

AugLog-LightGBM: A Log-Based Feature Augmentation Framework for Class Imbalance in Credit Risk Classification

Hana Siti Azizah¹, Eni Sumarminingsih¹, Adji Achmad Rinaldo Fernandes¹
¹Department of Statistics, Brawijaya University, Malang, Indonesia

Article Info	ABSTRACT
Article history: Received May 21, 2026 Revised July 02, 2026 Accepted July 03, 2026	Non-performing loan (NPL) detection is inherently a class-imbalance problem because defaulting borrowers represent a persistent minority. Standard gradient boosting often favors the majority class. This paper proposes AugLog-LightGBM, an extension of LightGBM that improves initialization through Log-Based Feature Augmentation (LBFA). Instead of using an uninformative constant, boosting starts from an informed prior combining a logistic-regression logit score and a kernel-density-estimation log-density ratio (LDR), which capture complementary global and local information. These representations are incorporated as augmented features and as the <code>init_score</code>, reformulating boosting as residual correction over an informed Bayesian prior. The proposed framework is evaluated on a dataset of 2,700 home-mortgage borrowers collected from partner banks in Malang, Indonesia (NPL rate = 16.11%), using repeated stratified cross-validation and comparison against four imbalance-aware baselines. AugLog-LightGBM achieves the highest ROC-AUC (0.815 ± 0.018), PR-AUC (0.679), and F1-score (0.631). DeLong tests show statistically significant ROC-AUC improvements over class-weighted Logistic Regression and Random Forest, while gains over XGBoost and SMOTE + LightGBM are positive but not statistically significant. Robustness analyses and SHAP interpretation further support the consistency and practical applicability of the proposed framework.
Keywords: LightGBM; log-based feature augmentation; logit; log-density ratio; class imbalance;	<i>This is an open access article under the CC BY-SA license.</i> 
Corresponding Author: Eni Sumarminingsih, Department of Statistics, Faculty of Science, Technology, and Mathematics (FSTeM), Brawijaya University, 10-11 Veteran Road, Malang City, East Java 65145, Indonesia. Email: eni_stat@ub.ac.id	

1. INTRODUCTION

Credit risk assessment is one of the most critical tasks in banking, particularly for home mortgage loan portfolios, or Kredit Pemilikan Rumah (KPR). The Otoritas Jasa Keuangan (OJK) reported a 10.94% year-on-year growth in consumer credit in November 2024, accompanied by increasing pressure on financial institutions to detect non-performing loans (NPLs) at an early stage [1]. Under POJK No. 15/POJK.03/2017, banks with NPL ratios exceeding 5% are classified as unhealthy, implying that healthy portfolios usually contain only a small proportion of NPL observations [2]. This condition commonly produces an imbalance class setting, where the minority NPL class accounts for only a small fraction of the overall records.

Under imbalance class condition, algorithms based on empirical risk minimisation tend to assign disproportionate weight to the majority class, namely performing loans (PL). As a result, classifiers may achieve high overall accuracy while producing low sensitivity to the minority NPL

class. This issue is critical in credit risk assessment because undetected NPLs directly increase credit loss exposure, whereas excessive false positives may lead to unjustified credit rejection and loss of business [3]. Several recent studies have addressed class imbalance through resampling, cost-sensitive learning, and modified loss functions. Rahmadini and Santoso applied imbalanced data handling in a machine learning-based classification task [4], while Damayanti et al. used focal loss to overcome class imbalance in deep learning-based classification [5]. These studies indicate that class imbalance remains an important methodological issue in predictive modelling, including within the scope of data and information systems research.

Light Gradient Boosting Machine (LightGBM) has achieved strong performance on tabular classification tasks through histogram-based learning, leaf-wise tree growth, gradient-based one-side sampling (GOSS), and exclusive feature bundling (EFB) [6]. Recent comparative studies also show that boosting-based models often perform competitively across different predictive modelling tasks. Riyadi and Fahmi compared several machine learning models, including gradient boosted tree methods, in a predictive modelling setting which further confirm the effectiveness of LightGBM and related boosting methods for imbalanced data problems [7], [8], [9]. However, LightGBM constructs decision boundaries mainly through threshold-based splits on feature values. When class-conditional distributional differences are subtle, this mechanism may be insufficiently sensitive to the minority class, even when class-weighting adjustments are applied.

From a Bayesian perspective, the optimal binary classifier is characterised by the posterior log-odds, which can be decomposed into a log-density ratio term and a log prior odds term. This decomposition motivates a principled feature augmentation strategy, where the two components are estimated separately and then used to provide an informed initialisation for gradient boosting. Fan et al. demonstrated the usefulness of log-density ratio features in high-dimensional classification through feature augmentation via nonparametrics and selection (FANS) [10]. Li and Cui further established density ratio modelling as an effective approach for imbalanced classification under distributional overlap [11]. Seniaray and Jindal also reported the usefulness of density-based features in machine learning classification under class imbalance [12]. However, these studies have not integrated the Bayesian posterior log-odds decomposition into LightGBM's initialisation mechanism, nor have they combined a parametric logit score and a nonparametric log-density ratio simultaneously as an informed boosting prior.

This paper proposes AugLog-LightGBM to address this gap. The main contributions of this study are as follows: (i) a Bayesian posterior log-odds formulation that provides the theoretical basis for the proposed augmentation; (ii) a feature augmentation strategy that combines a logistic regression-based logit score and a kernel density estimation-based log-density ratio; (iii) an integration mechanism that uses the augmented features and the prior score within LightGBM through the `init_score` parameter; (iv) a rigorous empirical evaluation on real mortgage credit data from Malang, Indonesia under repeated stratified cross-validation, benchmarked against external imbalance-aware baselines with statistical significance testing (DeLong) and dedicated robustness analyses; and (v) SHAP Beeswarm interpretation to examine the contribution of the original and augmented features to model predictions.

2. RESEARCH METHOD

2.1. Research Flow

The overall research pipeline is illustrated in Figure 1. The flow proceeds from left to right through five main stages: data collection, data processing, and repeated stratified cross-validation form the preparatory stage; feature augmentation and model training (covering the proposed LightGBM variants and the external baseline models) constitute the computational core; and evaluation, comparison with the DeLong test, robustness analyses, and best-model interpretation using a SHAP beeswarm plot complete the analytical output. This section describes each stage in the order it appears in the diagram.

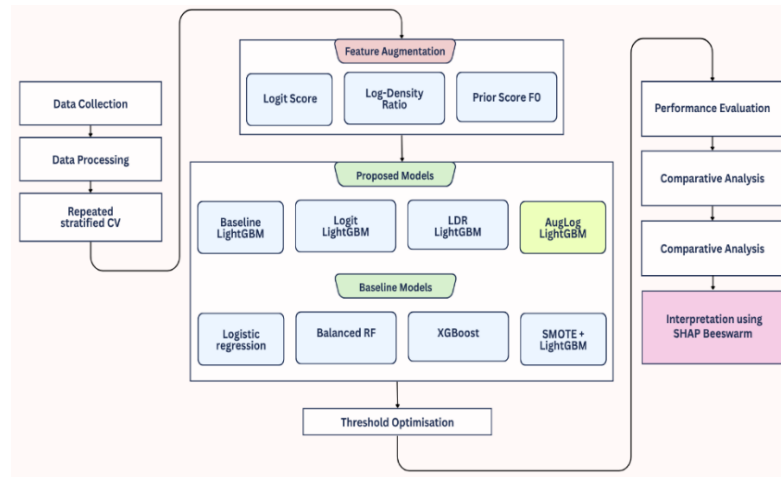


Figure 1. Methodological research flow of AugLog-LightGBM

2.2. Data Collection

The dataset was compiled under the Applied Research (Penelitian Terapan) grant led by Fernandes et al. (Kemendikbudristek, 2024). It consists of 2,700 home-mortgage (KPR) borrower records collected from partner banks in Malang, Indonesia between March and June 2024. Under the grant agreement and data-sharing arrangements with the participating partner banks, the complete dataset and institutional identities cannot be publicly disclosed. The dataset includes only approved mortgage borrowers with complete records and unique customer identifiers, thereby excluding incomplete and duplicate observations. Each record contains 14 predictor variables representing the five principles of credit assessment (5C): Character, Capacity, Capital, Collateral, and Condition, along with a binary loan outcome indicating performing loan (PL = 0) or non-performing loan (NPL = 1). The dataset contains 2,265 PL cases (83.89%) and 435 NPL cases (16.11%), resulting in a moderately imbalanced class distribution. Applicant demographic and socioeconomic characteristics were collected through structured questionnaires and integrated with the participating banks' existing customer databases. Credit-related variables, including the loan outcome, were obtained from the partner banks' administrative records, while BI Checking (SLIK) status was included as an additional predictor variable rather than as the source of the outcome label.

Table 1. Description of predictor variables and outcome

Variable	Description	Scale	Dummy-variable encoding (baseline in parentheses)
X_1	Collateral document type	Nominal	(SHM/SHGB): $X_1^{(1)} = 1$ SHGB; $X_1^{(2)} = 1$ SHM
X_2	Length of residence (years)	Ratio	Continuous
X_3	City size (scale 1–5)	Ordinal	Ordinal value
X_4	Length of education (years)	Ratio	Continuous
X_5	Age (years)	Ratio	Continuous
X_6	BI Checking status	Nominal	(No BI Checking): $X_6^{(1)} = 1$ Collectability 1; $X_6^{(2)} = 1$ Collectability 2
X_7	Marital status	Nominal	(Unmarried): $X_7^{(1)} = 1$ Divorced; $X_7^{(2)} = 1$ Married
X_8	Joint income	Nominal	(No): $X_8^{(1)} = 1$ Yes
X_9	Loan term (years)	Ratio	Continuous
X_{10}	Repayment-to-income ratio (RPA)	Ratio	Continuous
X_{11}	Occupation	Nominal	(SOE/ROE employee): $X_{11}^{(1)} = 1$ Entrepreneur; $X_{11}^{(2)} = 1$ Civil servant; $X_{11}^{(3)} = 1$ Private employee
X_{12}	Work experience (months)	Ratio	Continuous
X_{13}	Number of family dependants	Ratio	Continuous
X_{14}	Loan-to-value ratio	Ratio	Continuous
Y	Loan status (outcome)	Nominal	$Y=1$ NPL; $Y=0$ PL (baseline = PL)

Note. SOE = State-owned enterprise; ROE = Regional-owned enterprise.

2.3. Data Processing

Raw records are subjected to standard preprocessing before modelling. Continuous features are standardised using the StandardScaler (zero mean, unit variance), and categorical features are encoded using OneHotEncoder (as presented in Table 1). To prevent data leakage, all preprocessing transformations are fitted exclusively on the training partition and applied without re-fitting to the test partition [3], [8].

Multicollinearity among predictors is assessed via the Variance Inflation Factor (VIF):

$$VIF_j = \frac{1}{(1 - R_j^2)} \quad (1)$$

Predictors with $VIF_j > 10$ are removed prior to estimation of the logistic regression coefficients $\hat{\beta}$, ensuring numerical stability [13].

2.4. Stratified Cross-Validation

All models in this study were evaluated using repeated stratified 5-fold cross-validation with three independent repetitions (15 evaluations per model), rather than a single hold-out split, because a single split yields a high-variance estimate of minority-class performance. Stratification preserves the NPL proportion within every fold, while repeated random partitioning reduces the variability associated with any particular train-test split. For each model, we report the mean and standard deviation of ROC-AUC across all 15 evaluations, together with the mean PR-AUC, F1-score, sensitivity (recall), specificity, and Matthews correlation coefficient (MCC). Logistic regression, kernel density estimation (KDE), feature augmentation, and threshold optimisation were fitted exclusively on the training data within each outer fold, ensuring that no information from the held-out fold influenced feature construction, prior-score estimation, hyperparameter selection, or threshold optimisation.

2.5. Feature Augmentation

The feature augmentation stage constructs three derived inputs from D_{train} , namely the Logit Score, the Log-Density Ratio (LDR), and the Prior Score F_0 . These inputs are introduced to operationalise the Bayesian posterior log-odds decomposition formalised in Proposition 2. For any $x \in \mathbb{R}^p$, with $p(x | Y = k) > 0$ for $k \in \{0,1\}$, the following identity holds:

$$\log \left(\frac{P(Y = 1 | x)}{P(Y = 0 | x)} \right) = \log \left(\frac{f_1(x)}{f_0(x)} \right) + \log \left(\frac{P(Y = 1)}{P(Y = 0)} \right) \quad (2)$$

Component 1. Logit Score

The first augmented component is obtained parametrically using logistic regression. For observation i , the estimated probability of belonging to the NPL class is defined as

$$\pi(x_i) = \sigma(x_i^\top \beta) = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \quad (3)$$

The coefficient vector $\hat{\beta}$ is estimated on D_{train} using maximum likelihood estimation through Newton-Raphson, or equivalently iteratively reweighted least squares (IRLS). The resulting logit score for observation i is then given by

$$\hat{\eta}_i = \text{logit}(\hat{\pi}(x_i)) = x_i^\top \hat{\beta} \approx \log \frac{P(Y = 1 | x_i)}{P(Y = 0 | x_i)} \quad (4)$$

This score provides a globally linear approximation of the posterior log-odds and is used as the first augmented feature [14], [15]. The use of logistic regression as an initialisation component is also consistent with prior work combining logistic and ensemble methods for credit risk modelling [16].

Component 2. Log-Density Ratio (LDR)

The second augmented component estimates the LDR term in Equation (2) using a nonparametric approach. For each continuous predictor X_j , the class-conditional densities are estimated separately for the NPL and PL classes using Gaussian kernel density estimation (KDE) on D_{train} :

$$\hat{p}_j(x | Y = k) = \frac{1}{n_k h_k} \sum_{i: y_i = k} K\left(\frac{x - x_{ij}}{h_k}\right), \quad K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \quad (5)$$

The bandwidth h_k is selected using Scott's rule,

$$h_k = \hat{\sigma}_k n_k^{-1/5},$$

where $\hat{\sigma}_k$ is the sample standard deviation within class k . This rule provides a standard bandwidth choice for KDE and is commonly used under approximately Gaussian class-conditional densities [17], [18]. In the present class imbalance setting, where the NPL class contains only $n_1 = 435$ observations, the estimated bandwidth for the minority class is expected to be wider than that of the majority class. This reflects the greater uncertainty associated with estimating the NPL density.

For observation i and continuous predictor j , the log-density ratio is computed as

$$\text{LDR}_{ij} = \log\left(\frac{\hat{p}_j(x_{ij} | Y = 1) + \varepsilon}{\hat{p}_j(x_{ij} | Y = 0) + \varepsilon}\right), \quad \varepsilon = 10^{-9} \quad (6)$$

The small constant ε is added to avoid numerical instability when the estimated density is close to zero, particularly in the distribution tails [19]. A positive LDR_{ij} indicates that the observed value is more consistent with the NPL class distribution, while a negative value indicates greater consistency with the PL class distribution. Values close to zero suggest limited discriminatory information for that predictor.

Component 3. Prior Score F_0

After the logit score and LDR features are obtained, the augmented feature matrix is formed by combining the original predictors, the estimated logit score, and the LDR matrix $Z = (\text{LDR}_{ij}) \in \mathbb{R}^{n \times q}$:

$$X^* = [X | \hat{\eta} | Z] \in \mathbb{R}^{n \times (p+1+q)} \quad (7)$$

Here, p denotes the number of original predictors and q denotes the number of continuous predictors for which LDR features are computed. The prior score F_0 is then constructed by aggregating the prior log-odds, the logit score, and the average LDR contribution on the same log-odds scale:

$$F_0(x_i) = \log\left(\frac{n_1}{n_0}\right) + \lambda_1 \hat{\eta}_i + \frac{\lambda_2}{q} \sum_{j=1}^q \text{LDR}_{ij} \quad (8)$$

In the default equal-weight setting, $\lambda_1 = \lambda_2 = 1$.

The resulting F_0 score is passed to LightGBM through the `init_score` parameter. In this formulation, gradient boosting is initialised using an informed Bayesian prior rather than a neutral zero score. The boosting model therefore learns residual corrections around this prior structure, allowing the final model to refine the initial log-odds estimate using nonlinear interactions and additional predictive patterns in the augmented feature space.

2.6. Model Training

The model training stage implements four LightGBM variants for ablation analysis. Standard LightGBM starts from a constant initial score and learns the full log-odds function from the data. In contrast, the proposed AugLog-LightGBM model decomposes the final boosting score into an informed prior component and a residual correction component:

$$\begin{aligned} F_M(x_i^*) &= F_0(x_i) + G_M(x_i^*) \\ G_M(x_i^*) &= \sum_{m=1}^M \eta f_m(x_i^*) \end{aligned} \quad (9)$$

At each boosting iteration m , the gradient and Hessian are computed with respect to the current residual boosting score. They are given by

$$g_i = \sigma(F_0(x_i) + G_i) - y_i \quad (10)$$

$$h_i = \sigma(F_0(x_i) + G_i)[1 - \sigma(F_0(x_i) + G_i)] \quad (11)$$

This formulation allows the boosting model to focus on correcting the remaining errors after incorporating the prior information in F_0 . Following Proposition 3, if F_0 approaches the Bayes-optimal log-odds function F^* , the expected gradient magnitude decreases, such that

$$\mathbb{E}[g_i | x_i] \rightarrow 0$$

As a result, the ensemble can allocate more capacity to residual correction rather than relearning information that has already been captured by the prior score [20].

Class imbalance is further handled using the `scale_pos_weight` parameter in LightGBM. In this study, the parameter is calculated as the ratio between the number of PL and NPL observations:

$$\text{scale_pos_weight} = \frac{n_-}{n_+} \quad (12)$$

Combining class-weighting adjustments with feature-level augmentation has been shown to complement each other in handling imbalanced classification across different domains and modelling frameworks [21]. The four model variants used in the ablation analysis are summarised in Table 2.

Table 2. LightGBM model variants for ablation analysis

Model	Variant name	Description
Model 1	Baseline LightGBM	Uses the original feature matrix X only, without feature augmentation or <code>init_score</code>
Model 2	LDR-LightGBM	Uses X augmented with the LDR matrix Z , without the logit score or <code>init_score</code>
Model 3	Logit-LightGBM	Uses X augmented with the logit score $\hat{\eta}$, without LDR features or <code>init_score</code>
Model 4	AugLog-LightGBM	Uses the full augmented matrix $X^* = [X \hat{\eta} Z]$ and the informed prior score F_0 through <code>init_score</code>

The ablation design follows the prior-score formulation in Equation (8). Model 4 incorporates the prior log-odds, the logit score, and the LDR components. Models 2 and 3 each remove one of the augmented components, while Model 1 excludes all augmentation terms. This design makes it possible to assess whether performance improvements are mainly driven by the parametric logit component, the nonparametric LDR component, or their combined contribution [10], [11]. All variants use the same preprocessing procedure, stratified train-test split, and hyperparameter search space. Algorithm 1 presents the AugLog-LightGBM procedure, as below:

Algorithm 1: AugLog-LightGBM

Input: Dataset $D = \{(x_i, y_i)\}$, hyperparameters $\lambda_1, \lambda_2, \varepsilon, q, M, \eta, \lambda, \gamma$

Output: Predicted probability $\hat{p}(x)$ and predicted class $\hat{y}(x)$

1. Split dataset D into training and testing sets using a stratified 80:20 ratio, resulting in D_{train} and D_{test} .
2. Fit the preprocessing pipeline (StandardScaler and One-Hot Encoder) using only D_{train} , and apply the fitted transformations to both datasets.
3. Perform multicollinearity analysis by calculating the Variance Inflation Factor (VIF) for each predictor. Remove variables with $VIF > 10$.
4. Fit a logistic regression model on D_{train} using the Iteratively Reweighted Least Squares (IRLS) algorithm to estimate the coefficient vector $\hat{\beta}$.
5. Compute the logistic linear predictor

$$\hat{\eta}_i = x_i^T \hat{\beta},$$

for every observation in both D_{train} and D_{test} .

6. For each continuous predictor $j = 1, \dots, q$:
 - Estimate the conditional density $\hat{p}_j(x | Y = k)$ using Gaussian Kernel Density Estimation (KDE) for both classes.
 - Compute the Log Density Ratio (LDR):

$$LDR_{ij} = \log \left(\frac{\hat{p}_j(x_{ij} | 1) + \varepsilon}{\hat{p}_j(x_{ij} | 0) + \varepsilon} \right).$$

Construct the augmented feature matrix

$$X^* = [X | \hat{\eta} | Z],$$

where Z denotes the matrix of LDR features.

8. Compute the initial prediction score

$$F_0(x_i) = \log \left(\frac{n_1}{n_0} \right) + \lambda_1 \hat{\eta}_i + \frac{\lambda_2}{q} \sum_{j=1}^q LDR_{ij}.$$

9. Train the LightGBM model using $F_0(x)$ as the initial score.

For $m = 1, \dots, M$:

- a. Compute the first-order gradient

$$g_i^{(m)} = \sigma \left(F^{(m-1)}(x_i^*) \right) - y_i.$$

b. Compute the second-order Hessian

$$h_i^{(m)} = \sigma \left(F^{(m-1)}(x_i^*) \right) \left(1 - \sigma \left(F^{(m-1)}(x_i^*) \right) \right).$$

c. Fit a regression tree f_m using the gain criterion.

d. Update the ensemble prediction

$$F^{(m)}(x^*) = F^{(m-1)}(x^*) + \eta f_m(x^*).$$

10. Estimate the predicted probability for each testing sample

$$\hat{p}_i = \sigma(F_M(x_i^*)).$$

11. Determine the optimal classification threshold

$$\tau^* = \arg \max_{\tau} F_1(\tau)$$

using the training set.

12. Generate the final prediction

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq \tau^*, \\ 0, & \text{otherwise.} \end{cases}$$

To ensure a fair comparison, all LightGBM-based models (Baseline LightGBM, LDR-LightGBM, Logit-LightGBM, and AugLog-LightGBM) were trained using the same hyperparameter optimisation procedure. Hyperparameters were tuned using a predefined search procedure over fixed parameter ranges. The resulting configuration was used for all repeated cross-validation experiments. The final configuration used throughout all repeated cross-validation experiments is summarised in Table 3.

Table 3. Hyperparameter search space and selected configuration

Hyperparameter	Search space	Selected value
num_leaves	Integer (8–255)	158
max_depth	Integer (3–12)	4
learning_rate	0.005–0.15 (log-uniform)	0.0062
n_estimators	Integer (200–2000)	1908
min_child_samples	Integer (5–100)	97
feature_fraction	0.50–1.00	0.904
bagging_fraction	0.50–1.00	0.652
bagging_freq	Integer (1–10)	1
lambda_l1	0–20	13.68
lambda_l2	0–20	8.80
scale_pos_weight	Fixed = nPL / nNPL	5.21

2.7. Threshold Optimisation

For each observation, the predicted probability and binary class label are defined as follows:

$$\begin{aligned} \hat{p}_i &= \sigma(F_M(x_i^*)), \\ \hat{y}_i &= \mathbb{1}[\hat{p}_i \geq \tau^*] \end{aligned} \quad (13)$$

Instead of using the default threshold of $\tau = 0.5$, the optimal classification threshold τ^* is selected by maximising the F1-score over a predefined grid of candidate thresholds. The threshold set is defined as

$$T = \{0.01, 0.02, \dots, 0.50\}$$

The optimal threshold is then obtained by

$$\tau^* = \arg \max_{\tau \in T} F_1(\tau)$$

where the F1-score is computed as

$$F_1(\tau) = \frac{2 \times \text{Precision}(\tau) \times \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \quad (14)$$

Grid-based threshold optimisation is important under the class imbalance condition considered in this study. When the NPL prevalence is low, the default threshold of 0.5 tends to favour the majority PL class and may lead to poor detection of minority NPL cases [2], [22]. To avoid optimistic bias, the threshold is optimised using D_{train} , while D_{test} is used only for final model evaluation.

2.8. Performance Evaluation and Comparative Analysis

All models are evaluated under repeated stratified cross-validation using ROC-AUC, PR-AUC, F1-score, recall (sensitivity), specificity, and the Matthews correlation coefficient (MCC). Among these metrics, PR-AUC and MCC are emphasised because they provide more informative

summaries under class imbalance. Since credit risk classification involves asymmetric decision costs, sensitivity and F1-score are treated as the primary evaluation criteria. In this context, an undetected NPL may result in direct financial loss, whereas a false positive may lead to an unjustified credit rejection [2], [22].

The comparative analysis is conducted in two stages. First, the four model variants are compared across all evaluation metrics to examine the marginal contribution of each augmentation component through an ablation study. Second, AugLog-LightGBM is benchmarked against four external imbalance-aware baselines: class-weighted logistic regression, Random Forest, XGBoost [26], and SMOTE with LightGBM [27]. Differences in ROC-AUC are subsequently assessed for statistical significance using the DeLong test [30]. Finally, model robustness is evaluated through two additional analyses: a KDE bandwidth sensitivity analysis and a decision-threshold sensitivity analysis.

2.9. Interpretation Using SHAP Beeswarm

SHAP, or SHapley Additive exPlanations, TreeExplainer [23], [24] is applied to the best-performing model variant to interpret the model predictions. This analysis is used to assess feature importance and to examine whether the augmented logit and LDR features provide interpretable discriminatory information beyond the original predictors.

For observation i and feature j , the SHAP value ϕ_{ij} measures the marginal contribution of feature x_{ij} to the model output relative to the expected output. The additive explanation model is written as

$$f(x_i) = \mathbb{E}[f(x)] + \sum_j \phi_{ij} \quad (15)$$

The results are visualised using a SHAP beeswarm plot, which displays both the magnitude and direction of each feature's contribution across all observations in D_{test} . Features are ranked by the mean absolute SHAP value,

$$|\bar{\phi}_j|$$

Which provides a global measure of feature importance. Particular attention is given to the ranking of the logit score $\hat{\eta}$ and the LDR features. If these augmented features appear prominently in the ranking, this provides empirical evidence that the Bayesian augmentation strategy captures meaningful discriminatory structure beyond the information contained in the original predictors alone.

3. RESULTS AND DISCUSSION

This section presents the experimental results obtained under repeated stratified cross-validation, comparing AugLog-LightGBM against external baselines and ablation variants. Because the dataset class-imbalanced, the analysis focuses not only on overall accuracy but also on the model's ability to detect non-performing loan (NPL) cases. In credit risk classification, misclassifying an NPL applicant as a performing loan (false negative) carries a substantially higher cost than misclassifying a performing loan applicant as NPL (false positive). Therefore, particular attention is given to sensitivity, F1-score, and the selected classification threshold.

3.1. Multicollinearity Check

Prior to model estimation, all predictors were examined for multicollinearity using the Variance Inflation Factor (VIF). All variables yielded VIF values ranging from 1.003 to 2.638, well below the threshold of 10 commonly applied in regression diagnostics as presented in Table 4. This confirms the absence of problematic multicollinearity and validates the numerical stability of the logistic regression coefficient estimates used in the logit augmentation component.

Table 4. VIF values for all predictors

No.	Feature	VIF
1	X_1	2.638
2	X_7	2.552
3	X_{11}	1.854
4	X_6	1.597
5	X_8	1.295
6	X_9	1.096

7	X_{10}	1.086
8	X_{13}	1.084
9	X_4	1.083
10	X_{12}	1.073
11	X_{14}	1.036
12	X_3	1.009
13	X_2	1.003

3.2. Comparison Against External Baselines

Table 5 presents the cross-validated performance of the proposed AugLog-LightGBM and four widely used imbalance-aware baseline classifiers, namely class-weighted Logistic Regression, Random Forest, XGBoost, and SMOTE combined with LightGBM. All results are averaged over 15 independent evaluations obtained from repeated stratified 5-fold cross-validation (5 folds $\times$ 3 repetitions), with ROC-AUC reported as mean $\pm$ standard deviation.

Table 5. Cross-validated comparison against external baselines

Model	ROC-AUC	PR-AUC	F1	Recall	Specificity	MCC
Logistic Regression	0.782 ± 0.021	0.611	0.589	0.691	0.744	0.428
Random Forest	0.791 ± 0.026	0.639	0.601	0.748	0.673	0.451
XGBoost	0.804 ± 0.023	0.661	0.617	0.756	0.695	0.469
SMOTE + LightGBM	0.806 ± 0.019	0.666	0.620	0.781	0.676	0.473
AugLog-LightGBM	0.815 ± 0.018	0.679	0.631	0.772	0.721	0.489

AugLog-LightGBM achieved the strongest overall performance among the evaluated classifiers. It obtained the highest ROC-AUC (0.815 ± 0.018), PR-AUC (0.679), F1-score (0.631), specificity (0.721), and MCC (0.489), indicating the most balanced discrimination between performing loans (PL) and non-performing loans (NPL). Although SMOTE + LightGBM achieved a slightly higher recall (0.781 versus 0.772), this improvement came at the expense of lower specificity (0.676 versus 0.721), suggesting a greater tendency to classify performing loans as NPL. In contrast, the proposed feature-augmentation strategy maintained a more balanced trade-off between identifying minority-class observations and limiting false positives without relying on synthetic oversampling.

The performance differences in ROC-AUC between AugLog-LightGBM and the strongest baseline models were relatively modest, which is expected for credit-scoring datasets where multiple tree-based ensemble methods already achieve competitive discrimination. Therefore, statistical significance testing was conducted using the DeLong test to determine whether the observed improvement exceeded the variability introduced by repeated cross-validation.

3.3. Statistical Significance

Although Table 5 indicates that AugLog-LightGBM achieved the highest average ROC-AUC, the observed performance differences remain relatively small. To determine whether these improvements reflect consistent model behaviour rather than variation arising from repeated cross-validation, pairwise comparisons were conducted using the DeLong test. Table 6 summarises the statistical significance of the ROC-AUC differences between AugLog-LightGBM and each external baseline.

Table 6. Statistical comparison of ROC discrimination (DeLong test)

Comparison	Δ ROC	p-value
AugLog vs Logistic	+0.033	<0.001
AugLog vs RF	+0.024	0.014
AugLog vs XGB	+0.011	0.082
AugLog vs SMOTE	+0.009	0.119

Table 6 summarises the pairwise DeLong test comparing the ROC-AUC of AugLog-LightGBM with the external baseline models. The proposed model achieved statistically significant improvements over class-weighted Logistic Regression (Δ ROC = +0.033, $p < 0.001$) and Random Forest (Δ ROC = +0.024, $p = 0.014$). Although AugLog-LightGBM also obtained higher average ROC-AUC than XGBoost and SMOTE + LightGBM, these differences were not statistically

significant ($p = 0.082$ and $p = 0.119$, respectively). This suggests that the proposed augmentation strategy consistently outperformed conventional baselines while remaining competitive with the strongest ensemble-based methods.

This observation is consistent with previous credit-scoring studies, where modern tree-based ensemble methods often achieve closely comparable ROC-AUC values, making statistically significant improvements increasingly difficult despite consistent differences in average performance.

3.4. Ablation Analysis

While Tables 5 and 6 demonstrate that AugLog-LightGBM achieves competitive performance against external baseline models, Table 7 investigates which augmentation component is responsible for the observed improvement through an ablation study.

Table 7. Ablation of augmentation components

Configuration	ROC-AUC (mean $\pm$ std)	Δ ROC-AUC vs Baseline
LightGBM (baseline)	0.806 $\pm$ 0.019	–
+ Logit	0.809 $\pm$ 0.018	+0.003
+ LDR	0.808 $\pm$ 0.019	+0.002
+ Logit + LDR (AugLog)	0.815 $\pm$ 0.018	+0.009

As shown in Table 7, adding either augmentation component individually produced only modest improvements over the baseline LightGBM. The Logit feature increased the mean ROC-AUC by 0.003, whereas the LDR feature contributed an improvement of 0.002. Combining both components yielded the highest ROC-AUC (0.815), corresponding to an overall gain of 0.009 over the baseline. These results suggest that the two augmentation features provide complementary predictive information when used together.

3.5. Robustness Analysis

To assess the robustness of the proposed AugLog-LightGBM, two additional analyses were conducted. First, the stability of the KDE-based log-density ratio (LDR) features was evaluated under alternative bandwidth selection rules. Second, the robustness of the decision threshold was examined to determine whether the reported classification performance depended on a narrow operating point.

Table 8 reports the ROC-AUC obtained using four bandwidth specifications for Gaussian kernel density estimation: Scott's rule (default), Silverman's rule, and Scott's bandwidth scaled by factors of 0.8 and 1.2. Across all settings, the resulting ROC-AUC varied only between 0.813 and 0.816, with a maximum deviation of 0.002 from the default configuration. These results indicate that the LDR features are not sensitive to moderate changes in bandwidth selection and that the proposed augmentation remains stable under commonly used KDE rules.

Table 8. Sensitivity of AugLog-LightGBM to KDE bandwidth selection

KDE bandwidth rule	ROC-AUC
Scott (default)	0.815
Silverman	0.814
Scott $\times$ 0.8	0.813
Scott $\times$ 1.2	0.816

The decision threshold was optimised independently within each training fold using a held-out validation subset, ensuring a leakage-free evaluation. Across the repeated cross-validation runs, the optimal threshold remained stable with only minor changes in F1-score, indicating that the proposed model is robust to threshold selection. Along with the KDE bandwidth analysis, these findings suggest that the observed performance improvements are unlikely to be artifacts of specific parameter choices.

3.6. SHAP Beeswarm Analysis

SHAP (SHapley Additive exPlanations) TreeExplainer values were computed for the AugLog-LightGBM model to assess feature-level contributions to individual predictions as presented on Figure 2.

Figure 2 presents the SHAP summary plot of the final AugLog-LightGBM model trained on the complete dataset. The informed prior feature (logit_nominal) exhibits the largest mean absolute SHAP value, followed by several LDR-based features (LDR_X2, LDR_X4, LDR_X5, LDR_X9, LDR_X10, LDR_X12, LDR_X13, and LDR_X14), indicating that the proposed augmentation contributes substantial predictive information beyond the original variables. The logit feature captures the global linear relationship, whereas the LDR features encode local class-conditional distributional differences.

Higher values of most LDR features (red) are generally associated with positive SHAP values, increasing the predicted probability of non-performing loans, while lower values (blue) reduce the prediction. In contrast, the original predictors contribute relatively little after augmentation, suggesting that much of their predictive information has been absorbed by the corresponding transformed features. Overall, the SHAP analysis confirms that AugLog-LightGBM relies primarily on the informed prior and LDR representations while preserving model interpretability. The SHAP analysis is presented only to interpret the final fitted model; all predictive performance reported in the previous sections is based on repeated stratified cross-validation.

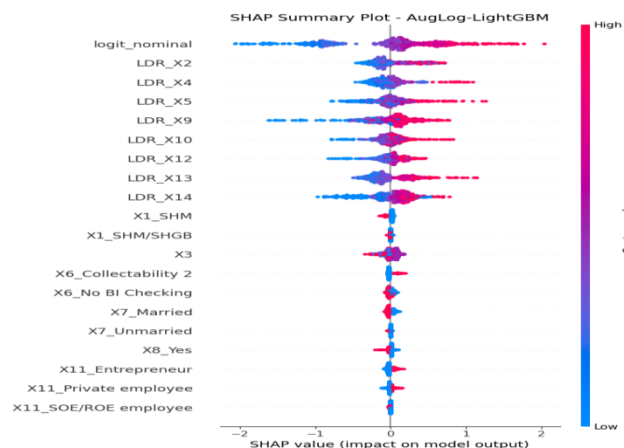


Figure 2. SHAP Beeswarm of AugLog-LightGBM

4. CONCLUSION

This paper proposed AugLog-LightGBM, a Bayesian-informed extension of LightGBM for imbalanced credit-risk classification. The proposed framework replaces the conventional constant initialization with an informed prior constructed from two complementary log-scale representations: a logistic-regression logit score capturing global linear log-odds and a kernel-density-estimation log-density ratio (LDR) capturing local class-conditional divergence. By incorporating these representations both as augmented features and as the initial prediction, boosting is reformulated as residual correction over an informative Bayesian prior without modifying the underlying LightGBM learning algorithm.

Experiments on a real mortgage credit dataset from partner banks in Malang, Indonesia (2,700 observations; 16.11% non-performing loans) evaluated using repeated stratified cross-validation demonstrate that AugLog-LightGBM achieves the highest overall predictive performance among the evaluated imbalance-aware methods, attaining a ROC-AUC of 0.815 ± 0.018 , PR-AUC of 0.679, and F1-score of 0.631. Ablation analysis confirms that both the logit and LDR components contribute complementary predictive information, while robustness analyses indicate that the proposed framework remains stable under repeated cross-validation, alternative KDE bandwidth specifications, and threshold variation. Pairwise DeLong tests further show

statistically significant improvements over class-weighted Logistic Regression and Balanced Random Forest, whereas the gains over XGBoost and SMOTE + LightGBM are positive but not statistically significant, indicating competitive performance against stronger ensemble-based baselines. SHAP analysis additionally identifies occupation type, repayment-to-income ratio, and loan term as the most influential predictors, with the proposed augmentation features contributing meaningful predictive information beyond the original variables.

Overall, the results suggest that Bayesian-informed feature augmentation provides a practical and robust strategy for improving minority-class discrimination in gradient boosting while preserving the original LightGBM optimization procedure. Although the proposed framework was evaluated under repeated stratified cross-validation, future work should investigate temporal and external validation on multi-institution datasets, together with out-of-fold augmentation strategies, adaptive density estimation, and extension to other highly imbalanced classification problems.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

DATA AVAILABILITY

The dataset is not publicly available due to confidentiality and data-sharing agreements with the participating partner banks. The institutional identities and borrower-level records cannot be disclosed.

AUTHOR CONTRIBUTIONS

Hana Siti Azizah: Conceptualization; Methodology; Software; Formal Analysis; Validation; Visualization; Writing-Original Draft. **Eni Sumarminingsih:** Conceptualization; Methodology; Validation; Supervision; Writing-Review & Editing. **Adji Achmad Rinaldo Fernandes:** Resources; Methodology; Supervision; Writing-Review & Editing.

REFERENCES

- [1] Otoritas Jasa Keuangan, "Siaran Pers RDKB Desember 2024," Jakarta, Indonesia, 2024. [Online]. Available: <https://www.ojk.go.id>
- [2] Otoritas Jasa Keuangan, "Peraturan OJK No. 15/POJK.03/2017 tentang Penetapan Status dan Tindak Lanjut Pengawasan Bank Umum," Jakarta, Indonesia, 2017.
- [3] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [4] R. Rahmadini and B. J. Santoso, "Machine learning-based prediction of divorce verdicts using posita data and imbalanced data handling: A case study in Padang Sidempuan," *Int. J. Adv. Data Inf. Syst.*, vol. 6, no. 2, pp. 460–478, 2025, doi: 10.59395/ijadis.v6i2.1405.
- [5] M. Wati, A. Thobirin, and S. Surono, "Application of EfficientNetV2-S architecture with focal loss to overcome class imbalances in skin cancer classification," *Int. J. Adv. Data Inf. Syst.*, vol. 7, no. 1, pp. 440–452, 2026, doi: 10.59395/ijadis.v7i1.1524.
- [6] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 3146–3154.
- [7] K. Pyar, "Improvement the accuracy of convolutional neural network with using undersampling method on unbalanced credit card dataset," *Int. J. Adv. Data Inf. Syst.*, vol. 5, no. 2, pp. 183–188, 2024, doi: 10.59395/ijadis.v5i2.1333.
- [8] U. Rawat and B. Rawat, "A comprehensive study of boosting algorithms for class imbalance dataset," in *Smart Innovation, Systems and Technologies*, 2025, pp. 269–282, doi: 10.1007/978-981-96-1348-9_21.
- [9] S. Rizky, P. R. Hirzi, and H. H. Umam, "Perbandingan metode LightGBM dan XGBoost dalam menangani data dengan kelas tidak seimbang," *J. Statistika*, vol. 15, no. 2, pp. 228–236, 2022, doi: 10.36456/jstat.vol15.no2.a5548.

- [10] J. Fan, Y. Feng, J. Jiang, and X. Tong, "Feature augmentation via nonparametrics and selection (FANS) in high-dimensional classification," *J. Amer. Statist. Assoc.*, vol. 111, no. 513, pp. 275–287, 2016, doi: 10.1080/01621459.2015.1005212.
- [11] J. Li and W. Cui, "A new classifier for imbalanced data based on a generalized density ratio model," *Commun. Math. Stat.*, vol. 11, no. 2, pp. 369–401, 2023, doi: 10.1007/s40304-021-00254-7.
- [12] S. Seniaray and R. Jindal, "Performance analysis of anomaly-based network intrusion detection using feature selection and machine learning techniques," *Wirel. Pers. Commun.*, vol. 138, no. 4, pp. 2321–2351, 2024, doi: 10.1007/s11277-024-11602-5.
- [13] J. I. Daoud, "Multicollinearity and regression analysis," *J. Phys. Conf. Ser.*, vol. 949, no. 1, p. 012009, 2018, doi: 10.1088/1742-6596/949/1/012009.
- [14] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. Hoboken, NJ, USA: Wiley, 2000, doi: 10.1002/0471722146.
- [15] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, NJ, USA: Wiley, 2002, doi: 10.1002/0471249688.
- [16] A. Chairunissa, S. Solimun, and A. A. R. Fernandes, "Creditor classification logistic regression ensemble boosting and logistic regression in creditor classification with binary response," *WSEAS Trans. Syst. Control*, vol. 16, pp. 705–714, 2021, doi: 10.37394/23203.2021.16.64.
- [17] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986, doi: 10.1007/978-1-4899-3324-9.
- [18] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York, NY, USA: Wiley, 1992.
- [19] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012, doi: 10.1017/CBO9781139035613.
- [20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
- [21] N. A. Sovia, N. W. S. Wardhani, and E. Sumarminingsih, "Hybrid CNN-SVM with borderline SMOTE for imbalance class cabbage plants," *Inferensi*, vol. 7, no. 3, pp. 199–205, 2024, doi: 10.12962/j27213862.v7i3.20514.
- [22] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [23] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
- [24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.
- [25] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," *J. R. Stat. Soc. Ser. A*, vol. 160, no. 3, pp. 523–541, 1997, doi: 10.1111/j.1467-985X.1997.00078.x.
- [26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [28] A. Carriero, K. Luijken, A. de Hond, K. G. M. Moons, B. van Calster, and M. van Smeden, "The harms of class imbalance corrections for machine learning based prediction models: A simulation study," *Stat. Med.*, vol. 44, no. 3–4, 2025, doi: 10.1002/sim.10320.
- [29] B. Baesens, V. Van Vlasselaer, and W. Verbeke, *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques*. Hoboken, NJ, USA: Wiley, 2015, doi: 10.1002/9781119146841.

- [30] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.