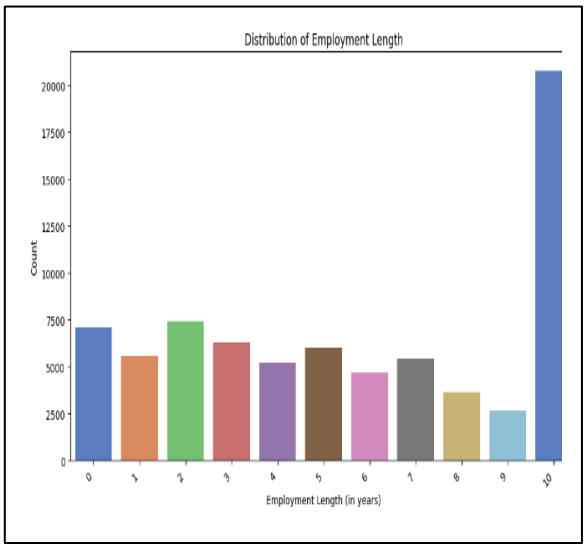
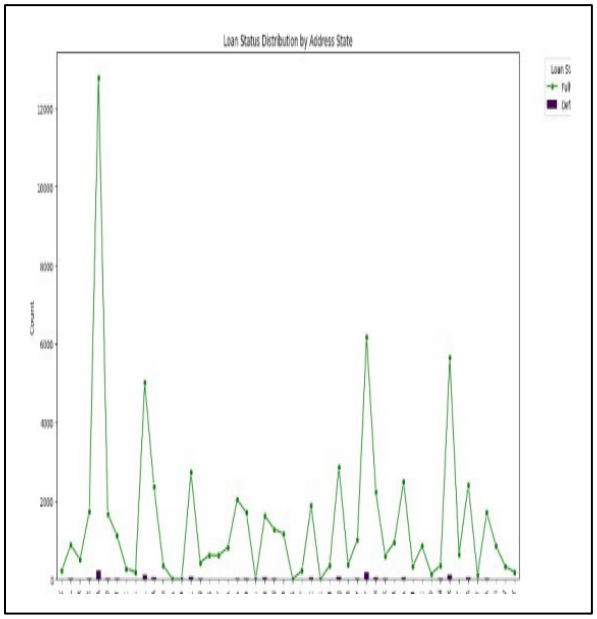


Figure 3 - Distribution of Employement



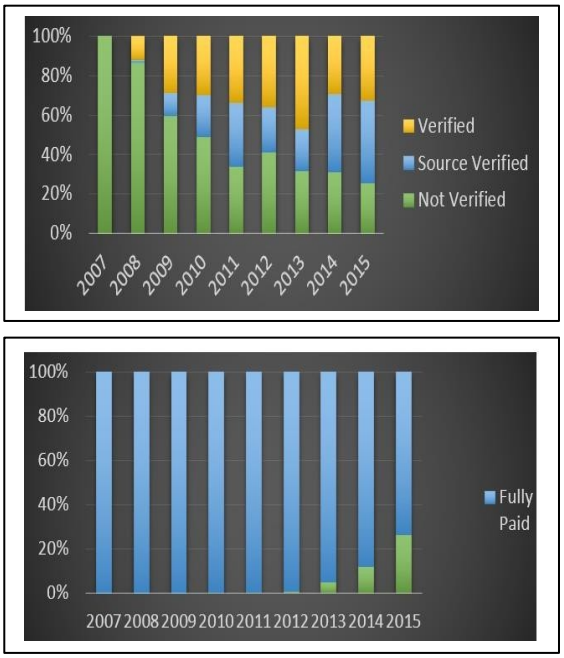


Figure 4- Analysis of Data

The analysis of the loan data reveals several key insights about borrower behavior and risk factors. Most loans were taken for debt consolidation, with many defaults in this category. A trend shows that higher delinquencies are associated with an increased likelihood of default, except for borrowers with consistent, timely payments. Over the years, there has been an increase in defaults, reaching over 25% in 2015. This is concerning, especially considering the reduced percentage of non-verified borrowers, indicating a stricter verification process. Examining borrower demographics, most of the bank's cream customers in CA, NY, TX, and FL are the highest contributors. The maximum length of employment for individuals seeking loans is ten years, suggesting workforce stability. A and B-grade customers show loyalty and have lower default rates. The prevalent loan term is 36 months, with fewer borrowers opting for 60 months. The verification status concerning annual income shows that most customers are verified, indicating a thorough verification process. Many borrowers have taken loans for mortgage purposes, suggesting a trend of using loans to address existing financial obligations.

V. MODEL BUILDING

Two crucial steps should be performed before fitting a model to the data: splitting and balancing the training dataset. For splitting the dataset, we use SMOTE for balancing the data. The dataset was systematically divided into an 80:20 ratio, with 80% allocated to the training set for model construction and the remaining 20% reserved as a validation or test set to evaluate model performance. Employing Ensembling techniques, we aimed to enhance model accuracy by conducting analyses with a 92% accuracy rate using the Random Forest model, While XGBoost gives 95% accuracy. Although both the

models performed well, XGBoost outperformed Random Forest.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.926934	0.927034	0.926934	0.926936
XGBoost	0.958453	0.95854	0.958453	0.958447

Following the completion of the model-building phase, we proceeded to conduct forecasting models to assess the potential credit risk associated with loans in the future. Leveraging daily data, we aggregated the information monthly and yearly, recognizing that both representations convey the same trends. The conversion of daily data into monthly data facilitated a more comprehensive

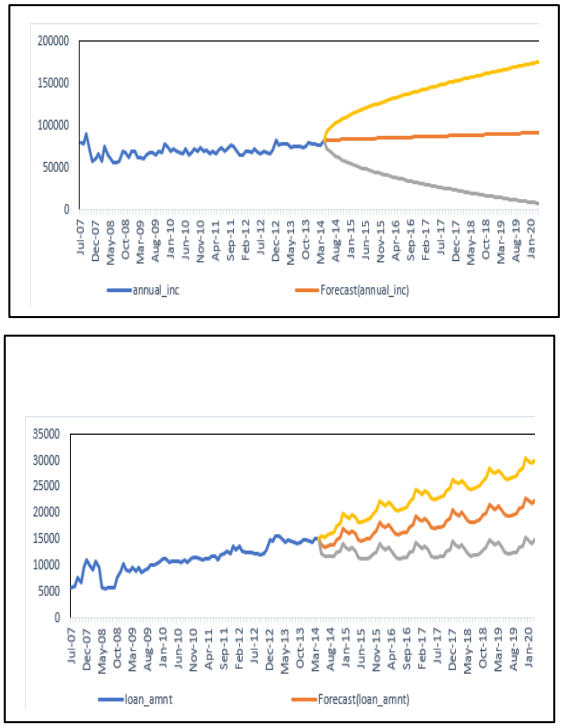


Figure 5 - Credit risk associated with loans

analysis. We selected five critical columns from the original dataset: loan amount, annual income of the customer, debt-to-income ratio, interest rate, and calculated debt (obtained by multiplying the debt-to- income ratio with annual income). The chosen variables are considered particularly significant in evaluating credit risk. Given the dataset's timeframe spanning from 2007 to 2015, our forecasting models enable predictions up to 2020. However, it's crucial to acknowledge that the model's predictive capacity is constrained by the historical data available. If additional data becomes accessible, incorporating it into the existing framework would extend the model's forecasting horizon to encompass upcoming years. This underscores the dynamic nature of the forecasting process, which is adaptable to evolving datasets for more accurate and insightful predictions. The LSTM model is built for forecasting.

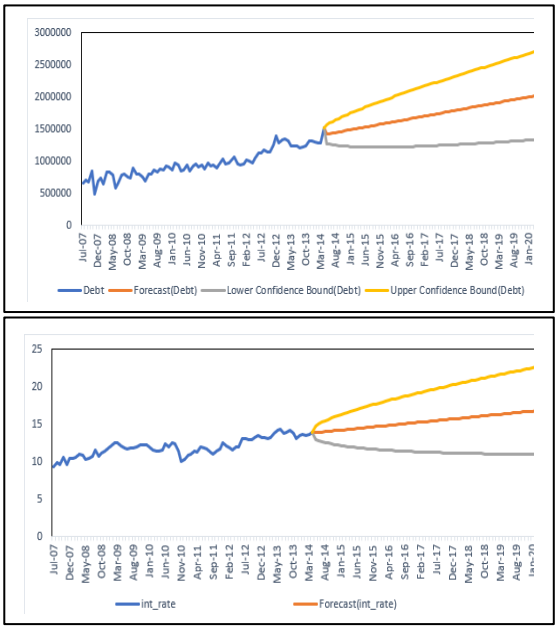


Figure 6 Relationship between the increase in loan amount

The relationship between the increase in loan amount, debt-to-income ratio, interest rate, loan amount, and calculated debt (the product of debt-to- income ratio and annual income) and their impact on the future credit risk of the bank is a critical aspect of financial analysis. Each of these factors plays a distinct role in shaping the overall credit risk profile of the bank.

1. Loan Amount: Higher loan amounts pose a greater risk to banks, as larger loans entail larger potential losses in the event of default. Increased loan amounts may lead to higher default rates if borrowers struggle to repay substantial sums, thus elevating credit risk.

2. Debt-to-income ratio: A higher debt-to-income ratio indicates that borrowers leverage a significant portion of their income to service existing debts. A high ratio suggests financial strain and raises concerns about the borrower's ability to manage additional debt obligations, thereby increasing credit risk.

3. Interest Rate: Rising interest rates can amplify credit risk by making loan repayment more expensive for borrowers. Higher interest rates increase the cost of borrowing, potentially straining borrowers' finances and increasing the likelihood of default, especially for variable-rate loans.

4. Debt: The relationship between loan amount and debt is crucial, as it determines the borrower's overall debt burden relative to their income. Increasing loan amounts contribute to higher total debt levels for borrowers. If borrowers' debt obligations exceed their income, it heightens the risk of default, adversely affecting the bank's credit risk profile.

Considering these factors collectively, an upward trend in loan amount, debt-to-income ratio,

interest rates, loan amounts, and debt can exacerbate credit risk for banks in the future. Banks must closely monitor these metrics and employ robust risk management strategies to mitigate potential risks and maintain a healthy credit portfolio.

VI. CONCLUSION

The comprehensive framework presented in this research enhances credit risk analysis and prediction, offering a proactive approach to minimize loan defaults and safeguard bank viability. By integrating machine learning and forecasting techniques, the framework provides a more accurate credit risk assessment, empowering financial institutions with actionable insights for informed lending decisions and effective risk management. This research demonstrates the effectiveness of ensemble learning techniques, particularly Random Forest and XGBoost, in enhancing prediction performance. The LSTM model for forecasting enables the identification of future credit trends, further strengthening the risk assessment process. Analyzing borrower behavior and risk factors provides valuable insights into credit risk dynamics, highlighting the importance of continuous monitoring and proactive risk mitigation strategies.

REFERENCES

[1]. Yin, Wei, Berna Kirkulak-Uludag, Dongmei Zhu, and Zixuan Zhou. "Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending." Applied Soft Computing 142 (2023): 110302.

[2]. Mehta, Amit, Max Neukirchen, Sonja Pfetsch, and Thomas Poppensieker. "Managing market risk: Today and tomorrow." McKinsey & Company McKinsey Working Papers on Risk 32, no. 1 (2012): 24-36.

[3]. Kokate, Shrikant, and Manna Sheela Rani Chetty. "Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model." International Journal of Safety and Security Engineering 11, no. 5 (2021): 565-572.

[4]. Shen, Feng, Xingchao Zhao, Gang Kou, and Fawaz E. Alsaadi. "A new deep learning ensemble credit risk evaluation model with an improved synthetic minoity oversampling technique." Applied Soft Computing 98 (2021): 106852..

[5]. Spuchľáková, Erika, Katarína Valašková, and Peter Adamko. "The credit risk and its measurement, hedging and monitoring." Procedia Economics and finance 24 (2015): 675-681.

[6]. Saha, Trishita, Saroj Kumar Biswas, Saptarsi Sanyal, Souvik Kumar Parui, and Biswajit Purkayastha. "Credit Risk Prediction using Extra Trees Ensemble Method." In 2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), pp. 1-8. IEEE, 2023..