American Journal of Data Science and Artificial Intelligence (AJDSAI)

ISSN: 3069-3632 (ONLINE)

VOLUME 2 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Data Science and Artificial Intelligence (AJDSAI)

Volume 2 Issue 2, Year 2026
ISSN: 3069-3632 (Online)
DOI: https://doi.org/10.54536/ajdsai.v2i2.8054
https://journals.e-palli.com/home/index.php/ajdsai

XGBoost versus LightGBM for Loan Approval Prediction: A Controlled, Like-for-Like Comparison

Sagar Poudel[1*], Bhoj Raj Ghimire[1]

Article Information

Received: April 12, 2026

Accepted: June 22, 2026

Published: September 04, 2026

Keywords

Credit Risk, Gradient Boosting, Lightgbm, Loan Approval Prediction, XGBoost

ABSTRACT

Deciding whether to approve a loan application is one of the highest-stakes decisions a lender makes: approving a risky applicant invites default, while turning away a sound one loses a good customer. Predicting the approval decision accurately is therefore central to comprehensive credit management. Gradient boosting has become the dominant approach for such tabular credit data, and two libraries, XGBoost and LightGBM, account for most of its practical use. Although both are widely applied to credit-risk and loan-approval prediction, they are seldom compared directly under matched conditions, which leaves little firm evidence for choosing between them; this study addresses that gap. A controlled, like-for-like comparison was carried out in which both models were trained on the same dataset of 45,000 loan applications using one pre-processing pipeline and a matched set of structural hyper-parameters, and were then evaluated with accuracy, precision, recall, F1-score and the area under the ROC curve. The comparison was re-examined with five-fold stratified cross-validation and repeated on a second, independent loan-approval dataset of a different character. On the primary dataset, XGBoost reached an accuracy of 93.5% and an ROC-AUC of 0.979, a small margin above LightGBM on every metric, and it retained the higher mean AUC under cross-validation; on the second, high-signal dataset both models exceeded 98% accuracy and XGBoost again held the higher cross-validated AUC, winning four of five folds. Across both dataset types, XGBoost therefore demonstrated a consistent, though modest, improvement in discrimination, an effect compatible with its level-wise tree growth and stronger regularization. The principal limitations are the use of a single matched configuration and two historical datasets. The findings indicate that both libraries are accurate and deployable, with XGBoost showing a marginal but reproducible edge in predictive discrimination.

INTRODUCTION

Lending sits at the centre of every financial system. Banks, microfinance institutions and savings cooperatives extend credit to households and firms, and how carefully they decide who receives that credit shapes both their own solvency and the wider availability of finance. The first and most consequential point in that process is the approval decision: from the limited information on an application, the lender must judge whether to grant the loan at all. An approval granted to an applicant who cannot repay erodes the institution's capital, while a sound applicant who is turned away is a paying customer needlessly lost. Predicting the approval decision accurately and consistently is therefore a problem of direct economic value, and it is the problem this study takes up.

For much of the history of consumer lending this judgement was made by hand, through fixed rule-based scorecards and an officer's reading of the applicant's credit file. Such methods are easy to explain, but they are slow, partly subjective, and poorly suited to capturing the non-linear interactions among income, indebtedness and credit history that often separate a reliable borrower from a risky one. As application records have moved into digital form, machine learning has emerged as a practical alternative that learns these patterns directly from past decisions and improves as more data accumulate (Barbaglia *et al.,* 2023; Yang, 2024). On the structured, mixed numeric-and-categorical tables that credit applications produce, tree-based ensembles, and gradient boosting above all, have repeatedly delivered the strongest predictive performance (Akinjole *et al.,* 2024; Zhu *et al.,* 2023).

Two implementations dominate the everyday use of gradient boosting. XGBoost grows its trees level by level and restrains their complexity through explicit first- and second-order regularization (Chen & Guestrin, 2016), whereas LightGBM grows trees leaf by leaf, bins continuous features into histograms, and adds sampling and feature-bundling techniques that make it remarkably fast on large data (Ke *et al.,* 2017). Both are staples of modern credit scoring, and broad benchmarks usually describe them as close in accuracy (Anghel *et al.,* 2018; Bentejac *et al.,* 2021). What is far less settled is which of the two a practitioner should actually choose for loan approval prediction, because the libraries are seldom placed on the same data under the same settings. Many studies pit boosting against weaker baselines, fold the two algorithms into a single ensemble (Zhu *et al.,* 2024), or average results across unrelated domains, none of which answers the practical question directly.

The question carries a particular edge in developing

[1] Faculty of Science, Health and Technology, Nepal Open University, Lalitpur, Nepal
[*] Corresponding author's e-mail: sagar.poudel255@gmail.com

financial systems. In many institutions, machine-learning models are still adopted carefully, partly because there is little local, like-for-like evidence on which algorithm to trust for a given task, and partly because a poor choice is expensive to undo once a scoring system is in production. A transparent and reproducible comparison of the two most widely used boosting libraries is therefore useful not only as an academic exercise but as practical guidance for teams that must commit to a single framework. By holding every part of the experiment fixed except the learning algorithm, this study aims to give that choice a safer evidential footing than the present, conflicting literature provides.

The study is accordingly guided by three research questions. First, under identical training conditions, which of the two frameworks, XGBoost or LightGBM, gives more accurate and reliable predictions of loan approval? Second, does any advantage that emerges survive when the comparison is repeated with cross-validation and on a second, independent loan-approval dataset? Third, which applicant and loan attributes weigh most heavily on the approval decision? From these questions follow the objectives: to build a fully matched experimental pipeline for the two frameworks, to evaluate them with several complementary metrics and two robustness checks, and to identify the most influential predictors. This study contributes to the literature by establishing a controlled protocol and providing an architectural interpretation with a controlled protocol in which only the boosting algorithm varies; an evaluation that combines five metrics, five-fold cross-validation and a second dataset rather than a single split; and an interpretation grounded in how the two libraries build their trees. This relevant literature, sets out the data and method, presents and discusses the results, and closes with the study's implications and limitations.

LITERATURE REVIEW

The work relevant to this study falls into five overlapping aspects: the adoption of machine learning for credit assessment, the move from single classifiers to troupes, head-to-head comparisons among gradient boosting libraries, the treatment of class imbalance in credit data, and the growing demand for explain-ability. Taken together, they describe a field that has largely agreed on which tools to use but not on how to choose among the closest of them, which is precisely the gap this study addresses.

That data-driven models now outperform static records is well documented. Yang (2024) compared a range of learners for loan approval and credit-risk assessment and reported clear gains over conventional scoring, and Barbaglia *et al.* (2023), working on European data, found ensemble methods markedly more accurate than econometric baselines. Comparable applied gains have been reported for general bank-loan settings; Haque and Hassan (2024), for instance, applied several machine-learning techniques directly to the bank-loan decision. Looking past raw accuracy, Alonso Robisco and Carbo Martinez (2022) evaluated several algorithms on a model-risk-adjusted basis and identified XGBoost and random forests as the most efficient once model risk is taken into account. Two recent reviews place these findings in perspective: Markov *et al.* (2022) survey the main families of credit-scoring methods and the trade-offs they involve, while Alvi *et al.* (2024), reviewing work from 2015 to 2024, document a rapid migration toward boosting and hybrid models but warn that weak comparability across studies makes their results difficult to reconcile.

A recurring lesson is that single classifiers rarely suffice. Logistic regression and single decision trees are transparent but miss the interactions that drive credit risk, which has pushed the field toward ensembles. Akinjole *et al.* (2024) combined several base learners and obtained higher accuracy and stability than any one of them on public loan data, and Emmanuel *et al.* (2024) reported comparable gains from a stacked classifier paired with filter-based feature selection. Closer to the approval task itself, Ozkurt (2024) benchmarked gradient boosting, XGBoost and AdaBoost for personal-loan eligibility and found XGBoost the most accurate of the three. Prior research consistently establishes that ensemble learning, and gradient boosting in particular, is the most effective approach to this family of credit-decision problems.

Within the boosting family, XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke *et al.,* 2017) are the reference implementations, and several studies have set them side by side. Broad benchmarks by Bentejac *et al.* (2021) and Anghel *et al.* (2018) found the variants close in accuracy, with the ranking turning on the dataset and on the care taken in tuning, while Hancock and Khoshgoftaar (2020) argued that robustness and the quality of probability estimates, rather than speed, should decide the choice. Domain-specific studies do not converge on a single winner. Chang *et al.* (2024) compared XGBoost, LightGBM and deep models on credit-card data; Nguyen and Ngo (2025) ranked four boosters on personal-loan data and found LightGBM strongest; Hlongwane *et al.* (2024a), adding alternative data, again placed LightGBM ahead; and Ko *et al.* (2022) reported LightGBM best for peer-to-peer lending. Ying *et al.* (2025), by contrast, extended LightGBM with an attention-based hybrid design to push it further in digital-finance credit risk. This spread of outcomes is exactly what makes a single, tightly controlled comparison on the same loan-approval data informative rather than redundant.

The robustness of XGBoost has also been demonstrated well outside credit scoring, which strengthens the case for treating it as a dependable baseline here. Working on the very different problem of predicting issue-resolution time in agile software projects, Kunwar and Ghimire (2026) found that XGBoost produced the most consistent hold-out results, with tighter and more reliable error distributions than the models it was compared against. In a further domain, Lawal *et al.* (2026) used an XGBoost-based ensemble to predict food insecurity across more

than seventy thousand United States census tracts and reported the highest accuracy and near-perfect ranking ability, together with stable and interpretable behaviour. That the same framework holds up so well across such different tasks indicates that its strength does not depend on any one dataset, and it is part of why XGBoost is retained as a firm reference point in the controlled loan-approval comparison that follows.

Because approved and rejected applications are seldom present in equal numbers, several studies concentrate on how imbalance is handled. Gu *et al.* (2024) paired SMOTE with XGBoost to build a credit scorecard for small and micro enterprises, and Hou *et al.* (2025) compared eight resampling techniques with XGBoost for financial-distress prediction, showing that the resampling choice can move the result as much as the model itself. The implication for a fair comparison is direct: the treatment of imbalance must be held identical for both learners, or it becomes impossible to tell whether a difference reflects the algorithm or the sampling scheme.

Concurrently, model transparency has become a critical requirement, since in a regulated setting a credit decision must be strong as well as accurate. Zhu *et al.* (2023) showed that explainable models can match the accuracy of opaque ones on loan data; Hossain *et al.* (2025) went further, bridging differential privacy and explainable machine learning in a secure bank-loan prediction system; and Hlongwane *et al.* (2024b) used Shapley values to build interpretable scorecards that rival logistic regression for clarity. Together these works suggest that interpretability and predictive performance are no longer in difficulty, and that case-level explanation is now expected alongside accuracy.

Read as a whole, the literature has settled on gradient boosting as its workhorse but not on a single implementation, and the evidence on which library is preferable is genuinely divided: depending on the dataset, the resampling scheme and the tuning effort, either XGBoost or LightGBM can come out ahead (Alonso Robisco & Carbo Martinez, 2022; Chang *et al.,* 2024; Hlongwane *et al.,* 2024a; Ko *et al.,* 2022; Nguyen & Ngo, 2025). Two gaps follow. First, most comparisons either bury the two algorithms within a larger benchmark or merge them into a single ensemble (Zhu *et al.,* 2024) rather than isolating them under matched conditions. Second, the factors that most affect the outcome, namely identical pre-processing, an identical split, identical hyper-parameters and identical imbalance handling, together with the handling of high-cardinality categorical features that Miljkovic and Wang (2025) show can materially shift credit-scoring accuracy, are rarely controlled at the same time (Hancock & Khoshgoftaar, 2020). A clean, like-for-like comparison of XGBoost against LightGBM on the same loan-approval data, confirmed by cross-validation and on a second dataset, therefore addresses a real and currently open question, and supplying it is the aim of this study.

MATERIALS AND METHODS

Research Design

The study uses a quantitative, experimental design based on secondary data. The most convincing way to decide which of two algorithms is preferable is to run both on real data under controlled conditions and measure the outcome. A single machine-learning pipeline, comprising data loading, exploratory analysis, cleaning, encoding, training and evaluation, was therefore built, and both XGBoost and LightGBM were passed through it without altering anything except the model itself. All work was carried out in Python within a Jupyter Notebook environment, which keeps the experiment transparent and reproducible. To guard against a result that holds only for one partition of the data, the comparison is repeated with five-fold stratified cross-validation and is run again on a second, independent loan-approval dataset that differs in size, balance and task.

Datasets

Two publicly available loan datasets were used. The primary dataset contains 45,000 loan records described by 13 predictors and a binary target, loan status, that records whether an application was approved (1) or rejected (0); its features span applicant demographics, financial information, loan characteristics and credit-history indicators, and about 22% of the applications were approved. The second dataset is a publicly available loan-approval dataset of 4,269 records described by 11 predictors, including a credit (CIBIL) score, income, requested loan amount, loan term and several asset values, with a binary target indicating whether each application was approved or rejected; about 62% of the applications were approved. It is cleaner and carries stronger predictive signal than the primary dataset, and is used solely to test whether the comparison generalizes to a different, but closely related, credit-decision task. The two datasets are summarized in Table 1.

The predictors of the primary dataset group naturally into four families: applicant demographics such as age, gender and education; financial standing such as annual income, home ownership and credit score; loan characteristics such as the amount requested, the interest rate, the stated purpose and the loan-to-income ratio; and credit-history indicators such as the length of credit history and any previous default on file. This mixture of demographic, financial and behavioural variables is typical of consumer credit data and is well suited to tree-based models, which can combine numeric thresholds with categorical splits without the analyst having to specify interaction terms by hand. The second dataset has a similar mixture but a stronger headline predictor in the credit score, and was chosen deliberately so that the comparison could be examined on two credit-decision tasks of different type, size and difficulty.

Table 1: Characteristics of the two datasets used in the study

Property	Dataset 1 (primary)	Dataset 2 (validation)
Labelled records	45,000	4,269
Predictor features	13	11
Prediction target	Loan approval	Loan approval
Target positive rate	22.2% approved	62.2% approved
Signal strength	High	High (clean)
Class imbalance handling	None required	None required

Data Preprocessing

Pre-processing was deliberately simple and, most importantly, identical for both models, so that any difference in performance could be attributed to the algorithm rather than to differing data treatment. Binary attributes such as gender and the prior-default indicator were mapped to {0, 1}; the ordered education variable was encoded with increasing integers; and the nominal attributes, home ownership and loan intent, were one-hot encoded with one category dropped to avoid redundancy. A small number of clearly impossible values, such as ages above the 99th percentile, were capped at the median so that a few outliers could not distort training. The numeric features were standardized with a robust scaler. This step is harmless but in fact unnecessary, because XGBoost and LightGBM are tree-based and split on thresholds; it was applied uniformly and therefore cannot influence the comparison. On the second dataset, the identifier column was removed and the two categorical fields (education and self-employment status) were binary-encoded, while the credit score, income and asset values were used as supplied; the data contained no missing values and required no record to be discarded.

Gradient Boosting Models

Both frameworks construct an additive ensemble of regression trees in which each new tree corrects the residual errors of the ensemble built so far. Formally, the prediction for an instance x is

$$\hat{y}_i = \Sigma\ (k = 1 \ldots K)\ f_k(x_i), \quad f_k \in F, \tag{1}$$

where K is the number of trees and F is the space of regression trees. The trees are learned by minimizing a regularized objective that balances goodness of fit against model complexity:

$$L = \Sigma_i\ l(y_i, \hat{y}_i) + \Sigma_k\ \Omega(f_k), \quad \text{with} \quad \Omega(f) = \gamma T + \tfrac{1}{2}\ \lambda\ \|w\|^2, \tag{2}$$

in which l is a differentiable loss, T is the number of leaves in a tree, w is the vector of leaf weights, and γ and λ are regularization parameters. The two libraries differ chiefly in how a tree is grown under this objective. XGBoost expands every level of a tree before going deeper (level-wise growth) and leans on the L1 and L2 penalties in Ω to keep the model conservative, which makes it resistant to over-fitting. LightGBM instead extends the leaf that offers the largest reduction in loss (leaf-wise growth), bins continuous features into histograms, and adds two techniques, Gradient-based One-Side Sampling, which keeps the instances with the largest gradients while sub-sampling the rest, and Exclusive Feature Bundling, which packs mutually exclusive sparse features together. These devices make LightGBM very fast and memory-efficient, but the leaf-wise strategy can also fit noise more readily on data of moderate size. Exposing exactly this difference in growth strategy is the purpose of the experiment.

Experimental Setup

Each dataset was split into 80% training and 20% test partitions, stratified on the target so that both partitions preserved the approval rate, and the same seeded split was reused for both models. The decisive rule of the experiment was fairness: XGBoost and LightGBM were given the same number of trees, the same learning rate, the same maximum tree depth, the same row- and feature-sampling rates and the same random seed, so that the structural settings the two libraries share were held identical. The differences that remained were those built into each framework, chiefly its tree-growth strategy together with the L2 regularization each applies by default, which is stronger in XGBoost than in LightGBM. The configuration for the primary dataset is listed in Table 2. Because the second dataset is reasonably balanced, no class weighting was required for it; both models were trained on it under their own matched configuration (500 trees, a learning rate of 0.03 and a maximum depth of 3, again identical across the two libraries), chosen to suit its smaller size. A fixed random seed was used throughout, so the experiment is fully reproducible.

Table 2: Hyper-parameters used for both frameworks on the primary dataset; the structural settings are identical, while the L2 penalty is left at each library's default

Hyper-parameter	XGBoost	LightGBM
Number of trees	400	400
Learning rate	0.05	0.05
Maximum tree depth	5	5
Row / feature subsample	0.9 / 0.9	0.9 / 0.9
L2 regularization (reg_lambda)	1.0	0.0
Random seed	42	42

Evaluation Metrics

Five metrics were computed from the confusion matrix of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Accuracy gives the overall share of correct decisions; precision and recall describe, respectively, how many of the records predicted to be approved are truly approved and how many of the truly approved records are identified; and the F1-score is their harmonic mean:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN), \quad (3)$$

$$\text{Precision} = TP / (TP + FP), \quad \text{Recall} = TP / (TP + FN), \quad (4)$$

$$F1 = 2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall}). \quad (5)$$

The fifth metric, the area under the receiver operating characteristic curve (ROC-AUC), measures how well a model ranks applications by their likelihood of approval across every decision threshold. For credit scoring this is the most informative measure, because it is threshold-independent and is linearly related to the Gini coefficient routinely used in practice, $\text{Gini} = 2 \cdot \text{AUC} - 1$. To ensure that the single-split results were not an artefact of a fortunate partition, every comparison was repeated with five-fold stratified cross-validation, and the mean fold AUC was used as the primary basis for choosing between the two closely matched models.

Reporting several metrics together is deliberate, because each answers a different question that matters in lending. Accuracy alone can be misleading on imbalanced data, since a model that simply predicted rejection for every application would already appear fairly accurate; precision speaks to the cost of approving applications that should have been declined, recall to the cost of turning away applications that ought to have been approved, and the F1-score to the balance between the two. The ROC-AUC sits above all of these because it summarizes the quality of the risk ranking independently of where the approval threshold is set, which is exactly the flexibility a lender needs when it adjusts its risk appetite over the economic cycle. Reading the metrics as a set, rather than optimizing any one of them in isolation, gives a more honest picture of which framework is better suited to the task.

Software and Implementation

All experiments were implemented in Python and executed from Jupyter notebooks. Data handling relied on pandas and NumPy; the exploratory plots and result figures were produced with Matplotlib and seaborn; and the train-test split, stratified cross-validation, pre-processing utilities and evaluation metrics were taken from scikit-learn, used together with the official XGBoost and LightGBM packages. Keeping the entire workflow within a single, widely available open-source stack means that the experiment can be reproduced exactly by any reader with access to the same data and code, which is important for a study whose central claim concerns a small but systematic difference between two models. Both frameworks were trained on the same machine, so the training-time observations reported in the next section reflect the algorithms themselves rather than differences in the computing environment.

RESULTS AND DISCUSSION

Exploratory Analysis

The primary dataset proved clean, with no missing values and no duplicated records, which is convenient because it removes data cleaning as a confounding factor. The target is imbalanced but not severely so: about 78% of applications were rejected and 22% approved, as shown in Figure 1. Because a model that simply predicted "rejected" for everyone would already score about 78% accuracy, the analysis leans on ROC-AUC and F1-score rather than on accuracy alone. Among the categorical features, applicants with a previous default on file were very rarely approved, and renters were approved at a different rate from those who owned a home or held a mortgage; among the numeric features, the loan-to-income ratio and the interest rate showed the clearest association with the approval outcome, while income played a smaller role. These patterns agree with ordinary banking intuition and, as Section 4.6 shows, with what the trained models came to rely on.

Predictive Performance

Both frameworks were trained and scored on the same held-out test set. The results in Table 3 and Figure 2 are consistent: on this dataset XGBoost is ahead of LightGBM on every metric. Its ROC-AUC of 0.9790

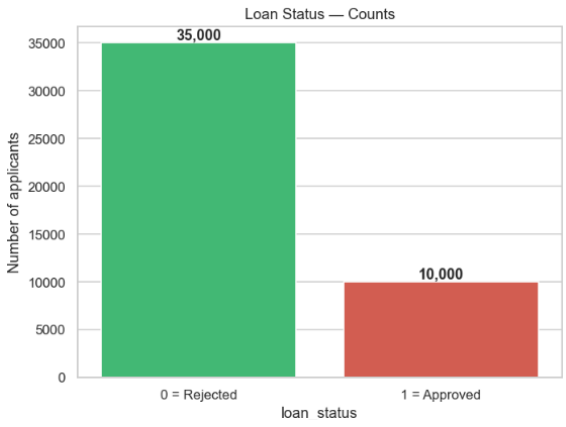

Figure 1: Distribution of the loan-status target in the primary dataset

exceeds the 0.9782 of LightGBM, and it also leads on accuracy, precision, recall and F1-score. The margins are small, as one would expect from two members of the same boosting family trained with identical settings, but the direction of the difference is the same for all five metrics. Importantly, XGBoost does not buy a gain on one measure with a loss on another; it is uniformly, if modestly, the stronger classifier, while LightGBM's only measurable advantage on this dataset was a slightly shorter training time.

The per-class behaviour is also informative. At the standard 0.5 decision threshold, XGBoost correctly identified about 80% of the truly approved applications while keeping precision close to 0.90, which means that most of the applications it predicted would be approved were in fact approved. This combination matters in practice, because the two kinds of error carry very different costs: a false rejection turns away a customer who would have repaid, whereas a false approval results in a loan that is unlikely to be recovered. A classifier that keeps both precision and recall high, as both frameworks do here, is therefore more useful than one that drives a single metric up at the expense of the other. That XGBoost achieves slightly higher values on both sides of this trade-off, rather than trading one against the other, is the clearest sign that its advantage is genuine rather than an artefact of the chosen threshold.

Table 3: Validation performance of the two frameworks on the primary dataset under identical settings

Metric	XGBoost	LightGBM	Difference
ROC-AUC	0.9790	0.9782	+0.0008
Accuracy	0.9353	0.9338	+0.0015
Precision	0.8974	0.8957	+0.0017
Recall	0.8005	0.7945	+0.0060
F1-score	0.8462	0.8421	+0.0041

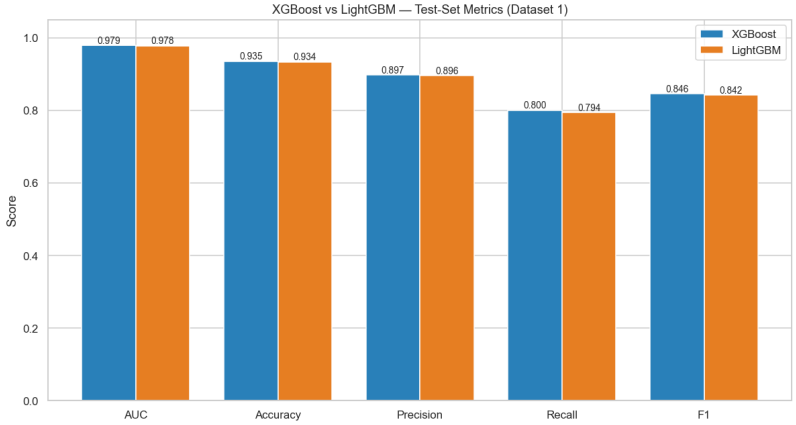

Figure 2: Validation metrics for XGBoost and LightGBM on the primary dataset

Discrimination

Because a credit model is ultimately judged on how well it ranks risk, the ROC curve is the fairest single view of its quality. Figure 3 plots both curves. They are very close, as the AUC values imply, but the XGBoost curve lies on or just above the LightGBM curve across most of the range, which is why its enclosed area is the larger of the two. Since the area under this curve is linearly related to the Gini coefficient, even a small but stable advantage in AUC means that XGBoost orders applicants by risk

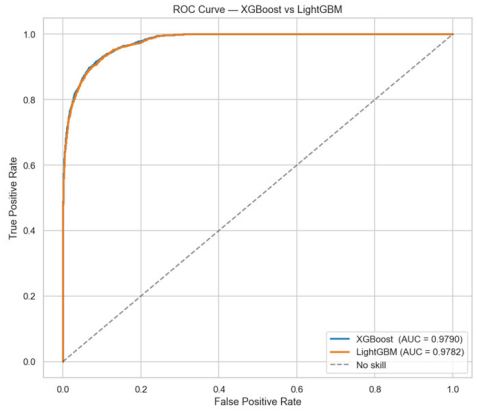

Figure 3: ROC curves for XGBoost and LightGBM on the primary dataset

slightly more reliably, which is the property a scorecard is designed to deliver.

Cross-Validation Robustness

A single split can flatter one model by chance, so the comparison was repeated with five-fold stratified cross-validation. Figure 4 shows the per-fold AUC and Table 4 reports the averages. XGBoost retains the higher mean AUC, 0.97819 against 0.97797, even though both models used identical settings, and it leads on the majority of folds. The persistence of the advantage when the experiment is averaged over five independent partitions is the strongest evidence in the study that the difference is systematic rather than accidental, and it is consistent with the view that XGBoost's heavier regularization yields marginally better generalization.

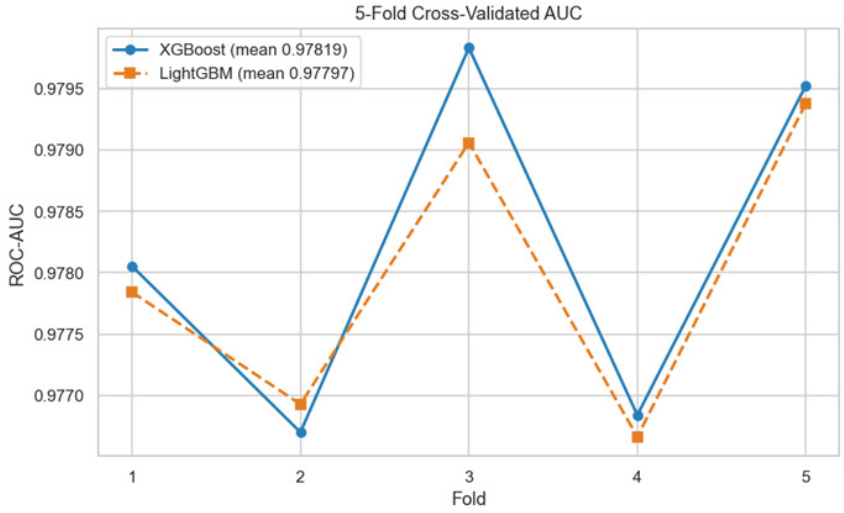

Figure 4: Five-fold cross-validated AUC on the primary dataset

Table 4: Mean five-fold cross-validated AUC on both datasets

Dataset	XGBoost (mean AUC)	LightGBM (mean AUC)
Dataset 1 (primary)	0.97819	0.97797
Dataset 2 (4 of 5 folds)	0.99812	0.99779

Cross-Dataset Validation

To rule out a result specific to one dataset, the whole comparison was repeated on the second dataset, a public loan-approval dataset of 4,269 records in which the task is again to predict the approval decision, on different applicants and features. This dataset is cleaner and carries far stronger signal, so both models are highly accurate on it, with ROC-AUC close to 0.998 and accuracy around 98%. The relative ranking nonetheless echoes the primary result: as Figure 5 shows, XGBoost obtains the marginally higher test-set AUC (0.9984 against 0.9980), and under five-fold cross-validation it again holds the higher mean AUC, winning four of the five folds (Table 4). At the fixed 0.5 threshold LightGBM recorded a slightly higher

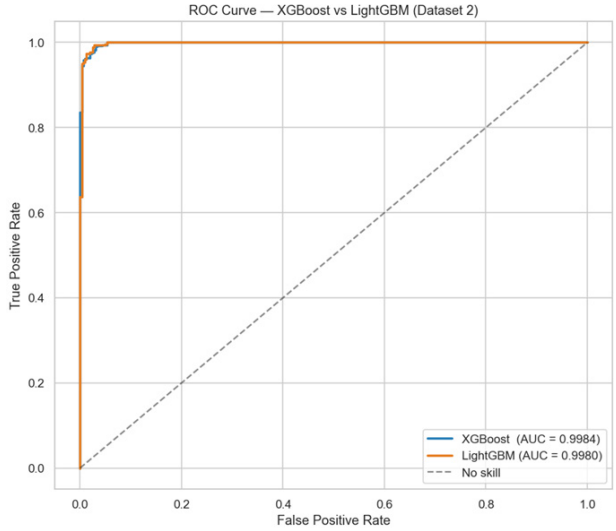

Figure 5: ROC curves for XGBoost and LightGBM on the second (loan-approval) dataset

accuracy (98.5% against 98.0%), but the difference is well within the noise expected on a dataset where both models exceed 98%, and the threshold-independent AUC continues to favour XGBoost. Observing the same direction of advantage on a second dataset of very different size, balance and task increases confidence that the result is not specific to a single dataset, while also confirming that the advantage, though consistent, is very small whenever the data carry strong signal.

Feature Importance

It is reassuring that the two models not only perform similarly but also rely on the same drivers. Figure 6 compares their leading features. Both place a prior default on file, the loan-to-income ratio, the interest rate and income among the most important variables, which agrees with the exploratory analysis and with banking intuition. This agreement indicates that both models learned genuine economic structure rather than noise; XGBoost merely combines that structure a little more effectively, as the predictive metrics show. The consistency of the important predictors across the two frameworks also lends face validity to the models, which matters when the results are used to support real lending decisions.

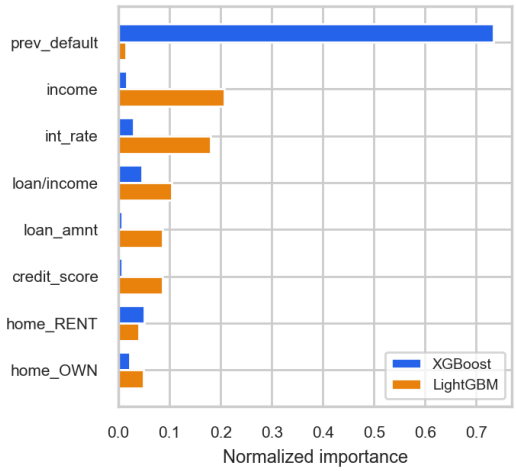

Figure 6: Normalized feature importance for XGBoost and LightGBM on the primary dataset

Discussion

The behaviour observed throughout the experiments has a clear cause in the way the two libraries build trees. The leaf-wise growth of LightGBM is fast and often accurate, but by always pursuing the largest immediate gain it is more willing to fit noise, which is plentiful in credit data. The level-wise growth of XGBoost, combined with stronger L1 and L2 regularization, produces a more cautious model that generalizes slightly better to unseen records. In lending, where the cost of approving a borrower who later defaults greatly exceeds the cost of a fraction of a second of additional computation, this caution is worth more than raw speed. LightGBM does keep a genuine advantage in training time, but on data of the size studied here both models finish in roughly a second, so speed never becomes the deciding factor.

These findings sit comfortably within the wider literature. Several domain studies that report LightGBM ahead of XGBoost (Chang *et al.,* 2024; Hlongwane *et al.,* 2024a; Ko *et al.,* 2022; Nguyen & Ngo, 2025) used different datasets, resampling schemes and, in some cases, separately tuned models, which is precisely the source of variability that a matched comparison removes. The present result does not contradict those studies; rather, it shows that when the two frameworks are placed under identical conditions on the same data, the small advantage in cross-validated discrimination falls to XGBoost on two datasets of very different character. This is consistent with the model-risk-adjusted finding of Alonso Robisco and Carbo Martinez (2022) and with the broader observation that the ranking of boosting libraries depends heavily on experimental choices (Bentejac *et al.,* 2021).

It is worth being precise about the computational side, since speed is LightGBM's principal selling point. On the primary dataset both frameworks completed training in roughly one second and produced predictions effectively instantaneously, so on data of this scale the speed difference, although real and in LightGBM's favour, has no practical consequence for a daily or even an hourly scoring process. The picture would change on datasets several orders of magnitude larger, where the histogram binning and gradient-based sampling of LightGBM become decisive; the present study simply shows that, at the scale most relevant to a single institution's loan book, predictive quality rather than speed is the binding constraint. A similar caution applies to generalizability: the consistent ranking across two datasets of very different size, dimensionality and imbalance is encouraging, but it remains a two-dataset result, and a fully general statement would require many more datasets drawn from different lending contexts and time periods.

Two limitations should be acknowledged. The comparison used a single matched configuration; although this is the fairest way to isolate the algorithm, it does not reveal how far each model could be pushed with exhaustive, individual tuning. The study also relied

on two historical datasets, which cannot capture shifting economic conditions or behavioural factors that were absent from the data, and on the high-signal second dataset both models are so accurate that the difference between them, while consistent in direction, is very small. These limitations temper the strength of the conclusion and point directly to the future work outlined below.

Practical Implications

For a lender choosing between the two libraries, the results invite a pragmatic reading rather than a categorical verdict. Where predictive accuracy and the stability of the risk ranking are the dominant concerns, and the data volume is in the tens of thousands rather than the hundreds of millions of records, XGBoost is the marginally safer default, because its small advantage in discrimination held on two datasets of different character and across cross-validation folds. Where models must be retrained very frequently, or trained on extremely large datasets, the speed and memory efficiency of LightGBM may well outweigh a difference in AUC of a few ten-thousandths. In either case the operating threshold should not be left at the textbook value of 0.5 but tuned to the institution's risk appetite, since the ROC analysis shows that both models support a continuum of trade-offs between approving sound applicants and rejecting risky ones. Finally, because both frameworks rest on the same interpretable drivers, identified in Section 4.6, they can be paired with feature-attribution methods such as SHAP to produce the case-level explanations that borrowers and regulators increasingly require, without any change to the underlying model. The broader lesson for practitioners is that the choice between these two libraries is rarely decisive on accuracy alone and should be made with deployment constraints, retraining frequency and explainability requirements explicitly in view.

CONCLUSION

This study compared XGBoost and LightGBM for loan approval prediction under strictly matched conditions, with the aim of providing the controlled, like-for-like evidence that the literature currently lacks. Both frameworks proved accurate and deployable, each reaching an accuracy of roughly 93% with a high AUC on the primary dataset, so neither can be dismissed for practical use. When the two were placed side by side, however, XGBoost demonstrated a small but consistent improvement: it led LightGBM on every metric on the primary dataset, retained the higher mean AUC under five-fold cross-validation, and reproduced this advantage on a second, independent loan-approval dataset, where, despite both models exceeding 98% accuracy, it still won four of five cross-validation folds. Across both dataset types, therefore, XGBoost showed marginally better and more stable discrimination, an effect compatible with its level-wise tree growth and stronger regularization. The difference is modest rather than decisive, and LightGBM remains competitive, particularly where training speed on very large data is a priority; the practical implication is that, for data of the scale examined here, XGBoost is a slightly more dependable choice for loan approval prediction while both frameworks are suitable in practice. The significance of the work lies less in crowning a winner than in showing how sensitive such comparisons are to experimental design, and in offering a reproducible protocol for making them fairly. For practitioners the message is reassuring in one respect and cautionary in another: either framework can be deployed with confidence on data of this kind, but the common assumption that the two are effectively interchangeable is not quite correct, and the small, repeatable edge in discrimination favours XGBoost when accuracy is the priority. This matters most for institutions, such as those in many developing financial systems, that are adopting machine-learning credit models for the first time and benefit from evidence-based guidance on which framework to standardize on.

Several directions would strengthen and extend the comparison. Tuning each framework independently, for example with Bayesian optimization, would test whether the advantage survives full, model-specific tuning. Extending the evaluation to larger and locally sourced banking datasets would weigh the speed advantage of LightGBM against the accuracy advantage of XGBoost under realistic conditions. Adding CatBoost and stacked or hybrid ensembles, incorporating probability calibration and cost-sensitive evaluation, and pairing the models with explainable-AI tools such as SHAP would allow the frameworks to be compared on transparency and on the economic cost of errors as well as on raw accuracy.

REFERENCES

Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., & Ogunleye, B. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. *Mathematics, 12*(21), Article 3423. https://doi.org/10.3390/math12213423

Alonso Robisco, A., & Carbo Martinez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation, 8*(1), Article 70. https://doi.org/10.1186/s40854-022-00366-1

Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. *Heliyon, 10*(21), Article e39770. https://doi.org/10.1016/j.heliyon.2024.e39770

Anghel, A., Papandreou, N., Parnell, T., De Palma, A., & Pozidis, H. (2018). *Benchmarking and optimization of gradient boosting decision tree algorithms*. arXiv. https://arxiv.org/abs/1809.04559

Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in Europe with machine learning. *Journal of Financial Econometrics, 21*(2), 569–596.

Bentejac, C., Csorgo, A., & Martinez-Munoz, G. (2021). A comparative analysis of gradient boosting algorithms.

Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. *Risks, 12*(11), Article 174. https://doi.org/10.3390/risks12110174

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Emmanuel, I., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data, 11*(1), Article 23. https://doi.org/10.1186/s40537-024-00882-0

Gu, Z., Lv, J., Wu, B., Hu, Z., & Yu, X. (2024). Credit risk assessment of small and micro enterprise based on machine learning. *Heliyon, 10*(5), Article e27096. https://doi.org/10.1016/j.heliyon.2024.e27096

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*(1), Article 94. https://doi.org/10.1186/s40537-020-00369-8

Haque, F. M. A., & Hassan, M. M. (2024). Bank loan prediction using machine learning techniques. *American Journal of Industrial and Business Management, 14*(12), 1690–1711. https://www.scirp.org/pdf/ajibm20241412_72123507.pdf

Hlongwane, R., Ramaboa, K. K. M., & Mongwe, W. (2024a). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE, 19*(5), Article e0303566. https://doi.org/10.1371/journal.pone.0303566

Hlongwane, R., Ramaboa, K. K. M., & Mongwe, W. (2024b). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLOS ONE, 19*(8), Article e0308718. https://doi.org/10.1371/journal.pone.0308718

Hossain, M. M., Mamun, M., Munir, A., Rahman, M. M., & Chowdhury, S. H. (2025). A secure bank loan prediction system by bridging differential privacy and explainable machine learning. *Electronics, 14*(8), Article 1691. https://doi.org/10.3390/electronics14081691

Hou, G., Tong, D. L., Liew, S. Y., & Choo, P. Y. (2025). Comparative analysis of resampling techniques for class imbalance in financial distress prediction using XGBoost. *Mathematics, 13*(13), Article 2186. https://doi.org/10.3390/math13132186

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154).

Ko, P.-C., Lin, P.-C., Do, H.-T., & Huang, Y.-F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy, 24*(6), Article 801. https://doi.org/10.3390/e24060801

Kunwar, S., & Ghimire, B. R. (2026). Leakage-free early prediction of issue resolution time in agile software projects: A comparative study on dataset quality. *American Journal of Data Science and Artificial Intelligence, 2*(1), 58–68. https://doi.org/10.54536/ajdsai.v2i1.7792

Lawal, O., Okolie, A., Obunadike, C., Akwabeng, P. M., Ikhifa, M. O., & Alumona, P. (2026). Predicting food insecurity across U.S. census tracts: A machine learning analysis using the USDA Food Access Research Atlas. *American Journal of Data Science and Artificial Intelligence, 2*(1), 1–14. https://doi.org/10.54536/ajdsai.v2i1.6285

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science, 8,* 180–201. https://doi.org/10.1016/j.jfds.2022.07.002

Miljkovic, T., & Wang, P. (2025). A dimension reduction assisted credit scoring method for big data with categorical features. *Financial Innovation, 11*(1), Article 29. https://doi.org/10.1186/s40854-024-00689-1

Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. *Cogent Economics & Finance, 13*(1), Article 2465971. https://doi.org/10.1080/23322039.2025.2465971

Ozkurt, C. (2024). Enhancing financial decision-making: Predictive modeling for personal loan eligibility with gradient boosting, XGBoost, and AdaBoost. *Information Technology in Economics and Business, 1*(1), 7–13. https://doi.org/10.69882/adba.iteb.2024072

Yang, C. (2024). Research on loan approval and credit risk based on the comparison of machine learning models. *SHS Web of Conferences, 181,* Article 02003. https://doi.org/10.1051/shsconf/202418102003

Ying, C., Shi, A., & Li, X. (2025). Hybrid boosted attention-based LightGBM framework for enhanced credit risk assessment in digital finance. *Humanities and Social Sciences Communications, 12*(1), Article 1036. https://doi.org/10.1057/s41599-025-05230-y

Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X., & Song, J. (2024). *Ensemble methodology: Innovations in credit default prediction using LightGBM, XGBoost, and LocalEnsemble* [Preprint]. arXiv. https://arxiv.org/abs/2402.17979

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management, 6*(3), 123–133. https://doi.org/10.1016/j.dsm.2023.04.003