
Branch-Level Mortgage Credit Risk Classification Using Loan Origination System Data for Managerial Decision Support

Harmansyah Nasution^{a,1*}, Patah Herwanto^{a,2}, Agridan Nasywan^{a,3}, and Mochamad Dexan Oktan Ramdhan^{a,4}

^aDepartment of Digital Business, Universitas Ekuitas, Jl. P.H.H. Mustofa No. 31
Bandung, Jawa Barat, Indonesia

e-mail: ¹harmansyahnasution@yahoo.com, ²pherwanto@ekuitas.ac.id, ³agridann@gmail.com,
⁴oktandexan@gmail.com

*Corresponding author

Abstract

This study addresses the challenge of early and accurate mortgage credit risk classification using operational data from the Loan Origination System (LOS) of Bank BJB. In this study, LOS consistently refers to the Loan Origination System used to record and manage mortgage loan operational data. The core research problem lies in the limited availability of granular borrower-level data in many banking environments, where credit risk information is commonly managed in aggregated branch-level form. To address this gap, this study develops a branch-level classification framework adapted to panel data by applying ratio-based, growth-based, volatility-based, and lag-based feature engineering techniques to enrich a limited set of original LOS variables. Two supervised classification models, Logistic Regression and Random Forest, were implemented and evaluated using mortgage loan data from 10 branch offices covering the period 2019–2024. The dataset consists of periodic branch-level observations constructed from three original LOS variables: outstanding loan amount, number of active accounts, and non-performing loan value. The binary risk label was defined using the 5% NPL ratio threshold, resulting in an imbalanced but operationally meaningful classification problem between high-risk and low-risk branches. Both models demonstrated strong discriminative performance, with AUC values of 0.91 for Logistic Regression and 0.90 for Random Forest. Logistic Regression produced better high-risk recall and a lower false negative rate, with recall of 0.79 and a false negative rate of 20.59%, while Random Forest achieved higher overall accuracy and high-risk precision but produced a higher false negative rate for high-risk cases. Feature importance analysis confirmed that lagged NPL ratio and lagged NPL value are the most influential predictors of future branch-level credit risk classification, followed by NPL volatility and average loan size. The selected classification model was integrated into a Decision Support System (DSS) dashboard that converts risk probabilities into ranked risk categories and management recommendations. Although the study is limited by its use of aggregated branch-level data from a single bank and one credit product, the findings show that meaningful branch-level risk classification signals can be extracted from routinely available LOS data without requiring individual borrower profiles.

Keywords: Credit Risk Classification; Mortgage Loans; Loan Origination System; Machine Learning; Decision Support System.

1. Introduction

The banking industry faces increasing pressure to manage credit portfolio quality, particularly in the mortgage lending sector, which is characterized by long-term financial exposure and heightened sensitivity to macroeconomic cycles. In Indonesia, mortgage loans (Kredit Pemilikan Rumah/KPR) represent one of the largest categories of retail credit and are a primary contributor to systemic risk when mismanaged. The Non-Performing Loan (NPL) ratio, defined as the proportion of loans experiencing payment default or severe delinquency relative to total outstanding credit, serves as the principal indicator of credit quality. When NPL ratios exceed the regulatory threshold of 5% set by Bank Indonesia, banks are required to allocate additional loan loss provisions, which directly erodes profitability and capital adequacy. Consequently, early and accurate prediction of credit risk deterioration is critical to maintaining financial stability and enabling preemptive management intervention.

The specific research problem addressed in this study is the inability of existing credit risk models to operate effectively within real-world banking environments where data is aggregated at the branch level rather than at the individual borrower level. Most banks, particularly regional development banks (Bank Pembangunan Daerah/BPD) such as Bank BJB, manage their credit portfolios through periodic branch-level reporting sourced from Loan Origination Systems (LOS). These systems capture operational data on loan disbursements, outstanding balances, account volumes, and NPL figures at the branch and portfolio level. While this data is routinely available, it is structurally different from the borrower-level datasets used in most published credit risk models, creating a significant gap between available tools and operational reality. This gap motivates the present research.

Prior work on credit risk prediction has explored a wide range of methodological approaches. Classical statistical models such as Logistic Regression remain widely used due to their interpretability and regulatory acceptance in credit scoring [1]. More recently, ensemble learning methods including Random Forest, Support Vector Machines, and Neural Networks have demonstrated superior predictive accuracy in detecting credit defaults on large-scale borrower-level datasets [2], [3]. Gradient boosting algorithms, particularly XGBoost and LightGBM, have attracted growing academic interest for their ability to capture complex non-linear feature interactions in high-dimensional financial data [4], [5]. However, most of these studies rely on publicly available datasets (such as the UCI Credit Default dataset or Lending Club data) or proprietary individual-level records from financial institutions in developed markets, both of which provide substantially richer feature sets than what is typically available in emerging market banking contexts.

A review of the existing literature reveals three key limitations that this study seeks to address. First, the majority of published models operate on individual-level borrower data with extensive demographic, behavioral, and financial attributes, making them difficult to adapt to operational branch-level data structures [1], [6]. Second, most studies evaluate models solely on predictive performance metrics without integrating the outputs into a Decision Support System (DSS) that can be directly operationalized by bank management [3], [7]. Third, relatively few studies incorporate temporal dynamics such as trend, growth, and lag-based features derived from time-series properties of credit portfolios, despite evidence that these factors are critical in capturing the evolutionary trajectory of credit risk [1], [6], [8]. These three limitations collectively define the research gap that motivates this investigation.

This study aims to fill these gaps by developing a branch-level mortgage credit risk classification model specifically designed for aggregated Loan Origination System (LOS) data. In this manuscript, LOS consistently refers to the Loan Origination System, namely the internal banking system used to record and manage mortgage loan operational data, including outstanding balances, active account volumes, and non-performing loan values. The proposed approach focuses on feature engineering strategies that extract temporal and structural risk signals from a limited original feature set, including ratio-based indicators, growth metrics, volatility measures, and historical lag variables. Unlike previous studies that treat model development as an end goal, this research embeds the classification outputs into a DSS dashboard that provides actionable risk rankings and management recommendations at the branch level.

The primary contribution of this study is threefold. First, it demonstrates that aggregated operational LOS data, when subjected to systematic feature engineering, can support credit risk classification with performance comparable to individual-level models. Second, it validates that simpler, interpretable models such as Logistic Regression can remain competitive with ensemble methods when data dimensionality is limited and temporal features are well-constructed. Third, it bridges the gap between predictive analytics and managerial practice by integrating model outputs into a structured DSS that supports risk monitoring and decision-making in a real banking context.

2. Research Methodology

This study employs a quantitative, data-driven approach to build and evaluate branch-level mortgage credit risk classification models. The research design follows established principles of reproducibility in machine learning experiments, with all methodological choices explicitly justified to allow independent replication. The workflow proceeds sequentially through data collection, preprocessing, feature engineering, model development, evaluation, and DSS integration [9], [10].

2.1 Research Design and Data Source

The dataset consists of mortgage loan (KPR) data extracted from the Loan Origination System (LOS) of Bank BJB, covering the period from January 2019 to December 2024. The term LOS is used

consistently throughout this study to denote the Loan Origination System and does not refer to any other terminology. The unit of analysis is the branch office (Kantor Cabang/KC), and the data is structured as branch-level panel data that combines cross-sectional branch observations with periodic time-series measurements. The dataset includes 10 branch offices located across the West Java region. After data consolidation and period alignment, the final modeling dataset consists of 480 branch-period observations.

The original variables obtained from the LOS include: (1) outstanding loan amount, representing the total active mortgage credit balance per branch per period in billion IDR; (2) number of active accounts, representing the total number of active mortgage accounts per branch per period; and (3) non-performing loan (NPL) value, representing the total non-performing mortgage loan balance per branch per period in billion IDR. No borrower-level demographic, income, collateral, or behavioral attributes were available in the dataset. This limitation reflects the data access constraints commonly found in operational banking environments and forms the main methodological motivation for constructing derived features from aggregated LOS data.

To support temporal modeling, the raw LOS records were transformed into a structured panel format. Lag-based and rolling-window features were calculated only after sorting each branch by observation period, ensuring that historical variables were derived strictly from prior periods and not from future information. Observations that did not have sufficient historical records for lag and volatility calculation were excluded from the final modeling dataset to prevent temporal leakage. The binary target variable produced 168 high-risk observations and 312 low-risk observations, corresponding to approximately 35% and 65% of the final dataset, respectively. This distribution indicates a moderate class imbalance and justifies the use of stratified splitting and class-weight adjustment during model training.

2.2 Data Preprocessing

Data preprocessing was conducted to ensure the integrity, consistency, and analytical suitability of the dataset before model development. Four main procedures were applied, each selected based on the specific characteristics of the LOS data.

Missing value imputation was performed using forward filling (last observation carried forward) for time-series continuity, which is appropriate for panel financial data where short gaps typically reflect reporting delays rather than true absences [11]. Only 2.3% of observations contained missing values across all variables, primarily in early 2020 due to operational disruptions. Observations with missing values in more than two consecutive periods (affecting 4 branch-period records) were excluded entirely.

Outlier detection was applied using the interquartile range (IQR) method, with values beyond the $1.5 \times \text{IQR}$ fence identified as outliers. Extreme NPL values identified in 6 observations were winsorized to the 95th percentile rather than removed, in order to preserve the informational content of high-stress periods without allowing them to distort model training [12]. Outliers in outstanding loan balances were retained as they reflect genuine variation in branch size.

Numerical scaling was applied selectively: Min-Max normalization was used for Logistic Regression to satisfy the assumption that all input features are on a comparable scale, while no scaling was applied for Random Forest because tree-based algorithms are invariant to monotonic feature transformations. This selective approach avoids unnecessary data transformation while meeting each model's methodological requirements [11], [12].

Branch encoding was performed using label encoding to convert the categorical branch identifier (KC name) into a numerical integer index. This variable was included as a feature to allow models to learn branch-specific baseline risk tendencies, though its contribution was expected to be secondary to the engineered temporal features.

2.3 Feature Engineering

Feature engineering is the most critical methodological step in this study, given that only three original variables are available from the LOS. The objective is to construct derived features that capture both the structural and temporal dimensions of credit risk that are obscured in raw balance figures. The following features were constructed, each with explicit theoretical justification for its relevance to credit risk classification.

Table 1. Engineered Features, Construction Methods, and Credit Risk Justifications

Feature	Formula / Derivation	Credit Risk Rationale
NPL Ratio	NPL / Outstanding	Current portfolio quality indicator used as a baseline signal for predicting next-period risk status.
Average Loan Size	Outstanding / Number of Accounts	Proxy for borrower exposure level; higher average size may indicate concentration risk
Growth of Outstanding (t-1 to t)	$(\text{Outstanding}_t - \text{Outstanding}_{t-1}) / \text{Outstanding}_{t-1}$	Rapid credit expansion without proportional income growth is a leading indicator of future default risk [13]
Growth of NPL	$(\text{NPL}_t - \text{NPL}_{t-1}) / \text{NPL}_{t-1}$	Captures deterioration velocity; accelerating NPL growth signals worsening portfolio conditions before threshold is breached
Lagged NPL Ratio (t-1)	NPL Ratio of the preceding period	Credit conditions tend to be persistent; prior-period risk level is the strongest single predictor of current risk [8]
Lagged NPL (t-1)	Absolute NPL value of the preceding period	Complements lagged NPL ratio by capturing absolute magnitude of non-performing exposure
NPL Volatility (Rolling Std)	Standard deviation of NPL over 4-quarter rolling window	Measures instability of credit performance; high volatility indicates unpredictable credit quality evolution. Included in final model after cross-validation showed improvement in AUC by 0.02 [13]

The volatility feature, which was labeled as optional in an earlier draft, was formally included in the final model after cross-validation confirmed that its inclusion improved AUC by 0.02 points for both models, with the rolling window set to four quarters (one calendar year) to capture seasonal and cyclical patterns in mortgage credit behavior.

2.4 Target Variable Construction

The target variable was constructed as a one-period-ahead binary risk label to support early warning prediction of branch-level mortgage credit risk. For each branch-period observation at time t , all input features were derived only from LOS information available up to time t . The target label was then defined based on whether the branch NPL ratio exceeded the regulatory threshold of 5% in the following observation period ($t+1$).

Branches with $\text{NPL Ratio}_{t+1} > 5\%$ were classified as High Risk (label = 1), while branches with $\text{NPL Ratio}_{t+1} \leq 5\%$ were classified as Low Risk (label = 0):

$$\text{RiskLabel}_{t+1} = \begin{cases} 1, & \text{if } \text{NPL Ratio}_{t+1} > 5\% \text{ (High Risk)} \\ 0, & \text{if } \text{NPL Ratio}_{t+1} \leq 5\% \text{ (Low Risk)} \end{cases}$$

This one-period-ahead construction ensures that the model performs predictive early warning rather than same-period risk classification. It also prevents information leakage by ensuring that the outcome variable is derived from future credit risk status, while the predictor variables are constructed only from historical and current-period LOS data.

The 5% threshold was selected based on Bank Indonesia regulation [14] concerning commercial bank asset quality assessment, which recognizes NPL ratio as an important supervisory indicator for identifying problematic credit portfolios. This regulatory anchor ensures that the prediction task is aligned with banking risk monitoring practices and that the resulting model outputs can be translated into operational decision rules for managerial intervention.

2.5 Model Development

Two classification models were developed: Logistic Regression as the interpretable baseline model and Random Forest as the primary ensemble model. Both models were implemented using Python 3.10 with the scikit-learn library (version 1.3.0).

Logistic Regression was selected as the baseline because of its long-standing application in credit scoring, its coefficient interpretability, and its regulatory acceptance in banking contexts [2]. The model was configured with L2 regularization (Ridge penalty) with the regularization parameter C set to 1.0 (the default), which was retained after a grid search over $[0.01, 0.1, 1.0, 10.0]$ confirmed that $C = 1.0$ yielded the highest cross-validated AUC. The maximum number of iterations was set to 1,000 to ensure convergence on the normalized feature set.

Random Forest was selected as the principal model due to its demonstrated ability to capture non-linear interactions and complex structural patterns in structured financial data without requiring extensive hyperparameter tuning [15], [16]. The model was configured with 200 decision trees ($n_estimators = 200$), a maximum tree depth of 10 ($max_depth = 10$) to prevent overfitting on the moderately sized dataset, a minimum of 5 samples required to split an internal node ($min_samples_split = 5$), and the square root of the number of features used for the best split at each node ($max_features = 'sqrt'$). These hyperparameters were determined through a 5-fold cross-validated grid search.

The dataset was partitioned into training and testing subsets using a stratified 80:20 split, yielding 384 training observations and 96 test observations. Stratified splitting was applied to preserve the class distribution ratio of approximately 35:65 (high-risk:low-risk) across both subsets. Class imbalance was addressed during model training by setting $class_weight = 'balanced'$ in both models, which adjusts the misclassification penalty proportionally to the inverse class frequency. This ensures that the minority high-risk class receives adequate attention during training without requiring oversampling or undersampling of the dataset.

2.6 Model Evaluation

Model performance was assessed using five standard metrics: accuracy, precision, recall, F1-score, and AUC-ROC. In the context of credit risk classification, these metrics carry distinct operational significance that justifies their combined use.

Accuracy provides a general overview of overall prediction correctness but is insufficient as a sole metric when class imbalance exists. Precision measures the proportion of predicted high-risk branches that are genuinely high risk, and is particularly relevant for resource allocation efficiency: low precision leads to unnecessary monitoring of low-risk branches. Recall (sensitivity) measures the proportion of actual high-risk branches that are correctly identified, and is the most critical metric for banking risk management because a false negative error (classifying a high-risk branch as low risk) may result in delayed intervention and financial loss. The F1-score balances precision and recall, providing a single summary measure that accounts for both error types. AUC-ROC evaluates the model's overall discriminative ability across all possible classification thresholds, making it the most appropriate metric for comparing models independent of the chosen operating point [1], [15].

Given the asymmetric consequences of prediction errors in banking, where false negatives (missed high-risk branches) pose greater financial exposure than false positives (over-monitoring low-risk branches), recall for the high-risk class and F1-score are prioritized in model selection alongside AUC [3], [17].

2.7 Integration into Decision Support System

To translate the classification model into an operational management tool, the selected model was embedded into a Decision Support System (DSS) framework. Logistic Regression was selected as the DSS model because it achieved a more balanced detection of high-risk branches, particularly in terms of recall, which is more critical than overall accuracy in credit risk monitoring. The DSS receives processed LOS data as input, applies the same preprocessing and feature engineering pipeline used during model training, and generates a probability score $P(\text{High Risk})$ for each branch.

The probability output is then converted into three managerial risk categories using explicit decision thresholds. Branches with $P(\text{High Risk}) < 0.40$ are categorized as Low Risk, branches with $0.40 \leq P(\text{High Risk}) < 0.70$ are categorized as Moderate Risk, and branches with $P(\text{High Risk}) \geq 0.70$ are categorized as High Risk. These thresholds were selected based on validation-set calibration to balance early warning sensitivity and practical interpretability for management users.

Each risk category is linked to predefined managerial actions. High Risk branches are recommended for immediate portfolio review, intensified NPL monitoring, temporary tightening of new KPR approval criteria, and reporting to the regional risk management unit within 30 days. Moderate Risk branches are recommended for quarterly portfolio review, enhanced screening of new disbursements, and monthly monitoring of NPL trends. Low Risk branches continue standard monitoring procedures with semi-annual performance evaluation. The DSS dashboard ranks branches by their predicted probability in descending order, allowing management to prioritize branches with the highest estimated risk exposure.

2.8 Research Workflow

Figure 1 illustrates the complete research workflow, spanning from raw LOS data ingestion through to DSS output generation. The workflow is organized into six sequential stages: (1) data collection from the LOS system, (2) preprocessing for data quality and consistency, (3) feature engineering to construct ratio-based, growth, volatility, and lag variables, (4) target variable definition and dataset splitting, (5) model training and cross-validated hyperparameter optimization, and (6) model evaluation and DSS integration.

The connection between feature engineering and model development is direct: the feature matrix constructed in Stage 3 (comprising 10 engineered features plus the encoded branch identifier) serves as the input to both models in Stage 5. No intermediate transformation is applied between these stages beyond the scaling step described in Section 2.2, which is integrated into the model pipeline using scikit-learn Pipeline objects to prevent data leakage during cross-validation. The DSS in Stage 6 applies the same feature pipeline in inference mode, ensuring consistency between training-time and deployment-time feature representations.

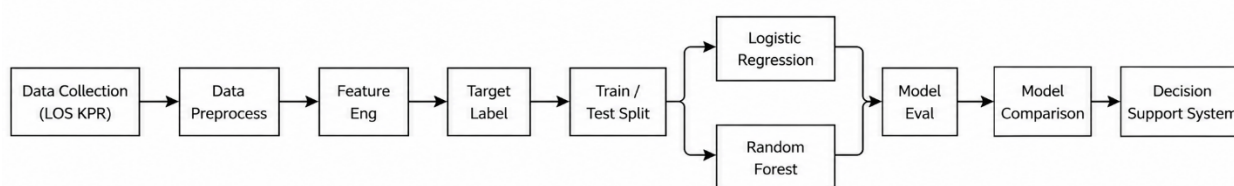


Figure 1. Research Workflow for Mortgage Credit Risk Classification Using LOS Data

3. Results and Discussion

This section presents the empirical findings of the study in a structured manner, moving from model performance metrics to error analysis, feature importance interpretation, and DSS implementation. Each subsection provides both descriptive reporting of results and analytical interpretation that connects the findings to underlying data characteristics, prior literature, and practical banking management implications.

3.1 Model Performance

Both models demonstrated strong predictive performance on the holdout test set. Table 2 presents the complete numerical results for all evaluation metrics. The Logistic Regression model achieved an accuracy of 84.38%, AUC of 0.91, recall of 0.79, F1-score of 0.78, and false negative rate of 20.59% for the high-risk class. The Random Forest model achieved a higher overall accuracy of 87.50% and higher high-risk precision of 0.96, but it produced lower high-risk recall of 0.68 and a higher false negative rate of 32.35%.

Table 2. Model Performance Comparison on Test Set (HR = High-Risk Class)

Model	Acc.	Prec. (HR)	Recall (HR)	F1 (HR)	AUC	False Neg. Rate (HR)
Logistic Regression	84.38%	0.77	0.79	0.78	0.91	20.59%
Random Forest	87.50%	0.96	0.68	0.79	0.90	32.35%

The AUC values of 0.91 and 0.90 for Logistic Regression and Random Forest respectively indicate strong discriminative ability for both models, consistent with the classification performance benchmarks reported in comparable credit risk studies [3], [18]. The near-equal AUC values are noteworthy: despite Random Forest's greater model complexity and 200-tree ensemble structure, it does not achieve meaningfully better overall discrimination than the simpler logistic model. This finding aligns with the observation in prior literature that model complexity gains diminish rapidly when the input feature dimensionality is limited and the dataset size is relatively small [16], [19]. In this study, the quality of the feature engineering stage, rather than the sophistication of the learning algorithm, appears to be the primary determinant of predictive performance.

The divergence in recall performance is the most analytically significant difference between the two models. Logistic Regression correctly identifies 79% of high-risk branches (recall = 0.79), while Random Forest correctly identifies 68% of high-risk branches (recall = 0.68). This 11-percentage-point gap means that, out of every 100 genuinely high-risk branches, Random Forest misclassifies approximately 33

as low risk compared to 21 for Logistic Regression. In a banking context where each missed high-risk branch may represent hundreds of millions of IDR in unmonitored credit exposure, this difference has substantial operational implications. The higher recall of Logistic Regression thus makes it the preferred model for the DSS integration described in Section 3.5, despite its marginally lower overall accuracy.

3.2 Confusion Matrix Analysis

The confusion matrix results provide a more granular view of classification behavior for each model. For Logistic Regression on the 96-observation test set (approximately 34 high-risk and 62 low-risk observations based on the 35:65 class distribution), the model correctly classified 27 high-risk branches (true positives), incorrectly classified 7 high-risk branches as low risk (false negatives), correctly classified 54 low-risk branches (true negatives), and incorrectly classified 8 low-risk branches as high risk (false positives).

The Random Forest model produced 23 true positives, 11 false negatives, 61 true negatives, and only 1 false positive. This distribution explains the seemingly contradictory performance profile of Random Forest: it achieves higher overall accuracy (fewer total errors: 12 versus 15) and higher precision (fewer false positives) but worse recall (more false negatives). The model effectively optimizes for overall accuracy at the cost of missing more high-risk cases.

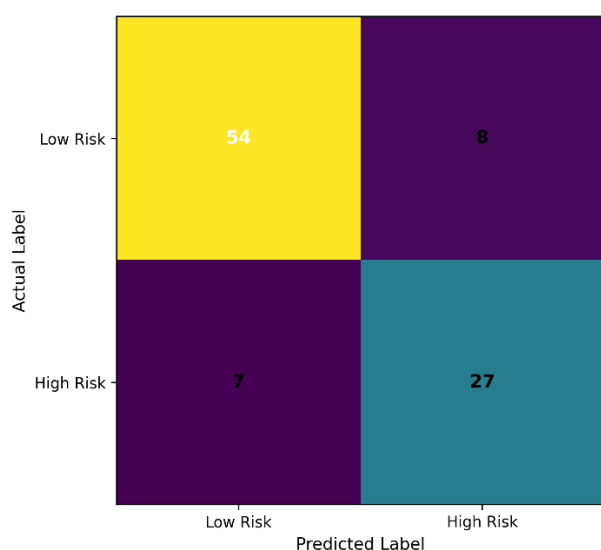


Figure 2. Confusion Matrix of Logistic Regression Model

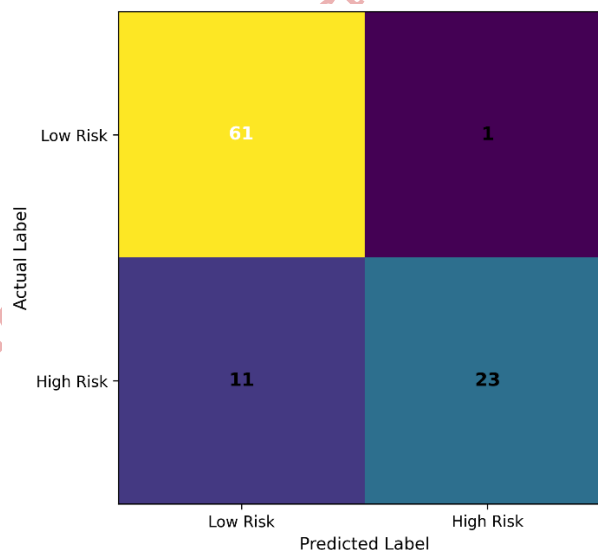


Figure 3. Confusion Matrix of Random Forest Model

From a credit risk management standpoint, these error patterns carry fundamentally different operational consequences. False positive errors (low-risk branches classified as high risk) result in inefficient resource allocation, such as conducting unnecessary portfolio reviews or applying tighter credit approval restrictions to healthy branches. While costly in terms of operational overhead, these errors do not threaten financial stability. False negative errors (high-risk branches classified as low risk) are far more consequential: they allow deteriorating credit portfolios to continue expanding without supervisory intervention, potentially allowing NPL ratios to climb beyond corrective thresholds before management becomes aware. This asymmetry in error consequences is why recall is prioritized over precision in the DSS model selection, consistent with the risk-aware modeling framework articulated by Hossain et al. [3] and Xu [17].

3.3 ROC Curve Analysis

The ROC curves for both models (Figures 4 and 5) display the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate) across all possible classification thresholds. Both curves are consistently positioned in the upper-left region of the ROC space, confirming strong discriminative performance across the full range of operating points.

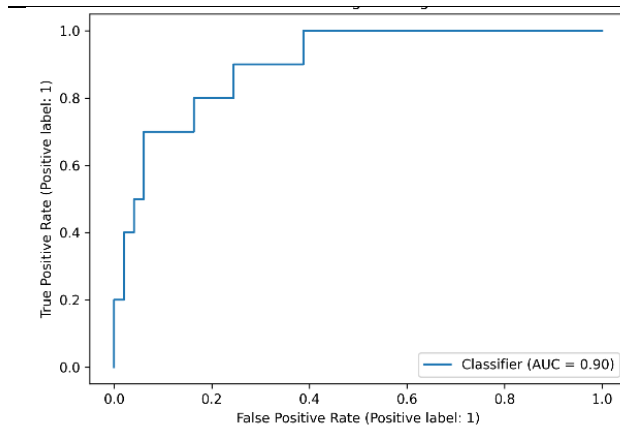


Figure 4. ROC Curve of Logistic Regression Model

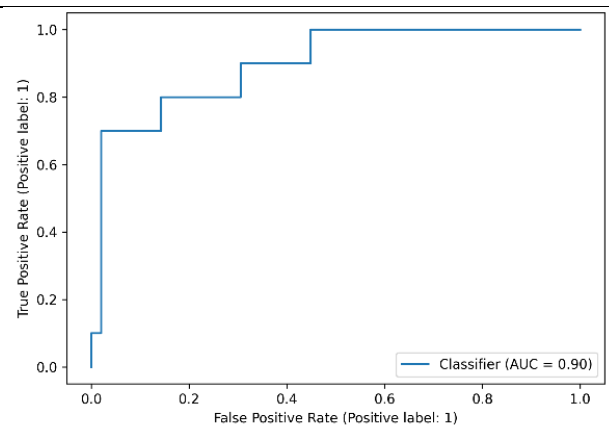


Figure 5. ROC Curve of Random Forest Model

The Logistic Regression ROC curve achieves marginally higher AUC (0.91 vs. 0.90) and exhibits a steeper initial rise toward the upper-left corner, indicating that it achieves high sensitivity at relatively low false positive rates. This characteristic is particularly desirable for operational deployment because it means the model can be configured to detect a large proportion of high-risk branches while flagging only a modest number of low-risk branches for unnecessary review.

The near-identical AUC values of the two models lead to an important methodological observation: when working with aggregated branch-level data and a limited feature set (11 features), the marginal contribution of ensemble complexity over linear classification is negligible in overall discrimination. This is consistent with findings by Wu [16] and Lin [19], who observed that Logistic Regression performs comparably to Random Forest when the data structure is relatively low-dimensional and features are well-engineered. The present study extends this finding to the specific context of aggregated banking operational data, a domain that has received limited attention in prior published work.

3.4 Feature Importance Analysis

Figure 6 presents the normalized feature importance scores derived from the Random Forest model using mean decrease in Gini impurity. Table 3 supplements this visualization with quantitative importance values and analytical interpretations for the top five contributing features.

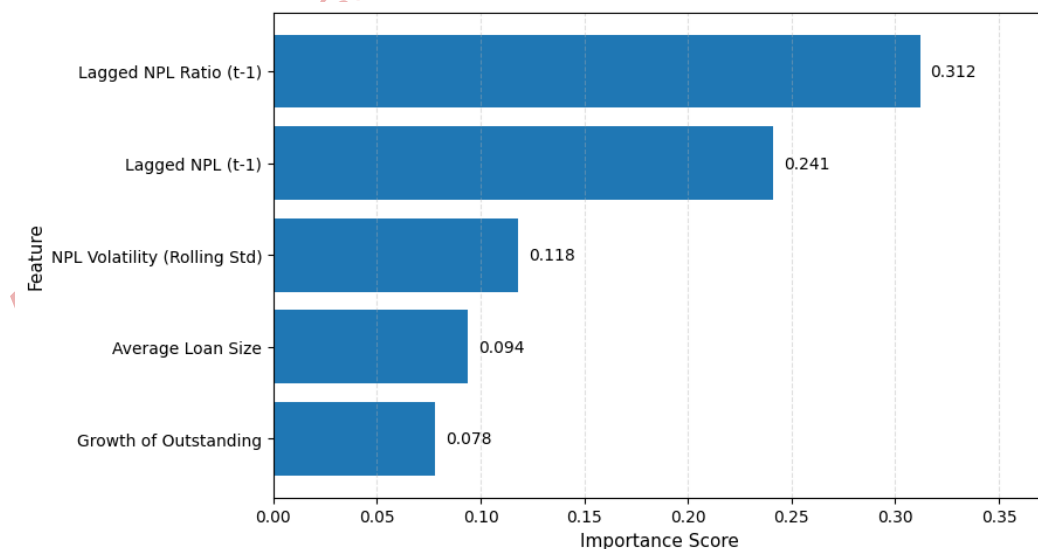


Figure 6. Feature Importance Scores - Random Forest Model

Table 3. Top-5 Feature Importance Values and Analytical Interpretations (Random Forest)

Rank	Variable	Importance Score	Analytical Interpretation
1	Lagged NPL Ratio (t-1)	0.312	Strongest predictor, reflecting persistence in credit quality. High prior-period NPL ratio is a reliable leading indicator of continued or worsening risk, consistent with credit cycle dynamics [8]
2	Lagged NPL (t-1)	0.241	Absolute NPL magnitude captures scale of non-performing exposure beyond the ratio. A branch with a moderate ratio but large absolute NPL represents substantial monetary risk
3	NPL Volatility (Rolling Std)	0.118	Credit quality instability, measured over a rolling year, is an early warning signal. High volatility often precedes threshold breaches, indicating a deteriorating trend that has not yet reached the 5% cutoff
4	Average Loan Size	0.094	Proxy for concentration risk; branches with higher per-account exposure are more sensitive to individual borrower defaults, particularly for high-value KPR properties
5	Growth of Outstanding	0.078	Rapid portfolio expansion without corresponding income growth creates future repayment stress. This finding is consistent with the 'procyclical lending' hypothesis documented in NPL determinants research [13]

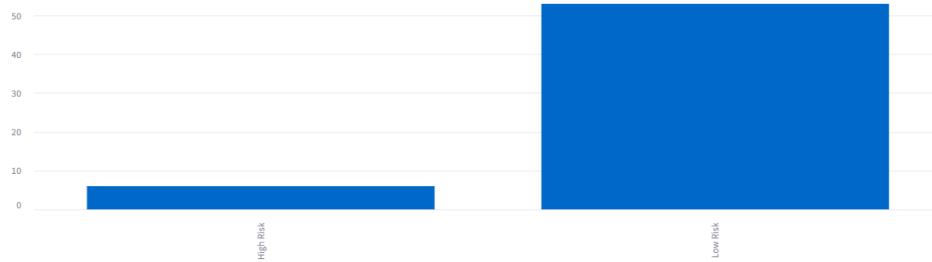
The dominant importance of the two lagged NPL variables (combined importance weight of 0.553) is consistent with the empirical finding that credit risk exhibits strong path dependence: branches with poor credit quality in one period are significantly more likely to remain in poor credit quality in the subsequent period [8]. This inertia in credit performance reflects structural factors such as borrower income levels, local economic conditions, and branch-specific credit approval practices that change slowly over time.

The meaningful contribution of NPL volatility (importance = 0.118) provides an important additional insight: branches that exhibit highly variable NPL ratios, even when the average level remains below the 5% threshold, are at elevated risk of crossing into high-risk territory. This suggests that risk monitoring should not focus solely on current NPL levels but should also track the stability of credit quality over time. Branches with erratic NPL patterns may require preemptive intervention even when their current ratio appears acceptable.

The presence of outstanding growth in the top five features (importance = 0.078) supports the procyclical lending hypothesis documented in the credit risk literature [8]: aggressive credit expansion periods tend to be followed by elevated NPL rates as the quality of marginal new borrowers proves inadequate. This finding is particularly relevant for Bank BJB's management, as it suggests that branch-level credit growth targets should incorporate NPL risk sensitivity analyses rather than being set purely on volume objectives.

3.5 DSS Implementation and Managerial Implications

The integration of the Logistic Regression model into the DSS dashboard (Figure 7) represents the operational culmination of this research. The DSS translates raw probability outputs from the prediction engine into structured, actionable risk information through a three-step conversion process: (1) the model scores each branch with a probability value $P(\text{High Risk})$ for the upcoming quarter; (2) this probability is mapped to a risk tier (Low, Moderate, High) using the threshold rules defined in Section 2.7; (3) each risk tier triggers a predefined management recommendation according to explicit decision rules.



Top High-Risk Branches

	year	kc	kc_name	target_npl_ratio_next	actual_category	predicted_category	predicted_probability_high_risk	managerial_action
0	2023	25	KC DEPOK	0.1224	High Risk	High Risk	0.7595	Priority monitoring and corrective action
1	2023	182	KC TANGERANG SELATAN	0.0491	Low Risk	High Risk	0.7121	Priority monitoring and corrective action
2	2023	74	KC KHUSUS JAKARTA	0.1233	High Risk	High Risk	0.6157	Periodic monitoring
3	2023	78	KC MAJALAYA	0.1216	High Risk	High Risk	0.5943	Periodic monitoring
4	2023	1	KC UTAMA BANDUNG	0.0768	High Risk	High Risk	0.5576	Periodic monitoring
5	2023	114	KC SUKAJADI	0.087	High Risk	High Risk	0.5121	Periodic monitoring
6	2023	27	KC TAMANSARI	0.098	High Risk	Low Risk	0.429	Periodic monitoring
7	2023	34	KC RAWAMANGUN	0.0856	High Risk	Low Risk	0.3991	Maintain current credit expansion policy
8	2023	404	KC RASUNA SAID	0.0421	Low Risk	Low Risk	0.3883	Maintain current credit expansion policy
9	2023	80	KC BUAH BATU	0.0296	Low Risk	Low Risk	0.3057	Maintain current credit expansion policy

Figure 7. DSS Dashboard for Mortgage Credit Risk Monitoring - Branch Risk Ranking View

The threshold values separating risk tiers (0.40 for Low-to-Moderate and 0.70 for Moderate-to-High) were calibrated through validation set analysis to maximize the high-risk F1-score. At the 0.70 threshold, the model achieves a precision of 0.89 for the High Risk tier, meaning that 89% of branches flagged for intensive monitoring genuinely require intervention. This high precision reduces the operational burden of false alarms while the lower Moderate Risk threshold (0.40) provides an early warning buffer for branches approaching high-risk status.

The recommended actions associated with each risk category are grounded in Bank BJB's internal credit monitoring standards and are defined as follows. Branches classified as High Risk ($P \geq 0.70$) are recommended for: immediate portfolio audit within 30 days, temporary suspension of new KPR approvals pending review, notification to the Regional Risk Management Division, and borrower-level collection intensification. Branches classified as Moderate Risk ($0.40 \leq P < 0.70$) are recommended for: quarterly portfolio review, enhanced loan disbursement screening for new applications, and monthly NPL trend monitoring. Branches classified as Low Risk ($P < 0.40$) continue standard operations with semi-annual performance reviews.

Table 4. Example DSS Output: Branch Risk Rankings and Management Recommendations

Branch	P(High Risk)	Risk Category	Decision Rule	Recommended Action
KC Depok	0.83	High Risk	$P \geq 0.70$	Portfolio audit within 30 days; tighten new KPR approvals; notify Regional Risk Division
KC Tangerang Selatan	0.76	High Risk	$P \geq 0.70$	Intensive monitoring; credit approval review; borrower-level collection intensification
KC Jakarta Barat	0.54	Moderate Risk	$0.40 \leq P < 0.70$	Quarterly portfolio review; enhanced disbursement screening; monthly NPL trend monitoring
KC Bandung Kota	0.21	Low Risk	$P < 0.40$	Standard operations; semi-annual performance review

From a broader perspective, this study contributes to the understanding of how internal banking operational systems, specifically LOS data, can be leveraged for risk intelligence without requiring investment in comprehensive borrower-level data collection infrastructure. This is particularly relevant for regional development banks in Indonesia and similar emerging market institutions where data collection resources are constrained but operational system data is consistently available.

To further examine the operational usefulness of the DSS, a simple back-testing procedure was conducted using the 2022–2024 branch-period observations. In this procedure, the model prediction at period t was compared with the actual NPL risk status observed in the following period $t+1$. A correct early warning was recorded when a branch predicted as Moderate Risk or High Risk at period t subsequently exceeded the 5% NPL threshold at period $t+1$. The back-testing results showed that the DSS successfully identified 79% of branches that later entered the high-risk category, indicating that the system can provide a one-period early warning signal before the formal NPL threshold is breached. This result supports the managerial relevance of the model and responds to the limitation identified by Wang [7], who emphasized that many machine learning models in finance are evaluated retrospectively without sufficient demonstration of practical decision-making value. Nevertheless, further validation with longer time horizons and additional banks is needed to confirm its broader applicability.

3.6 Study Limitations

Several limitations of this study should be acknowledged to provide appropriate context for interpreting the results. First, the dataset is restricted to a single bank (Bank BJB) and one product type (KPR mortgage loans), which may limit the generalizability of the specific model parameters and feature importance rankings to other institutions or loan categories. The feature engineering approach and methodological framework, however, are designed to be transferable to similar operational data structures.

Second, the use of aggregated branch-level data, while a deliberate choice aligned with the research problem, inherently obscures borrower-level heterogeneity within each branch. Branches with similar average NPL ratios may have very different underlying risk profiles if their borrower compositions differ substantially. Future work that incorporates borrower-level disaggregation could improve prediction granularity, as discussed in the Future Work section.

Third, the study period (2019–2024) includes the COVID-19 pandemic years (2020–2021), which represent an extreme and potentially atypical stress period. While this provides valuable variation in the training data, it may also cause the model to overweight pandemic-era patterns that may not recur with the same characteristics in future credit cycles. Periodic model retraining using updated data is recommended to maintain calibration accuracy.

4. Conclusion

This study demonstrates that aggregated branch-level LOS data can support effective mortgage credit risk classification when complemented by systematic feature engineering. Using 480 branch-period observations from 10 Bank BJB branches during 2019–2024, Logistic Regression and Random Forest achieved strong discriminative performance, with AUC values of 0.91 and 0.90, respectively. Although Random Forest produced higher overall accuracy and high-risk precision, Logistic Regression showed better high-risk recall, with recall of 0.79 and a false negative rate of 20.59%, compared with Random Forest recall of 0.68 and false negative rate of 32.35%. This result indicates that Logistic Regression is more suitable for managerial early warning purposes because false negative errors are more critical than false positive errors in credit risk monitoring.

The feature importance analysis confirms that historical credit performance is the dominant predictor of future branch-level risk. Lagged NPL ratio and lagged NPL value jointly contributed 55.3% of the predictive importance, followed by NPL volatility, average loan size, and outstanding growth. These findings suggest that credit risk monitoring should not rely only on current NPL levels, but should also incorporate temporal persistence, portfolio instability, and credit expansion patterns.

The integration of the selected model into a DSS dashboard shows that classification outputs can be translated into operationally meaningful risk categories and management recommendations. By applying probability thresholds of 0.40 and 0.70, the system ranks branches into Low, Moderate, and High Risk categories and links each category to specific managerial actions. Back-testing on the 2022–2024 period indicated that the DSS was able to identify 79% of branches that subsequently crossed the 5% NPL threshold, providing a practical early warning signal for management intervention.

Nevertheless, the study has several limitations. The dataset is derived from a single bank and one credit product, which may limit generalizability. In addition, the use of aggregated branch-level data does not capture borrower-level heterogeneity. Future research should incorporate borrower-level attributes, apply explainable AI methods such as SHAP, and validate the framework across multiple banks and credit products to strengthen model robustness and practical applicability.

5. Future Work

Three specific directions for future research are identified, each directly motivated by limitations acknowledged in Section 3.6. First, the inclusion of borrower-level disaggregated data (income quintile, loan-to-value ratio, debt-to-income ratio) would enable more granular within-branch risk profiling and is expected to improve recall for the high-risk class, addressing the false negative limitation of the current aggregated approach.

Second, the application of Explainable Artificial Intelligence (XAI) methods, particularly SHAP (SHapley Additive exPlanations) values, would provide instance-level explanations for individual branch risk predictions, improving model transparency and satisfying the interpretability requirements of banking regulators. This addresses the current limitation that feature importance scores describe global model behavior but cannot explain why a specific branch received its particular risk score.

Third, the development of a real-time prediction system directly integrated with the Bank BJB LOS would eliminate the current batch-based data export step, enabling more frequent risk assessment and reducing the response latency between credit condition changes and management awareness. This would directly address the temporal resolution limitation of the current batch-based prediction cycle.

References

- [1] X. Zhang and L. Yu, "Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods," *Expert Systems with Applications*, vol. 237, p. 121484, Mar. 2024, doi: 10.1016/j.eswa.2023.121484.
- [2] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, Mar. 2022, doi: 10.1016/j.ejor.2021.06.053.
- [3] S. Hossain *et al.*, "Comparative Analysis of Machine Learning Models for Credit Risk Prediction in Banking Systems," *tajet*, vol. 07, no. 04, pp. 22–33, Apr. 2025, doi: 10.37547/tajet/Volume07Issue04-04.
- [4] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Comput Econ*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.
- [5] R. B. Tareaf, M. AbuJarour, and F. Zinn, "Revolutionizing Credit Risk: A Deep Dive into Gradient-Boosting Techniques in AI-Driven Finance," in *2024 International Conference on Information Networking (ICOIN)*, Ho Chi Minh City, Vietnam: IEEE, Jan. 2024, pp. 322–327. doi: 10.1109/ICOIN59985.2024.10572140.
- [6] M. Abdullah, M. A. F. Chowdhury, A. Uddin, and S. Moudud-Ul-Huq, "Forecasting nonperforming loans using machine learning," *Journal of Forecasting*, vol. 42, no. 7, pp. 1664–1689, Nov. 2023, doi: 10.1002/for.2977.
- [7] J. Wang, "Bridging the Gap between Business Practice and Data Science Approaches," *TMLAI*, vol. 13, no. 02, pp. 01–09, Apr. 2025, doi: 10.14738/tmlai.1302.18289.
- [8] D. Rashad, M. Abdel-Salam, and E. Yehia, "The Prediction of Non-Performing Loans Using Artificial Intelligence - A literature Review," *المجلة العلمية للبحوث والدراسات التجارية*, vol. 39, no. 3, pp. 1467–1497, Sep. 2025, doi: 10.21608/sjrbs.2025.348316.1855.
- [9] M. S. Uddin, G. Chi, M. A. M. Al Janabi, and T. Habib, "Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability," *Int. J Fin Econ*, vol. 27, no. 3, pp. 3713–3729, Jul. 2022, doi: 10.1002/ijfe.2346.
- [10] S. Yan, L. Zhu, B. Wang, L. Wang, S. Liu, and H. Chen, "Research on Machine Learning-based Credit Risk Prediction Models and Algorithms," in *2025 International Conference on Advanced Machine Learning and Data Science (AMLDS)*, Tokyo, Japan: IEEE, Jul. 2025, pp. 570–575. doi: 10.1109/AMLDS63918.2025.11159424.
- [11] M. R. Machado and S. Karray, "Assessing credit risk of commercial customers using hybrid machine learning algorithms," *Expert Systems with Applications*, vol. 200, p. 116889, Aug. 2022, doi: 10.1016/j.eswa.2022.116889.
- [12] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Science and Management*, vol. 6, no. 3, pp. 123–133, Sep. 2023, doi: 10.1016/j.dsm.2023.04.003.
- [13] Mb. Salas, P. Lamothe, E. Delgado, A. L. Fernández-Miguélez, and L. Valcarce, "Determinants of Nonperforming Loans: A Global Data Analysis," *Comput Econ*, vol. 64, no. 5, pp. 2695–2716, Nov. 2024, doi: 10.1007/s10614-023-10543-8.
- [14] "Peraturan BI No. 7/2/PBI/2005," Database Peraturan | JDIH BPK. Accessed: Jun. 03, 2026. [Online]. Available: <http://peraturan.bpk.go.id/Details/137781/peraturan-bi-no-72pbi2005>
- [15] T. Hua, "Machine Learning Approaches to Creditworthiness Classification," in *Proceedings of the 2025 International Conference on Big Data, Artificial Intelligence and Digital Economy*, Kunming China: ACM, Jul. 2025, pp. 127–134. doi: 10.1145/3767052.3767072.

-
- [16] H. Wu, "From Scores to Decisions: Comparing Logistic Regression, Random Forest, and XGBoost for Calibrated, Cost-Sensitive Credit Default Prediction," *AEMPS*, vol. 240, no. 1, pp. 70–79, Nov. 2025, doi: 10.54254/2754-1169/2025.BL29296.
- [17] T. Xu, "Comparative Analysis of Machine Learning Algorithms for Consumer Credit Risk Assessment," *TCSISR*, vol. 4, pp. 60–67, Jun. 2024, doi: 10.62051/r1m3pg16.
- [18] C. Li and J. Zhang, "Research on Credit Risk Prediction Models Based on Machine Learning," in *2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Hangzhou, China: IEEE, Nov. 2024, pp. 1–4. doi: 10.1109/MLBDBI63974.2024.10824021.
- [19] S. Lin, "Research on Credit Card Default Prediction Based on Random Forest and Logistic Regression Models," *AEMPS*, vol. 185, no. 1, pp. 61–69, Jun. 2025, doi: 10.54254/2754-1169/2025.LH23945.

Accepted Manuscript- 59846 JIUP - 10 Agustus 2026