JURNAL INFOTEL
VOL. 18, NO. 3, AUGUST 2026, PP. 438–454. DOI:10.20895/INFOTEL.V18I3.10791

RESEARCH ARTICLE

Optimizing Credit Risk Models: A Comparative Study of LightGBM and XGBoost on Imbalanced Big Data

Irfan Budiyanto[1] and Arief Hermawan[2,*]

[1,2]Program Studi Teknologi Informasi, Program Magister, Universitas Teknologi Yogyakarta, Yogyakarta, 55285, Indonesia

*Corresponding email: ariefdb@uty.ac.id

Received: March 10, 2026; Revised: July 29, 2026; Accepted: August 05, 2026.

Abstract: Predicting loan defaults is a critical challenge in banking, compounded by a severe class imbalance between default and non default loans. This study evaluates LightGBM and XGBoost in predicting credit defaults using the Lending Club dataset of 2,260,668 records. A rigorous preprocessing pipeline eliminating data leakage attributes and applying Recursive Feature Elimination with Cross Validation (RFECV) reduced the feature space from 145 to 58 representative attributes. Two class imbalance strategies were benchmarked using 5 fold Stratified Cross Validation: Scenario A (internal algorithmic cost-weighting) and Scenario B (synthetic oversampling via ADASYN). The results show that internal weighting significantly outperformed synthetic oversampling in both algorithms. Specifically, XGBoost under Scenario A achieved the highest default detection capability, with a Recall of 0.69 and an AUC-PR of 0.3917, compared to a Recall of 0.53 and AUC-PR of 0.3534 when paired with ADASYN. Similarly, LightGBM under Scenario A yielded a Recall of 0.30 and AUC-PR of 0.3905, while ADASYN degraded its Recall to 0.25 and AUC-PR to 0.3603. Rather than clarifying decision boundaries, synthetic sampling introduced noise in regions with heavy class overlap. These findings demonstrate that built in cost weighting is substantially more robust than synthetic data generation for large scale, high dimensional credit datasets.

Keywords: credit scoring, loan default prediction, LightGBM, XGBoost, imbalanced data, ADASYN

© Jurnal Infotel ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

In the banking industry, particularly for institutions where credit lending is a core business, the ability to accurately predict loan defaults is not merely a technical support function but a critical determinant of corporate sustainability [1], [2]. Precise predictions ensure healthy business cycles by channeling capital to the right debtors, whereas a failure to detect risks early can destabilize operations and lead to substantial financial losses due to non-performing loans [3], [4]. However, a significant challenge encountered in practice is the prevalence of imbalanced datasets, where the number of customers who fail (minority class) is much lower than the number of those who repay their loans (majority class), making it difficult for risk assessment models to accurately recognize default patterns [5]. This imbalance triggers a systematic bias where models tend to prioritize the majority class to achieve accuracy figures that appear impressive, but are ultimately misleading [6], [7]. Consequently, the model may not recognize high-risk debtors, which is a critical oversight, as the economic cost of not identifying a defaulter is significantly higher than rejecting a potentially good customer [8], [9]. Therefore, this research moves beyond conventional accuracy to focus on evaluation metrics such as the F1-Score and the Area Under the Precision-Recall Curve (AUC-PR), which are much more informative for skewed distributions. These metrics are selected because they provide an objective assessment of the model's ability to balance prediction exactness (Precision) with risk detection sensitivity (Recall) regarding the minority class [10], [11].

Various solutions have been developed to address these issues through the implementation of oversampling techniques and the deployment of robust algorithms. Gradient boosting algorithms, specifically XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine), have emerged as industry standards due to their superior performance on large-scale datasets [12]. XGBoost is renowned for its strong control over overfitting through regularization mechanisms L1/L2, while LightGBM offers higher computational efficiency via histogram-based algorithms and leaf-wise growth strategies [13]. On the other hand, the Adaptive Synthetic Sampling (ADASYN) technique serves as an adaptive data balancing method that focuses specifically on generating synthetic samples in regions that are the most difficult for the model to learn [14].

The technical challenges in implementing these methods are further complicated by the characteristics of credit data, which often exhibit a high degree of class overlap. Based on preliminary analysis using Principal Component Analysis (PCA), it is evident that the statistical profiles of defaulting customers frequently intertwine with those of non-defaulting customers within the available feature space [15]. This phenomenon of overlap creates a "blind spot" for the model, where even the addition of synthetic data carries the risk of increasing noise if not managed by an algorithm capable of delineating precise decision boundaries [16], [17].

Although these components have been extensively researched, a review of the literature reveals that most studies either analyze them in isolation or rely on small-scale datasets that fail to fully capture real-world complexity. A significant research gap remains regarding a direct comparison between LightGBM and XGBoost when specifically combined with ADASYN on large-scale datasets, a configuration that has rarely been explored in depth. The primary novelty of this research lies in conducting a rigorous head-to-head evaluation of these two distinct Gradient Boosting Decision Tree (GBDT) architectures using the mas-

sive Lending Club dataset, which comprises approximately two million records, following treatment with ADASYN.

The selection of ADASYN as the data balancing technique is predicated on its adaptive mechanism, which specifically assigns higher weights to minority samples that are difficult to learn [16], [18]. While modern gradient boosting algorithms like LightGBM are equipped with internal balancing mechanisms through class weight parameters, the application of external resampling techniques such as ADASYN remains crucial for actively refining data distribution along complex decision boundaries [19]. This approach is substantiated by findings from prior studies using similar datasets, where the implementation of ADASYN demonstrated a capability to significantly elevate model performance scores compared to baseline models that lacked imbalance handling strategies [20].

The primary objective of this research is to comprehensively analyze and compare the performance of LightGBM and XGBoost in predicting loan defaults. Specifically, this study aims to assess how these models interact with ADASYN to identify high-risk customers while also evaluating computational efficiency and the effectiveness of evaluation metrics such as F1-Score and AUC-PR. The findings are expected to provide data-driven guidance for financial institutions in selecting the optimal combination of models and data handling techniques to mitigate risk of credit default.

Although prior studies have explored oversampling methods like ADASYN or gradient boosting models independently, most evaluations rely on modest dataset sizes or treat resampling techniques in isolation without direct cross-algorithm benchmarking. Consequently, a critical research gap persists regarding how distinct Gradient Boosting Decision Tree (GBDT) architectures, specifically LightGBM's leaf-wise growth versus XGBoost's level-wise growth, interact with synthetic oversampling in big data containing heavy class overlap. This study addresses this gap by conducting a direct head-to-head performance comparison of LightGBM and XGBoost in 2.2 million Lending Club credit records. The primary novelty lies in evaluating whether external synthetic resampling (ADASYN) offers any tangible advantage over native, loss-function-based class weighting when classifying high-dimensional, highly imbalanced credit default data.

2 Research Method

This research follows a systematic series of steps to create a reliable credit-default prediction model. The approach includes processing large-scale data, using advanced feature-selection methods, and testing ways to handle imbalanced data to reduce the bias toward the majority class that often occurs in banking risk assessments.

2.1 Research Framework

As illustrated in Figure 1, the research methodology is executed through several interconnected phases to ensure a robust and valid predictive framework. The process begins with Data Acquisition, where the dataset available at https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv is loaded and subjected to Data Understanding to analyze variable characteristics and class distributions. This is followed by a Preprocessing stage, which involves rigorous Data Cleaning, specifically removing features with unique values ≤ 1, handling missing values exceeding 50%, and discarding data leakage variables

before performing Data Transformation via label encoding to convert categorical variables into numerical formats. In the Experimental Setup, the data is partitioned through a Data Split into training and test sets to prevent information leakage, followed by a Data Scale using a StandardScaler to normalize the value range according to Eq.(1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where: x is the original value, μ is the mean, and σ is the standard deviation of the feature.

Subsequently, Feature Selection is performed using Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most informative feature subset while eliminating computational redundancy. The core Modeling phase tests two distinct branches: Scenario A (Without Resampling) and Scenario B (With Resampling: ADASYN), both of which are implemented using LightGBM and XGBoost algorithms. Finally, the model's performance is assessed in the Evaluation stage using AUC-PR and F1-Score metrics to provide a balanced measure of precision and recall through their harmonic mean.

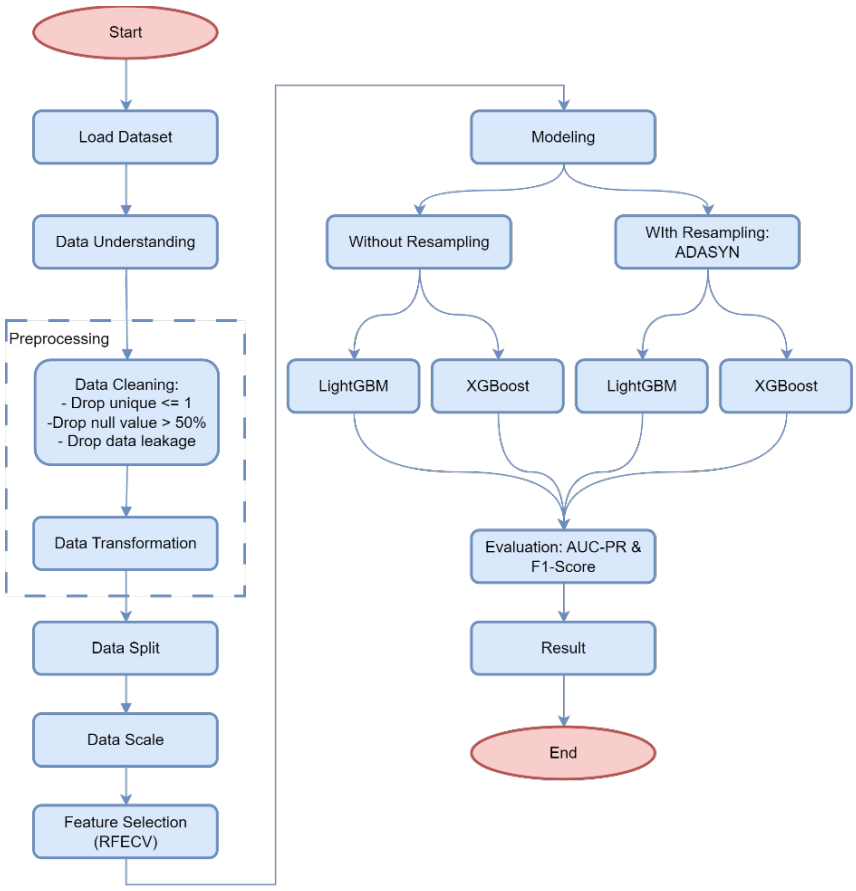

Figure 1: Research flowchart.

2.2 Data Pre-processing and Transformation

The pre-processing phase is a crucial step in establishing a robust model that is resilient against bias and noise, which could otherwise undermine the stability of the prediction [21]. The process begins with Data Cleaning, specifically targeting the elimination of variables with unique values ≤ 1 and features containing null values $> 50\%$ to maintain input quality [22], [23]. Subsequently, features classified as data leakage—variables containing information only available after the credit status is determined—are identified and removed to prevent misleading accuracy inflation [24].

High-cardinality categorical attributes (`emp_title`, `title`, `zip_code`, and `addr_state`) were excluded because their excessive number of unique categories risks forcing tree-based models to overfit localized noise rather than learning generalizable credit risk patterns [23]. Although these variables contain a granular context, removing them reduced the complexity of the model without degrading the predictive stability. Data transformation is executed using the Label Encoding technique for categorical features to ensure they can be processed effectively by tree-based algorithms such as LightGBM and XGBoost [25]. To standardize the scale of numerical features, a StandardScaler is employed to transform the data, resulting in a mean (μ) of 0 and a standard deviation (σ) of 1, as defined in Eq. (1). This step is critically performed after the Data Split phase to ensure that the information from the test set does not "leak" into the training process, thereby preserving the objectivity of the results [26].

2.3 Feature Selection Using RFECV

In this study, feature selection was executed using Recursive Feature Elimination with Cross-Validation (RFECV). Prior to running RFECV, feature pairs exhibiting a Pearson correlation coefficient $|r| \geq 0.9$ were evaluated. For each redundant pair, the attribute with the lower LightGBM Gain Importance score was discarded. Eliminating multicollinearity beforehand prevents feature importance dilution, allowing RFECV to evaluate true feature contributions more reliably [27]. This process is performed iteratively by training the model on the entire initial feature set, calculating feature importance scores, and gradually eliminating the features that contribute the least to the model's output [28]. The primary objective of this stepwise elimination is to identify the optimal feature subset (S^*) that yields the maximum performance score (M) based on the specified evaluation metrics, as expressed in (2).

$$S^* = \arg\max_{S \subseteq F} \text{Score}(M, S) \tag{2}$$

where S^* represents the optimal feature subset, $\arg\max$ denotes the optimization operation used to identify the subset that yields the highest performance score, F represents the initial collection of features, S denotes the candidate feature subset evaluated during each iteration, and $\text{Score}(M, S)$ represents the cross-validation performance of model M based on the specified evaluation metrics. The implementation of Cross-Validation at this stage is critical as it ensures that the selected feature subset performs well not just on a single data segment, thereby avoiding sampling bias, but also possesses strong generalization capabilities on previously unseen data [27].

2.4 Modeling Scenarios

This research contrasts two distinct approaches for handling class imbalance to assess the effectiveness of balancing techniques in gradient boosting models, as shown in Table 1. To establish a controlled baseline for comparing the two Gradient Boosting Decision Tree (GBDT) architectures, both models were implemented using their respective default hyperparameter configurations. This approach ensures that the performance variance observed across scenarios primarily reflects the inherent capability of each algorithm and the impact of the internal weighting mechanism rather than specific hyperparameter tuning. While hyperparameter tuning could further optimize absolute metric scores, utilizing default settings provides an unbiased baseline structural evaluation of how each GBDT algorithm natively interacts with class imbalance and synthetic oversampling.

Table 1: Differences in model parameter configuration

Component	**Scenario A**	**Scenario B**
Imbalance Handling Method	Loss function weight adjustment	Synthetic sample generation for the minority class
Training Data	Original dataset	Resampled dataset
LightGBM Parameter	`is_unbalance=True`	`is_unbalance=False`
XGBoost Parameter	`scale_pos_weight=[ratio]`	`scale_pos_weight=1`
Key Characteristics	Assigns higher penalties to minority-class prediction errors	Generates synthetic samples in sparse feature spaces to clarify decision boundaries

Scenario A (Internal Weighting) trains the model in the original distribution without physical manipulation, achieving balance by modifying the loss function through internal parameters such as `is _unbalance = True` in LightGBM or `scale _pos _weight` in XGBoost. Mathematically, this method assigns a larger weight (w) to the minority class based on the ratio, as expressed in (3).

$$w = \frac{N_{\text{majority}}}{N_{\text{minority}}} \tag{3}$$

where w denotes the weight assigned to the minority class, N_{majority} represents the number of observations in the majority class, and N_{minority} denotes the number of observations in the minority class, thereby imposing a higher penalty for misclassifying defaults without altering the data structure. Methodologically, Scenario A assumes that modifying loss gradients via cost-weighting effectively rebalances misclassification penalties without distorting the underlying empirical feature distribution or altering natural feature correlation structures.

Scenario B (Resampling ADASYN) utilizes the Adaptive Synthetic (ADASYN) technique to physically equilibrate class ratios prior to training. ADASYN operates based on data density, focusing the generation of synthetic samples on low-density areas populated by "hard-to-learn" examples. This strategy aims to refine the data distribution along complex decision boundaries, a task made challenging by PCA analysis revealing significant overlap between compliant and defaulting customers. Methodologically, Scenario B assumes that minority instances occupy distinct, interpolable feature neighborhoods where

k-nearest neighbor (k-NN) synthesis can generate valid debtor profiles without injecting artificial noise into majority class decision boundaries.

The entire modeling process was validated using Stratified K-Fold Cross-Validation with five folds (n_split=5), which ensures that the data partitions are representative of the overall dataset structure [29]. The application of this stratified technique is critical for maintaining a consistent minority class ratio across every data fold, thereby ensuring that the resulting model performance estimation remains stable and unbiased. As primary indicators of success, this study employs the F1-Score and AUC-PR metrics to evaluate the classifiers' effectiveness. In commercial credit risk management, False Negatives (FNs) approving loans for borrowers who subsequently default—result in direct principal capital losses, whereas False Positives (FPs) rejecting creditworthy applicants incur only the opportunity cost of foregone interest margin. Because principal default losses typically dwarf interest income margins, financial risk models must prioritize Recall (minimizing FNs) over Precision. Conventional accuracy is avoided as it frequently yields misleading results on datasets with extreme class disparity, whereas AUC-PR offers a more robust assessment of the model's ability to predict the default class by calculating the Area Under the Precision-Recall Curve, as expressed in (4).

$$\text{AUC-PR} = \int_0^1 p(r)\, dr \tag{4}$$

where $p(r)$ denotes precision as a function of recall r, and the integral is evaluated over the recall range from 0 to 1.

2.5 Evaluation Metrics

The selection of appropriate evaluation metrics is critical in this study, particularly given the highly imbalanced nature of the credit dataset used. Relying solely on standard accuracy metrics often yields results that are heavily biased toward the majority class, thereby failing to accurately reflect the model's true performance in detecting financial risk [30]. Consequently, this research emphasizes metrics that specifically evaluate the classification performance regarding the minority class, such as loan defaults.

2.5.1 Precision and Recall

These two metrics offer distinct yet interconnected perspectives within credit risk management. Precision (Exactness) evaluates how accurately the model predicts that a customer will default, and maintaining high precision is vital in banking to minimize the risk of rejecting actually worthy customers, which represents a significant opportunity cost [31].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

where TP denotes cases in which the model correctly predicts that a customer will default on their loan, while FP represents cases in which the model predicts a default but the customer actually repays the loan successfully.

Recall (Detection Power) measures the proportion of actual defaulting customers successfully captured by the model [14]. Recall is a top priority in risk mitigation strategies as

it assists banks in avoiding direct financial losses resulting from bad debts [29].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

where FN denotes cases in which the model predicts that a customer will repay, but in reality, they default.

2.5.2 F1-Score

Given the inherent trade-off between Precision and Recall, the F1-Score is employed as a balancing metric . The F1-Score utilizes the harmonic mean of these two values, thereby imposing a significant penalty if either metric drops to a very low level.

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

2.5.3 AUC-PR vs. ROC-AUC

This research specifically selects the Area Under the Precision-Recall Curve (AUC-PR) over the more conventionally used ROC-AUC metric. The primary justification for this selection is based on the condition of the Lending Club dataset, which exhibits a high degree of overlap between compliant and defaulting customers.

Weakness of ROC-AUC: This metric incorporates True Negatives (TN) in its calculation through the False Positive Rate axis. In datasets with a massive majority class, a high volume of TNs can automatically inflate the ROC-AUC score to impressive levels (approaching 1.0), even if the model fails to detect the minority class effectively.

Advantage of AUC-PR: This metric focuses exclusively on the performance regarding the minority class (Positive Class) without allowing the count of correctly classified compliant customers to dilute the results. AUC-PR provides a more "honest" and objective depiction of the model's ability to draw decision boundaries in dense, hard-to-distinguish regions.

2.6 Code Availability and Reproducibility

To ensure the reproducibility of the comparative analysis between LightGBM and XGBoost, the complete implementation code, including the ADASYN resampling and RFECV procedures, is hosted on GitHub at https://github.com/irfanbudianto/LGBM_vs_XGB.git

3 Results

3.1 Data Preprocessing and Transformation

The initial phase of this research involved the acquisition of the Lending Club dataset, comprising $2{,}260{,}668$ observations and 145 original attributes. A strategic preprocessing framework was implemented, focusing on dimensionality reduction to enhance computational efficiency and model precision. The process began by filtering the `loan_status` attribute to establish a binary target. The majority class consisted of `Fully Paid` and `Does`

not meet the credit policy. Status: Fully Paid, while the minority class consisted of Charged Off, Default, and Does not meet the credit policy. Status: Charged Off.

Data purification was conducted systematically by eliminating zero-variance columns and features with missing values exceeding the 50% threshold. Furthermore, high-cardinality categorical features, such as emp_title, title, zip_code, and addr_state, were removed due to their limited predictive contribution and potential to increase model complexity and overfitting. To maintain the integrity of the predictive model, 14 data leakage attributes, such as total_pymnt, recoveries, and last_pymnt_amnt, were excluded because this information becomes available only after loan issuance and subsequent repayment activity. The remaining categorical variables were transformed using LabelEncoder. Finally, an iterative feature selection process was performed for attribute pairs with a correlation coefficient of $|r| \geq 0.9$. For each highly correlated pair, the feature with the lower LightGBM Gain Importance score was discarded. This procedure successfully reduced the feature set from 145 to 58 representative attributes.

Following the reduction of attributes, the data were prepared to ensure a valid and consistent comparison across the four experimental treatments: the baseline and ADASYN scenarios for both LightGBM and XGBoost. Remaining missing values were handled using the dropna() method. This approach was considered appropriate given the large scale of the dataset, which allowed the class distribution to remain relatively stable, shifting only slightly from 79.9:20.1 to 79.5:20.5.

3.2 Optimal Feature Identification via RFECV

The dataset was partitioned using a stratified split to preserve the target ratio and subsequently normalized through StandardScaler. Streamlining the dataset from 145 to 58 features yielded substantial computational efficiency gains, cutting training time per fold by over 60% and enabling full 5-fold cross-validation on 2.2 million rows to complete in under 30 seconds for baseline models. The top-ranked predictors identified by RFECV strongly reflect credit risk fundamentals. The sub_grade feature emerged as the most dominant predictor (importance score: 379,459), capturing the lender's comprehensive risk rating. Attributes such as credit_duration and dti (debt-to-income ratio) represent long-term credit history stability and immediate debt-servicing capacity, respectively. Higher dti values signal elevated debt distress, directly increasing default risk. The relative importance of the top ranked predictors is summarized in Table 2.

3.3 Topological Analysis: PCA and Class Separability

To diagnose class separability before modeling, a 2D PCA visualization was conducted on the optimal features. As illustrated in Figure 2, the first two principal components account for only 21.32% of the total variance, confirming that default risk patterns are deeply entangled across a high-dimensional, non-linear manifold. This heavy overlap explains why linear classifiers struggle with credit datasets and highlights why non-linear GBDT algorithms are necessary to establish complex decision boundaries. This inherent complexity justifies the use of advanced GBDT architectures over simpler linear models.

Table 2: Top 10 feature importance rank

Feature	Score	Description
sub_grade	379,459.95	Specific loan subgrade (e.g., A1, B3)
credit_duration	54,695.85	Years of credit history length
term	30,976.22	Number of monthly payments (36 or 60)
loan_amnt	23,580.07	Original requested loan amount
dti	19,951.07	Debt-to-income ratio
avg_cur_bal	17,155.03	Average current balance across all accounts
emp_length	11,252.87	Employment length in years
acc_open_past_24mths	11,156.66	Accounts opened in the last 24 months
annual_inc	10,470.31	Borrower's reported annual income
home_ownership	10,438.95	Home ownership status (RENT, OWN, MORTGAGE, OTHER)

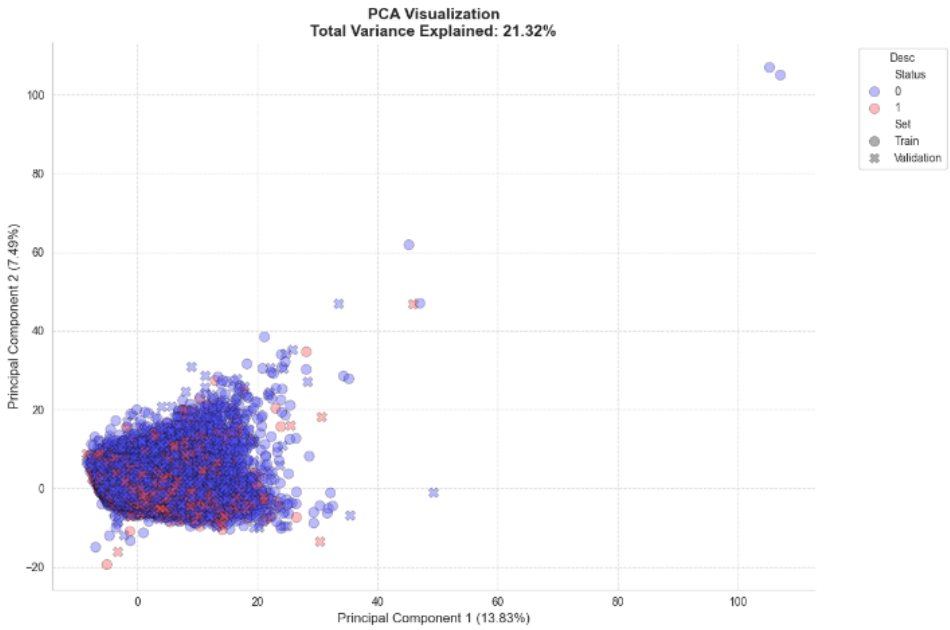

Figure 2: PCA visualization.

3.4 Performance Comparison across Scenarios

To ensure a robust and unbiased performance estimation, all models were evaluated using 5-fold Stratified Cross-Validation with shuffling. This approach maintains a consistent class distribution across all folds, which is critical for the stability of the model's performance on imbalanced datasets. The comparative results of the experimental treatments are summarized in Table 3.

Table 3: Performance comparison across modeling scenarios

	Scenario A		Scenario B	
Metric	**LightGBM**	**XGBoost**	**LightGBM**	**XGBoost**
CV F1-Score (Avg)	0.3597	**0.4405**	0.3125	0.4151
CV AUC-PR	0.3905	**0.3912**	0.3603	0.3535
Final AUC-PR (Validation)	0.3909	**0.3917**	0.3576	0.3534
Recall (Class 1)	0.3000	**0.6900**	0.2500	0.5300
F1-Score (Class 1)	0.3600	**0.4400**	0.3100	0.4100

Table 4: Average confusion matrix across 5-fold cross-validation

Scenario	Model	Avg TP	Avg TN	Avg FP	Avg FN
A	LightGBM	10,223	118,019	12,808	23,593
A	XGBoost	23,236	82,371	48,456	10,580
B	LightGBM + ADASYN	8,613	118,142	12,686	25,203
B	XGBoost + ADASYN	18,651	93,441	37,386	15,165

The average confusion matrix counts in Table 4 reveal the operational impact of each model. XGBoost under Scenario A achieved the lowest number of False Negatives (10,580), missing significantly fewer default cases compared to LightGBM's 23,593. This confirms XGBoost's superior sensitivity in capturing high-risk debtors, which is essential for minimizing credit losses. Conversely, the introduction of ADASYN in Scenario B led to an increase in False Negatives for both architectures, indicating that synthetic samples tended to obscure actual default patterns rather than clarify them.

3.4.1 Scenario A: Internal Weighting (Baseline)

LightGBM: demonstrated exceptional stability across the CV folds, with a mean F1-score of 0.3597 (± 0.0016). A validation AUC-PR of 0.3909 indicates strong generalization. While highly reliable in identifying performing loans (0.90 recall), it struggled with default detection (0.30 recall).

XGBoost: emerged as a highly risk-sensitive model. With a class 1 recall of 0.69, it proved significantly more effective at capturing potential defaults than LightGBM. However, this sensitivity resulted in a trade-off: lower precision (0.32) and a higher volume of false positives (48,456), reflecting a more aggressive risk-mitigation stance.

3.4.2 Scenario B: Synthetic Augmentation (ADASYN)

LightGBM + ADASYN: while maintaining stability (CV F1-score 0.3125), this combination showed a decrease in sensitivity (0.25 recall) compared to the internal weighting scenario. This suggests that for LightGBM, synthetic data points may have complicated the decision boundary rather than clarified it.

XGBoost + ADASYN: exhibited solid generalization with a validation AUC-PR (0.3534) nearly identical to the CV mean (0.3535), confirming that the pipeline effectively prevented data leakage. Although it captured 53% of the minority class, its overall performance remained below the results achieved in Scenario A.

4 Discussion

4.1 The Efficacy of Internal Weighting (Scenario A)

The experimental findings indicate that Scenario A, which utilizes internal algorithmic weighting, consistently outperforms the synthetic augmentation approach of Scenario B. XGBoost under Scenario A emerged as the most risk-sensitive model, achieving a significant recall of 0.69 and the highest final AUC-PR of 0.3917. The superior risk sensitivity of XGBoost compared to LightGBM in this study can be technically attributed to its level-wise tree growth strategy [32], [33]. While LightGBM's leaf-wise approach focuses on minimizing global loss and can sometimes overlook sparse minority patterns in high-overlap datasets, XGBoost's level-wise growth maintains a more balanced tree structure [32], [33], [34]. This allows the model to capture broader decision boundaries within the high-dimensional feature space, effectively increasing its power to detect default patterns (Recall) even at the expense of a higher false-positive rate. From a credit risk management perspective, this high recall is critical as it demonstrates the model's superior ability to capture potential defaults, thereby minimizing direct financial losses from non-performing loans. To illustrate the practical financial impact, consider a validation fold containing 33,816 default cases. XGBoost under Scenario A correctly detected 23,236 defaults (Recall 0.69, FN: 10,580), whereas LightGBM under Scenario B detected only 8,613 defaults (Recall 0.25, FN: 25,203). Assuming an average loan exposure of $10,000, the high-recall XGBoost model prevents approximately 14,600 additional defaults, representing over $146 million in mitigated credit exposure compared to the ADASYN-augmented LightGBM model. While this sensitivity leads to a higher volume of false positives, the economic cost of missing a default is typically much higher than the opportunity cost of rejecting a potentially good customer.

4.2 The Synthetic Noise Paradox: ADASYN Limitations (Scenario B)

Implementation of ADASYN in Scenario B resulted in a noticeable degradation of performance metrics across both algorithms. Although ADASYN aims to balance the class distribution, it appears to introduce significant noise into the decision boundaries [16]. This is particularly evident in the decline of XGBoost's recall from 0.69 to 0.53 when resampling was applied. A similar detrimental trend was observed in LightGBM, where the recall dropped from 0.30 to 0.25, confirming that synthetic data points complicated the decision boundary rather than clarifying it for the leaf-wise algorithm.

The failure of Scenario B can be directly attributed to the high degree of class overlap visualized in the PCA analysis. In regions where defaulting and compliant customers are statistically intertwined, indicated by the fact that the first two principal components explain only 21.32% of the total variance, ADASYN's mechanism generates synthetic samples in dense, hard-to-distinguish areas. These synthetic points likely blurred the decision boundaries further, acting as computational noise that misled both GBDT architectures [35].

Consequently, the Precision-Recall (PR) Curve for Scenario B models consistently maintained a lower trajectory than the baseline models. To mitigate such limitations in future implementations, alternative resampling strategies particularly hybrid cleaning techniques (e.g., SMOTE-ENN or SMOTE-Tomek Links) that explicitly remove borderline noisy samples, or cost-sensitive gradient adjustments such as Focal Loss should be evaluated. These results suggest that for large-scale financial datasets with high overlap, leveraging built-in algorithmic cost-weighting (Scenario A) is a more robust and effective strategy for risk detection than external synthetic resampling [36]. Furthermore, while these empirical evaluations rely on Lending Club records, the underlying principle that native loss-function cost weighting is superior to synthetic resampling in dense, high-dimensional feature spaces is highly generalizable to other credit scoring, retail banking, and financial fraud domains characterized by severe class imbalance and heavy feature overlap [29].

4.3 Robustness of AUC-PR as an Evaluation Metric

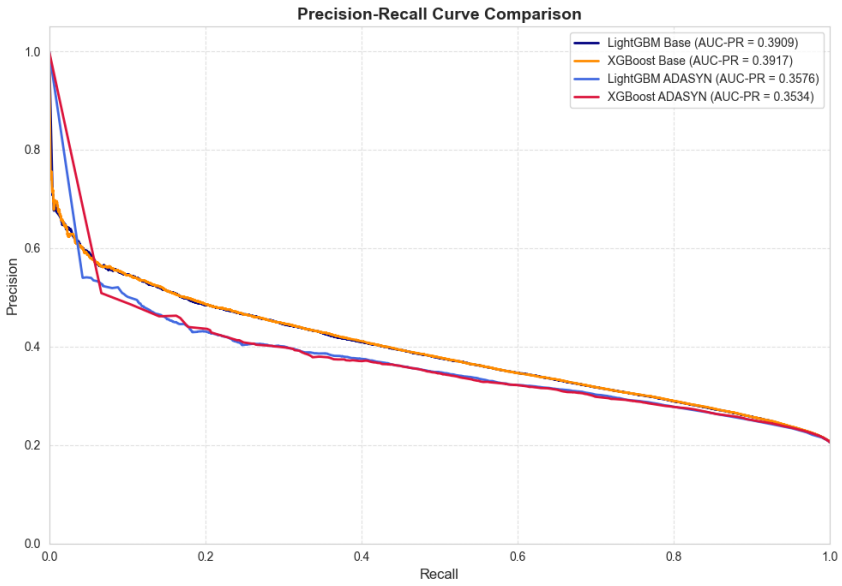

Figure 3: Precision recall curve comparison.

The Precision-Recall (PR) Curve comparison, as illustrated in Figure 3, provides visual confirmation of the models' effectiveness. Scenario A models maintained a higher trajectory than those in Scenario B. This research prioritizes AUC-PR over the conventional ROC-AUC because AUC-PR offers a more "honest" assessment of model performance on imbalanced data. By focusing exclusively on the minority class, AUC-PR ensures that the results are not artificially inflated by the massive count of correctly classified compliant customers.

5 Conclusion

This study demonstrates that while LightGBM and XGBoost effectively handle large-scale financial data, their default-detection capability depends heavily on the chosen class-imbalance strategy. Feature reduction via RFECV optimized the attribute set from 145 to 58, boosting computational efficiency while sharpening predictive focus. Crucially, native algorithmic cost-weighting (Scenario A) consistently surpassed synthetic oversampling via ADASYN (Scenario B). XGBoost under Scenario A emerged as the most optimal model for risk mitigation, delivering a strong Recall of 0.69 and an AUC-PR of 0.3917. In high-overlap credit datasets, ADASYN introduced synthetic noise that degraded precision-recall performance.

For financial risk managers, these findings indicate that external oversampling methods can diminish default-detection sensitivity. Institutions should instead leverage built-in algorithmic cost-weighting to maximize default detection. A limitation of this study is its reliance on historical, static loan records from a single platform. Future work should integrate automated hyperparameter optimization (e.g., Bayesian Optimization) and explicit cost-sensitive loss functions calibrated to institutional Loss Given Default (LGD) parameters to further refine the precision-recall trade-off.

Declarations

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Generative AI Usage Statement

During the preparation of this work, the author(s) used AI tools (such as Gemini) solely for language translation, grammatical correction, and polishing the text clarity. The author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

List of Contributions

1. Irfan Budianto: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Writing - Original Draft.
2. Arief Hermawan: Conceptualization, Supervision, Writing - Review & Editing.

References

[1] I. Aruleba and Y. Sun, "An improved ensemble method with data resampling for credit risk prediction," *IEEE Access*, vol. 13, pp. 71275–71287, 2025.

[2] N. Uddin, M. K. Uddin Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder, and S. Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 327–339, June 2023.

[3] E. Owusu, R. Quainoo, S. Mensah, and J. K. Appati, "A deep learning approach for loan default prediction using imbalanced dataset," *International Journal of Intelligent Information Technologies*, vol. 19, no. 1, 2023.

[4] B. G. Odunlami and B. Nwonu, "Credit risk prediction using ensemble and linear machine learning models," *Journal of Advanced Artificial Intelligence*, vol. 2, pp. 1–8, Aug. 2025.

[5] S. M. Fati, "A loan default prediction model using machine learning and feature engineering," *ICIC Express Letters*, vol. 18, pp. 27–37, Jan. 2024.

[6] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022.

[7] N. H. Cahyana, Y. Fauziah, Wisnalmawati, A. S. Aribowo, and S. Saifullah, "The evaluation of effects of oversampling and word embedding on sentiment analysis," *Jurnal Infotel*, vol. 17, no. 1, pp. 54–67, 2025.

[8] T. Yu, W. Huang, X. Tang, and D. Zheng, "A hybrid unsupervised machine learning model with spectral clustering and semi-supervised support vector machine for credit risk assessment," *PLOS ONE*, vol. 20, no. 1, pp. 1–25, 2025.

[9] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3559–3579, 2023.

[10] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng, "Advanced user credit risk prediction model using LightGBM, XGBoost and TabNet with SMOTEENN," 2024.

[11] C. Yu, Y. Xu, J. Cao, Y. Zhang, Y. Jin, and M. Zhu, "Credit card fraud detection using advanced transformer model," in *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*, pp. 343–350, Aug. 2024.

[12] N. K. C. Pratiwi, "In-depth exploration and comparison of machine learning performances for early-stage diabetes risk prediction," *Jurnal Infotel*, vol. 16, pp. 353–368, May 2024.

[13] H. Du, L. Lv, H. Wang, and A. Guo, "A novel method for detecting credit card fraud problems," *PLOS ONE*, vol. 19, pp. 1–26, Mar. 2024.

[14] A. M. Khani, A. Mohaghar, A. Rezasoltani, and S. H. Hosseinian, "Advanced hyperparameter optimization and adaptive synthetic sampling in machine learning for predictive maintenance of industrial machinery," *International Journal of Research in Industrial Engineering*, vol. 14, pp. 607–629, Dec. 2025.

[15] D. Datta, P. K. Mallick, A. V. N. Reddy, M. A. Mohammed, M. M. Jaber, A. S. Alghawli, and M. A. A. Al-qaness, "A hybrid classification of imbalanced hyperspectral images using ADASYN and enhanced deep subsampled multi-grained cascaded forest," *Remote Sensing*, vol. 14, pp. 1–32, Oct. 2022.

[16] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels," *Technologies*, vol. 13, p. 88, Mar. 2025.

[17] M. Song, H. Ma, Y. Zhu, and M. Zhang, "Credit risk prediction based on improved ADASYN sampling and optimized LightGBM," *Journal of Social Computing*, vol. 5, pp. 232–241, Sept. 2024.

[18] L. H. Chia, "Finding the sweet spot: Optimal data augmentation ratio for imbalanced credit scoring using ADASYN," Oct. 2025.

[19] E. Ileberi and Y. Sun, "Advancing model performance with ADASYN and recurrent feature elimination and cross-validation in machine learning-assisted credit card fraud detection: A comparative analysis," *IEEE Access*, vol. 12, pp. 133315–133327, 2024.

[20] A. Salehpour, M. Norouzi, M. A. Balafar, and K. SamadZamini, "A cloud-based hybrid intrusion detection framework using XGBoost and ADASYN-augmented random forest for IoMT," *IET Communications*, Dec. 2024.

[21] A. Ogungbire and S. S. Pulugurtha, "Effectiveness of data imbalance treatment in weather-related crash severity analysis," *Transportation Research Record*, vol. 2678, no. 11, 2024.

[22] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-based machine learning algorithm for loan default risk prediction," *Mathematics*, vol. 12, pp. 1–32, Oct. 2024.

[23] S. B. Paudel, B. Devkota, and S. Timilsina, "Multi-class credit risk analysis using deep learning," *Journal of Engineering and Sciences*, vol. 2, pp. 82–87, Nov. 2023.

[24] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, p. 1829, Feb. 2024.

[25] L. Hoanh-Su *et al.*, "Default prediction in the finance industry based on ensemble learning: Combining machine learning and deep learning," *Business Systems Research*, vol. 16, pp. 198–218, June 2025.

[26] H. Dong, R. Liu, and A. W. Tham, "Accuracy comparison between five machine learning algorithms for financial risk evaluation," *Journal of Risk and Financial Management*, vol. 17, p. 50, Jan. 2024.

[27] A. I. Adler and A. Painsky, "Feature importance in gradient boosting trees with cross-validation feature selection," *Entropy*, vol. 24, p. 687, May 2022.

[28] H. Chen, C. Yang, M. Du, and Y. Zhang, "Research on credit risk prediction under unbalanced dataset based on ensemble learning," *Mathematical Problems in Engineering*, vol. 2023, 2023.

[29] F. Zhao, G. Li, Y. Lyu, H. Ma, and X. Zhu, "A cost-sensitive ensemble deep forest approach for extremely imbalanced credit fraud detection," *Quantitative Finance*, vol. 23, pp. 1397–1409, Oct. 2023.

[30] J. Alvi, I. Arif, and K. Nizam, "Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management," *Heliyon*, vol. 10, p. e39770, Nov. 2024.

[31] Z. K. Kang, S. Y. Teh, S. Y. G. Tan, and W. C. Ng, "Loan default prediction using machine learning algorithms," *Journal of Informatics and Web Engineering*, vol. 4, no. 3, 2025.

[32] P. Florek and A. Zagdański, "Benchmarking state-of-the-art gradient boosting algorithms for classification," May 2023.

[33] B. So, "Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM," *Scandinavian Actuarial Journal*, vol. 2024, no. 10, pp. 1013–1035, 2024.

[34] N. Nguyen and D. Ngo, "Comparative analysis of boosting algorithms for predicting personal default," *Cogent Economics & Finance*, vol. 13, p. 2465971, Feb. 2025.

[35] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, p. 122778, June 2024.

[36] O.-A. Ampomah, E. Agyemang, K. Acheampong, and L. Agyekum, "Enhancing credit default prediction using boruta feature selection and DBSCAN algorithm with different resampling techniques," Sept. 2025.