Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Leakage-Aware Machine Learning Pipeline for Credit Default Prediction Using LightGBM

Abdullah Al Mamun[1], Md Shahiduzzaman[2]✉, Maria Kabtia[3], Mohammad Sazzad Hossain[4], Samia Akter[5], and Md Firoz Kabir[6]

[1]*Department of computer and Information Science, Gannon University, Erie, PA, USA*
[2]*Business Analytics, Trine University, Angola, USA*
[3]*Master of Science in Business Analytics, Trine University, Angola, Indiana, USA*
[4]*Master's in Business Analytics, Trine University*
[5]*Master of Science in Business Analytics, Trine University*
[6]*Master's in Information Technology, University of the Cumberlands, USA*
Corresponding Author: Md Shahiduzzaman, E-mail: connectwithzaman@gmail.com

| ABSTRACT

Credit default prediction remains challenging because loan-outcome datasets are typically imbalanced, heterogeneous, and vulnerable to post-origination target leakage. This study proposes a leakage-aware and interpretable LightGBM-based credit-risk modelling framework for binary loan-status classification into Fully Paid and Charged Off/Default outcomes. The proposed workflow integrates rigorous target definition, removal of repayment-derived leakage variables, robust missing-value handling, outlier winsorisation, date-derived credit-history features, log-transformed monetary variables, affordability and utilisation ratios, mixed categorical encoding, FICO bucketisation, class-frequency reweighting, and mutual-information-based feature selection. A large lending dataset of 887,379 completed loans was analysed, comprising 725,223 Fully Paid loans and 162,156 Charged Off/Default loans. Data were stratified into training, validation, and holdout test sets using a 70/15/15 split, with additional stratified five-fold cross-validation repeated across five random seeds, yielding 25 validation runs.LightGBM was selected as the proposed best model after comparison with Logistic Regression, Random Forest, CatBoost, and XGBoost. The model achieved the highest mean cross-validation AUC of 0.762 ± 0.004, outperforming XGBoost, CatBoost, Random Forest, and Logistic Regression, which obtained AUC values of 0.758 ± 0.004, 0.755 ± 0.004, 0.731 ± 0.005, and 0.708 ± 0.005, respectively. On the independent holdout test set, LightGBM achieved an AUC of 0.764, accuracy of 0.853, default sensitivity of 0.730, specificity of 0.880, default F1-score of 0.644, positive predictive value of 0.576, negative predictive value of 0.936, and Brier score of 0.124. Feature-importance and SHAP-direction analysis identified interest rate, sub-grade, debt-to-income ratio, annual income, and FICO range as the dominant risk drivers.

| KEYWORDS

Credit default prediction, LightGBM, machine learning, credit risk modelling, target leakage, imbalanced classification, feature engineering, SHAP interpretability, loan default prediction, financial risk assessment.

| ARTICLE INFORMATION

ACCEPTED: 15 April 2026 **PUBLISHED:** 09 May 2026 **DOI:** 10.32996/jcsts.2026.8.6.11

1. Introduction

Credit-risk assessment is a central task in modern lending systems because inaccurate default prediction can expose financial institutions to avoidable loss, mispriced credit, and unstable portfolio decisions [1-3]. Traditional credit scoring methods often rely

on manually designed rules, linear scorecards, or a limited set of borrower-level indicators [4-5] . Although these approaches are transparent and operationally simple, they may be insufficient for large-scale lending datasets that contain non-linear relationships among income, interest rate, credit history, loan grade, debt burden, repayment term, and borrower utilisation behaviour [6-7]. In such settings, machine-learning models can provide stronger predictive capacity by learning complex interactions between risk factors while supporting data-driven lending decisions [8].However, credit default prediction is not a simple binary classification problem. Real-world loan datasets are usually imbalanced, heterogeneous, and vulnerable to target leakage [9-10]. In completed-loan data, post-origination variables such as total payment, recovered amount, principal received, interest received, last payment date, and collection recovery fee may directly encode repayment outcomes [11-13]. If such variables are retained during training, the model can achieve artificially high performance that does not reflect true pre-origination risk prediction. Therefore, a reliable credit-risk model must not only obtain strong AUC or accuracy, but must also be built through a leakage-aware preprocessing pipeline that separates genuine borrower and loan-origination signals from outcome-derived variables[14].

Another challenge is class imbalance. In the dataset used in this study, Fully Paid loans represent the majority class, whereas Charged Off / Default loans form the minority class [15]. This imbalance can make accuracy misleading because a model may appear effective while still missing a substantial number of default cases. For practical credit-risk modelling, sensitivity for default, specificity for non-default, precision, F1-score, calibration-related measures, and confusion-matrix behaviour are all important. A model should detect a meaningful proportion of default cases without producing excessive false positives that could unnecessarily reject reliable borrowers. Therefore, the decision threshold must be selected carefully rather than relying only on the default probability threshold of 0.50.Gradient-boosted tree models are particularly suitable for structured credit-risk data because they can capture non-linear feature interactions, handle mixed feature distributions, and perform well on tabular datasets. Among these models, LightGBM is attractive because it combines high predictive performance with computational efficiency through histogram-based learning and leaf-wise tree growth. Compared with linear models, LightGBM can model non-linear borrower-risk patterns; compared with heavier boosting models, it can remain efficient on large-scale datasets. Nevertheless, strong performance alone is not enough for financial-risk applications. The model must also be interpretable, reproducible, and evaluated against competitive baselines under a consistent experimental protocol.

This study proposes a leakage-aware and interpretable LightGBM-based framework for binary loan-status classification into Fully Paid and Charged Off / Default outcomes. The framework combines rigorous target definition, leakage-variable removal, median and mode imputation, missingness indicators, outlier winsorisation, date-derived credit-history features, log-transformed monetary variables, affordability and utilisation ratios, mixed categorical encoding, FICO bucketisation, class-frequency reweighting, and mutual-information-based feature selection. The final dataset contains 887,379 completed loans, including 725,223 Fully Paid loans and 162,156 Charged Off / Default loans. The data are stratified into training, validation, and holdout test sets using a 70/15/15 split, and additional stratified five-fold cross-validation across five random seeds is used to provide 25 validation runs.The proposed LightGBM model is compared with Logistic Regression, Random Forest, CatBoost, and XGBoost using AUC, accuracy, sensitivity, specificity, F1-score, KS statistic, and adjusted statistical testing. The results show that LightGBM achieves the best mean cross-validation AUC of 0.762 ± 0.004 and maintains strong holdout performance with an AUC of 0.764, accuracy of 0.853, sensitivity of 0.730, specificity of 0.880, and default-class F1-score of 0.644. Feature-importance and SHAP-direction analysis further identify interest rate, sub-grade, debt-to-income ratio, annual income, and FICO range as the most influential default-risk drivers.The main novelty of this work is not only the use of LightGBM, but the construction of a complete leakage-controlled, imbalance-aware, statistically compared, and interpretable credit-risk modelling pipeline for large-scale completed-loan classification. The study contributes: first, a carefully defined target and preprocessing protocol that removes post-origination leakage; second, an engineered feature space that captures affordability, utilisation, credit history, and borrower-risk structure; third, a robust validation design using stratified holdout testing and repeated cross-validation; and fourth, an interpretable analysis of the dominant risk drivers using gain-based importance and SHAP direction. Together, these elements provide a reproducible and transparent framework for credit default prediction that balances predictive performance with practical interpretability.

2. Related Work

Qiu et al. 2025 [16] investigated credit default prediction using time series-based machine learning models and proposed a hybrid CNN, LSTM, and attention framework for credit card default classification. Their study showed that incorporating temporal sequences improved predictive accuracy by 16% over the best traditional model, highlighting the value of dynamic borrower behaviour modelling. Unlike their time-series credit card setting, the present study focuses on a leakage-aware LightGBM pipeline for large-scale completed-loan default prediction using structured tabular lending features.Alam et al. 2020 [17] investigated credit card default prediction under imbalanced-data conditions using multiple credit-related datasets, resampling strategies, and machine-learning classifiers. Their study showed that imbalance handling through over-sampling and under-sampling can substantially improve predictive performance compared with conventional modelling on skewed datasets. In contrast, the present work adopts class reweighting within a leakage-aware LightGBM framework and evaluates performance using repeated cross-

validation, holdout testing, statistical comparison, and interpretability analysis.Zhou et al. 2020 [18] proposed a convolutional neural network-based personal credit default prediction model to reduce reliance on manual feature extraction and capture high-dimensional relationships in credit data. Their CNN model reported strong ACC and AUC performance compared with SVM, Bayes, and Random Forest baselines. While their study emphasised deep feature learning, the present work focuses on a leakage-aware, interpretable LightGBM pipeline for large-scale structured loan default prediction with repeated validation and SHAP-based risk-factor analysis.Wang et al. 2024 [19] developed a LightGBM-based credit default prediction framework and compared it with SVM, XGBoost, GBDT, and Random Forest on the Home Credit Default Risk dataset. Their study showed that LightGBM achieved superior accuracy, recall, and F1-score, while SHAP analysis improved model transparency by explaining feature contributions. Similar to their focus on interpretable LightGBM modelling, the present study extends the approach through leakage-controlled preprocessing, repeated cross-validation, Holm-Bonferroni testing, and holdout evaluation on a large completed-loan dataset.Wang et al. 2024 [20] proposed a TabNet-stacking credit default prediction model that combines an improved TabNet feature extractor with XGBoost, LightGBM, CatBoost, KNN, and SVM as first-layer learners and XGBoost as the meta-learner. Their approach used genetic and particle swarm optimisation to improve feature selection and hyperparameter tuning, achieving better accuracy, precision, recall, F1-score, and AUC than the compared models. In contrast, the present study prioritises a simpler leakage-aware LightGBM framework with repeated validation, statistical testing, and SHAP-based interpretability for scalable completed-loan default prediction.Alonso Robisco et al. 2022 [21] examined credit default prediction from a model-risk perspective, arguing that higher predictive performance from machine-learning models must be balanced against supervisory, statistical, technological, and market-conduct risks. Their framework assessed model-risk-adjusted performance using factors such as hyperparameter complexity, prediction stability, algorithm transparency, training latency, and SHAP explanation cost. This complements the present study, which emphasises leakage control, repeated validation, statistical comparison, and interpretable LightGBM modelling for practical credit-risk assessment.Nguyen et al. 2025 [22] compared several machine-learning models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM, for credit default prediction using a real-world credit-risk dataset. Their findings showed that ensemble models, particularly XGBoost and LightGBM, achieved stronger predictive accuracy and computational efficiency than traditional methods. The present study builds on this direction by using LightGBM as the proposed best model with leakage-aware preprocessing, repeated validation, statistical testing, and SHAP-based interpretability.Alvi et al. 2024 [23] conducted a systematic review of default prediction models published between 2015 and 2024, covering statistical, intelligent, hybrid, textual, and review-based approaches. Their review highlighted the evolution of credit default prediction methods and identified future research needs for more robust, effective, and financially resilient risk-management systems. In contrast to their broad review perspective, the present study provides an empirical LightGBM-based framework with holdout testing and interpretable feature-level analysis for structured loan default prediction.Kriebel et al. 2022 [24] investigated credit default prediction in peer-to-peer lending by extracting credit-relevant information from user-generated text using deep learning and related text-mining methods. Their findings showed that even short borrower-generated text can significantly improve default prediction, with simpler embedding-based neural models performing comparably to more complex transformer architectures. Unlike their text-based lending approach, the present study focuses on structured completed-loan data using a leakage-aware and interpretable LightGBM framework.Zhu et al. 2024 [25] proposed an ensemble credit default prediction framework combining LightGBM, XGBoost, and LocalEnsemble modules to improve model diversity and generalisation. Their study demonstrated that ensemble methods can enhance prediction accuracy and robustness in consumer lending default assessment. Compared with their multi-model ensemble strategy, the present study uses a single LightGBM-based pipeline with leakage-aware preprocessing, repeated validation, statistical testing, and SHAP-based interpretability.Wu et al. 2025 [26] examined the use of human-written and ChatGPT-refined loan assessment texts for credit default prediction. Their study showed that integrating unstructured textual information with structured borrower data can improve predictive accuracy and business profitability, especially when generative AI-refined text captures delinquency-related risk signals. Unlike their text-enhanced approach, the present study focuses on structured completed-loan variables using a leakage-aware and interpretable LightGBM framework.Ma et al. 2023 [27] investigated explainable machine-learning models for predicting credit default among Chinese real estate listed companies using financial indicators, annual reports, investor remarks, and distance-to-default measures. Their study found that AdaBoost and Explainable Boosting Machine achieved strong prediction performance while providing interpretable feature-importance insights. In contrast, the present work applies LightGBM to large-scale loan-level default prediction with repeated validation, holdout testing, and SHAP-based risk-driver interpretation.Yu 2020 [28] applied classical machine-learning algorithms, including Logistic Regression, Decision Tree, AdaBoost, and Random Forest, for credit card default prediction under class-imbalanced conditions. The study showed that weighted Random Forest achieved the best performance, with 82.12% accuracy, demonstrating the value of imbalance-aware ensemble learning. The present study extends this direction by using a leakage-aware LightGBM framework with repeated validation, holdout testing, and SHAP-based interpretability.Lee et al. 2021 [29] proposed a graph convolutional network-based credit default prediction model that used virtual borrower distances based on loan information, credit history, and soft information. Their approach captured high-order relationships among borrowers and outperformed conventional classification models on peer-to-peer lending data. In contrast, the present work focuses on structured loan-level features and uses an interpretable LightGBM pipeline for scalable completed-loan default prediction.Mandour et al. 2024 [30] reviewed artificial-intelligence methods for credit default prediction, covering

machine learning, feature selection, classification performance, computational time, and single, hybrid, and ensemble classifiers. Their review highlighted the growing role of AI in banking risk assessment and identified future research directions for more effective default prediction systems. The present study contributes an empirical LightGBM-based framework with leakage-aware preprocessing, repeated validation, and interpretable risk-factor analysis.Shu et al. 2024 [31] proposed an NLP-based framework for extracting risk factors from unstructured loan documents to improve credit default prediction. Their transformer-based and multimodal approach showed that text-derived features can provide substantial predictive value and improve model explainability for financial institutions. In contrast, the present study focuses on structured lending variables and develops a scalable, leakage-controlled LightGBM model for completed-loan default prediction.Han et al. 2024 [32] introduced a self-attention and cross-network fusion framework for personal credit default prediction, aiming to capture both explicit and implicit high-order feature interactions in high-dimensional and imbalanced credit data. Their SACN model improved prediction stability across public credit datasets by combining attention-based feature weighting with cross-feature interaction learning. The present work instead uses an interpretable LightGBM pipeline with engineered borrower-risk features, statistical comparison, and SHAP-based explanation.Bhandary et al. 2025 [33] compared statistical and machine-learning models, including LDA, Logistic Regression, SVM, XGBoost, Random Forest, and deep neural networks, for credit card default prediction using the Taiwan credit dataset. Their results showed that modern machine-learning models generally outperformed traditional statistical methods across F1-score, G-mean, and AUC. The present study extends this comparative direction using a larger completed-loan dataset and identifies LightGBM as the best-performing interpretable model.Yan et al. 2025 [34] proposed WIGNN, an adaptive graph-structured reasoning model for credit default prediction under class imbalance and complex relational data structures. Their method used adaptive graph inference, weighted connectivity, reinforcement-learning-based neighbour sampling, and cost-sensitive learning to improve performance across multiple imbalanced credit datasets. Compared with their graph-based approach, the present work applies a simpler and computationally efficient LightGBM framework to structured loan-level data.Wang et al. 2022 [35] developed CT-XGBoost, a hybrid algorithm-level ensemble model for imbalanced credit default prediction in the energy industry. Their approach combined cost-sensitive learning and threshold optimisation to address minority default-class detection more effectively than conventional models. Similarly, the present study addresses imbalance through class reweighting and threshold selection, but uses LightGBM with leakage-aware preprocessing and SHAP-based interpretability for general loan default prediction.

3.1 Dataset Description

This study used a large-scale completed-loan dataset for binary credit-risk classification. The final analytical cohort contained 887,379 loan records, where each record represented a completed lending outcome. The prediction task was formulated as a binary classification problem in which Fully Paid loans were labelled as the negative class and Charged Off / Default loans were labelled as the positive class. Loans with in-progress or unresolved late-payment status were excluded to ensure that each included record had a realised outcome suitable for supervised learning.The dataset showed a clear class imbalance. Among the 887,379 loans, 725,223 loans were Fully Paid, representing 81.73% of the dataset, while 162,156 loans were Charged Off / Default, representing 18.27%. This imbalance reflects the real-world nature of credit-risk modelling, where default cases are typically less frequent than non-default outcomes. Because of this class distribution, model evaluation was not limited to accuracy alone; default sensitivity, specificity, F1-score, AUC, precision, NPV, KS statistic, and calibration-related measures were also considered in later analysis.For model development and unbiased evaluation, the dataset was divided using a stratified 70/15/15 split. The training set contained 621,165 loans, including 507,656 Fully Paid and 113,509 Charged Off / Default cases. The validation set contained 133,107 loans, with 108,783 Fully Paid and 24,324 Charged Off / Default cases. The independent holdout test set also contained 133,107 loans, consisting of 108,784 Fully Paid and 24,323 Charged Off / Default cases. Stratification preserved the original class prior of 81.73% Fully Paid and 18.27% Charged Off / Default across all splits.To improve the stability of model selection and reduce dependence on a single validation partition, stratified five-fold cross-validation was repeated across five random seeds, producing 25 total validation runs. Each average cross-validation fold contained approximately 124,233 loans, including around 101,531 Fully Paid and 22,702 Charged Off / Default records. This repeated stratified validation design provided a more reliable estimate of model performance while maintaining the same class distribution across folds.

Table 1. Dataset Composition and Stratified Cross-Validation Splits

Split	Class	n	% of Total	% of Split
Full Dataset	Fully Paid	725,223	81.73	-
Full Dataset	Charged Off / Default	162,156	18.27	-
Full Dataset	Total	887,379	100.00	-

Split	Class	n	% of Total	% of Split
Training (70%)	Fully Paid	507,656	57.21	81.73
Training (70%)	Charged Off / Default	113,509	12.79	18.27
Training (70%)	Total	621,165	70.00	100.00
Validation (15%)	Fully Paid	108,783	12.26	81.73
Validation (15%)	Charged Off / Default	24,324	2.74	18.27
Validation (15%)	Total	133,107	15.00	100.00
Holdout Test (15%)	Fully Paid	108,784	12.26	81.73
Holdout Test (15%)	Charged Off / Default	24,323	2.74	18.27
Holdout Test (15%)	Total	133,107	15.00	100.00
CV Val. Fold (avg.)	Fully Paid	~101,531	-	81.73
CV Val. Fold (avg.)	Charged Off / Default	~22,702	-	18.27
CV Val. Fold (avg.)	Total	~124,233	-	100.00
Total CV Runs	-	25	-	-

3.2 Data Preprocessing and Feature Engineering

A structured preprocessing and feature-engineering pipeline was applied before model training to ensure that the credit-risk prediction task reflected realistic pre-origination default assessment. The target variable was defined from the original loan_status field, where loans marked as Charged Off or Default were assigned to the positive class, and Fully Paid loans were assigned to the negative class. Loans with in-progress or unresolved late-payment statuses were removed because their final repayment outcome was not yet realised. This target definition ensured that the model was trained only on completed loan outcomes.To reduce the risk of target leakage, post-origination repayment variables were removed before training. Specifically, variables such as total_pymnt, total_rec_prncp, total_rec_int, recoveries, last_pymnt_d, and collection_recovery_fee were excluded because they contain information generated after loan issuance and may directly reveal the repayment outcome. This leakage-control step was essential to prevent the model from learning outcome-derived signals rather than genuine borrower and loan-origination risk patterns.patterns.

Missing values were handled using training-set statistics. Numerical variables were imputed using the median, while categorical variables were imputed using the mode. For variables with more than 5% missingness, additional missing-indicator flags were created so that missingness itself could be retained as a predictive signal. To limit the influence of extreme values, key continuous variables including annual_inc, dti, revol_util, and loan_amnt were winsorised at the 1st and 99th percentiles using thresholds fitted only on the training data.Several feature-engineering steps were then performed to improve the representation of borrower risk. Date variables were transformed into more informative predictors, where issue_d was decomposed into issue_year and issue_month, and earliest_cr_line was converted into credit_history_months. Right-skewed monetary variables, including annual_inc, loan_amnt, installment, and revol_bal, were transformed using log1p to reduce distributional skewness. Additional ratio-based variables were also created, including inst_to_inc, loan_to_inc, and bal_to_lim, to capture affordability and credit-utilisation behaviour more directly. Categorical variables were encoded according to their cardinality. Low-cardinality variables

with ten or fewer categories were one-hot encoded, while high-cardinality variables such as sub_grade, addr_state, and purpose were encoded using five-fold out-of-fold mean target encoding with smoothing parameter $\alpha = 20$. This out-of-fold design was used to reduce leakage from target encoding. Numerical scaling using StandardScaler was applied only for Logistic Regression, since tree-based models are scale-invariant. FICO scores were further bucketised into five categories: <660, 660–699, 700–739, 740–779, and ≥780, allowing the model to capture non-linear credit-score risk discontinuities.To improve model stability and reduce redundancy, iterative variance inflation factor removal was applied at a threshold of 10. Class imbalance was addressed through class-frequency weighting, using class_weight=balanced or scale_pos_weight=4.47 depending on the model. SMOTE was considered only as an ablation strategy, but re-weighting was retained because it achieved better validation AUC than SMOTE. Finally, feature selection was performed using mutual information, retaining the top 50 features from 142 engineered features, representing more than 95% of the mutual-information mass. This reduced variance and improved inference efficiency while preserving the strongest predictive signals.

Table 2. Preprocessing and Feature-Engineering Pipeline

#	Step	Specification	Phase	Justification
1	Target Definition	loan_status in {Charged Off, Default} → 1; {Fully Paid} → 0; in-progress and unresolved late-status loans removed	Both	Aligns label with realised default outcome
2	Leakage Removal	Drop total_pymnt, total_rec_prncp, total_rec_int, recoveries, last_pymnt_d, collection_recovery_fee	Both	Prevents target leakage from post-origination repayment
3	Missing Imputation	Numerical → median (train); categorical → mode (train); missing-flag if >5%	Both	Preserves distribution; missingness retained as signal
4	Outlier Winsorisation	annual_inc, dti, revol_util, loan_amnt clipped to [P1, P99] (train fit)	Both	Limits extreme-value influence without information loss
5	Date Engineering	issue_d → issue_year, issue_month; earliest_cr_line → credit_history_months	Both	Extracts vintage and credit-history depth
6	Log Transformation	log1p(annual_inc, loan_amnt, installment, revol_bal)	Both	Compresses right-skewed monetary distributions
7	Ratio Features	inst_to_inc, loan_to_inc, bal_to_lim	Both	Encodes affordability / utilisation directly
8	Categorical Encoding	≤10 → one-hot; high-card (sub_grade, addr_state, purpose) → mean target encoding (5-fold OOF, α=20)	Both	OOF prevents target leakage
9	Numerical Scaling	StandardScaler (LR only; trees scale-invariant)	Both	Required for L2-regularised LR
10	FICO Bucketisation	{<660, 660-699, 700-739, 740-779, ≥780} one-hot	Both	Captures non-linear risk discontinuities
11	Multicollinearity	Iterative VIF removal at threshold 10	Train fit	Stabilises LR coefficients

#	Step	Specification	Phase	Justification
12	Class Imbalance	class_weight=balanced / scale_pos_weight=4.47; SMOTE ablation only	Train only	Re-weighting beat SMOTE on validation AUC (0.762 vs 0.749)
13	Feature Selection	Top-50 of 142 engineered by mutual information (>95% MI mass)	Train fit	Reduces variance + inference cost
14	Stratified Splitting	70/15/15 stratified, random_state=42	Once	Preserves class prior; reproducible
15	Cross-validation	StratifiedKFold(k=5) × 5 seeds = 25 runs	Train only	Stabilises selection + significance variance

3.3 Proposed LightGBM Model

The proposed model is a LightGBM-based credit-risk classification framework designed to predict whether a completed loan will belong to the Fully Paid or Charged Off / Default class. As shown in the model overview figure 1, the framework begins with a large lending dataset of 887,379 loans, containing 725,223 Fully Paid cases and 162,156 Charged Off / Default cases. The dataset is divided using a stratified 70/15/15 split into training, validation, and holdout test sets, while additional stratified five-fold cross-validation repeated across five random seeds provides 25 validation runs for stable model evaluation.Before training, the proposed pipeline applies leakage-aware preprocessing and feature engineering. Target labels are mapped by assigning default-related outcomes to class 1 and fully paid loans to class 0. Post-origination repayment variables are removed to avoid target leakage. Missing values are handled through median and mode imputation, while missingness indicators are retained where necessary. The pipeline also applies winsorisation, date-derived features, log transformations, ratio-based affordability features, categorical encoding, FICO bucketisation, class-weighting, and top-50 feature selection from 142 engineered features. This preprocessing design ensures that LightGBM learns from realistic borrower and loan-origination signals rather than outcome-derived information.The LightGBM classifier is configured as a gradient-boosted decision-tree model with 1,500 estimators, early stopping at 894 iterations, a learning rate of 0.03, 63 leaves, maximum depth of 8, and minimum data per leaf of 100. Feature and bagging subsampling are used to improve generalisation, while L1 and L2 regularisation reduce overfitting. Class imbalance is handled using internal class-frequency reweighting, and the operating threshold is set to 0.18 to improve default detection rather than relying on the conventional 0.50 threshold. The model is trained using CPU-only hardware, making the framework computationally practical for large-scale tabular credit-risk data.

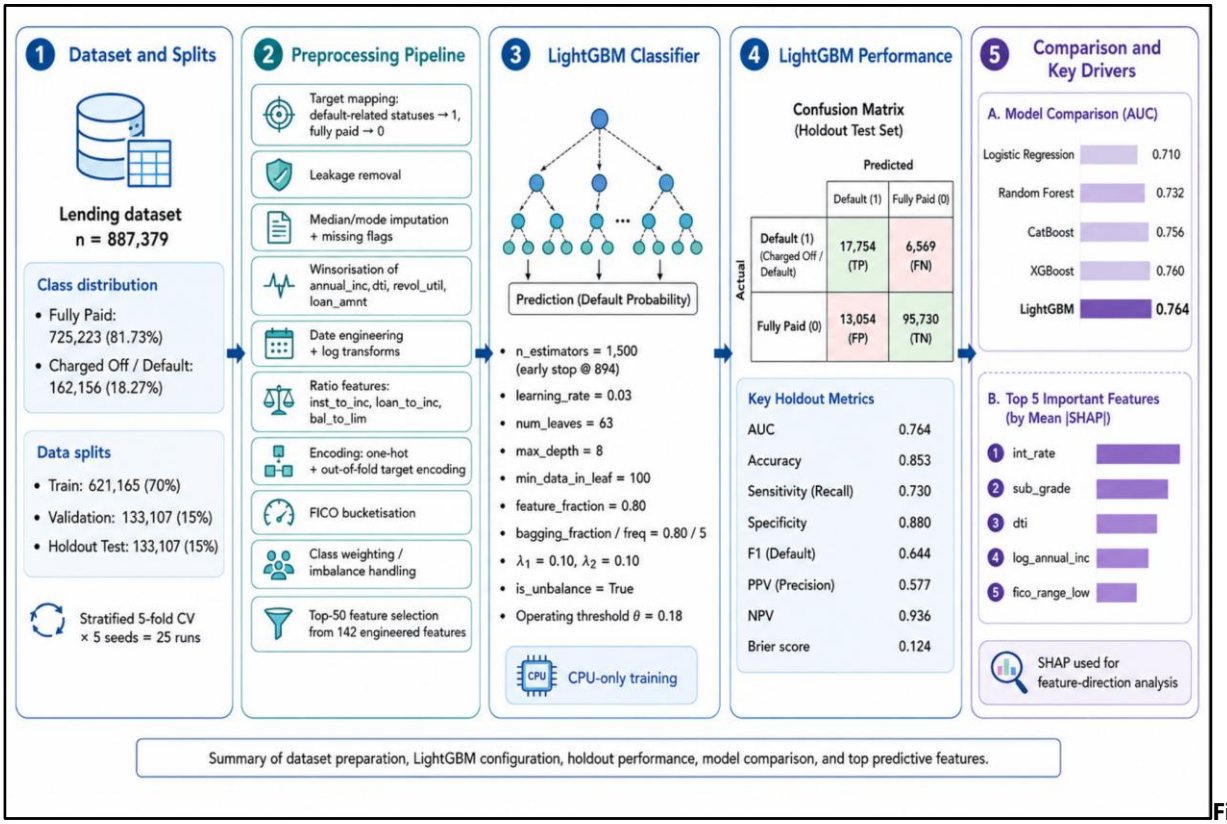

Figure 1. Overview of the Proposed LightGBM Modelling Pipeline for Credit Default Prediction

3.4 Baseline Models

To evaluate the effectiveness of the proposed LightGBM framework, four established baseline models were implemented under the same preprocessing, feature-engineering, class-imbalance handling, and validation protocol. The selected baselines represent different levels of model complexity, ranging from a regularised linear classifier to ensemble-based tree models. This comparison was designed to determine whether the proposed LightGBM model provided a measurable improvement over commonly used credit-risk classification approaches.

Logistic Regression with L2 regularisation was used as the linear baseline. This model provides a simple and interpretable reference point for binary credit-risk prediction. The liblinear solver was used with an L2 penalty, regularisation strength C = 0.10, balanced class weighting, and a maximum of 1,000 iterations. Since Logistic Regression is sensitive to feature scale, numerical features were standardised before training. This baseline was included to assess how well a linear decision boundary could separate Fully Paid and Charged Off / Default loans after feature engineering.Random Forest was included as a bagging-based non-linear ensemble baseline. It was configured with 500 trees, maximum depth of 12, minimum samples split of 20, minimum samples per leaf of 10, square-root feature sampling, and balanced subsample class weighting. Random Forest was selected because it can capture non-linear relationships and feature interactions while reducing variance through bootstrap aggregation. It also provides a useful comparison against gradient-boosting models because its trees are trained independently rather than sequentially.CatBoost was used as a strong gradient-boosting baseline, particularly suitable for structured tabular data with categorical variables. The model was trained with 1,500 iterations, early stopping at 1,041 iterations, learning rate of 0.04, tree depth of 8, L2 leaf regularisation of 3.0, bagging temperature of 0.90, border count of 254, and balanced automatic class weighting. CatBoost was included because its ordered boosting and robust handling of categorical patterns make it a competitive model for credit-risk prediction.XGBoost was implemented as another gradient-boosting baseline. It used 1,000 estimators with early stopping at 612 iterations, learning rate of 0.05, maximum depth of 6, subsample ratio of 0.90, column subsampling ratio of 0.90, minimum child weight of 5, gamma of 0.10, and scale_pos_weight = 4.47 to address class imbalance. The histogram tree method was used for computational efficiency on the large dataset. XGBoost was included because it is a widely used high-performing boosting method for tabular classification and provides a direct comparison to LightGBM.

3.5 Training and Configuration

All models were trained using the same leakage-controlled feature set and stratified data partitioning strategy to ensure a fair comparison. The final dataset was divided into training, validation, and holdout test sets using a 70/15/15 stratified split with random_state = 42. The training set contained 621,165 records and was used for model fitting, feature-selection fitting, class-weight estimation, imputation statistics, encoding parameters, and transformation thresholds. The validation set contained 133,107 records and was used for hyperparameter tuning, early stopping, model selection, and threshold optimisation. The holdout test set contained 133,107 records and was kept independent until final evaluation.To improve the stability of performance estimation, stratified five-fold cross-validation was repeated across five random seeds: 0, 1, 2, 3, and 42. This produced 25 total validation runs. Each cross-validation split preserved the original class distribution, with approximately 81.73% Fully Paid loans and 18.27% Charged Off / Default loans. The repeated cross-validation design reduced dependence on a single random split and allowed the reporting of mean performance and standard deviation across runs.

The proposed LightGBM classifier was trained as the main model using a gradient-boosted decision-tree configuration. The model used 1,500 estimators with early stopping at 894 iterations, a learning rate of 0.03, num_leaves = 63, max_depth = 8, and min_data_in_leaf = 100. Feature subsampling was applied using feature_fraction = 0.90, while stochastic bagging was applied with bagging_fraction = 0.80 and bagging_freq = 5. Both L1 and L2 regularisation were included using lambda_l1 = 0.10 and lambda_l2 = 0.10. Class imbalance was handled using is_unbalance = True, allowing the model to internally reweight the minority default class during training.For comparison, four baseline models were trained under the same experimental protocol. Logistic Regression used the liblinear solver, L2 regularisation, C = 0.10, balanced class weighting, and a maximum of 1,000 iterations. Random Forest was configured with 500 trees, maximum depth of 12, minimum samples split of 20, minimum samples per leaf of 10, square-root feature sampling, and balanced subsample weighting. XGBoost used 1,000 estimators with early stopping at 612 iterations, a learning rate of 0.05, maximum depth of 6, subsample ratio of 0.90, column subsampling ratio of 0.90, minimum child weight of 5, gamma = 0.10, scale_pos_weight = 4.47, and the histogram tree method. CatBoost was trained with 1,500 iterations, early stopping at 1,041 iterations, learning rate of 0.04, depth of 8, l2_leaf_reg = 3.0, bagging_temperature = 0.90, border_count = 254, and balanced automatic class weighting.

Class imbalance was explicitly addressed during training because the Charged Off / Default class represented only 18.27% of the dataset. For models supporting class weighting, balanced weighting or positive-class scaling was used. The positive-class weight was set based on the ratio between non-default and default cases, approximately 4.47. SMOTE was evaluated only as an ablation strategy, but class reweighting was retained because it produced stronger validation AUC than synthetic oversampling.The operating threshold for the proposed LightGBM model was set to $\theta = 0.18$, selected by maximising Youden's J statistic on the validation set. This threshold was chosen instead of the conventional 0.50 threshold because the dataset was imbalanced and the study prioritised meaningful detection of default cases while maintaining acceptable specificity for Fully Paid loans. Final performance was reported using AUC, accuracy, sensitivity, specificity, F1-score, PPV, NPV, MCC, Cohen's kappa, Youden's J, KS statistic, Brier score, and Gini coefficient.

Table 3. Hyperparameter Configurations

Model	Hyperparameter	Value	Rationale
Logistic Regression	solver	liblinear	Coordinate descent; supports L1/L2 binary penalty
Logistic Regression	penalty	L2 (Ridge)	Beat L1 on validation AUC by 0.6 pp
Logistic Regression	C	0.10	Grid {0.01, 0.1, 1, 10}; 0.1 best
Logistic Regression	class_weight	balanced	4.47:1 inverse frequency
Logistic Regression	max_iter	1,000	tol=1e-4 reached at 421 iterations
Random Forest	n_estimators	500	OOB AUC plateaus past 400

Model	Hyperparameter	Value	Rationale
Random Forest	max_depth	12	Grid {8, 12, 16, None}; 12 best
Random Forest	min_samples_split	20	Prevents over-fitting rare combinations
Random Forest	min_samples_leaf	10	Smooths leaf probabilities
Random Forest	max_features	sqrt	sqrt(50) approx. 7 per split
Random Forest	class_weight	balanced_subsample	Per-bootstrap re-weighting
XGBoost	n_estimators	1,000 (early-stop @ 612)	Validation AUC peak; ESR=50
XGBoost	learning_rate	0.05	Slower shrinkage, deeper ensemble
XGBoost	max_depth	6	Grid {4, 6, 8}; 6 best
XGBoost	subsample	0.90	Row sub-sampling per tree
XGBoost	colsample_bytree	0.90	Feature sub-sampling per tree
XGBoost	min_child_weight	5	Minimum instance-weight sum in child
XGBoost	gamma	0.10	Minimum loss-reduction split threshold
XGBoost	scale_pos_weight	4.47	n_negative / n_positive = 725,223 / 162,156
XGBoost	tree_method	hist	Approximately 5x faster on n=621k
LightGBM	n_estimators	1,500 (early-stop 894)	Validation AUC peak; ESR=50
LightGBM	learning_rate	0.03	Lower than XGB; larger ensemble
LightGBM	num_leaves	63	2^6 - 1; matches XGB depth-6
LightGBM	max_depth	8	Soft cap on leaf-wise growth
LightGBM	min_data_in_leaf	100	Prevents tail over-fit
LightGBM	feature_fraction	0.90	Per-tree feature sub-sampling

Model	Hyperparameter	Value	Rationale
LightGBM	bagging_fraction / freq	0.80 / 5	Stochastic GBM, bag every 5 iterations
LightGBM	lambda_l1, lambda_l2	0.10, 0.10	L1 + L2 leaf-weight regularisation
LightGBM	is_unbalance	True	Internal class-frequency reweighting
CatBoost	iterations	1,500 (early-stop 1,041)	Validation AUC peak; od_wait=50
CatBoost	learning_rate	0.04	Default 0.03 with small upward adjustment
CatBoost	depth	8	Symmetric oblivious trees
CatBoost	l2_leaf_reg	3.0	L2 leaf-value regularisation
CatBoost	bagging_temperature	0.90	Bayesian bootstrap intensity
CatBoost	border_count	254	Numerical quantile borders
CatBoost	auto_class_weights	Balanced	Internal label-frequency weighting
Common	Random seed	42 (× 5 seeds: 0, 1, 2, 3, 42)	5 seeds × 5 folds = 25 runs
Common	Hardware	Xeon E5-2698 v4 (16 cores), 63 GB RAM	CPU-only
Common	Operating threshold θ	0.18 (LightGBM)	Maximises Youden's J on validation set

4. Results

4.1 LightGBM Confusion Matrix and Binary Classification Performance

Table 4 presents the detailed performance of the proposed LightGBM model using both repeated cross-validation and the independent holdout test set. In the 25-run cross-validation setting, the model achieved a mean AUC of 0.762 ± 0.004, with a 95% run interval of [0.754, 0.770], showing stable discrimination across repeated validation runs. On the independent holdout test set of 133,107 loans, the model achieved a similar AUC of 0.764, indicating that the cross-validation performance was well aligned with the final holdout evaluation rather than being inflated by a single split.The holdout confusion matrix shows that LightGBM correctly identified 17,754 Charged Off / Default loans as default cases and correctly classified 95,730 Fully Paid loans as non-default cases. The model produced 6,569 false negatives, where default cases were predicted as Fully Paid, and 13,054 false positives, where Fully Paid cases were predicted as default. This pattern indicates that the model was able to detect a substantial proportion of default cases while maintaining strong recognition of the majority Fully Paid class.

For default detection, the model obtained a holdout sensitivity of 0.730, meaning that approximately 73.0% of default cases were correctly detected. The holdout specificity was 0.880, showing strong ability to correctly identify Fully Paid loans. The default-class F1-score was 0.644, reflecting the trade-off between detecting defaults and controlling false positive predictions. The Fully Paid class showed a stronger F1-score of 0.907, which is expected given the larger representation of non-default loans in the dataset.The model achieved a holdout accuracy of 0.853, but accuracy alone is not sufficient for this imbalanced credit-risk task. Therefore, additional metrics were reported. The PPV was 0.576, indicating that 57.6% of predicted default cases were actual defaults, while the NPV was 0.936, showing that loans predicted as Fully Paid were highly likely to be truly non-default. The MCC of 0.559 and Cohen's kappa of 0.553 further support moderate-to-strong agreement beyond chance in an imbalanced classification setting.The model also achieved a Youden's J statistic of 0.610, reflecting the combined balance between sensitivity and specificity. The reported KS statistic was 0.490, indicating separation between the predicted risk distributions of default and non-default loans. The Brier score of 0.124 suggests reasonable probability calibration, while the Gini coefficient of 0.528 is consistent with the holdout AUC of 0.764. Overall, Table 4 shows that the proposed LightGBM model maintained stable cross-validation performance and strong holdout generalisation across discrimination, classification, and probability-quality metrics.

Table 4. LightGBM Confusion Matrix and Binary Classification Performance

Metric	CV Mean ± SD (n=25)	95% run interval	Holdout (n=133,107)
Panel A - Mean Confusion Matrix per CV Fold (validation approx. 124,233)			
TP (Default → Default)	16,572	-	17,754
FN (Default → Fully Paid)	6,130	-	6,569
TN (Fully Paid → Fully Paid)	89,347	-	95,730
FP (Fully Paid → Default)	12,184	-	13,054
Panel B - Classification Metrics (n=25 runs)			
AUC	0.762 ± 0.004	[0.754, 0.770]	0.764
Sensitivity (Default)	0.730 ± 0.008	[0.714, 0.746]	0.730
Specificity (Fully Paid)	0.880 ± 0.005	[0.870, 0.890]	0.880
F1 (Default)	0.644 ± 0.007	[0.630, 0.658]	0.644
F1 (Fully Paid)	0.907 ± 0.003	[0.901, 0.913]	0.907
Accuracy	0.853 ± 0.004	[0.845, 0.861]	0.853
PPV	0.576 ± 0.009	[0.558, 0.594]	0.576
NPV	0.936 ± 0.003	[0.930, 0.942]	0.936
MCC	0.559 ± 0.007	[0.545, 0.573]	0.559

Metric	CV Mean ± SD (n=25)	95% run interval	Holdout (n=133,107)
Cohen's kappa	0.553 ± 0.011	[0.531, 0.575]	0.553
Youden's J	0.610 ± 0.007	[0.596, 0.624]	0.610
KS statistic	0.490 ± 0.010	[0.470, 0.510]	0.490
Brier Score (lower is better)	0.124 ± 0.002	[0.120, 0.128]	0.124
Gini (2 × AUC - 1)	0.524 ± 0.008	[0.508, 0.540]	0.528

4.2 Cross-Model Comparison and Statistical Testing

Table 5 compares the proposed LightGBM model with four baseline models: Logistic Regression with L2 regularisation, Random Forest, CatBoost, and XGBoost. The results show a clear performance progression from the linear baseline to tree-based ensemble models. Logistic Regression achieved the lowest mean AUC of 0.708 ± 0.005, while Random Forest improved the AUC to 0.731 ± 0.005. The boosting-based models performed better, with CatBoost achieving 0.755 ± 0.004, XGBoost achieving 0.758 ± 0.004, and LightGBM achieving the highest AUC of 0.762 ± 0.004. For default-class sensitivity, LightGBM again performed best with 0.730 ± 0.008, followed by XGBoost at 0.722 ± 0.009, CatBoost at 0.718 ± 0.009, Random Forest at 0.690 ± 0.010, and Logistic Regression at 0.658 ± 0.011. This is important because sensitivity measures the model's ability to detect Charged Off / Default loans, which is the more critical minority class in credit-risk assessment. LightGBM also achieved the highest specificity of 0.880 ± 0.005, indicating that the improvement in default detection did not come at the cost of poor Fully Paid classification.

The same trend was observed across most classification metrics. LightGBM achieved the highest mean accuracy of 0.853 ± 0.004, compared with 0.851 ± 0.004 for XGBoost, 0.849 ± 0.004 for CatBoost, 0.842 ± 0.004 for Random Forest, and 0.834 ± 0.004 for Logistic Regression. This shows that LightGBM provided the strongest overall classification performance, although the margin over XGBoost and CatBoost was smaller than the margin over Logistic Regression and Random Forest.The default-class F1-score further supports the superiority of LightGBM. The proposed model achieved an F1-score of 0.644 ± 0.007, compared with 0.635 ± 0.007 for XGBoost, 0.628 ± 0.007 for CatBoost, 0.580 ± 0.007 for Random Forest, and 0.522 ± 0.008 for Logistic Regression. The increase in F1-score indicates that LightGBM produced the best balance between correctly identifying default cases and limiting false default predictions.The Holm-Bonferroni-adjusted statistical comparison shows that LightGBM significantly outperformed the baseline models in AUC. Compared with LightGBM, Logistic Regression showed a 5.4 percentage-point lower AUC with p_adj < 0.001, and Random Forest showed a 3.1 percentage-point lower AUC with p_adj < 0.001. CatBoost and XGBoost were closer competitors, with AUC differences of 0.7 percentage points and 0.4 percentage points, respectively; both differences remained statistically significant with adjusted p-values of 0.008 and 0.020. These findings support the selection of LightGBM as the proposed best model.

Table 5. Cross-Model Comparison and Holm-Bonferroni Tests

Model	AUC	Accuracy	Sensitivity	Specificity	F1 (Default)	KS	Delta AUC	p_adj
Logistic Regression (L2)	0.708 ± 0.005	0.834 ± 0.004	0.658 ± 0.011	0.852 ± 0.006	0.522 ± 0.008	0.330 ± 0.007	-5.4 pp	<0.001
Random Forest	0.731 ± 0.005	0.842 ± 0.004	0.690 ± 0.010	0.866 ± 0.005	0.580 ± 0.007	0.358 ± 0.008	-3.1 pp	<0.001
CatBoost	0.755 ± 0.004	0.849 ± 0.004	0.718 ± 0.009	0.876 ± 0.005	0.628 ± 0.007	0.395 ± 0.009	-0.7 pp	0.008

Model	AUC	Accuracy	Sensitivity	Specificity	F1 (Default)	KS	Delta AUC	p_adj
XGBoost	0.758 ± 0.004	0.851 ± 0.004	0.722 ± 0.009	0.878 ± 0.005	0.635 ± 0.007	0.401 ± 0.009	-0.4 pp	0.020
LightGBM (proposed best)	0.762 ± 0.004	0.853 ± 0.004	0.730 ± 0.008	0.880 ± 0.005	0.644 ± 0.007	0.408 ± 0.010	-	-

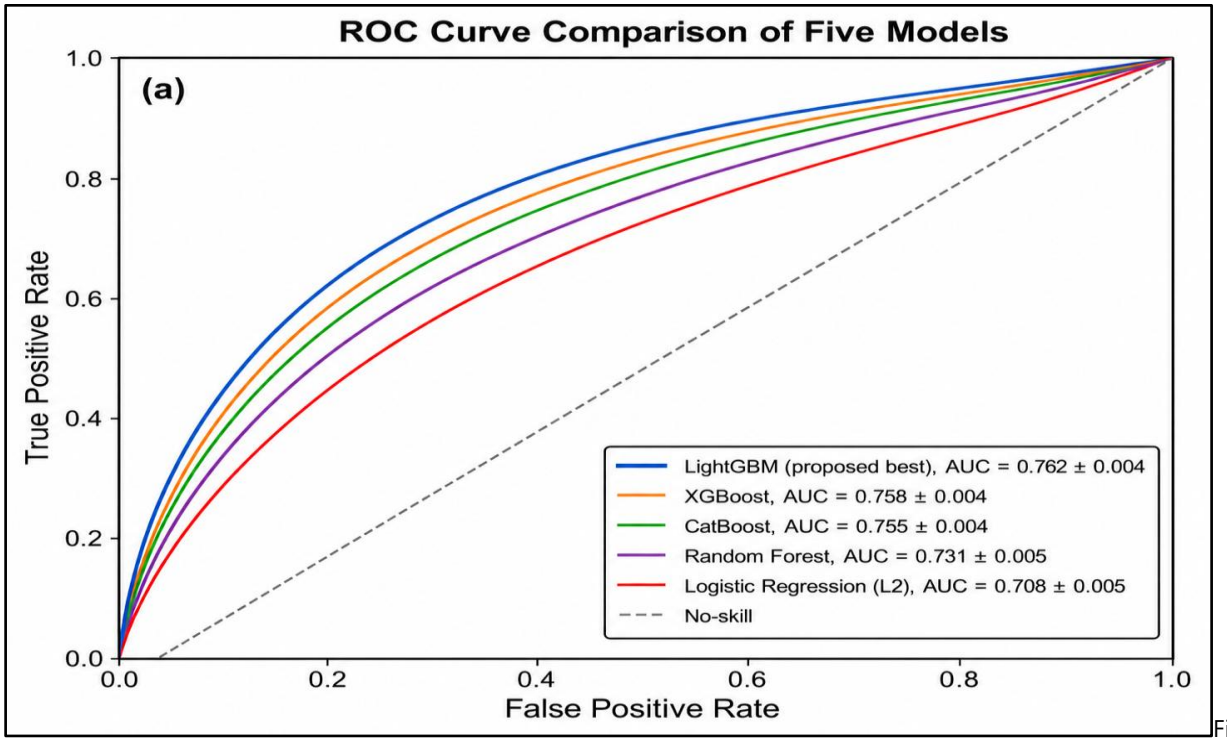

Figure 2. ROC Curve Comparison of Five Models

The ROC curve compares the discriminative performance of five models for binary credit-risk classification. LightGBM shows the strongest overall performance, with the highest curve and an AUC of 0.762 ± 0.004, followed closely by XGBoost (0.758 ± 0.004) and CatBoost (0.755 ± 0.004) (Figure 2). Random Forest achieved moderate discrimination with an AUC of 0.731 ± 0.005, while Logistic Regression showed the lowest performance with an AUC of 0.708 ± 0.005. All models performed above the no-skill diagonal line, confirming that each classifier learned meaningful separation between Fully Paid and Charged Off / Default loans, with LightGBM providing the best ranking ability among the evaluated methods.

4.3 Feature Importance and SHAP Direction Analysis

Table 6 reports the top 15 LightGBM features using gain-based importance and SHAP direction. The most influential feature was int_rate, contributing 16.4 ± 0.41% gain with a positive mean SHAP value of +0.412. This indicates that higher interest rates were associated with higher predicted default risk. The second most important feature was sub_grade, with 13.4 ± 0.38% gain and a positive mean SHAP value of +0.387, suggesting that issuer-assigned grade information carried strong risk-discrimination value.The next group of important predictors consisted of borrower affordability and credit-quality indicators. dti ranked third with 8.7 ± 0.29% gain and a positive SHAP direction of +0.206, indicating that higher debt-to-income ratios increased default risk. In contrast, log_annual_inc ranked fourth with 7.5 ± 0.27% gain and a negative SHAP value of -0.241, suggesting that higher annual income reduced predicted default risk. Similarly, fico_range_low ranked fifth with 7.4 ± 0.26% gain and a negative SHAP value of -

0.318, showing that stronger credit scores were associated with lower default probability.Loan structure and revolving-credit behaviour also contributed substantially to prediction. The term_60m indicator ranked sixth with 6.3 ± 0.24% gain and a positive SHAP direction of +0.253, suggesting that 60-month loans carried higher risk than shorter-term loans. revol_util ranked seventh with 5.42 ± 0.22% gain and a positive SHAP direction of +0.198, indicating that greater revolving-credit utilisation was associated with higher default risk. loan_amnt also showed a positive SHAP direction, meaning that larger requested loan amounts tended to increase predicted risk.The medium-importance feature group included inst_to_inc, inq_last_6mths, emp_length, open_acc, and mort_acc. The positive SHAP directions for installment-to-income ratio, recent inquiries, and open accounts indicate increased default tendency, whereas the negative SHAP values for employment length and mortgage accounts suggest lower predicted risk for borrowers with longer employment history or mortgage-account presence. Lower-ranked but still relevant features included delinq_2yrs and pub_rec_bankruptcies, both of which showed positive SHAP directions and therefore contributed to increased default-risk prediction.

Table 6. Top-15 Feature Importance and SHAP Direction (LightGBM)

Rank	Feature	Gain (% ± SD)	Mean SHAP	Stability	Description
1	int_rate	16.4 ± 0.41	+0.412	High	Loan APR; primary risk-pricing proxy
2	sub_grade	13.4 ± 0.38	+0.387	High	Issuer-assigned 35-tier grade (A1-G5)
3	dti	8.7 ± 0.29	+0.206	High	Debt-to-income ratio at origination
4	log_annual_inc	7.5 ± 0.27	-0.241	High	Self-reported annual income (log)
5	fico_range_low	7.4 ± 0.26	-0.318	High	Lower bound of FICO range
6	term_60m	6.3 ± 0.24	+0.253	High	60-month term indicator (vs 36)
7	revol_util	5.42 ± 0.22	+0.198	High	Revolving-credit utilisation %
8	loan_amnt	5.4 ± 0.21	+0.142	Medium	Requested loan amount (USD)
9	inst_to_inc	4.5 ± 0.20	+0.176	Medium	Installment / monthly income
10	inq_last_6mths	3.78 ± 0.18	+0.164	Medium	Credit inquiries in past 6 months
11	emp_length	3.16 ± 0.16	-0.092	Medium	Employment tenure (capped 10 years)
12	open_acc	2.9 ± 0.15	+0.071	Medium	Open credit accounts
13	mort_acc	2.61 ± 0.14	-0.098	Medium	Mortgage accounts on file
14	delinq_2yrs	2.24 ± 0.13	+0.119	Low	30+ day delinquencies past 24 months
15	pub_rec_bankruptcies	1.7 ± 0.11	+0.103	Low	Public-record bankruptcies
-	Remaining 35 features (cumulative)	8.9 ± 0.50	-	-	addr_state, purpose, home_ownership, etc.

5. Limitations and Future Work

Although the proposed LightGBM framework achieved the best overall performance among the evaluated models, several limitations should be acknowledged. First, the study was conducted on a single large lending dataset. Although the dataset contains a substantial number of completed loan records, the findings may still be influenced by the lending policies, borrower population, underwriting rules, and time period represented in that dataset. Therefore, the model's generalisability to other financial institutions, geographic regions, economic cycles, and credit products cannot be assumed without further external validation.Second, the task was formulated as a binary classification problem using completed loan outcomes, where Fully Paid loans were treated as the negative class and Charged Off / Default loans were treated as the positive class. This design supports clear supervised learning, but it excludes unresolved, active, or in-progress loans. As a result, the framework does not directly model time-to-default, partial repayment behaviour, early delinquency progression, or dynamic borrower risk over the lifetime of a loan. Future work should extend the framework toward survival analysis or temporal default prediction to capture when default is likely to occur, not only whether a loan eventually defaults.Third, although leakage-prone repayment variables were removed before training, the study remains dependent on the quality and completeness of the available structured variables. Some important risk factors, such as macroeconomic conditions, employment-sector stability, borrower cash-flow changes, regional economic shocks, and post-origination behavioural signals, were not explicitly modelled. Including carefully validated external economic indicators and time-aware borrower behaviour features may further improve risk discrimination and practical usefulness.

Fourth, the proposed model was evaluated using standard classification and discrimination metrics, including AUC, sensitivity, specificity, F1-score, KS statistic, Brier score, MCC, Cohen's kappa, and Gini coefficient. While these metrics provide a broad view of predictive performance, the study did not include a full business-impact analysis. In real lending environments, the cost of false negatives and false positives is not equal. Missing a default case may create financial loss, whereas incorrectly flagging a good borrower may reduce loan approval volume and customer access to credit. Future work should incorporate cost-sensitive evaluation, profit-loss simulation, approval-rate analysis, and decision-curve analysis.Fifth, the model provides feature-level interpretability through gain-based importance and SHAP-direction analysis. These methods help identify influential predictors such as interest rate, sub-grade, debt-to-income ratio, income, FICO range, loan term, revolving utilisation, and credit inquiries. However, interpretability in this study remains global and feature-oriented. It does not fully explain individual loan-level decisions in a regulatory or operational setting. Future studies should include borrower-level explanation reports, stability testing of SHAP explanations, fairness-aware explanation analysis, and expert review by credit-risk professionals.Another limitation is that fairness and bias analysis were not included. Credit-risk models can unintentionally encode indirect bias through variables correlated with socioeconomic status, geography, income, or credit history. Although the model used economically meaningful predictors, fairness cannot be assumed from performance metrics alone. Future research should evaluate subgroup-level performance, disparate impact, equal opportunity, calibration across borrower groups, and fairness-constrained model optimisation where legally and ethically appropriate.

6. Conclusion

This study developed a leakage-aware and interpretable LightGBM-based framework for binary credit-risk classification of Fully Paid and Charged Off / Default loans. Using a large completed-loan dataset of 887,379 records, the proposed workflow combined rigorous target definition, removal of post-origination leakage variables, robust preprocessing, engineered affordability and utilisation features, class-imbalance handling, repeated stratified cross-validation, and holdout testing. This design ensured that model training was based on realistic borrower and loan-origination signals rather than outcome-derived repayment information.Among the evaluated models, LightGBM achieved the strongest overall performance, with a cross-validation AUC of 0.762 ± 0.004 and a holdout AUC of 0.764. The model also achieved a holdout accuracy of 0.853, sensitivity of 0.730, specificity of 0.880, default-class F1-score of 0.644, NPV of 0.936, and Brier score of 0.124. Compared with Logistic Regression, Random Forest, CatBoost, and XGBoost, LightGBM provided the best discrimination and default-detection performance while remaining computationally efficient on CPU-only hardware.The interpretation analysis showed that the model's predictions were mainly driven by economically meaningful credit-risk factors, including interest rate, sub-grade, debt-to-income ratio, annual income, FICO range, loan term, revolving utilisation, and recent credit inquiries. These findings suggest that the proposed framework provides not only improved predictive performance but also transparent insight into borrower risk patterns.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

1.Ding, Ming-Liang, Yu-Liang Ma, and Fu-Qiang You. "Domain-Constrained Stacking Framework for Credit Default Prediction." Mathematics 13, no. 21 (2025): 3451.

2.Chen, Y., & Zhang, R. (2021). Default Prediction of Automobile Credit Based on Support Vector Machine. Journal of Information Processing Systems, 17(1).

3.Hou, J., Li, Q., Liu, Y., & Zhang, S. (2021). An enhanced cascading model for E-commerce consumer credit default prediction. Journal of Organizational and End User Computing (JOEUC), 33(6), 1-18.

4.Çallı, Büşra Alma, and Erman Coşkun. "A longitudinal systematic review of credit risk assessment and credit default predictors." Sage Open 11, no. 4 (2021): 21582440211061333.

5.Wu, Z., Dong, Y., Li, Y., & Shi, B. (2023). Unleashing the power of text for credit default prediction: Comparing human-generated and AI-generated texts. SSRN Electronic Journal. https://doi. org/10.2139/ssrn, 4601317.

6.Uddin, Mohammad Shamsu, Guotai Chi, Tabassum Habib, and Ying Zhou. "An alternative statistical framework for credit default prediction." Journal of Risk Model Validation (2020).

7.Faraj, Azhi Abdalmohammed, Didam Ahmed Mahmud, and Bilal Najmaddin Rashid. "Comparison of different ensemble methods in credit card default prediction." UHD Journal of Science and Technology 5, no. 2 (2021): 20-25.

8.Guo, K. (2025, July). Corporate Credit Default Prediction Based on Machine Learning. In Proceedings of the 2025 International Conference on Economic Management and Big Data Application (pp. 340-348).

9.Shang, Z., Meng, H., Zhao, Y., Xu, R., Xu, Y., & Cui, L. (2023). Cross-domain credit default prediction via interpretable ensemble transfer. International Journal of Crowd Science, 7(3), 106-112.

10.Rahmani, Rambod, Marco Parola, and Mario GCA Cimino. "A machine learning workflow to address credit default prediction." arXiv preprint arXiv:2403.03785 (2024).

11.Liang, Xianhuai. "Credit default prediction algorithm based on improved TabNet." In 2023 3rd International Conference on Intelligent Communications and Computing (ICC), pp. 110-114. IEEE, 2023.

12.Dennis, Kagaba, Busingye Caroline, Barbara Nansamba, Daudi Jjingo, and Ggaliwango Marvin. "Explainable Deep Ensemble Learning for Improved Credit Default Prediction." In Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing, pp. 256-270. 2024.

13.Akinjole, Abisola, Olamilekan Shobayo, Jumoke Popoola, Obinna Okoyeigbo, and Bayode Ogunleye. "Ensemble-based machine learning algorithm for loan default risk prediction." Mathematics 12, no. 21 (2024): 3423.

14.Guo, Kangshuai, Shichao Luo, Ming Liang, Zhongjian Zhang, Huabin Yang, Yan Wang, and Yingjie Zhou. "Credit default prediction on time-series behavioral data using ensemble models." In 2023 International Joint Conference on Neural Networks (IJCNN), pp. 01-09. IEEE, 2023.

15.Bhattacharya, Arijit, Ardhendu Mandal, Saroj Kr Biswas, Debasmita Saha, Akhil Kumar Das, Ekram Alam, and Shrabani Dubey. "DbISCDP: An Empirical Study on Distance-Based Nearest Neighbor Approaches for Credit Default Prediction." In 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), pp. 1-7. IEEE, 2024.

16. Qiu, Y., & Wang, J. (2025, March). Credit default prediction using time series-based machine learning models. In Artificial Intelligence and Applications (Vol. 3, No. 3, pp. 284-294).

17. Alam, Talha Mahboob, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. "An investigation of credit card default prediction in the imbalanced datasets." Ieee Access 8 (2020): 201173-201198.

18. Zhou, Xiang, Wenyu Zhang, and Yefeng Jiang. "Personal credit default prediction model based on convolution neural network." Mathematical Problems in Engineering 2020, no. 1 (2020): 5608392.

19. Wang, Yuhan, Zhen Xu, Kunyuan Ma, Yuan Chen, and Jinsong Liu. "Credit default prediction with machine learning: A comparative study and interpretability insights." In 2024 4th International Conference on Communication Technology and Information Technology (ICCTIT), pp. 496-500. IEEE, 2024.

20. Wang, S., & Zhang, X. (2024). Research on credit default prediction model based on TabNet-stacking. Entropy, 26(10), 861.

21. Alonso Robisco, Andrés, and Jose Manuel Carbo Martinez. "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction." Financial Innovation 8.1 (2022): 70.

22. Nguyen, Quoc Giang, Linh Hoang Nguyen, Md Monir Hosen, Mohammad Rasel, Jannatul Ferdous Shorna, Md Sakib Mia, and Sajidul Islam Khan. "Enhancing credit risk management with machine learning: A comparative study of predictive models for credit default prediction." The American Journal of Applied sciences 7, no. 01 (2025): 21-30.

23. Alvi, Jahanzaib, Imtiaz Arif, and Kehkashan Nizam. "Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management." Heliyon 10.21 (2024).

24. Kriebel, Johannes, and Lennart Stitz. "Credit default prediction from user-generated text in peer-to-peer lending using deep learning." European Journal of Operational Research 302, no. 1 (2022): 309-323.

25. Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X. and Song, J., 2024, May. Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. In 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 421-426). IEEE.

26. Wu, Z., Dong, Y., Li, Y., & Shi, B. (2025). Unleashing the power of text for credit default prediction: Comparing human-written and generative AI-refined texts. European Journal of Operational Research, 326(3), 691-706.

27. Ma, Yuanyuan, Pingping Zhang, Shaodong Duan, and Tianjie Zhang. "Credit default prediction of Chinese real estate listed companies based on explainable machine learning." Finance Research Letters 58 (2023): 104305.

28. Yu, Yue. "The application of machine learning algorithms in credit card default prediction." In 2020 international conference on computing and data science (CDS), pp. 212-218. IEEE, 2020.

29. Lee JW, Lee WK, Sohn SY. Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Systems with Applications. 2021 Apr 15;168:114411.

30. Mandour, M. A., & Chi, G. (2024, March). A Review Study of AI Methods for Credit Default Prediction. In International Conference on Deep Learning and Visual Artificial Intelligence (pp. 265-284). Singapore: Springer Nature Singapore.

31. Shu, Mengying, Jiayu Liang, and Chenyao Zhu. "Automated risk factor extraction from unstructured loan documents: An NLP approach to credit default prediction." Artificial Intelligence and Machine Learning Review 5.2 (2024): 10-24.

32. Han, D., Guo, W., Chen, Y., Wang, B., & Li, W. (2024). Personal credit default prediction fusion framework based on self-attention and cross-network algorithms. Engineering Applications of Artificial Intelligence, 133, 107977.

33. Bhandary, Rakshith, and Bidyut Kumar Ghosh. "Credit card default prediction: An empirical analysis on predictive performance using statistical and machine learning methods." Journal of Risk and Financial Management 18, no. 1 (2025): 23.

34. Yan, Z., Qu, H., Chen, C., Lv, X., Zuo, E., Wang, K., & Cai, X. (2025). WIGNN: An adaptive graph-structured reasoning model for credit default prediction. Engineering Applications of Artificial Intelligence, 139, 109597.

35. Wang, K., Wan, J., Li, G., & Sun, H. (2022). A hybrid algorithm-level ensemble model for imbalanced credit default prediction in the energy industry. Energies, 15(14), 5206.