International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-15 Issue-1, December 2025

# Predicting Bad Debt Risk in Banks using Machine Learning

**Chi Quynh Nguyen, Ngoc Thi Bich Do**

***Abstract*: This study presents the construction of a credit risk prediction model to improve the effectiveness of risk management at credit institutions. The urgency of the study is underscored by the internal bad-debt ratio of the Vietnamese banking system increasing by nearly 3.4 times by the end of 2023, while the cost of credit risk provisioning rose by 40% compared to 2022. The key challenge is to address a severe data imbalance (bad-debt accounts for 1-5%). Advanced data preprocessing techniques are applied, including handling missing values with the miceforest library and feature selection using Mutual Information combined with Correlation. The key experimental solution is the Mixture of Experts (MoE) Model, using Stratified K-Fold to train experts on 1:1-balanced data. The results show that the MoE model achieves the highest performance with a Recall of 0.87 and an F1-score of 0.79, outperforming the classical Machine Learning models. Applying the model achieves 85-90% forecasting accuracy, optimises the credit process, reduces appraisal time by 25-30%, and supports the sustainable development of the financial system.***

***Keywords: Bad Debt, Banking Finance, Credit Risk Prediction, Machine Learning.***

***Nomenclature:***
CAR: Capital Adequacy Ratio
ANN: Artificial Neural Network
SVM: Support Vector Machine
PD: Probability of Default
MICE: Multiple Imputation by Chained Equations
AUC: Area Under the Curve
SMOTE: Synthetic Minority Over-Sampling Technique

## I. INTRODUCTION

### A. Necessity of Research

Banking credit activities are key in providing capital for the economy. However, bad-debt risks remain a constant threat to the sustainable development of credit institutions. Several studies have focused on predicting bad debt risk in banks [7,8].

This study is based on several specific motivations and urgencies:

**Manuscript received on 27 November 2025 | First Revised Manuscript received on 05 December 2025 | Second Revised Manuscript received on 07 December 2025 | Manuscript Accepted on 15 December 2025 | Manuscript published on 30 December 2025.**

*Correspondence Author(s)
**Chi Quynh Nguyen***, Department of Computer Science, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Email ID: chinq@ptit.edu.vn, ORCID ID: 0009-0007-6197-2486
**Ngoc Thi Bich Do**, Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Email ID: ngocdtb@ptit.edu.vn, ORCID ID: 0009-0004-3250-0154, 0009-0006-2956-3132

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

*i. Increasing Bad Debt Ratio and Pressure on Risk Management:* By the end of 2023, the total bad debt on the balance sheet of the Vietnamese banking system was about 6.82% (equivalent to VND 682,621 billion), an increase of nearly 3.4 times compared to the end of 2022 (2.0%). The fact reflects serious challenges in credit risk management.

*ii. Pressure on Cost Management and Regulatory Compliance:* In banks, the trend of competition requires credit institutions to improve their risk management. Data from the Vietnam Banks Association shows that commercial banks' credit risk provisioning costs in 2023 increased by an average of 40% compared to 2022. In addition, Vietnamese banks must meet a minimum capital adequacy ratio (CAR) of 8%, as set out in Circular 08/2020/TT-NHNN and Basel II regulations. By the end of 2023, many banks faced pressure to maintain CAR amid rising bad debts.

*iii. Opportunities from Big Data and Digital Transformation:* The development of digital technology and the ability to process big data open up new opportunities in assessing and forecasting credit risks. According to the report, the big data analytics market in the global banking and finance industry is expected to grow at a CAGR of 23.5% from 2021 to 2026, from 29.87 billion USD in 2020 to 86.68 billion USD in 2026. According to a 2023 KPMG survey, 78% of banks in Vietnam are investing in data analytics and early warning systems for credit risk, up from 55% in 2022.

*iv. Expected Effectiveness of Forecasting Models:* Research and development of data-based risk forecasting models bring apparent effects, including:

- Reducing credit appraisal time by 25-30%.
- Save 15-20% of operating costs.
- Improve the accuracy of credit risk forecasting by up to 85-90%.
- Reduce the rate of newly arising bad debt by about 20-25%

With actual data and digital transformation trends, this research is of great significance both in theory and practice, meeting the urgent needs of credit institutions.

### B. Research Objectives

The research aims to develop an intelligent bad-debt forecasting model based on advanced machine learning and data analysis techniques to automate and optimise the credit risk assessment process. Specific objectives include:

*i.* Improve the accuracy in early identification of loans at risk of becoming bad debt.
*ii.* Reduce the time and effort of staff in appraisal and decision-making.


*© Copyright: All rights reserved.*

iii. Develop an automated decision support system capable of processing and analysing multiple data sources (financial, transactional, macro) simultaneously.

iv. *Support Long-Term Strategic Planning:* Create a powerful support tool in planning long-term risk management strategies for credit institutions. By forecasting early and accurately, the model will help managers gain an overview of the bad debt situation, thereby enabling timely and effective prevention and treatment measures. The application of technology in bad-debt management helps ensure the sustainable development of the financial-banking system, especially in volatile economic conditions.

### C. Research Objects and Scope

i. *Research Object:* The research object focuses on two main aspects: The first focuses on a dataset of loans with 124 attributes, including:

- Attributes on loan characteristics (loan value, term, interest rate, loan purpose, loan type, repayment method).
- Attributes on customer information (demographic index, personal finance and customer behaviour).
- Attributes on credit history (payment history, number of late payments, credit limit utilization rate).
- Financial indicators (debt-to-income ratio, solvency, cash flow).
- Derivative variables are created from combining and transforming original attributes to exploit the potential relationships in the data further.

The second aspect of the study is building and evaluating bad debt forecasting models, including:

- Machine learning algorithms used to build forecasting models (from traditional methods to advanced deep learning techniques).
- Pre-processing and selecting of important attributes from 124 initial attributes, to optimize model performance.
- Methods for evaluating model performance through indicators such as accuracy, sensitivity, specificity and area under the ROC curve.
- The ability to explain and interpret model results, ensuring transparency and reliability in the decision-making process.
- Practical application of the model in the credit risk assessment process at credit institutions.

ii. *Scope of Research*

This study analyses a dataset of 124 attributes from disbursed loans and provides the actual bad-debt status for each loan. We have built a bad-debt prediction model using machine-learning algorithms, focusing on the risk of personal-loan defaults.

We have applied effective machine learning methods to financial analysis, focusing on data pre-processing techniques and attribute selection that align with the dataset's characteristics. Then we have created a model that can be integrated into the current credit risk assessment process, proposing effective implementation and operation solutions.

### D. Research Methods

The research focuses on a theoretical overview, dataset analysis and processing (cleaning, handling missing values, attribute selection), design and construction of forecasting models, analysis and interpretation of forecasting results, and, finally, proposing solutions for deploying the model in practice.

The research uses a combination of three primary methods:

i. *Document Research Method:* Collecting and synthesizing domestic and foreign studies on credit risk forecasting, bad debt assessment models, and theories on credit risk management. This method helps to study fundamental theories, data analysis methods and refer to reports and statistics.

ii. *Mathematical Statistical Method:* Plays a key role in collecting, processing data and applying quantitative analysis methods such as logistic regression, decision trees, and neural networks to build forecasting models. Indicators such as accuracy, sensitivity, and specificity are used to assess the model's reliability.

iii. *Expert Method:* Implemented by consulting experts in the banking and finance sector on factors affecting bad debt risk. Experts will provide comments on the forecasting model and the variables used, and evaluate the model's feasibility and effectiveness in practice.

All the above research methods are systematically combined and complement each other to ensure the comprehensiveness and reliability of the research results. In particular, quantitative methods focus on building a forecasting model, while qualitative methods play a complementary role in understanding the problem and evaluating the model's feasibility. This combination helps the research not only have a solid theoretical basis but also ensures high practical applicability.

## II. BACKGROUND AND PROBLEM STATUS

### A. Theories Related to Credit Risk

i. *Concept of Credit Risk*

Financial institutions, especially banks, face credit risk when a borrower cannot fulfil their commitments to pay principal and/or interest under a debt contract. The reasons may stem from a reduction in the customer's financial capacity, economic fluctuations, or adverse legal and political factors.

Credit risk management plays a vital role in ensuring the safety and sustainable development of financial institutions. Sound credit risk management not only helps banks protect capital and optimise profits but also improves competitiveness. For the economy, it helps stabilise the financial markets and mitigate potential risks such as inflation or economic recession.

ii. *Related Theories [5]*

We mention here some of the critical economic and financial theories based on which credit risk can be studied as follows:

- Asymmetric information theory indicates that lenders often lack complete


© Copyright: All rights reserved.

information about borrowers. This leads to two main problems: Adverse selection and Moral hazard. Adverse selection occurs when high-risk customers are more easily accepted for loans under conditions of incomplete information. At the same time, moral hazard refers to the possibility that customers may change their behaviour, increasing their risk level after receiving a loan.

- Portfolio Management Theory emphasizes the importance of diversifying the credit portfolio to minimize overall risk. Banks can allocate loans to many customers and industries, limiting losses when one or several customers encounter financial difficulties.
- Signalling Theory has also been applied, in which borrowers can send signals about their creditworthiness through factors such as credit history, collateral or credit ratings.

### B. Credit Risk Mitigation Models

Many models and methods have been used to manage and mitigate credit risk, including traditional models and modern machine learning methods.

*i. Traditional Models:* Credit Scoring model, which uses quantitative and qualitative factors to assess the customer's ability to repay debt, while the Probability of Default (PD) model helps banks predict the possibility of customers not being able to fulfil their financial obligations within a specific period of time.

*ii. Modern Machine Learning Methods:* Logistic regression, Support Vector Machine (SVM), Artificial Neural Network (ANN) and XGBoost are increasingly applied to build credit risk prediction models.

These models can process large volumes of data and detect potential risk patterns with high accuracy, thereby helping automate credit risk management (thereby reducing operating time and costs).

This theoretical basis serves as a foundation for analysing the current state of credit risk management and for providing specific recommendations and solutions to improve its effectiveness in practice.

### C. Current Status

*i. The Limitations of Statistical Methods*

Statistical methods work well when relationships between independent variables are linear and straightforward. However, the data explosion of credit histories, real-time transactions, and customer profiles often involves many variables, interactions, nonlinearities, and complex relationships. As a result, statistical methods may not achieve high accuracy in credit risk forecasting [1]. In addition, when working with high-dimensional, large-scale data, they often require long processing times and low computational efficiency.

*ii. Data Imbalance Problem*

Credit risk forecasting faces a significant challenge due to severely imbalanced data [4]. The proportion of problem loans (bad debts) usually accounts for only a tiny part, about 1-5% of the total loan data.

Current methods for dealing with imbalanced data have many limitations:

- Under-sampling techniques risk losing important information from the majority group [9].
- Over-sampling methods [6] can lead to over-fitting, affecting the generalizability of the model.
- SMOTE techniques [3] do not always reflect the true nature of the bad debt problem.
- As a result, forecasting models often tend to be biased towards the majority group, leading to low accuracy in identifying high-risk loans.

## III. PROPOSED METHODOLOGY AND EXPERIMENTATION

The research solution is conducted on a real-world data set comprising 100,000 customer records from a credit institution. Each record corresponds to a customer who has been using credit products and services. The target variable is encoded as binary: value 1 indicates customers with bad debt, and value 0 indicates customers without bad debt.

### A. Data Preprocessing and Feature Selection

*i. Missing Values*

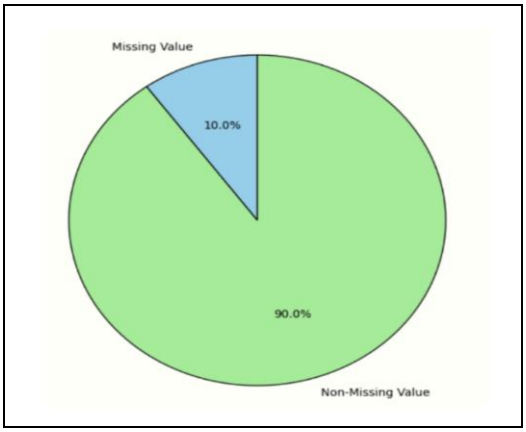
The original dataset has a missing values rate of 10%. To address this situation, the study uses the mice forest library (based on Multiple Imputation by Chained Equations (MICE) and Random Forests) to impute missing values. The mechanism of Miceforest makes it suitable for many types of data, including those with non-linear relationships, helping improve the model's accuracy (an average increase of 5-6%).

*ii. Feature Selection*

The original dataset contains 124 features. The selection of the most essential features is performed using Mutual Information in combination with Correlation. Mutual Information assesses the degree of dependence between variables, including complex relationships, not just simple linear relationships. By combining with Correlation, the study selects variables that are low-correlated with one another yet retain important information, thereby optimising the model's generalizability to unobserved data.

### B. Data Analysis and Visualization

Data visualisation analysis is used to identify trends and bad-debt thresholds by filtering features.

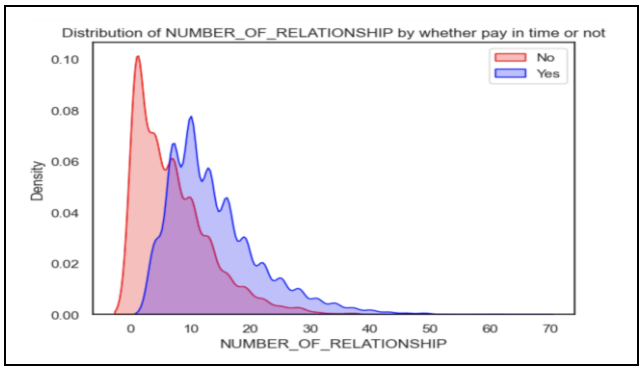



**[Fig.1: Ratio Between Bad Debt and Non-Bad Debt Customers in the Data Set]**

*i. Number of Financial Relationships:* The more financial relationships a customer has, the more

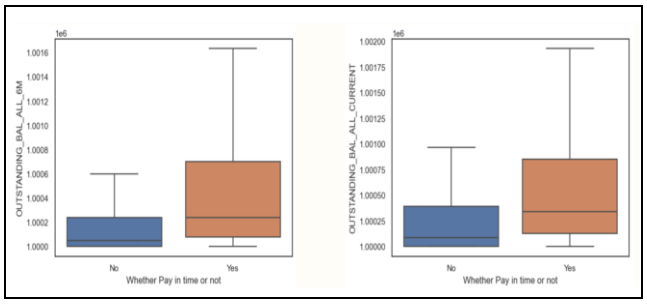

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

likely they are to pay on time. These customers are often responsible for managing their finances and have access to multiple sources of backup funding.




**[Fig.2: Number of Financial Relationships]**

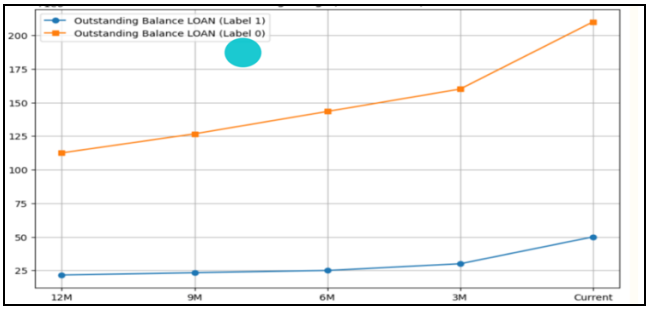
ii. *Total Outstanding Balance of Financial Products:* Outstanding Financial Products in the banking credit sector refers to the total amount of money that a customer (individual or organization) owes to a bank or financial institution for specific financial products. In both time periods (past 6 months and present), the larger the outstanding balance, the higher the likelihood of timely repayment. This requires sound financial management to maintain a credit rating and to secure easier access to new loans or other credit products. At the same time, we also see a rule in credit management: the higher the outstanding balance and the customer's stable credit history, the more likely they are to benefit from better credit conditions, such as lower interest rates or access to more loans.




**[Fig.3: Financial Out Standing Balance in the Previous 6 Months and Current]**

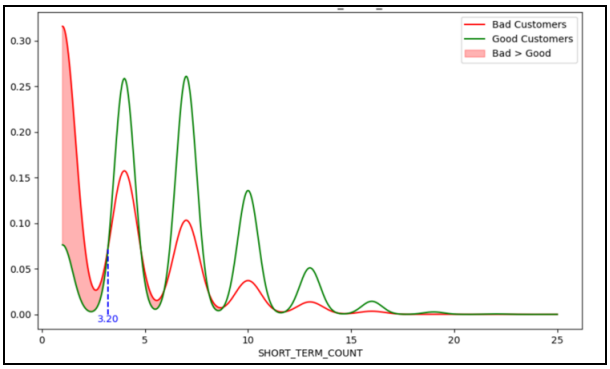
iii. *Average Outstanding Balance Loan:* The group that pays its debts on time usually has a high average debt balance and a strong growth trend, showing its reputation and high demand for loans.

The group that has overdue debts usually borrows smaller amounts of money. Their debt balance increases slowly; they have difficulty repaying the debt.



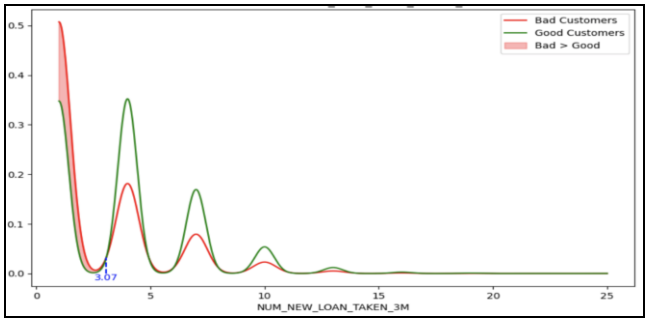

**[Fig.4: Value of Outstanding Balance Loan Over Time]**

iv. *Number of Short-Term Loans* (*SHORT_TERM_COUNT*): Short-term loans are generally defined as loans with a term of less than one year. These loans can include consumer credit and those that meet customers' temporary needs. The fact that customers have more short-term loans and can repay them on time indicates good debt management and financial stability. Their flexibility in managing short-term loans not only reflects an understanding of credit instruments but also demonstrates the ability to control personal spending and cash flow. Customers with a good repayment history on many short-term loans are often considered more trustworthy by financial institutions and can receive greater confidence in future credit decisions. From Figure 5, the bad debt threshold for this variable is≥1 and ≤3.2.




**[Fig.5: Distribution of Values of SHORT_ TERM_ COUNT in the Dataset]**

v. *Number of New Loans Taken 3 Months* (*NUM_NEW_LOAN_TAKEN_3M*): According to Figure 6, the newer loans borrowers have, the more likely they are to repay on time, which shows their ability to manage cash flow effectively. Having many new loans is not a sign of credit risk but rather a reflection of personal financial management, allowing customers to be flexible in using and reallocating financial resources. They may have developed solid personal financial strategies, such as closely monitoring their repayment schedules, using short-term loans to balance expenses, and having emergency plans in place. These factors not only help them ensure timely repayments but also build trust with financial institutions, which can lead to more favourable loan conditions in the future.



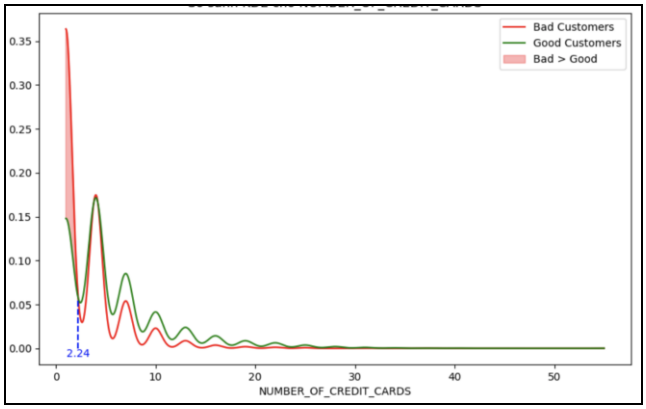

**[Fig.6: Value Distribution of New Customer Loans in the Dataset]**

*Retrieval Number: 100.1/ijitee.A120515011225*
*DOI: 10.35940/ijitee.A1205.15011225*
*Journal Website: www.ijitee.org*



*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*
*© Copyright: All rights reserved.*

vi. *Number of Credit Cards (NUMBER_OF_CREDIT_CARDS):* It seems that the more credit cards you have, the more likely you are to pay your debts on time. Contrary to popular belief that multiple credit cards indicate greater credit risk, when managed well, they are a positive indicator of an individual's financial capacity. Many credit institutions consider multiple cards a sign of financial reliability and stability, as maintaining them requires strict cash flow management and the ability to track expenses. Financial institutions can consider this factor when evaluating customers' credit more comprehensively, thereby making beneficial decisions, such as increasing credit limits or reducing loan interest rates.

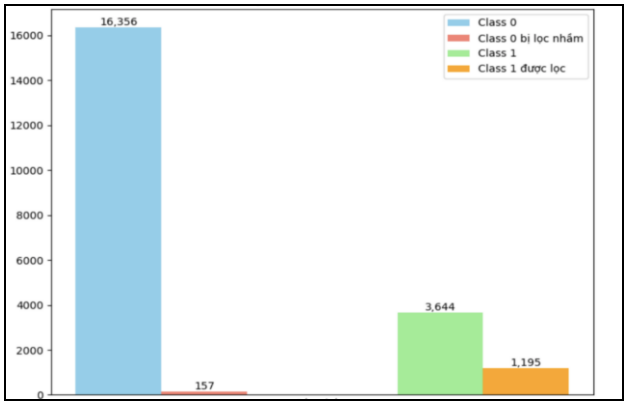



**[Fig.7: Value Distribution of Credit Card Numbers Owned by Customers in the Dataset]**

vii. *Preliminary Method of Filtering Bad Debt Customers:* Based on the analysis, the study has used some attributes to filter out customers with a high risk of bad debt according to the following criteria:

$1.00 \leq \text{NUM_NEW_LOAN_TAKEN_3M} \leq 3.07$,
$1.00 \leq \text{NUMBER_OF_CREDIT_CARDS} \leq 2.24$, and
$1.00 \leq \text{SHORT_TERM_COUNT} \leq 3.20$

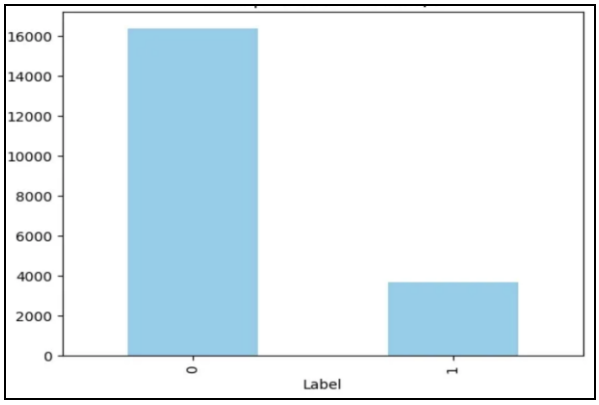
This method filtered out 33% of the total number of bad debt customers. The error rate for customers without bad debt was less than 1% of the total customer base. Customers without bad debt who were mistakenly filtered out are considered a high-risk group and should be included in the early warning and monitoring list.




**[Fig.8: Results Obtained After Performing the Filtering Method]**

## C. Machine Learning Model for Predicting Bad Debt

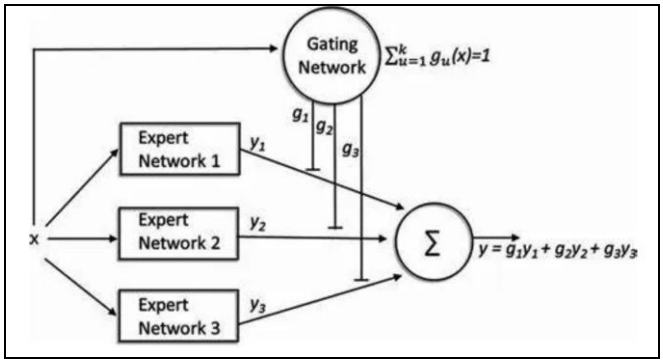
The goal is to detect label 1 (bad debt), which is outnumbered by label 0 (4 times).




**[Fig.9: Number of 2 Groups of Customers with Bad Debt and Without Bad Debt]**

i. *Mixture of Experts (MoE) Model*

The MoE model is used to address data imbalance. The MoE model architecture is shown in Figure 10. It consists of Expert Networks (individual neural networks, each expert processes a part of the input space) and a Gating Network (Coordinating Network, takes input x and outputs probability distributions g1, g2, ..., gn). The Gating Network decides which expert has more influence in decision-making.




**[Fig.10: Model Architecture]**

Then we apply MoE to address the imbalance. The data is divided into 4 parts, each containing 1 labelled sample paired with 0 labelled samples in a 1:1 ratio. Four Expert Networks are trained on these 4 balanced data sets.

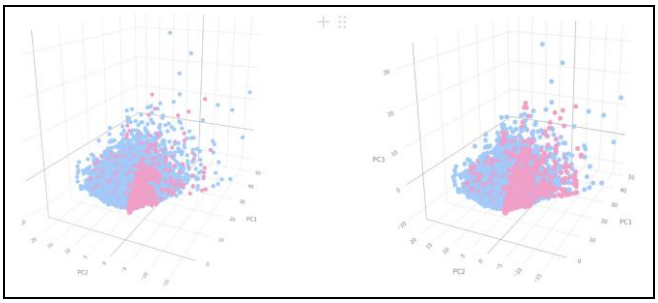
When training MoE, the process uses the Stratified-KFold method to divide the original data into 4 balanced parts. During MoE training, the weights of the Expert Networks are kept the same; only the weights of the Gating Network are updated to select the appropriate Expert model for the input data.

ii. *Combining Data Generation and Machine Learning Model*

This method is applied to overcome the limitations of the "black box" nature and high hardware resource requirements of the MoE model.

The study uses the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class (bad debt). SMOTE works by selecting a sample, identifying the K nearest neighbours, and then using an interpolation algorithm to generate new data points. Using SMOTE generates a more balanced dataset, allowing the model to learn features from both classes efficiently.


Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.



**[Fig.11: Original Data (Left) And Data After Generation (Right)]**

Then the machine learning models SVM, Logistic Regression, and XGBoost were trained on both datasets (original and SMOTE-generated). Finally, the optimal classification threshold was determined based on the trade-off between Recall and Specificity, aiming to maximise Recall (ability to detect bad-debt customers) while keeping overall accuracy at an acceptable level.

## IV. RESULTS ANALYSIS AND CONCLUSION

### A. Model Results Evaluation

Standard metrics used to measure the performance of a predictive model include: Accuracy, Precision, Recall, Area Under the Curve (AUC), and Gini Coefficient.

i. *Accuracy:* Measures the ratio of correct predictions to the total number of predictions.

ii. *Precision:* Measures the ratio of correct optimistic predictions to the total number of optimistic predictions.

iii. *Recall:* Measures the ratio of correct positive cases to the total number of positive cases (ability to detect bad debt).

iv. *AUC:* Measures the classification ability of the model, expressed as the area under the ROC curve.

v. *Gini Coefficient [2]:* Measures the deviation of the prediction distribution from a completely random distribution. A higher Gini value indicates a better classification model.

**Table I: The Index Results of the Prediction Models**

| Model | SMOTE | Accuracy | Precision | Recall | AUC | Gini | F1-Score |
|---|---|---|---|---|---|---|---|
| SVM | No | 0.81 | 0.73 | 0.39 | 0.82 | 0.64 | 0.51 |
| SVM | Yes | 0.85 | 0.77 | 0.66 | 0.87 | 0.74 | 0.71 |
| Logistic Regression | No | 0.83 | 0.75 | 0.41 | 0.83 | 0.66 | 0.53 |
| Logistic Regression | Yes | 0.86 | 0.73 | 0.76 | 0.89 | 0.78 | 0.74 |
| XGBoost | No | 0.90 | 0.80 | 0.60 | 0.91 | 0.82 | 0.68 |
| XGBoost | Yes | 0.91 | 0.82 | 0.71 | 0.93 | 0.86 | 0.76 |
| MoE | No (K-Fold Balanced) | 0.90 | 0.72 | 0.87 | 0.94 | 0.88 | 0.79 |

### B. Discuss the Results

i. *Impact of SMOTE*: The recall of all models increased significantly after applying SMOTE. This improvement was especially evident in SVM and Logistic Regression. The SVM recall increased sharply from 0.39 to 0.66, and the Logistic Regression recall rose from 0.41 to 0.76. This proves that SMOTE really helps improve the model's ability to detect bad-debt cases by adding minority-class samples to the training set, thereby reducing the model's bias towards the majority class.

ii. *GINI Index [2]:* The GINI index of SVM and XGBoost also increased when applying SMOTE, showing that the ability to distinguish between customers with and without bad debt risk was improved. This is especially important in financial risk problems.

iii. *MoE Performance:* The MoE model showed the highest Recall (0.87) and highest F1-score (0.79) among the compared models. MoE also achieved the highest Gini index (0.88) and highest AUC (0.94). This proves that MoE is highly effective at detecting bad-debt cases, especially when the data are unbalanced.

iv. *Model Conclusion*: The data balancing using SMOTE helped to reduce the impact of data imbalance, creating a more even dataset, thereby helping to overcome some of the limitations of the Mixture of Experts model, but the accuracy is still at an acceptable level. Therefore, it is possible to consider choosing between the two processing methods to suit the specific problem requirements. However, MoE remains the model with the highest predictive performance.

### C. Credit Scoring Model

In this section, we present a credit scoring method based on the output probability of the prediction model. After testing and consideration, the SVM model has been selected to develop a credit scoring system.

i. *Scoring Method*

To efficiently convert the model's output probability into a credit score, the team normalised the probability to a fixed scale ranging from 300 to 850. The normalization formula used is: $\text{Score} = 300 + (850-300) \cdot (1-P(\text{label}=1))$

Where:

- P is the default probability (probability of label 1) predicted from the model.
- (850−300) is the distance between the highest score (b) and the lowest score (a), allowing the probability to be converted into the desired credit score range.
- (1−P): The higher the P (higher risk of default), the lower the value of (1−P). Therefore, the credit score will decrease, reflecting the customer's higher risk. Conversely, the lower the P (lower risk of default), the higher the credit score will be.

ii. *Credit Rating*

Based on the distribution of the obtained scores, customers are divided into groups to have appropriate policies.

**Table II: Customer Score Distribution Results After Applying the Scoring**

| Point Range | Class 0 Rate | Class 1 Rate | Comments |
|---|---|---|---|
| 300.0 - 410.0 | 5% | 49.04% | Highest risk |
| 410.0 - 520.0 | 7.38% | 18.49% | High risk |
| 520.0 - 630.0 | 14.78% | 15.68% | Medium risk |
| 630.0 - 740.0 | 42.55% | 9.77% | Good credit |
| 740.0 - 850.0 | 35.29% | 22.7% | Excellent credit |

*Retrieval Number: 100.1/ijitee.A120515011225*
*DOI: 10.35940/ijitee.A1205.15011225*
*Journal Website: www.ijitee.org*



*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*
*© Copyright: All rights reserved.*

### D. Analyze the Differences Between the Two Groups

i. *Distribution of Customers in Class 0 (No bad debt):* The most significant proportion of Class 0 is in the range of 630.0 to 740.0 with 42.55%, followed by the range of 740.0 to 850.0 with 35.29%. This shows that the majority of customers in Class 0 have pretty good credit scores (above 630).

ii. *Distribution of Customers in Class 1 (Bad debt):* Class 1 has the highest proportion of customers in the range of 300.0 to 410.0, with 49.04%. This shows that the majority of customers in Class 1 have very low credit scores, indicating the possibility of default or poor credit.

iii. Class 0 is more concentrated in the high credit score ranges (a total of 77.84% above 630), while Class 1 is mainly focused in the low credit score range (below 410).

iv. The analysis above shows that Class 0 tends to represent customers with good credit scores, while Class 1 tends to include more customers at higher risk of default.

## V. CONCLUSION

We have successfully demonstrated that advanced machine learning models can significantly improve the ability to predict bad-debt risk (also credit risk) compared to traditional methods. With SMOTE to balance the data, the prediction accuracy is even higher. The evaluation results for Logistic Regression, SVM, XGBoost, and MoE indicate that the models achieve high forecasting accuracy, ranging from 85% to 90%.

It leads to some critical benefits to banks as follows:

A. Machine learning models can improve to early detect customers at risk of becoming bad debt in the future, thereby allowing banks to have time to implement timely interventions.

B. Help to minimize errors, increase efficiency and consistency of decisions in credit processes.

C. Better credit risk management will help banks increase their ability to pay bad debts, while increasing their competitiveness and profitability in the market.

D. However, implementing machine learning requires carefulness and adherence to data, security, ethics and regulatory principles.

## RECOMMENDATIONS

Based on the research results, we propose the following recommendations to help banks deploy machine learning in dire debt/credit risk prediction effectively and sustainably:

A. Banks need to prioritize investing in collecting, integrating and systematically managing an extensive and quality credit database, as this is a key factor for building accurate machine learning models.

B. Instead of using available models, banks should invest resources to develop customized machine learning models specifically designed for their business operations and data characteristics, to improve accuracy and practical applicability for dire debt prediction.

C. It is necessary to build processes to control the quality of input data, monitor the performance of models, and periodically evaluate to ensure safety, compliance and efficiency in operation.

D. Banks need to train employees on the basic concepts, advantages and limitations of machine learning. Hence, they have a better understanding of how to apply and manage models safely and effectively.

E. Banks should establish risk management policies, procedures and mechanisms that cover aspects such as data, technology, security and ethics, ensuring that machine learning applications are carried out responsibly and in compliance with regulations.

## DECLARATION STATEMENT

As the article's author, I must verify the accuracy of the following information after aggregating input from all authors.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

## REFERENCE

1. R. Bhandary, B.K. Ghosh. "Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods". J. Risk Financial Manag. vol 18, 23, (2025), DOI: https://doi.org/10.3390/jrfm18010023
2. V. Charles, T. Gherman, J.C. Paliza. "The Gini Index: A Modern Measure of Inequality". In Charles, V., Emrouznejad, A. (eds.), Modern Indices for International Economic Diplomacy. Palgrave Macmillan, Cham. (2022), DOI: https://doi.org/10.1007/978-3-030-84535-3_3
3. D. Elreedy, A.F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance". Information Sciences. Vol 505, pp 32-64,(2019), DOI: https://doi.org/10.1016/j.ins.2019.07.070
4. B. Krawczyk. "Learning from imbalanced data: open challenges and future directions". Prog Artif Intell, vol 5, pp. 221–232 (2016), DOI: https://doi.org/10.1007/s13748-016-0094-0
5. P. Koulafetis. "Modern credit risk management: Theory and practice". Publisher Palgrave Macmillan London, (2017), DOI: https://doi.org/10.1057/978-1-137-52407-2
6. M. Mujahid, E. Kına, F. Rustam et al. "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering". Journal of Big Data 11, 87 (2024), DOI: https://doi.org/10.1186/s40537-024-00943-4
7. A. Noriega, A. Rivera, và A. Herrera, "Machine Learning Models for Credit Risk Prediction: A Systematic Review". *Journal of Financial Engineering, 35(1)*, pp. 76-89, (2023), DOI: https://doi.org/10.3390/data8110169
8. B. Siddharth, L. Mohan, and Y.R. Reddy. "Machine learning techniques for credit risk evaluation: a systematic literature review". Journal of Banking and Financial Technology 4, no. 1, pp. 111-138, (2020),


*© Copyright: All rights reserved.*

DOI: https://doi.org/10.1007/s42786-020-00020-3
9. Z. Sun, W. Ying, W. Zhang, S. Gong. "Undersampling method based on minority class density for imbalanced data". Expert Systems with Applications, Vol 249, Part A, 123328, (2024), DOI: https://doi.org/10.1016/j.eswa.2024.123328

## AUTHOR'S PROFILE



**Chi Quynh Nguyen** graduated with a Bachelor of Science in Computer Science from Hanoi University of Technology, Vietnam, in 1999, with summa cum laude, and then received a Vietnamese Government Fellowship to earn a Master of Science in Computer Science at the University of California, Davis, USA, in 2004. Then she became Ph.D Candidate in Computer Science at the same University in 2006. Since 2008, she has been a senior lecturer in the Faculty of Information Technology at the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Her primary research focuses on data warehousing, data mining and bioinformatics, Mobility prediction, self-configuration of MANETs, and data aggregation methods in sensor networks.



**Ngoc Thi Bich Do,** in 2004, earned her Bachelor of Science in Information Technology from the University of Science and Technology. Subsequently, she successfully earned her Master's degree in Computer Science from the University of Hanoi in 2007. In 2010, she successfully earned a PhD degree from the Japan Advanced Institute of Science and Technology, specialising in Information Science. Since 2013, she has been an esteemed lecturer within the Faculty of Information Technology at the Posts and Telecommunications Institute of Technology. Her research interests are software testing, formal methods, numerical analysis, data mining, and machine learning.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.