Jurnal Perspektif
Vol. 10 No. 1 Mei 2026
Page 40-54

A Comparative Analysis of Boosting and Transformers Models For Loan Default Risk Prediction

Sandi Salvan Nuraliyudin[1*], Wiranto Herry Utomo[2]

[1,2]*Master of Informatic, President University*

Jl. Ki Hajar Dewantara, Kota Jababeka, Cikarang Baru, Bekasi, Indonesia

**sandi.nuraliyudin@student.president.ac.id*, wiranto.herry@president.ac.id

Received: 25 Agustus 2025; Accepted: 20 Januari 2026; First Available Online 23 Januari 2026; Published: 20 Mei 2026

DOI:10.15575/jp.v10i1.394

Abstrak

Pertumbuhan pesat pasar kredit global dan nasional meningkatkan akses pembiayaan bagi konsumen dan UMKM, namun juga memperbesar risiko gagal bayar yang dapat mengancam stabilitas keuangan. Di Indonesia, rasio kredit bermasalah naik dari 2,5% (2022) menjadi 3,1% (2024), sementara perkembangan fintech lending turut memperluas risiko tersebut. Penelitian ini bertujuan membandingkan kinerja tiga algoritma boosting (XGBoost, LightGBM, CatBoost) dan model deep learning berbasis Attention (Transformer) dalam memprediksi risiko gagal bayar pinjaman. Dataset terdiri dari 255.347 baris dan 18 variabel, melalui tahap pra-pemrosesan berupa pembersihan data, penanganan nilai hilang, deteksi outlier, serta penyeimbangan kelas menggunakan SMOTE, TomekLinks, dan kombinasi keduanya. Evaluasi dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC. Hasil menunjukkan model boosting memiliki akurasi tinggi (hingga 88,68% pada CatBoost dengan TomekLinks), tetapi bias terhadap kelas mayoritas. Sebaliknya, Transformer unggul pada data tidak seimbang, dengan Recall 70,22% dan F1-Score 31,49% pada SMOTE-TomekLinks. Analisis SHAP mengidentifikasi usia, suku bunga, lama bekerja, pendapatan, dan jumlah pinjaman sebagai fitur paling berpengaruh. Kesimpulannya, Transformer dengan SMOTE-TomekLinks merupakan model paling efektif dalam mendeteksi debitur berisiko gagal bayar.
Kata Kunci: Resiko gagal bayar Pinjaman; SMOTE; TomekLinks; Boosting; Transformer.

Abstract

The rapid growth of global and national credit markets has increased access to financing for consumers and MSMEs, but has also increased the risk of default, which can threaten financial stability. In Indonesia, the non-performing loan ratio rose from 2.5% (2022) to 3.1% (2024), while the development of fintech lending has also increased this risk. This study aims to compare the performance of three boosting algorithms (XGBoost, LightGBM, CatBoost) and an Attention-based deep learning model (Transformer) in predicting loan default risk. The dataset consists of 255,347 rows and 18 variables, and it underwent preprocessing stages such as data cleaning, handling missing values, outlier detection, and class balancing using SMOTE, TomekLinks, and a combination of both. Evaluation was carried out using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. The results show that the boosting model achieves high accuracy (up to 88.68% with CatBoost and TomekLinks), but is biased towards the majority class. In contrast, Transformer excels on imbalanced data, with a Recall of 70.22% and an F1-Score of 31.49% over SMOTE-TomekLinks. SHAP analysis identified age, interest rate, length of employment, income, and loan amount as the most influential features. In conclusion, Transformer with SMOTE-TomekLinks is the most effective model in detecting debtors at risk of default.
***Keywords**: Loan Default risk; SMOTE; TomekLinks; Boosting; Transformer.*

A. Introduction

Rapid global credit growth over the past two decades has increased access to financing but also raised the risk of default. According to (Fitch Ratings, 2025), The global leveraged loan market peaked at USD 1.337 trillion in 2024, up from USD 985 billion in 2017. Repricing and refinancing accounted for 84% of total transaction volume in that year. Furthermore, there was a sharp increase in sustainable loans, reaching USD 1.740 billion, up 12% year on year. (Sharpe, 2025). In Indonesia, despite relatively stable credit growth of around 8% per year, the non-performing loan (NPL) ratio increased from 2.5% in 2022 to 3.1% in 2024. Fintech lending is also growing rapidly in Indonesia. More than 100 online lending platforms have been registered and supervised by the Financial Services Authority (OJK), as small, unsecured, and short-term loans tend to be higher risk. This situation underscores the urgency of developing a more accurate and adaptive credit risk prediction system. (OJK, 2024).

A loan default occurs when a borrower fails to make their installment payments according to the agreed schedule. This condition significantly impacts the liquidity and solvency of financial institutions, leading to higher borrowing costs and eroding market confidence. At the macro level, high default rates can slow economic growth and create systemic risks that threaten international financial stability. Amid digital disruption and the increasing complexity of financial data, traditional credit assessment approaches are becoming less effective. Therefore, financial institutions are now adopting machine learning-based approaches to improve the accuracy of credit risk predictions, particularly default risk. This technology can efficiently process large volumes of data and identify risk patterns that conventional statistical approaches cannot detect. (Bello, 2024)(Soomro et al., 2024).

Although XGBoost, LightGBM, and CatBoost have been shown to be effective in many studies, very few have systematically compared the three in the context of credit risk, accounting for imbalanced data handling and hyperparameter optimization. (Poernamawatie et al., 2024). Most studies use only one model, without considering optimal configurations or data-balancing techniques that suit the dataset's characteristics. A comprehensive comparison of these three models is essential to provide objective insights into their advantages and disadvantages in real-world scenarios. The use of machine learning, especially boosting algorithms such as XGBoost, LightGBM, and CatBoost, has opened up new opportunities to improve prediction accuracy. These three algorithms have superior performance and high efficiency in processing complex data. (Nguyen & Ngo, 2025)(Akinjole et al., 2024) (Noriega et al., 2023).

Recent developments in deep learning have introduced the transformer architecture, originally designed for natural language processing but now being adapted for tabular data and financial applications. With its self-attention mechanism, the transformer can capture complex feature interactions and offers competitive performance compared to boosting methods. (Hu, n.d.) (Korangi et al., 2023). However, the application of transformers for predicting defaults is still limited and has not been studied in depth.

One of the main challenges in predicting default risk is data imbalance, where the number of customers who default is much smaller than those who pay on time. (Zhao et al., 2024) (Aftab & Matloob, 2019). This problem leads to bias in model training, where algorithms tend to ignore minority classes, even though these groups are crucial for accurate risk detection. Furthermore, the complexity of

financial data, which often contains missing values, multicollinearity between variables, and the dominance of categorical features, necessitates the use of adaptive and robust predictive models. (Dube & Verster, 2023)(Zhang et al., 2023).

The research gap is evident in the limited number of studies that directly compare boosting algorithms with transformer models for default prediction. This approach is crucial for providing a more objective overview of the advantages and limitations of each model, particularly when dealing with complex and unbalanced financial data. The novelty of this research lies in the comparative analysis of the boosting algorithms XGBoost, LightGBM, and CatBoost with transformers, complemented by the SMOTE and TomekLinks data balancing techniques. This study aims to conduct a comparative analysis of the performance of the three Boosting algorithms and Attention-Based Models (Transformers) in predicting loan default risk, considering the aspect of handling unbalanced data, and to provide recommendations for the best and most optimal model for use in credit risk assessment systems in financial institutions, particularly fintech.

The novelty of this research lies in a comprehensive comparison of three boosting algorithms (XGBoost, LightGBM, CatBoost) and an Attention-based deep learning model (Transformer) for predicting loan default risk using Indonesian credit data. This study also integrates the hybrid data-balancing technique SMOTE–TomekLinks into an end-to-end pipeline to improve minority-class detection. Furthermore, this study applies SHAP-based model interpretability to identify the most influential factors in default risk, thereby making novel contributions to the development of more accurate, transparent, and adaptive credit risk prediction systems.

B. Methods

This stage details the framework and methodological steps taken to achieve the research objectives. The methodology is designed to ensure that each stage, from experimental design, data collection and preparation, imbalance management, and model training and evaluation, is reproducible and the results are well-interpretable.

1. Research design

This section describes the comparative quantitative experimental framework used to assess the performance of three boosting algorithms and Attention-Based Models (Transformers) in predicting loan default risk. The research design is shown in Fig. 1.

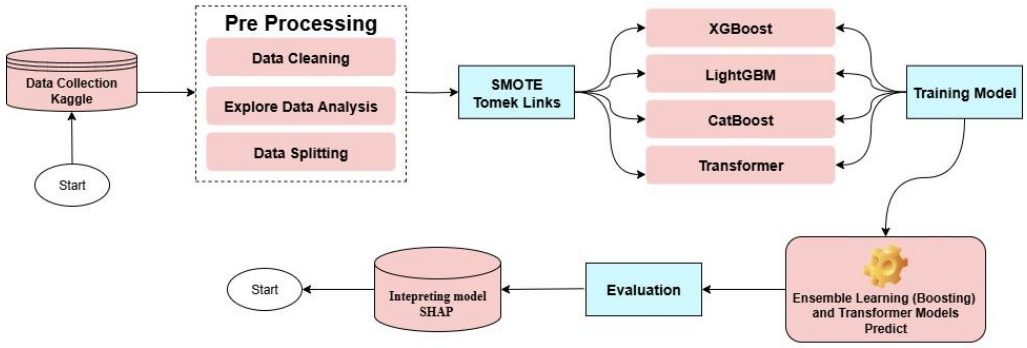

Figure 1. Flow Chart Research Design

The figure above illustrates a structured flow for loan default risk prediction. The process begins with data collection from Kaggle, followed by a pre-processing stage that includes data cleaning,

exploratory data analysis, and data splitting. Next, data balancing is performed using the SMOTE and Tomek Links methods to address class imbalance. Balanced data is used for model training with four algorithms: XGBoost, LightGBM, CatBoost, and Transformer. After the model is trained, performance evaluation and interpretation of results are performed using SHAP to identify the features most influential on default risk

2. Data Collection and deskription

This dataset, taken from Coursera's Loan Default Prediction Challenge, will provide authors with the opportunity to solve one of the most relevant machine learning problems in the industry with a unique dataset that will test their modeling skills. The dataset contains a total of 255,347 rows and 18 columns, or variables. Each row represents a single loan application, including borrower demographics, financial condition, loan parameters, and final repayment status.

Table 1. Deskription Variables

Variables	Definition
LoanID	A unique code for each loan application, serves as the primary key in data analysis.
Age	The borrower's age at the time of applying for the loan, measured in years reflects financial maturity and potential risk.
Education	The highest level of education attained, such as High School, Diploma, Bachelor's, Master's, or Doctorate. This variable helps describe the customer's financial literacy skills.
MartalStatus	The customer's marital status, including Single, Married, Divorced, and Widow/Widower, which can affect family responsibilities and financial stability
EmploymentType	Employment status category, such as Full-time, Part-time, Self-Employed, or Unemployed, which reflects the stability of income sources
MonthsEmployed	The length of time employed by the current company or business, measured in months the higher the number, the more consistent the employment history
Income	The borrower's annual income in millions of rupiah, which is the basis for calculating repayment capacity and determining interest rates
LoanAmount	The principal amount requested, measured in millions of rupiah illustrates the scale of the need for funds
InterestRate	The annual interest rate charged, expressed as a percentage, reflects the cost of borrowing
LoanTerm	The repayment period of the loan, measured in months; the longer the term, the greater the risk of changing economic conditions
LoanPurpose	The purpose for which the funds are used, such as Debt Consolidation, Home Improvement, Education, or Business, which may affect repayment priority
CreditScore	A standard credit score (e.g., 300–850) from a credit bureau, which reflects the borrower's historical repayment reputation.
NumCreditLines	The number of active lines of credit (credit cards, other loans) held the more, the more complex the debt management.
DTIRatio	Debt to-Income Ratio, which is the percentage of total monthly payments to monthly income, indicating the debt burden relative to income
HasMortgage	An indicator of mortgage ownership (1 = yes, 0 = no), which adds collateral to the loan.

Variables	Definition
HasDependents	An indicator of family dependents (1 = yes, 0 = no), which affects repayment capacity
HasCoSigner	An indicator of whether there are co-guarantors (1 = yes, 0 = no), which reduces the lender's risk.
Default	The loan outcome label (1 = default, 0 = non deafault), which is the primary dependent variable in the model.

3. Data Pre Processing

In any data analysis or machine learning project, newly collected raw data is rarely perfect. Data often contains incompleteness, inconsistencies, and noise, which can significantly degrade model performance. Therefore, data preprocessing is a crucial step that lays the foundation for building accurate and reliable models. (Gupta et al., 2024).

a) Data Cleaning

The first stage of preprocessing is data cleaning. This is a crucial step in the machine learning project lifecycle, ensuring data quality and reliability. Raw data from any source often contains errors, inconsistencies, missing values, or non-standard formats, all of which can significantly degrade the performance of predictive models.

- **Handling Missing Values**

  Missing values are a common problem in real-world datasets. Their presence can introduce bias into analysis, reduce the sample size available for training, and even cause some machine learning algorithms to fail. (Hidayaturrohman & Hanada, 2024).

- **Deleting Duplicates**

  Duplicate data is an exact copy of an existing data entry. The presence of duplicates can introduce bias into the model, overweighting certain observations and resulting in overly optimistic estimates of model performance. (Mintoo, 2025).

- **Handling Outliers**

  An outlier is a data point that differs significantly from the majority of other observations in a dataset. While not necessarily an error, outliers can significantly impact descriptive statistics (e.g., the mean) and are sensitive to certain model assumptions, potentially impairing the performance of a machine learning model. Use Z-scores (for normally distributed data) or quartile-based methods such as the Interquartile Range (IQR). (Chang et al., 2024).

b) Explore Data Analysis

Exploratory Data Analysis (EDA) is a fundamental phase in machine learning research methodology, which aims to understand the main characteristics of a data set through various visualization techniques and descriptive statistics. (Huang et al., 2025).

- **Descriptive Statistics**

  he first step in EDA is to calculate summary statistics for each feature in the data set. (Chang et al., 2024). For numeric features, the metrics to be calculated include:

  - Mean: The most common measure of central tendency.
  - Median: The middle value of sorted data, less sensitive to outliers than the mean.
  - Mode: The value that appears most frequently in the data.
  - Standard Deviation: Measures the spread or dispersion of data relative to the mean.
  - Quartiles (Q1, Q2/Median, Q3): Divides the data into four equal parts, providing insight into the distribution and detection of outliers through the interquartile range (IQR).
  - Minimum and Maximum Values: Shows the full range of each feature.

- **Data Visualization**

  Visualization is a powerful tool in EDA because it allows researchers to see patterns and anomalies that might not be apparent from statistical figures alone. Various types of graphs and plots will be used : Bar Charts and Pie Charts. (Mallinguh & Zoltan, 2022).

c) Data Splitting

Data splitting is the process of dividing a data set into distinct subsets with the goal of separating the data used in model training (the training set) from the data used in model performance evaluation (the validation set and the test set). The ultimate goal is to train a model on one data set and then test its performance on another data set the model has never seen before. This is crucial for measuring how well the model can generalize to new data and for avoiding a common problem called overfitting (when the model is too familiar with the training data but poorly able to predict new data). Data set splits typically follow a certain ratio:

- Training Set: The majority of the data (e.g., 80%) is used to train the model. The model will learn patterns, relationships, and features from this data.
- Test Set: The remaining data (e.g., 20%) is used for final evaluation after the model completes training. The model's performance on this set provides an estimate of how well the model will perform in the real world.

d) Handling Data Imbalance

In the context of loan default risk prediction, the "defaulting" class is typically a minority, often less than 15% of the total customer population, while the "non-defaulting" class dominates. This imbalance can bias machine learning models toward the majority class, resulting in low-risk borrower detection capabilities. (Sun, 2025).

- **Synthetic Minority Over Sampling Technique (SMOTE)**

  Creates synthetic minority samples by interpolating between nearby minority points. For each minority sample Xi, selected one of the nearest neighbors Xzi then synthetic point Xnew produced. (Imani et al., 2025)

$$\mathbf{Xnew = Xi + \lambda \times (Xzi - Xi),\ \lambda \sim u(0,1)}$$

- **TomekLinks**

  Tomek Link (T-Link) is an under-sampling method developed by Tomek. Tomek Link is considered a refinement of the Nearest-Neighbor Rule (NNR). (AT et al., 2016). The T-Link method can be used as a directional under-sampling method, removing observations from the majority class.

4. Forecasting Methods Based On Boosting And Transformer Algorithms Models

The modeling phase is a key element of this Research process, in which algorithms are used to develop predictive models of loan default risk. This Research adopts a boosting algorithm, an ensemble learning method that combines multiple weak learners into a robust predictive model. It also adopts a deep learning approach, using the transformer architecture, originally designed for natural language processing but now being adapted for tabular data and financial applications. With its self-attention mechanism, the transformer can capture complex interactions between features and offers significant potential.

a) XGBoost

XGBoost is one of the most popular and widely used boosting algorithms due to its high efficiency and ability to handle large datasets. XGBoost is an implementation of the Gradient Boosting Decision

Tree (GBDT) with several enhancements, including regularization, parallelization, and memory optimization. This algorithm builds decision trees iteratively, with each new tree attempting to minimize the error of the previous one. (Yang et al., 2025)

b) LightGBM

LightGBM is a boosting algorithm developed by Microsoft that aims to achieve high computational efficiency and good accuracy, especially for large-scale and industrial datasets. LightGBM builds trees based on leaves (as opposed to XGBoost which is based on levels), making it faster but more susceptible to overfitting. LightGBM also uses Gradient Based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce the amount of data and features processed.(S. Li et al., 2024).

c) CATBoost

CatBoost is a gradient-based boosting algorithm developed by Yandex, specifically designed to efficiently handle categorical data without the need for preprocessing such as one-hot encoding or label encoding. The algorithm uses an Ordered Boosting approach to address the problems of overfitting and predictive bias. CatBoost also builds trees with a symmetric structure and utilizes specialized statistical techniques to safely and accurately encode categorical features. (Anande et al., 2025).

The goal of this target modeling is to predict the probability of a borrower defaulting or not and to address the problem of glaring data imbalance, where the number of borrowers who do not default is much higher than those who do.

d) Transformer

Transformer is a deep learning architecture that uses a self-attention mechanism to process the entire sequence of data simultaneously to learn the relationships between elements, in contrast to previous sequential models that process data one by one.(Wang et al., 2024). The main formula in the self-attention mechanism is as follows:

$$\boldsymbol{Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V}$$

Information:

Q : Query matrix of all inputs.

K : Key matrix of all inputs.

V : Value matrix of all inputs.

d_k : dimension of the key vector. This is used to weight and stabilize the gradient.

softmax : This function converts attention scores into probabilities.

QK^T: The dot product of the Query and Key matrices that measures the similarity score.

V : The Value Matrix that will be multiplied by the probability weights.

5. Models Evaluation

After completing the modeling process using various boosting algorithms and Transformers, the next step is to evaluate the model's performance in predicting default risk. Given the imbalanced data (a much higher number of borrowers who did not default), the evaluation metrics used are not only accuracy but also more relevant metrics, such as:

a) Accuracy

Accuracy is an evaluation metric in machine learning that measures how often a predictive model produces correct output overall. (Emmanuel et al., 2024). Mathematically, accuracy is calculated using the following formula:

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

Information:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

b) Precision

Precision is an evaluation metric in machine learning that measures how accurate a model's positive predictions are. (Z. Li & Yao, 2024). In other words, it measures the model's accuracy in predicting borrower default. The formula is as follows:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Information:
TP = True Positive
FP = False Positive

c) Recall (Sensitivity)

Recall or Sensitivity is an evaluation metric that measures the extent to which a model is able to capture all true positive cases, that is, how well the model detects all default events that actually occur. (Emmanuel et al., 2024) The formula is as follows:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

Information:
TP = True Positive
FN = False Negative

d) F1-Score

The F1-Score is the harmonic mean of precision and recall, which is used to provide a balance between the two, especially when the data used is imbalanced (for example, the number of default cases is much less than the number of non-default cases). (Emmanuel et al., 2024) The formula is as follows:

$$\textbf{F1} - \textbf{Score} = 2 \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

e) AUC-ROC

AUC-ROC (Area Under the Curve - Receiver Operating Characteristic) is an evaluation metric used to measure the model's ability to distinguish between two classes, namely in this context: defaulters (positive) and non-defaulters (negative). (Emmanuel et al., 2024)

6. Intepreting model with SHAP (SHapley Additive exPlanation) Values

SHAP (SHapley Additive Explanations) is a framework for explaining the output of machine learning models. Simply put, SHAP provides a way to measure how much each feature (variable) contributes to the model's predictions. (Lundberg & Lee, 2017)

C. Result and Discussion

1. Result on Analysis Data

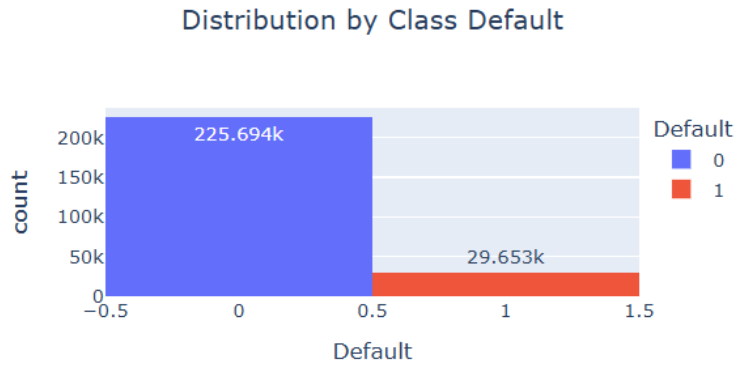

Figure 2. Ditribution by class default

Based on Figure 2, the class distribution of the target variable "Default" in the dataset shows that Default = 0 (customers who did not default) accounted for approximately 225,694 data, and Default = 1 (customers who did default) accounted for approximately 29,653 data. From the figure, it can be seen that the number of non-default data is much larger than the number of default data. This indicates a class imbalance: the majority class (non-default) is around 88.4%, while the minority class (default) is around 11.6%. This condition can cause the machine learning model to be biased towards the majority class, leading to predictions for customers who actually default being ignored. Therefore, imbalance data handling techniques such as SMOTE, TomekLinks, or class weighting need to be used so that the model can better detect minority classes.

Table 2. Descriptive statistics

	Age	Income	LoanAmount	CreditScore	Months Employed	Num CreditLines	InterestRate	LoanTerm	DTIRatio
count	255347.0	255347.0	255347.0	255347.0	255347.0	255347.0	255347.0	255347.0	255347.0
mean	43.5	82499.3	127578.9	574.3	59.5	2.5	13.5	36.0	0.5
std	15.0	38963.0	70840.7	158.9	34.6	1.1	6.6	17.0	0.2
min	18.0	15000.0	5000.0	300.0	0.0	1.0	2.0	12.0	0.1
25%	31.0	48825.5	66156.0	437.0	30.0	2.0	7.8	24.0	0.3
50%	43.0	82466.0	127556.0	574.0	60.0	2.0	13.5	36.0	0.5
75%	56.0	116219.0	188985.0	712.0	90.0	3.0	19.2	48.0	0.7
max	69.0	149999.0	249999.0	849.0	119.0	4.0	25.0	60.0	0.9

Based on Table 2, the loan data characteristics are as follows: **Age**: The age range of customers is quite wide, from 18 to 69 years, with a median of 43 years. **Income & Loan Amount**: There is significant variation in income and loan amounts, as evidenced by the high standard deviation. This indicates that customers have diverse financial profiles. **Credit Score**: The credit score ranges from 300 to 849. The median value of 574 indicates that most customers are in the average credit score category. **Loan Term**:

The loan term has a minimum value of 12 months, a median of 36 months, and a maximum of 60 months. **DTIRatio**: The debt-to-income ratio ranges from 0.1 to 0.9, with a median of 0.5. This indicates that the average customer has a debt ratio of 50% of their income.

2. Result on Pre Processing Data

Results: In the preprocessing stage, the initial data was split into features (X) and targets (y), with a training set of 16 rows and 16 features. However, the exploration results indicated a class imbalance in the target variable, so several data-balancing techniques were applied. The following are the results of each data balancing method applied to the dataset:

SMOTE-TomekLinks: This combination aims to balance the data by adding synthetic samples to the minority class (SMOTE) and removing outliers from the majority class (TomekLinks). As a result, the training dataset swells to (334232, 16). This indicates significant sample addition to balance the class distribution.

SMOTE (Synthetic Minority Over-sampling Technique): This method focuses solely on adding synthetic samples to the minority class. As a result, the training data set increases to (361110, 16), the largest among all methods. This figure reflects the results of an aggressive oversampling process to equalize the number of samples between classes.

TomekLinks: This method simply reduces the number of samples in the majority class. The total number of samples in the dataset is reduced to (192,251, 16). This reduction occurs because majority samples considered "TomekLinks" (adjacent pairs with the minority class) are removed to create a clearer boundary. Overall, this preprocessing stage successfully creates three balanced versions of the training data. Each dataset will be used to train the model separately, allowing for comparison of model performance on data processed in different ways.

In this section, the authors compare this with previous research. Based on a document review, (Nguyen & Ngo, 2025) study did mention using the SMOTE–TomekLinks technique to address class imbalance, but did not report the final number of observations after the balancing process. This is because the study's primary focus was on evaluating the performance of the boosting algorithm rather than presenting details of the data preprocessing stage. Furthermore, (Nguyen & Ngo, 2025) study used a dataset of approximately 7,200 rows with 12 attributes, while this study uses a much larger dataset, 277,000 rows with 18 features. This significant difference in dataset size and complexity could potentially impact the model performance of each study, making the results obtained between the two studies impossible to compare directly without considering the characteristics of the data used.

3. Performance Comparison Between Models

In this section, the authors compare the performance of three boosting models (XGBoost, LightGBM, and CatBoost) and Attention-Based Models (Transformers) models on three training datasets that have been processed to address class imbalance with the SMOTE and TomekLinks techniques. The main metric used for evaluation is Accuracy.Precssion, Recall, F1-Score, dan ROC-AUC, The comparison results can be seen in the table 3

Table 3. Comparison of models performance

Methods/Models	Accuracy%	Precission%	Recall%	F1-Score%	ROC-AUC%
XGBoost (SMOTE-TomekLinks)	83.72	27.59	24.75	26.09	70.42
XGBoost (SMOTE)	84.00	27.54	23.15	25.15	70.19
XGBoost (TomekLinks)	88.63	61.58	5.51	10.12	75.67
LightGBM (SMOTE-TomekLinks)	83.20	27.23	26.71	26.96	70.13
LightGBM (SMOTE)	83.76	27.40	24.14	25.67	70.26
LightGBM (TomekLinks)	88.65	62.21	5.80	10.61	75.69
CatBoost (SMOTE-TomekLinks)	82.18	25.16	27.06	26.08	68.63
CatBoost (SMOTE)	82.58	25.23	25.46	25.34	68.42
CatBoost (TomekLinks)	88.68	63.82	5.80	10.63	75.66
Transformer (SMOTE-TomekLinks)	64.51	20.29	70.22	31.49	71.36
Transformer (SMOTE)	88.55	69.34	2.48	4.79	74.38
Transformer (TomekLinks)	88.59	64.93	3.78	7.14	74.68

Based on the comparison of the table above in the evaluation of metrics **Accuracy** Model **CatBoost (TomekLinks)** had the highest accuracy of 88.68%. However, accuracy on imbalanced data can be misleading. This high accuracy indicates that the model is very good at predicting the majority class (e.g., customers will not default), but is not necessarily effective at predicting the minority class. Based on metric evaluation **Precision** Model **Transformer (SMOTE)** also excelled in the precision metric, scoring 63.82%. Precision measures how often a model's positive predictions are correct. This value is particularly important when the cost of false positive predictions is high, such as in loan approvals where you don't want to lend to customers who are likely to default. Based on metric evaluation **Recall** Model **Transformer (SMOTE-TomekLinks)** demonstrated a very dominant performance in recall with a score of 70.22%. Recall measures how many positive cases the model successfully detects. This value is especially important when the cost of failing to detect a false positive case is very high, such as in fraud or disease detection. The model successfully detected a large proportion of positive cases, although at the expense of precision.

Based on metric evaluation **F1-Score**, model **Transformer (SMOTE-TomekLinks)** excels with a score of 31.49%. The F1-Score is a metric that combines precision and recall into a single value, making it an ideal choice for evaluating models on imbalanced data. This value indicates that this model has the best balance between the two metrics among all the models tested. Based on metric evaluation **ROC-AUC** Model **LightGBM (TomekLinks)** has the highest ROC-AUC score of 75.75%. ROC-AUC measures the model's overall ability to distinguish between the two classes, regardless of the threshold used. This high score indicates that the model has excellent discrimination ability. Therefore, based on

the performance comparison table of models between the boosting algorithm and deep learning self-attention transformer, the one that is superior in predicting the risk of default is Transformer by combining the imbalanced data handling technique of SMOTE and TomekLinks, because it has advantages in 2 evaluation metrics, namely Recall and F1-Score.

In this section, the author will also compare with previous research, based on data with a high level of class imbalance. The evaluation results using the most relevant Recall and F1-Score metrics, the best model in this study is Transformer with the SMOTE–TomekLinks technique, which obtained a Recall of 70.22% and an F1-Score of 31.49%. In contrast to the study of (Nguyen & Ngo, 2025), which did not use the Transformer architecture in their analysis, this finding shows an element of novelty through the application of a self-attention-based model to predict the risk of default. For the boosting algorithm category, the model with the best performance is CatBoost using SMOTE–TomekLinks, with a Recall score of 27.06% and an F1-Score of 26.08%, which shows the most stable performance in identifying minority classes compared to LightGBM and XGBoost. These results differ from the research of (Nguyen & Ngo, 2025), who reported that LightGBM was the best-performing boosting algorithm on balanced data using the hybrid SMOTE–TomekLinks technique, with a sensitivity score of 95.46% and an F1-score of 95.45%.

This difference is primarily due to variations in data imbalance handling techniques, hyperparameter tuning strategies, and the characteristics of the dataset, which in this study had a more extreme level of imbalance. These conditions make CatBoost more adaptive in learning minority class patterns than LightGBM. Therefore, differences in preprocessing methods, class distribution, and model configuration are the main factors causing the difference in boosting performance between the two studies.

4. Result on Interpreting models SHAP

In this stage, we interpret our best model using the SHAP (SHapley Additive exPlanations) framework to understand the key factors behind the model's decisions. This analysis reveals the contribution of each feature to the model's predictions, both overall and for individual predictions. Below are the results of the model interpretation using SHAP (SHapley Additive exPlanations).

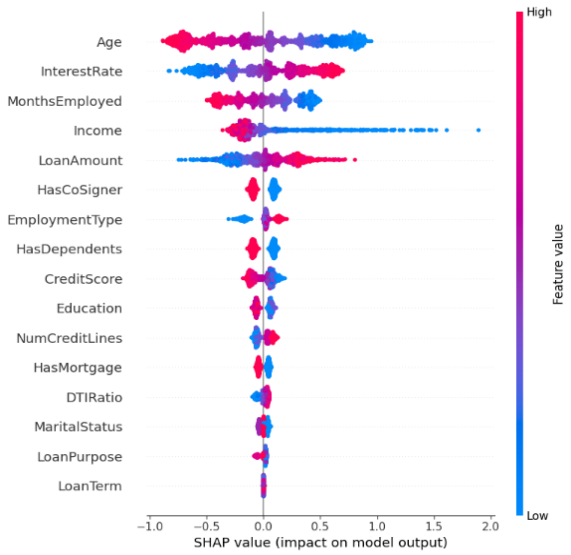

Figure 3. Interpretasi SHAP Models

Before analyzing specific features, it's important to understand how to read the SHAP Summary Plot above:

Y-axis (Features): Features are sorted from top to bottom by importance. The higher a feature's position, the greater its influence on the model's prediction.

X-axis (SHAP Values): Shows the impact of a feature on the model's output. Positive values (to the right) indicate a contribution that increases the chance of a "No Default" (safe), while negative values (to the left) increase the risk of a "Default."

Color (Feature Values): Red represents high feature values, while blue represents low feature values.

Point Density: The thickness or "bubbles" on the graph indicate the distribution of the data; the thicker the area, the greater the number of data samples at that point.

Therefore Based on the SHAP summary plot above. This plot illustrates how much each feature contributes to the model's prediction, the most important of which is **Age**: This is the most influential feature. Red dots (high age) have positive SHAP values, indicating that the older the customer, the less likely they are to default. Conversely, blue dots (low age) are on the negative side, indicating that younger age increases the risk of default. **InterestRate**: The second important feature: The red dots (high interest rates) have negative SHAP values, meaning the higher the loan interest rate, the greater the risk of default. **MonthsEmployed**: The red dots (high tenure) have positive SHAP values, indicating that longer tenure reduces default risk. **Income**: The red dots (high income) have positive SHAP values, meaning higher income tends to reduce default risk. **LoanAmount**: This feature exhibits a complex pattern. Some blue dots (low loan amounts) have negative SHAP values, but the highest positive SHAP values come from the blue and red dots, indicating that very high or very low loan amounts can significantly impact predictions, depending on their combination with other features.

This study has four main points that align with the research objectives, namely comparing the performance of three boosting algorithms (XGBoost, LightGBM, and CatBoost) with the Attention-Based (Transformer) model in predicting default risk; analyzing model performance using comprehensive evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC; testing the effectiveness of the SMOTE–TomekLinks hybrid data balancing technique in improving minority class detection; and providing recommendations for the best model, namely Transformer with SMOTE–TomekLinks. The benefits of this study for readers, especially for academics and financial industry practitioners, are providing a deep understanding of the implementation of accurate, efficient, and adaptive credit risk prediction models to data imbalances, and emphasizing the importance of model interpretability through the SHAP method as a basis for supporting more transparent, accountable, and data-based decision making in the context of credit risk management in financial institutions and the fintech sector in Indonesia.

D. Conclusion

Conclusion from the results of this study, it can be concluded that class imbalance is a major problem in the default risk prediction dataset, with the minority class accounting for only about 11.6% of the total data. This condition can bias the model and cause it to fail to detect true default cases. The performance comparison results show that the transformer model excels in 2 evaluation metrics, namely Recall and F1-Score, these two evaluation metrics are important for imbalanced data. Recall is to maximize the detection of default cases and F1-Score to get the best performance balance then Transformer (SMOTE-TomekLinks) is the most superior model. With a value **70.22% (Recall)** dan **31.49% (F1-Score).** In the analysis of model interpretation using SHAP validated that the model learns from logically relevant features and can conclude that Features such as Age (older age reduces risk), InterestRate (higher interest

rates increase risk), and MonthsEmployed (longer employment reduces risk) are the most significant factors influencing the model's predictions. Overall, this study not only succeeded in building a well-performing model, but also provided in-depth insights into the factors driving the model's decisions, which is crucial for informed decision-making. This study confirmed that the transformer with the SMOTE-TomekLinks balancing technique is the most effective model in detecting debtors at risk of default. This finding has important implications both academically, as it fills a comparative research gap between boosting and transformers, and practically, in supporting the decision-making of financial institutions and fintechs in mitigating credit risk.

Suggestions for further research include exploring advanced data balancing methods: This study has tested several data balancing techniques. For future Research, you could try more advanced balancing methods or different combinations, such as SMOTE-ENN or class weighting techniques in the model. You could also experiment with ensemble methods for imbalanced data, such as BalanceCascade or EasyEnsemble, to see if they provide more stable performance.

References

Aftab, A. I. S., & Matloob, F. (2019). Performance Analysis of Resampling Techniques on Class Imbalance Issue in Software Defect Prediction. *International Journal of Information Technology and Computer Science*, *11*(11), 44–53. https://doi.org/10.5815/ijitcs.2019.11.05

Akinjole, A., Shobayo, O., Popoola, J., & Okoyeigbo, O. (2024). *Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction*.

Anande, T., Alsaadi, S., & Leeson, M. (2025). Enhanced Modelling Performance with Boosting Ensemble Meta-Learning and Optuna Optimization. *SN Computer Science*, *6*(1). https://doi.org/10.1007/s42979-024-03544-3

AT, E., M, A., F, A.-M., & M, S. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Global Journal of Technology and Optimization*, *01*(S1). https://doi.org/10.4172/2229-8711.s1111

Bello, O. (2024). *Citation : Bello O . A . (2023) Machine Learning Algorithms for Credit Risk Assessment : An Economic and Financial Analysis Machine Learning Algorithms for Credit Risk Assessment : An Economic and Financial*. *10*(January 2023). https://doi.org/10.37745/ijmt.2013/vol10n1109133

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks*, *12*(11). https://doi.org/10.3390/risks12110174

Dube, L., & Verster, T. (2023). Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. *Data Science in Finance and Economics*, *3*(4), 354–379. http://www.aimspress.com/article/doi/10.3934/DSFE.2023021

Emmanuel, I., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00882-0

Fitch Ratings, 2025. (2025). *Pasar Pinjaman Berleverage Pecahkan Rekor dengan Capaian Triliunan Dolar di Tahun 2024*. Fitch Ratings, Inc. https://www.fitchratings.com/research/corporate-finance/leveraged-loan-market-breaks-records-with-trillion-dollar-milestone-in-2024-31-01-2025

Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2024). An implementation of machine learning on loan default prediction based on customer behavior. *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020*, *14*(01), 423–426. https://doi.org/10.54209/infosains.v14i01

Hidayaturrohman, Q. A., & Hanada, E. (2024). Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure. *BioMedInformatics*, *4*(4), 2201–2212. https://doi.org/10.3390/biomedinformatics4040118

Hu, Z. (n.d.). *A Transformer-based Neural Network to Predict Credit Card Default*.

Huang, H., Li, J., Zheng, C., Chen, S., Wang, X., & Chen, X. (2025). Advanced Default Risk Prediction in Small and Medum-Sized Enterprises Using Large Language Models. *Applied Sciences (Switzerland)*, *15*(5), 1–23. https://doi.org/10.3390/app15052733

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, *13*(3), 1–40. https://doi.org/10.3390/technologies13030088

Korangi, K., Mues, C., & Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, *308*(1), 306–320. https://doi.org/10.1016/j.ejor.2022.10.032

Li, S., Jin, N., Dogani, A., Yang, Y., Zhang, M., & Gu, X. (2024). Enhancing LightGBM for Industrial Fault Warning: An Innovative Hybrid Algorithm. *Processes*, *12*(1). https://doi.org/10.3390/pr12010221

Li, Z., & Yao, L. (2024). Multi-view GCN for loan default risk prediction. *Neural Computing and Applications*, *36*(20), 12149–12162. https://doi.org/10.1007/s00521-024-09695-x

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Mallinguh, E., & Zoltan, Z. (2022). Financial Institution Type and Firm-Related Attributes as Determinants of Loan Amounts. *Journal of Risk and Financial Management*, *15*(3). https://doi.org/10.3390/jrfm15030119

Mintoo, A. A. (2025). *BLOCKCHAIN IN BANKING : A REVIEW OF DISTRIBUTED LEDGER APPLICATIONS IN LOAN PROCESSING , CREDIT HISTORY , AND*. *04*(01), 101–138. https://doi.org/10.63125/gp61va54

Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. *Cogent Economics and Finance*, *13*(1). https://doi.org/10.1080/23322039.2025.2465971

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, *8*(11), 1–17. https://doi.org/10.3390/data8110169

OJK. (2024). *Statistik Perbankan Indonesia* (Vol. 22, Issue 12).

Poernamawatie, F., Susipta, I. N., & Winarno, D. (2024). Sharia Bank of Indonesia Stock Price Prediction using Long Short-Term Memory. *Journal of Economics, Finance And Management Studies*, *07*(07), 4777–4782. https://doi.org/10.47191/jefms/v7-i7-94

Sharpe, W. (2025). *Sustainable Debt in Focus: 2024 Summary and 2025 Outlook*. Natixis CIB. https://gsh.cib.natixis.com/our-center-of-expertise/articles/sustainable-debt-in-focus-2024-summary-and-2025-outlook

Soomro, A., Zakariyah, H., Aftab, S. M. A., Muflehi, M., Shah, A., & Meraj, S. (2024). Loan Default Prediction Using Machine Learning Algorithms: A Systematic Literature Review 2020–2023. *Pakistan Journal of Life and Social Sciences*, *22*(2), 6234–6253. https://doi.org/10.57239/PJLSS-2024-22.2.00469

Sun, X. (2025). Application of an improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis. *Scientific Reports*, *15*(1), 2569. https://doi.org/10.1038/s41598-025-86014-x

Wang, Y., Xu, Z., Yao, Y., Liu, J., & Lin, J. (2024). Leveraging Convolutional Neural Network-Transformer Synergy for Predictive Modeling in Risk-Based Applications. *2024 4th International Conference on Electronic Information Engineering and Computer Communication, EIECC 2024*, 1565–1570. https://doi.org/10.1109/EIECC64539.2024.10929474

Yang, S., Huang, Z., Xiao, W., & Shen, X. (2025). *Interpretable Credit Default Prediction with Ensemble Learning and SHAP*. http://arxiv.org/abs/2505.20815

Zhang, J., Wang, T., Wang, B., Chen, C., & Wang, G. (2023). Hyperparameter optimization method based on dynamic Bayesian with sliding balance mechanism in neural network for cloud computing. *Journal of Cloud Computing*, *12*(1). https://doi.org/10.1186/s13677-023-00482-y

Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A. G. (2024). Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction. *Mathematics*, *12*(5), 1–27. https://doi.org/10.3390/math12050701