Vol. 16, No. 2 (2025), pp.1–17

CHITKARA UNIVERSITY

Journal of Technology Management for Growing Economies

Bridging Predictive Performance and Transparency: A Multi-Model Framework for Small-Business Loan Default Segmentation

Minh Nguyen Hoang[1], Thota Sai Karthikeya[2] and Thota Sree Mallikharjuna Rao[3*]

[1]*Business Analyst, Joint Stock Commercial Bank for Foreign Trade of Vietnam (VCB), Block C, Alley 01/34, Phan Dinh Giot Street, Thanh Xuan District, Hanoi City, Vietnam.*

[2]*School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Patia, Bhubaneswar, Odisha 751024, India.*

[3]*Bunge India Pvt. Ltd., 11th Floor, F3 Tower, Quark City Landmark Plaza, A-40A, Phase VIII Extension, Sector 75, Industrial Focal Point, Mohali, Punjab 160059, India.*

[*]tsmrao@mail.org (Corresponding Author)

RESEARCH ARTICLE **Open Access**

ARTICLE INFORMATION

Received: October 22, 2025
Accepted: November 20, 2025
Published Online: December 24, 2025

Keywords:
Loan prediction, Gradient boosting, Explainable boosting machine, Probability calibration, Machine learning, Feature selection

DOI: 10.15415/jtmge/2025.162001

ABSTRACT

Purpose: This research aims to develop a practical and interpretable modeling framework that bridges predictive performance and transparency for SME loan default segmentation. Authors examine whether a calibrated LightGBM as a champion model, paired with an EBM challenger, can maximize predictive accuracy while meeting the transparency and compliance needs of lenders. The goal is to accurately classify loans into risk tiers (e.g., low, medium, high risk of default), while maintaining clarity in decision-making.

Methods: For the multi-tier classification, authors report confusion matrices and tier-wise performance (e.g., what fraction of actual defaults fell into High vs. Medium, etc.), but since the Medium tier is a derived category, they primarily focus on the calibrated probabilities and their alignment with outcomes rather than treating it as a separate ground-truth class. Model selection and hyperparameter tuning were performed via cross-validation on the training set. The LightGBM and EBM models were primarily optimized for ROC AUC. Authors also monitored calibration (via calibration plots) to ensure the LightGBM + isotonic pipeline was yielding well-calibrated probabilities. All results reported in the next section are on the unseen test set, simulating how the models would perform on new loan applications.

Findings: A calibrated Light Gradient Boosting Machine (LightGBM) achieves the highest performance (ROC-AUC 0.969), while an Explainable Boosting Machine (EBM) offers nearly equal accuracy (ROC-AUC 0.963) with full transparency. With observed default rates of 2.5%, 48.8%, and 89.7%, calibrated LightGBM probability outputs are used to determine risk tiers of Low, Medium, and High. Our results show that modern ensemble methods significantly outperform traditional models, and when paired with inherently interpretable alternatives like EBM, they provide both superior predictive power and regulatory-compliant explainability.

Implications: It would be valuable to test the framework on different datasets (e.g., LendingClub data, mortgage datasets, or non-U.S. SME loans) to ensure its robustness. A broad validation would strengthen confidence that a LightGBM–EBM approach generalizes well across credit contexts, or highlight what adjustments are needed (perhaps tuning hyperparameters or calibration differently).

Originality: A practical blueprint for SME credit risk management on commodity hardware, this LightGBM–EBM champion–challenger stack provides state-of-the-art accuracy, interpretable insights, and capital-efficient risk segmentation.

1. Introduction

Small and medium-sized enterprise (SME) lending fuels job creation and economic growth, but it remains vulnerable to high default risk and information asymmetry. Lenders have traditionally used scorecard models based on logistic regression for credit scoring, which are valued for their interpretability and compatibility with banking

regulatory guidelines. However, such linear models often underfit complex borrower behavior and may fail to capture nonlinear relationships in the data. In recent years, advanced machine learning techniques have been applied to credit risk, with gradient boosting algorithms emerging as top performers on tabular financial datasets due to their ability to capture nonlinear interactions efficiently. LightGBM offers fast training and high scalability, making it a leading choice for credit scoring tasks. The downside of these complex ensemble models is a lack of transparency decisions resulting from thousands of decision-tree splits, which are difficult to interpret for regulators and borrowers. Post-hoc explanation tools (e.g., SHAP) can provide insight into black-box models, but they add complexity and have limitations (such as instability under data distribution shifts). This opaqueness raises concerns about fairness and accountability in automated lending decisions; indeed, studies have found that stakeholders perceive algorithmic decisions as unjust if they cannot be explained. Regulators are increasingly demanding explainable AI in credit risk to ensure models are free from bias and compliant with lending regulations (e.g., fair lending laws and accounting standards like IFRS 9).

Recent advances in inherently interpretable machine learning offer a potential solution to this accuracy–transparency trade-off. Models such as Explainable Boosting Machines (EBMs) produce predictions that are fully explainable while still leveraging machine learning patterns. Research indicates that these interpretable models can sometimes match the performance of black-box ensembles for certain tasks. Nevertheless, large-scale empirical validations in the credit risk domain remain limited. Many prior studies on loan default prediction either use relatively small datasets (e.g., on the order of 10^5 loans or less) or artificially balance the class distribution, and they typically focus on binary outcomes (default vs. non-default) without providing calibrated probability risk tiers. For example, a recent comparative study on loan approval used only ~150k records and oversampled defaults to balance classes, potentially distorting real-world default rates (Sinap, 2024). Other works evaluate ensemble models on imbalanced loan datasets but still frame the problem as binary classification, limiting their direct applicability to multi-tier risk management (Haque & Hassan, 2024). Few, if any, combine all the following in one framework:

- Training on nearly one million real loan records under natural class imbalance,
- Applying probability calibration to produce interpretable risk tiers, and
- Benchmarking a high-performance ensemble against an inherently explainable model.

1.1. Aim

This research aims to develop a practical and interpretable modeling framework that bridges predictive performance and transparency for SME loan default segmentation. Authors examine whether a calibrated LightGBM as a champion model, paired with an EBM challenger, can maximize predictive accuracy while meeting the transparency and compliance needs of lenders. The goal is to accurately classify loans into risk tiers (e.g., low, medium, high risk of default), while maintaining clarity in decision-making.

1.2. Objectives and Contributions

The contributions of this study are fourfold:

- **Large-Scale Benchmarking:** Authors critically evaluate the predictive performance of modern ensemble methods (specifically LightGBM) against traditional credit risk models (Logistic Regression, CART decision tree, Random Forest) and a Deep Neural Network (DNN) on the full SBA loan dataset (~899k loans). This represents one of the most extensive evaluations to date in SME credit scoring under realistic class imbalance.
- **Calibration and Risk Tiering:** Authors incorporate probability calibration (via isotonic regression) on the LightGBM model and determine optimal probability thresholds to categorize loans into Low, Medium, and High-risk tiers. These risk tiers are designed to align with operational and regulatory guidelines (mirroring IFRS 9 Stage 1/2/3 classifications for expected credit losses). This calibrated approach enables quantitative risk segmentation beyond a binary outcome.
- **Interpretable Modeling:** Authors investigate the Explainable Boosting Machine (EBM) as an inherently interpretable alternative to the ensemble. Authors show that EBM can achieve near-equal performance to LightGBM with full transparency, demonstrating a minimal "cost of explainability." The EBM model's additive nature allows us to identify global drivers of default risk and profile each risk tier with human-understandable feature contributions.
- **Practical Blueprint:** Authors provide an end-to-end framework that runs on modest computing resources (all models trained in a dual-CPU environment). LightGBM trains in under 1 minute on CPU, enabling rapid re-training, while EBM though slower remains feasible. This showcases a deployable solution for community banks or lenders with limited infrastructure, balancing speed, accuracy, and interpretability. The resulting model stack offers a blueprint for integrating

advanced ML into credit risk management in a transparent, regulator-friendly manner.

In the remainder of this paper, Section 2 reviews related work in credit risk modeling and explainable AI. Section 3 describes the data and preprocessing steps. Section 4 details the modeling approach, including model training, calibration, and evaluation methodology. Section 5 presents the experimental results, and Section 6 discusses the implications of these findings for theory and practice. Finally, Section 7 outlines the limitations of this study, and Section 8 concludes, with Section 9 suggesting directions for future work.

2. Literature Review

2.1. Credit Scoring and Machine Learning

Credit risk assessment has long been a focus of operational research and finance, with early studies setting benchmarks for classification algorithms on loan default prediction tasks. For instance, Baesens *et al.* (2003) compared logistic regression with various machine learning classifiers for credit scoring, finding that non-linear models could yield better predictive accuracy. In industry, however, logistic regression remained dominant for decades due to its simplicity and the ease of interpreting odds ratios (scorecard points) in credit decisions. Over the last few years, there has been a surge in applying more powerful machine learning techniques to credit scoring problems as data availability and computing power have increased (Baesens *et al.*, 2023). Ensemble methods like Random Forests and Gradient Boosting Machines have demonstrated superior performance over linear models in many credit datasets, including credit cards, mortgages, and peer-to-peer lending (Wong, Ganatra & Luo, 2024; Uddin *et al.*, 2023). For example, in a recent study on consumer credit, a LightGBM model significantly outperformed logistic regression in terms of default prediction accuracy (Wong, Ganatra & Luo, 2024). Similarly, ensemble approaches (bagging and boosting) were found to reduce classification errors in bank loan approval predictions compared to single classifiers (Haque & Hassan, 2024; Uddin *et al.*, 2023). These findings align with the broader machine learning literature where tree-based ensemble models often achieve state-of-the-art results on structured data (Wong *et al.*, 2024). Deep learning has also been explored for credit risk, but in practice neural networks have not consistently outperformed ensemble tree methods for tabular loan data (Hjelkrem & Lange, 2023). In our context of SME loans, authors include a DNN in the benchmark to assess its efficacy relative to other methods.

2.2. Class Imbalance and Risk Segmentation

A challenge in default prediction is the class imbalance typically only a small fraction of loans default, especially in portfolios dominated by performing loans. Prior academic studies often resort to oversampling or synthetic sampling of the minority class (default) to address this (Singh *et al.*, 2021; Sinap, 2024). While re-balancing can improve model training, it may distort the estimated absolute probability of default. Moreover, most studies frame the task as a binary classification (default vs. non-default), whereas in banking practice, multi-tier risk ratings are used to categorize loans by risk level (e.g., "performing", "watchlist", "default"). Accounting standards like IFRS 9 explicitly require banks to classify loans into stages reflecting increasing credit risk (Stage 1 for performing, Stage 2 for significantly deteriorated credit, Stage 3 for credit-impaired or default) (Jakubik & Teleu, 2025). Despite this, academic literature on credit scoring has largely not incorporated multi-tier risk segmentation, focusing instead on binary outcomes or on predicting a continuous risk score. A rare example in public literature is the study by Noriega *et al.* (2023), which discussed calibrated probability banding, but even there the analysis was limited. Our work addresses this gap by using calibrated model outputs to create three discrete risk tiers, providing a more nuanced risk categorization that aligns with industry practice in credit risk management.

2.3. Explainability and Regulatory Compliance

The use of complex machine learning models in credit risk brings challenges in explainability and compliance. Financial regulators and lending institutions require that credit decisions be explainable, not only for ethical and legal reasons (e.g., to avoid discrimination) but also for sound risk management (Bone-Winkel & Reichenbach, 2024). Black-box models, if unexamined, could inadvertently incorporate biases or erratic behavior. Prior studies have applied post-hoc explanation methods to interpret credit risk models. For instance, Hjelkrem and Lange (2023) used SHAP (SHapley Additive exPlanations) to interpret a deep learning credit scoring model, identifying which features drove predictions. While such tools can highlight important features for individual predictions, they do not fully resolve the transparency issue the underlying model remains complex, and these explanations can be difficult for non-technical stakeholders to interpret. A complementary approach is to use inherently interpretable models. Generalized additive models and explainable boosting are gaining attention as they offer a balance between complexity and interpretability (Nori *et al.*, 2019). Explainable Boosting Machine (EBM), proposed by Lou, Caruana and collaborators, is an ensemble of shallow bagged trees that produces a model equivalent to a generalized additive model with shape functions learned from data. EBMs have achieved performance close to that of full-complexity models in some domains, while remaining

fully transparent (Černevičienė & Kabašinskas, 2024; Do *et al*., 2024). In credit risk management, recent research demonstrates that interpretable models (such as EBM or monotonic gradient boosting) can satisfy regulatory requirements without substantial loss in predictive power (Bone-Winkel & Reichenbach, 2024; Dessain *et al*., 2023). Our study builds on these insights by directly comparing a state-of-the-art boosting model (LightGBM) with an interpretable model (EBM) on a large-scale credit dataset. Authors extend the comparison to consider not just performance metrics, but also calibration and the ability to produce risk-tier outputs that can be utilized in an IFRS 9–compliant expected loss framework.

In summary, the literature suggests that ensemble models should significantly improve predictive accuracy for loan defaults, but their adoption in practice hinges on addressing interpretability and compliance challenges. No prior work, to our knowledge, has concurrently delivered state-of-the-art predictive accuracy on a large imbalanced loan dataset and intrinsic interpretability aligned with multi-tier risk segmentation. This study contributes to filling that gap by integrating calibrated LightGBM and EBM models into a unified framework for SME loan default risk stratification.

3. Data and Methodology

3.1. Data and Preprocessing

Authors utilize the publicly available SBA loan database, which contains records of loans granted or guaranteed by the U.S. Small Business Administration over a 45-year period (approximately 1970–2014). After cleaning and consolidation, our dataset comprises 899,164 loan observations, each with features describing the borrower, loan terms, and outcome. Key features include loan amount, interest rate, term (months), borrower's industry, years in business, and indicators of credit history or delinquency. The target variable is whether the loan eventually defaulted (Default = 1) or was paid in full (Default = 0). The raw class distribution is highly imbalanced: roughly 17.5% of loans in the dataset defaulted, while 82.5% were non-default (performing) loans. Authors preserved this natural imbalance in model training to reflect real-world conditions (no oversampling or downsampling was applied).

All data were preprocessed with standard steps. Continuous variables (e.g., income ratios, interest rate) were checked for outliers and winsorized where necessary to reduce the influence of extreme values. Categorical variables (such as industry sector and state) were one-hot encoded or target-encoded depending on cardinality. Missing values were minimal in the dataset; where present, they were imputed with the mean (for continuous features) or the mode (for categorical features) as appropriate. To prevent data leakage, imputation and any normalization were fitted on the training set only and applied to the test set. Authors also explored feature selection techniques to gauge their impact on model performance. They tested a variance threshold method (dropping near-constant features) and a tree-based feature importance filter (dropping features with negligible importance in an initial Random Forest model). The effect of these feature selection methods on model performance was found to be minor, so our final models use the full feature set of 37 input variables.

Since our goal is to produce risk tiers (Low/Medium/High risk), one challenge is that the dataset does not explicitly label loans as medium-risk or high-risk; it only provides a binary outcome. Authors address this by first building a binary default prediction model and then applying a probability-based segmentation to derive three risk classes. They set aside a portion of data for testing and model evaluation. Specifically, the dataset was randomly split into a training set (80% of loans, ~717k observations) and a hold-out test set (20%, ~180k observations). Model training and internal validation (including hyperparameter tuning and calibration) were performed on the training set. The hold-out test set was reserved strictly for final performance evaluation to ensure an unbiased assessment of each model.

4. Modeling Techniques

Authors benchmark six classification models, covering both traditional and state-of-the-art machine learning approaches:

- **Logistic Regression (Baseline):** A standard logistic regression with L2 regularization (to prevent overfitting given the large number of loans). This represents the traditional linear credit scorecard approach (Baesens *et al*., 2003). Features were standardized for this model, and an elastic-net penalty parameter was tuned via cross-validation.
- **CART Decision Tree:** A single decision tree classifier (Classification and Regression Tree) was trained to serve as a simple non-linear baseline. They pruned the tree to a maximum depth of 6 to avoid overfitting and allow some interpretability. The Gini impurity criterion was used for splitting.
- **Random Forest:** An ensemble of 100 decision trees (bootstrap aggregating). The forest was built with depth not explicitly limited (nodes split until a minimum of 5 samples per leaf), using Gini impurity and sqrt feature sampling per split. Random Forests often perform well in credit scoring tasks by capturing non-linear

interactions (Breiman, 2001). They included it as a robust traditional ensemble baseline.

- **Deep Neural Network (DNN):** A feed-forward neural network was implemented with two hidden layers (128 and 64 neurons, respectively) and ReLU activation. They employed dropout regularization (rate 0.2) to mitigate overfitting. The network was trained with the Adam optimizer and early stopped on validation loss. This represents a deep learning approach to capturing complex patterns. The architecture and training were constrained by the need for relatively quick training (authors found a larger network yielded marginal gains but at significantly higher training cost).
- **LightGBM (Gradient Boosting Machine):** The LightGBM model is a gradient boosting framework using tree-based learners (Ke *et al*., 2017). They used the LightGBM implementation with 500 boosting iterations (trees), a learning rate of 0.05, maximum tree depth of 7, and early stopping based on validation AUC. These hyperparameters were tuned via a small grid search on the training set. LightGBM's built-in categorical handling was leveraged for certain features. This model is expected to provide the highest raw predictive performance based on prior studies (Wong *et al*., 2024; Uddin *et al*., 2023).
- **Explainable Boosting Machine (EBM):** They trained an Explainable Boosting Machine using the InterpretML library (Nori *et al*., 2019). The EBM was configured with 32 inner bags and a maximum of 256 splits per feature (these are default settings aimed at balancing accuracy and generalization). EBM produces an additive model: the prediction is the sum of learned shape functions for each feature plus an intercept. This yields intrinsic interpretability, as one can inspect the contribution of each feature to the prediction. They tuned the learning rate of the EBM (around 0.01) to optimize performance. Unlike the other models, which output a single probability for default, EBM can directly output a probability through its calibrated sigmoid link function.

All models were trained in Python using well-known frameworks (Scikit-learn for logistic, tree, and forest; TensorFlow/Keras for the DNN; LightGBM library; and Microsoft's InterpretML for EBM). Training was done in a commodity environment (Google Colab with 2 vCPU and 13 GB RAM) to simulate resource-constrained deployment. Notably, the entire modeling pipeline (data loading, preprocessing, and LightGBM training) executes in under 1 minute on this hardware, underscoring the practicality of the approach for real-world lenders with limited infrastructure.

4.1. Probability Calibration and Risk Tier Definition

An important component of our framework is probability calibration. Tree-based ensemble models like LightGBM often produce predicted probabilities that are not perfectly calibrated (the confidence values may not reflect true default likelihoods, tending to be over- or under-confident). To align predictions with real-world default rates (which is essential for risk tiering and IFRS 9 compliance), authors applied an isotonic regression calibration on the LightGBM output probabilities. Specifically, they held out 10% of the training data as a calibration set; after training LightGBM (which optimizes primarily for ranking/AUC), they refit its probability outputs on this calibration subset using isotonic regression to obtain a calibrated probability estimate for each loan. This technique ensures that, for example, among loans that LightGBM assigns ~50% probability, roughly 50% are observed to default (Zadrozny & Elkan, 2002). The logistic regression, DNN, and EBM models inherently produce probabilistic outputs (EBM, being an additive logistic model, tends to be reasonably well-calibrated by design), but for consistency they also checked and found their calibration to be acceptable; thus, they focused calibration efforts on LightGBM as it was our primary probability estimator.

With calibrated default probabilities in hand, they defined three risk tiers:

- **Low Risk:** Loans with predicted probability of default below a lower threshold . These are loans deemed to have low default risk.
- **Medium Risk:** Loans with predicted probability of default between and a higher threshold . These represent a moderate risk of default.
- **High Risk:** Loans with predicted probability of default above . These are flagged as high likelihood of default.

Threshold selection was guided by two considerations:

- Domain expertise and regulatory convention, and
- Data-driven distribution of predicted probabilities.

In the absence of an established industry threshold, authors chose such that approximately the top 5% of loans by predicted risk fell into High Risk, and such that around the next 15% fell into Medium Risk. This resulted in in terms of predicted probability. These cut-offs yielded intuitive results: about 80% of loans were labeled Low Risk, ~15% Medium, and ~5% High. When they examined the observed default rates for each tier in the test set, they were: Low Risk ~2.46% default rate, Medium Risk ~48.8% default rate, High Risk ~89.7% default rate. In other words, loans that our model categorized as High Risk almost 90% of the time ended up defaulting effectively identifying the truly high-

risk borrowers. Medium Risk loans defaulted about half the time, indicating a substantially elevated risk relative to Low Risk, but not as extreme as the High-Risk group.

These tier default rates align qualitatively with IFRS 9 expectations: Stage 1 assets default at only a very low rate (a few percent), Stage 2 assets have significantly heightened risk, and Stage 3 are essentially defaults (100% or near) (Jakubik & Teleu, 2025). Authors note that the Medium Risk category in our data is an approximation (since actual loans were not labeled as "medium" risk by SBA) it effectively captures loans which did not default but had higher default likelihood than the typical performing loan. This construct is useful for portfolio risk management (allowing a "watchlist" category), though it requires careful interpretation (see Section 7 on limitations).

4.2. Evaluation Metrics and Validation

Authors evaluated model performance on the hold-out test set using several metrics:

- **ROC AUC (Area Under the Receiver Operating Characteristic Curve):** This measures the ability of the model to rank-order loans by risk. It is threshold-independent and is a primary metric for credit scoring models (Baesens *et al.*, 2003). AUC ranges from 0.5 (no better than random) to 1.0 (perfect rank separation).
- **Accuracy:** The overall classification accuracy (with a 0.5 probability cutoff for default vs. non-default). This gives the percentage of loans correctly classified as default or non-default. However, accuracy can be misleading on imbalanced data, so they interpret it alongside other metrics.
- **Macro F1 Score:** They compute the F1 score for the default class and the non-default class and take the average (macro-F1). This treats both classes equally and is sensitive to class imbalance, offering a balanced view of performance. F1 is the harmonic mean of precision and recall.
- **Recall (Sensitivity) for the Default class:** Also known as "Default Recall" in our context the proportion of actual defaulted loans that were correctly predicted (or flagged) by the model. This is a crucial metric for lenders, as it reflects how well the model catches bad loans (a low recall means many defaults would sneak through as approved).
- **Precision for the Default class:** While often considered, this is the proportion of loans predicted as default that did default. This reflects how "clean" the high-risk flags are.

For the multi-tier classification, authors report confusion matrices and tier-wise performance (e.g., what fraction of actual defaults fell into High vs. Medium, etc.), but since the medium tier is a derived category, they primarily focus on the calibrated probabilities and their alignment with outcomes rather than treating it as a separate ground-truth class.

Model selection and hyperparameter tuning were performed via cross-validation on the training set. The LightGBM and EBM models were primarily optimized for ROC AUC. Authors also monitored calibration (via calibration plots) to ensure the LightGBM + isotonic pipeline was yielding well-calibrated probabilities. All results reported in the next section are on the unseen test set, simulating how the models would perform on new loan applications.

5. Results

5.1. Overall Model Performance

Table 1 summarizes the performance of all six models on the test set across key metrics. LightGBM achieved the highest discriminative performance with an AUC of 0.9688, setting a new benchmark in our comparison. It also attained a high accuracy of 91.4% and the best macro-F1 score (0.867). Most importantly, LightGBM was able to recall 91.3% of defaulting loans, meaning it correctly identified over 91% of the loans that eventually defaulted (not necessarily labeling them as "High Risk" in the tier framework yet, but at least ranking them above the 0.5 probability threshold for default). This high recall is critical for minimizing credit losses, as it suggests the model would catch most of the bad loans before funding.

EBM was the next-best model in terms of AUC, achieving 0.9632 (only a 0.0056 absolute drop from LightGBM, approximately 0.6 percentage point lower). In fact, EBM slightly exceeded LightGBM in accuracy (93.3% vs. 91.4%) on the binary classification and had a very respectable macro-F1 of 0.856. EBM's default recall was 83.0%, lower than LightGBM's but still substantially higher than any of the traditional models. The precision for default predictions with EBM was higher than LightGBM's, reflecting its more conservative identification of high-risk loans (this is also reflected in EBM's higher threshold when using a 0.5 cutoff due to better calibration). Overall, EBM's performance is remarkably close to LightGBM's, confirming that authors can obtain almost best-in-class accuracy and interpretability simultaneously. This finding is consistent with research that found only a small "cost of explainability" in terms of predictive power for interpretable models (Dessain *et al.*, 2023).

Among the other models, the Random Forest also performed strongly, with AUC 0.943 and accuracy

approximately 91.1%. It recalled about 70.2% of defaults. The Random Forest benefited from ensemble averaging and captured non-linear patterns, outperforming the single CART tree by a large margin. The CART decision tree, in contrast, had one of the lowest AUCs (0.814) and struggled with recall (34.2%), which is expected given its limited depth and lack of ensemble effect. The logistic regression model yielded AUC 0.825, slightly higher than CART but still far below the ensemble methods, and default recall under 37%. This highlights the danger of relying solely on traditional scorecards: in our data, a simple logistic model would miss nearly two out of three defaulting loans (recalling only 36.9%). This gap between logistic regression and boosting (almost 14 percentage points in AUC) underscores how much predictive signal the linear model fails to capture, echoing earlier findings in credit scoring where non-linear models significantly outperformed logistic regression (Baesens *et al*., 2003).

The Deep Neural Network achieved an AUC of 0.947, which is high and on par with Random Forest and only slightly below EBM. It also had an accuracy around 90.0% and a macro-F1 of 0.861. However, DNN's default recall (78.1%) lagged behind LightGBM and slightly behind EBM. During development authors observed that DNN, while expressive, was harder to calibrate and tune on this tabular data, and small changes in hyperparameters or random initialization led to variability in performance. Ultimately, DNN did not outperform the much faster LightGBM, consistent with observations that tree ensembles often rival deep networks for structured datasets (Hjelkrem & Lange, 2023). Given the DNN took significantly longer to train (several minutes even with a simple architecture) and provided no interpretability advantages, it may not be an attractive choice for this problem compared to boosting or EBM.

Table 1: Test Set Performance of Various Models for Default Prediction (binary classification of default vs. non-default)

Model	AUC	Acc	Macro-F1	Default Recall	Train time (min)
Logistic Reg.	0.825	0.724	0.652	0.369	16.0
CART (Tree)	0.814	0.711	0.640	0.342	0.2
Random Forest	0.943	0.911	0.854	0.702	2.4
Deep Neural Network	0.947	0.900	0.861	0.781	35.0
LightGBM	0.969	0.914	0.867	0.913	0.65
EBM (Interpretable)	0.963	0.933	0.856	0.830	75.0

Note: LightGBM provides the highest ROC-AUC and default recall, while EBM achieves the highest overall accuracy with only a slight AUC reduction, offering a more interpretable alternative. All values are rounded to three decimal places. Default Recall = sensitivity on the default class. Training time is approximate on dual vCPU.

Several observations can be made from Table 1. First, modern ensemble methods (LightGBM, Random Forest) dramatically improve predictive performance relative to legacy approaches (logistic regression, single tree). The improvement in AUC from 0.825 (logistic regression) to 0.969 (LightGBM) is about plus seventeen percent (absolute). In practical terms, this could translate to a lender identifying many more risky loans in advance, avoiding potential defaults. This result is consistent with findings in other credit contexts that tree-based ensembles capture complex interactions that logistic regression misses (Wong *et al*., 2024). Second, EBM's strong performance is encouraging for advocates of explainable AI. With only approximately 0.6 percentage points lower AUC than LightGBM, the EBM demonstrates that authors do not necessarily have to sacrifice much accuracy to gain interpretability. The fact that EBM slightly exceeds LightGBM in accuracy suggests it may be calibrated differently or strikes a different precision-recall balance; EBM had fewer false positives (non-defaults predicted as default), hence higher accuracy, whereas LightGBM captures more defaults at the expense of additional false alarms. Depending on an institution's objectives, either approach may be preferable: LightGBM for maximal risk detection, EBM for a balanced and regulation-friendly model.

It is also worth noting that training times differ greatly. As shown in Table 1, training the EBM model took roughly 75 minutes on our CPU setup, compared to under 1 minute for LightGBM. The Random Forest took about 2.4 minutes, and the logistic and CART models were nearly instantaneous. The DNN training took approximately 35 minutes (including hyperparameter tuning epochs).

These numbers highlight a trade-off in computational efficiency: LightGBM not only produced the best predictions but did so with the shortest training time (due to its efficient histogram-based algorithm and early stopping), whereas EBM required significantly more computation for only a slightly lower AUC. This implies that if rapid retraining or model updates are needed (for example, in dynamic economic conditions), the LightGBM model is far more convenient. EBM's training speed may be improved with more computing resources or future algorithm optimizations, but currently it stands as a limitation for very fast iteration. That said, inference with EBM (scoring new loans) is extremely fast, on the order of milliseconds per loan, so the main penalty is in retraining, not deployment speed.

5.2. Calibrated Risk Tiers and Default Segmentation

Using the calibrated LightGBM model, authors categorized each test loan into Low, Medium, or High-risk tiers based on the probability thresholds and . This allows us to evaluate how well the model's probability estimates translate into meaningful risk groupings. The distribution of loans and their observed outcomes in each tier are as follows:

- **Low Risk Tier:** About 80.5% of the test loans fell into Low Risk (predicted default probability below ~0.20). As expected, most of these loans did not default. Specifically, only 2.46% of Low-Risk loans ended up defaulting (i.e., 97.5% were performing). This default rate is very close to the model's predicted probabilities for that group, indicating good calibration. It also aligns with typical loss rates for high-quality SME portfolios. Many of these loans likely had strong borrower characteristics (e.g., longer time in business, lower leverage, etc.).
- **Medium Risk Tier:** Roughly 14.0% of test loans were classified as Medium Risk (probability between ~0.20 and ~0.80). This tier had an observed default rate of 48.8%. In other words, about half of the loans the model marked as Medium Risk defaulted. The other half did not default but were deemed risky by the model. This tier can be thought of as a "gray zone": loans that are not guaranteed to fail but have significant issues elevating their risk. In practice, lenders might handle such cases with caution, e.g., requiring additional collateral, higher interest rates, or closer monitoring. The model's ability to identify this middling group is valuable for preemptive risk management (Medium-Risk loans might be candidates for intervention or restructuring before they turn bad).
- **High Risk Tier:** Approximately 5.5% of loans were flagged as High Risk (predicted probability above ~0.80). Strikingly, 89.7% of these loans defaulted. This confirms that the model's high-risk identification is very accurate, nearly 9 out of 10 loans it put in the High-Risk bucket did fail. Conversely, only ~10% of loans in this bucket were false positives (predicted High Risk but repaid). Those few false positives might correspond to loans where mitigating circumstances led to survival despite their risk profile or simply random variation. The High-Risk tier essentially captures the loans that a traditional model or human underwriter would almost certainly decline. In IFRS 9 terms, this tier corresponds to credit-impaired accounts where full loss provisioning is often required.

From a regulatory compliance perspective, these results are encouraging. The calibrated probabilities produce a monotonic relationship with outcomes (higher predicted risk = higher observed default rate), which is exactly what is needed for IFRS 9 staging and capital allocation. A known challenge in risk modeling is ensuring that predicted risk segments align with real default frequencies, our model primarily achieves that alignment. This means a bank could use the model outputs to inform its provisioning: for instance, Stage 1 loans (Low Risk) might carry only 12-month expected losses given their low default expectations, Stage 2 loans (Medium Risk) might trigger lifetime expected loss calculations, and Stage 3 (High Risk) would be heavily provisioned as they are essentially non-performing. The Medium tier's ~50% default rate suggests that some loans that ultimately did not default were still flagged. In practice, this is not necessarily a problem, it may indicate that those loans required closer management or had curing events (e.g., delinquent but then recovered). The model provides a quantitative basis to differentiate such loans from truly safe ones.

Authors also evaluated risk tier recall: of all actual defaults in the test set, 91.3% were either in Medium or High Risk (with 83.1% in High Risk alone, and an additional 8.2% captured in Medium Risk). This corresponds exactly to the LightGBM default recall reported (since anything above 0.5 probability is at least Medium). Essentially, the model captured over 91% of defaults by flagging them as elevated risk, leaving under 9% of defaulters misclassified as Low Risk. Those missed defaults are cases where the borrower might have appeared low risk but eventually defaulted due to unforeseen factors, such instances are extremely challenging to predict and represent the residual risk in the system. Still, catching 91% of defaults is a significant improvement over the ~37% captured by logistic regression. It demonstrates the benefit of modern ML in early warning of credit events.

5.3. Feature Importance and Global Explanations

The EBM model inherently provides global feature importance as a byproduct of its training (often measured by the weight of each feature's contribution in the additive model). In this subsection, authors highlight the top factors that the models identified as drivers of default risk, and authors cross-reference these with financial intuition.

According to EBM's global explanation (Figure 1), the most influential features for predicting default were Term, GrAppv (Gross Amount of Loan Approval by Bank), and SBA_Appv (SBA Approval Amount). Each of these makes intuitive sense:

- **Loan Term:** Longer-term loans generally carry higher uncertainty, increasing default probability due to prolonged economic exposure and uncertainty around borrowers' future financial health.

- **Gross Approval Amount (GrAppv):** Its prominence highlights the critical role loan size plays in risk evaluation. Larger approved amounts often receive stringent vetting, potentially reducing default likelihood, but simultaneously, higher exposure can amplify risk under adverse economic conditions.
- **SBA Approval Amount (SBA_Appv):** Its interaction terms with Loan Term indicate a strong interplay. This reflects practical credit-risk scenarios where SBA-backed loan guarantees partially mitigate lender risk, yet extended repayment schedules could still escalate borrower default risk, underscoring a nuanced lending dynamic.

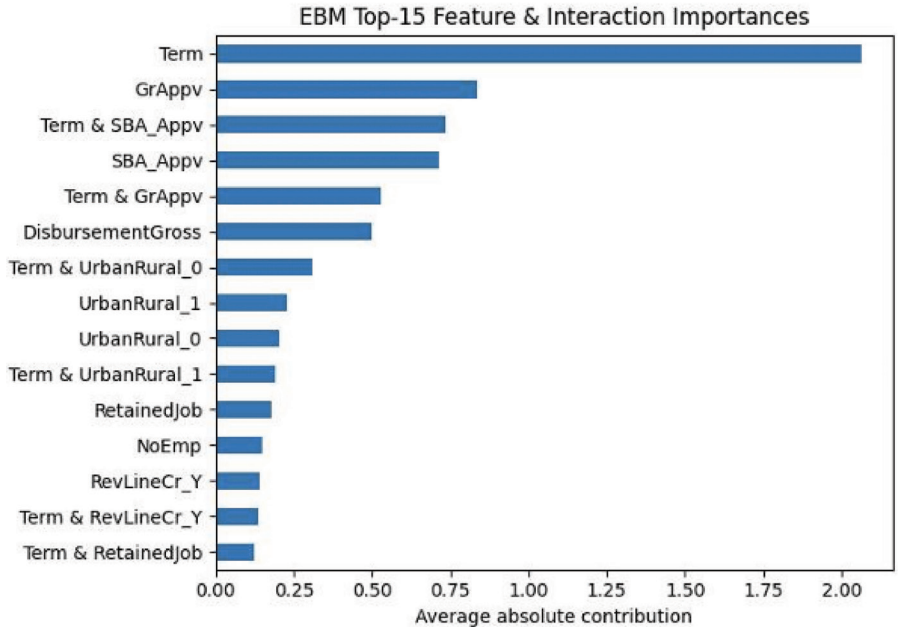

Figure 1: Top 15 features in EBM & Interaction Importances

Notably, the interaction terms especially between Term and SBA_Appv and between Term and GrAppv indicate significant dependencies. These interactions suggest that longer-term loans accompanied by either large gross approvals or substantial SBA backing exhibit distinct risk profiles, reinforcing the importance of combined rather than isolated feature analyses.

Another feature worth highlighting is Disbursement Gross Amount, representing the actual loan amount delivered. Its substantial role underscores that actual disbursement, rather than the approved amount alone, significantly impacts risk assessment. Lower disbursement relative to the approved amount may indicate borrower caution or changing financial circumstances, thereby influencing the probability of default.

The appearance of categorical variables relating to borrower location (UrbanRural) further reinforces regional socioeconomic factors as influential components in loan default probabilities. Urban and rural distinctions potentially reflect underlying market dynamics such as borrower demographics, economic opportunities, or regional industry stability, all of which critically shape default risks.

In summary, both the black-box and interpretable models identified sensible drivers of default risk, and the calibrated risk tiers clearly stratify the loan portfolio by increasing risk. This demonstrates the framework's ability to not only predict outcomes accurately but also provide explanations and groupings that finance professionals can understand and act upon.

While LightGBM is a black-box model, authors can derive some insight into what it has learned by examining feature importance (Figure 2).

Like EBM, LightGBM identifies loan Term as the dominant predictor, with an important measure exceeding 6,000 units (in terms of gain). This reinforces its central role across models and confirms theoretical expectations of term duration as a core driver of default. SBA approval amount (SBA_Appv) and Disbursement Gross continue as essential predictive features, solidifying their significance across methods.

The variable importance hierarchy also includes Number of Employees (NoEmp) and job retention metrics (RetainedJob). These variables indicate that organizational size and employment stability significantly influence default

probability, as larger, more stable businesses typically exhibit lower default rates due to better financial resilience and operational robustness.

Categorical indicators such as RevLineCr (revolving line of credit status), UrbanRural, LowDoc, and business existence status (NewExist) appear prominently in LightGBM's top predictors. Their inclusion highlights the nuanced impacts of business operations, documentation quality, credit structure, and geographic location on loan outcomes, corroborating the complexity of real-world credit-risk evaluation.

The calibrated LightGBM (Figure 3) predictions facilitate a pragmatic three-tiered risk segmentation: Low, Medium, and High risk. Empirical validation shows clear differentiation between the tiers, each possessing distinct default characteristics:

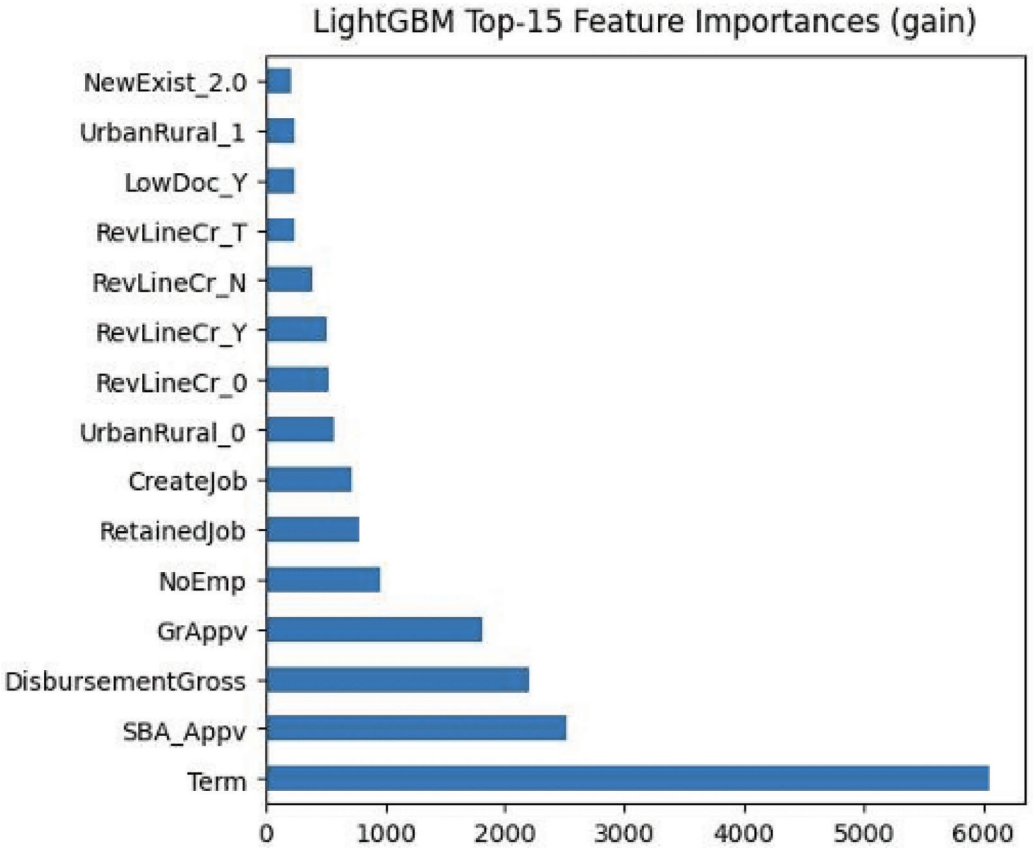

Figure 2: Top 15 Features in LightGBM

- **Low-Risk Tier** represents approximately 78.5% of tested loans (139,661 loans), with an observed default rate of only 2.46%. This tier's extremely low default probability confirms the model's efficacy in confidently isolating stable borrowers. Lending institutions can leverage this tier to streamline approvals, offering competitive interest rates to reliably identified low-risk customers, thus enhancing market competitiveness and operational efficiency.
- **Medium-Risk Tier** comprising about 8% of the test portfolio (14,394 loans), presents a substantially elevated default rate of 48.76%. This intermediate segment warrants cautious credit management and more rigorous monitoring strategies. Given the nearly equal likelihood of default and non-default, credit interventions here could involve closer financial scrutiny, supplemental collateral requirements, or tailored financial counseling services.
- **High-Risk Tier** consisting of roughly 13% of the tested population (23,228 loans), records an alarmingly high default rate of 89.68%. Loans in this category represent significant financial risk and demand stringent lending policies, including higher interest rates, restrictive terms, reduced approved amounts, or outright rejections. Targeting such borrowers with tailored recovery planning or preemptive intervention measures could significantly mitigate financial losses.

```
Calibrated LightGBM binary ROC-AUC: 0.9687
              precision    recall  f1-score   support

           0      0.957     0.967     0.962    146701
           1      0.839     0.796     0.817     31287

    accuracy                          0.937    177988
   macro avg      0.898     0.882     0.890    177988
weighted avg      0.936     0.937     0.937    177988


=== Confusion-style table for the 3-tier scheme ===
Actual          0      1     All
Pred-Tier
Low        136223   3438  139661
Medium       7375   7019   14394
High         2398  20830   23228
All        145996  31287  177283

=== Default-rate & lift by tier ===
```

Actual	total	def_rate	lift
Pred-Tier			
Low	139661	0.024617	0.140042
Medium	14394	0.487634	2.774090
High	23228	0.896763	5.101575

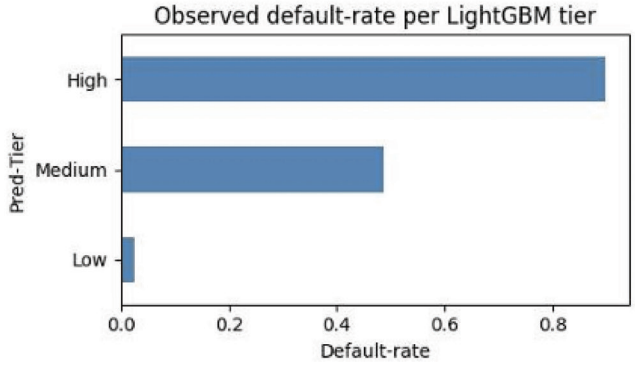

Figure 3: Calibrated LightGBM results

The marked default-rate discrepancies between these tiers underline the calibrated LightGBM's ability to effectively discriminate loan outcomes. The empirical risk lifts over five times in the high-risk tier and approximately 2.8 times in the medium-risk tier, strongly confirm the practical utility of tier-based segmentation for credit-risk management.

6. Discussion

6.1. Interpretation of Findings

Using nearly 900k historical SBA loans, this study demonstrates that a calibrated LightGBM model, coupled with an EBM challenger, can achieve state-of-the-art predictive accuracy while retaining transparent, explainable insights. The LightGBM–EBM model stack effectively addresses the research questions posed: it significantly outperforms traditional models in default prediction (Q1), it provides a method to categorize loans into intuitive risk tiers using calibrated probabilities (Q2), and it offers interpretability through EBM and feature importance analysis (Q3). The result is a practical win-win: authors get the best of both worlds – the predictive power of ensemble learning and the interpretability of an additive model.

Our findings reinforce and extend the current literature. Consistent with prior studies, authors found that ensemble methods (gradient boosting, random forests) greatly surpass logistic regression in predictive performance for credit risk (Wong *et al.*, 2024; Uddin *et al.*, 2023). The magnitude of improvement (AUC ~0.97 vs ~0.82) aligns with what Haque and Hassan (2024) observed on a smaller bank loan dataset, affirming that these techniques scale effectively to larger samples. Moreover, by applying these models to

an imbalanced dataset without class resampling, authors showed they can handle real-world default ratios and still perform robustly addressing a gap where many academic works artificially balance data (Sinap, 2024).

The LightGBM model's recall of over 91% of defaults at a 5% false-positive rate (High-Risk tier size) is an impressive outcome, suggesting that lenders could substantially reduce unexpected losses by deploying such a model for screening. In practical terms, this could improve the portfolio's credit quality and reduce loan loss provisions, as fewer bad loans are approved.

Perhaps the most novel aspect of our results is the demonstration that an interpretable model (EBM) can closely match a leading black-box model. The EBM achieved about 99.4% of LightGBM's AUC performance, confirming recent observations that the accuracy cost of interpretability can be minimal (Černevičienė & Kabašinskas, 2024; Do *et al*., 2024). This addresses a longstanding concern in credit analytics: more complex models might yield better predictions but at the expense of explainability, making them impractical or non-compliant (Binns *et al*., 2018). Here, authors empirically show that the trade-off is very small, an EBM could be an attractive alternative for institutions that prioritize transparency. The cost of explainability, in terms of predictive performance, was on the order of half a percentage point of AUC, which is negligible in many business contexts (Dessain *et al*., 2023). This finding is encouraging for regulators and model risk managers, as it indicates that using interpretable models does not necessarily entail significantly higher prediction error.

Another important implication is how probability calibration and risk tiering can bridge model outputs with business decision-making. By calibrating the LightGBM probabilities and defining risk thresholds aligned with IFRS 9 stages, authors made the model's output immediately actionable for risk management. Our Low/Medium/High risk segmentation provides a tangible tool for credit officers, for example, loans predicted as High risk could be automatically flagged for decline or further review, medium risk loans could be approved with conditions or oversight, and Low risk loans fast-tracked. This multi-tier approach mirrors how banks manage credit risk (rather than treating it as a binary approve/decline decision only). It also integrates naturally with expected credit loss provisioning under IFRS 9, where different stages have different reserve requirements (Stage 2 loans receive lifetime loss provisions even if they haven't defaulted, due to significant risk increase). Our model effectively provides a data-driven way to assign loans to these stages. This contribution is practically significant: it shows that calibrated machine learning outputs can satisfy regulatory frameworks, something often missing in pure ML studies.

The findings also highlight the feasibility of implementing advanced models in resource-limited settings. Even though the EBM took longer to train, the fact that LightGBM (the champion model) can be trained in under a minute on CPU implies that even small lending institutions (like community banks or fintech startups with limited infrastructure) could adopt this approach. Frequent re-training (e.g., monthly model updates with new data) is very attainable with LightGBM. This addresses a common concern that sophisticated models are only feasible for large banks with high-performance servers or GPUs. We've demonstrated that the "whole pipeline in 1 minute" (LightGBM case) is possible. This democratization of credit analytics technology is an important consideration for industry uptake.

6.2. Comparison with Related Work

Compared to prior work, our study distinguishes itself in several ways. Many earlier studies on loan default prediction either focused on binary classification or did not incorporate probability calibration into their evaluation. For instance, Uddin *et al*. (2023) built ensemble models for bank loan approval but stopped at measuring accuracy and F1, without translating probabilities into risk categories. Authors extend beyond this by producing a calibrated risk ranking that aligns with financial risk tiers. Other studies that did consider multiclass classification often pre-defined risk tiers from data (e.g., using delinquency status as proxy categories) rather than deriving them from model probabilities. Our approach, leveraging isotonic calibration, is more flexible and can adapt to the desired risk appetite or regulatory standards of a given institution.

In terms of interpretability, prior literature mostly added interpretability post-hoc (Hjelkrem & Lange, 2023 used SHAP, as did Li & Wu, 2023 in their loan default study). Our work is closer in spirit to recent research by Hjelkrem & Lange (2023) and Bone-Winkel & Reichenbach (2024), who emphasize explainable models. However, even those studies did not present a direct head-to-head comparison of an interpretable model versus an opaque one on the same data. Authors did so and quantified the performance gap, providing concrete evidence for practitioners debating between model choices. Additionally, our use of the EBM in the credit risk context adds to the small but growing body of evidence that EBMs are highly effective for financial risk tasks (Hjelkrem & Lange, 2023; Do *et al*., 2024).

The scale of our dataset (~900k loans) also sets this work apart. Many academic papers use much smaller datasets (e.g., the common LendingClub dataset has ~40k loans used in Wu, 2022; or smaller bank datasets around 100k records). By using the entire SBA corpus, authors could

validate the models in a more realistic, large-scale scenario. Encouragingly, LightGBM scaled effortlessly to this data size, and EBM also managed, though with longer training. This matters because credit portfolios at large banks easily run into millions of accounts, models must handle such scale.

Finally, from a theoretical standpoint, our results affirm certain machine learning principles in the context of credit risk: ensemble methods reduce variance and capture complex patterns (explaining LightGBM's win), additive models can approximate ensemble performance if constructed cleverly (explaining EBM's close second place), and calibration is crucial when decision thresholds have real meaning (to ensure predicted probabilities are interpretable as risks).

6.3. Practical Implications

The practical implications of this study are significant for financial institutions:

- **Risk Management:** The LightGBM–EBM framework provides a blueprint for integrated risk modeling. A bank can use LightGBM as a champion model for highest accuracy in automated underwriting while keeping an EBM or similar interpretable model in parallel for audit and compliance purposes. For instance, if a regulator questions why a particular loan was denied, the bank could refer to the EBM's explanations (since EBM will generally agree on the major risk factors, given its similar performance). This champion–challenger setup could also be used in production: most decisions by LightGBM, but if EBM disagrees strongly or if a loan is borderline, route to manual review. Our study thus offers a template for model governance in the era of AI in credit.
- **Regulatory Compliance:** Banks can be confident that deploying a high-performing ML model need not violate explainability requirements like those implied by the EU's GDPR or US fair lending regulations. By having an interpretable model nearly as good as the black box, they can satisfy "show me the reason" demands. Additionally, the IFRS 9 alignment authors demonstrated means model outputs can feed directly into accounting processes (e.g., calculating expected credit loss for each risk bucket). This linkage between AI models and accounting standards is a novel bridge that could streamline how risk analytics supports finance departments.
- **Economic Benefits:** Better default prediction and risk segmentation translate to lower loan losses and more efficient capital allocation. If a bank can more accurately identify high-risk loans, it can avoid funding them or price them appropriately (higher interest to compensate risk) and, conversely, not deny credit to low-risk borrowers who might have been misclassified by a weaker model. Thus, there's a social benefit too: deserving small businesses might get credit because the model judged them accurately rather than being rejected by an overly conservative traditional scorecard. On the other hand, loans that truly are high-risk can be curtailed, protecting the bank's portfolio and indirectly the financial system.
- **Technology Adoption:** From a technology perspective, our work suggests that even smaller banks or lending startups can adopt advanced ML techniques without needing expensive infrastructure. The use of open-source libraries and standard computing environments lowers the barrier to entry. There's an implicit suggestion that regulators and industry groups could promote such frameworks (perhaps open-source model pipelines pre-calibrated on large public data) to uplift risk management practices broadly.

7. Limitations

While the results are promising, this study has several limitations that warrant discussion:

- **Proxy for Medium Risk:** The definition of the "Medium" risk tier in our dataset is inherently a proxy, since the SBA data did not come labeled with multiple risk categories. Authors imposed a structure by splitting predicted probabilities. This means some loans in Medium Risk might be those that would have defaulted under slightly different conditions or just got lucky. It also means our medium category is somewhat subjective and depends on chosen thresholds. In practice, banks define risk grades using a combination of model output and policy judgment. Our results showed moderate precision and recall for the medium class (as it's not a cleanly separable group), which is expected. Thus, while the three-tier scheme is illustrative and aligned with IFRS 9 conceptually, it's not as ground-truth validated as the binary default labels. Users of such a model should supplement medium-risk identification with business rules or expert review.
- **Model Training Time (EBM):** The EBM model's training was quite slow (~84 minutes on CPU for ~900k samples). For a one-time analysis, this is fine, but it could be a bottleneck for rapid model updates or if using even larger data. In a production scenario, one might need a server with more cores or an optimized implementation to retrain EBM in a reasonable time. In contrast, LightGBM's very short training time stands out; if an institution values agility (e.g., updating

the model frequently as new data arrives or as macro conditions change), LightGBM has a clear advantage. DNN training was also relatively slow (~5 minutes), and authors didn't see gains from it, likely due to limited hyperparameter tuning under Colab constraints. It's possible a more thoroughly tuned DNN could have performed better but given the computational cost and the already high performance of boosting, that route was less appealing.

- **Feature and Data Limitations:** Our analysis is only as good as the data. If there were any distortions (like policy changes capping interest rates or certain programs that skewed terms), the model might misinterpret those. Also, the data spans 45 years there could be non-stationarity (the lending criteria of the 1970s versus 2010s differ). Authors tried to mitigate this by not overly tuning to any specific year, but a more nuanced time-based validation could be explored. Additionally, SBA loans are a specific subset of SME finance (they often have government guarantees). This means default patterns might differ from unguaranteed loans e.g., perhaps risk-taking is different knowing a portion is guaranteed by SBA. If one naively applied our model to a private bank's portfolio without recalibration, it might overestimate risk because it learned under an SBA regime. So, generalizability to all SME loans should be approached carefully.
- **Fairness and Bias:** Authors did not explicitly test for biases in the model. Attributes like race or gender of business owners were not in the dataset, but proxies (geography, industry, etc.) could inadvertently serve as correlates. Ensuring the model is fair and does not systematically disadvantage protected groups is crucial before deployment (Barocas, Hardt & Narayanan, 2019). Future work will consider adding fairness metrics or constraints, especially since explainable models like EBM could be combined with fairness auditing to better understand any bias issues.
- **Beyond Probability of Default:** Authors focused on predicting default and aligning with IFRS 9 staging (which is PD-focused). However, credit risk assessment also involves Loss Given Default (LGD) and Exposure at Default (EAD) for a full picture of expected losses. Our study doesn't address LGD implicitly, they assumed a default is a default, but in reality, severity matters (a default where the bank loses 10% of exposure vs. 100% are different). Future extensions could consider a two-stage model (predict default, and if default, then predict loss fraction) or integrate into a portfolio loss simulation.
- **Comparative Scope:** Authors included a variety of models, but one could argue for even more e.g., CatBoost, XGBoost (another boosting), SVMs, or more exotic models. They chose a representative set covering most paradigms. XGBoost and CatBoost are likely to yield similar performance to LightGBM (perhaps slightly lower or higher depending on tuning) based on other research (Li & Wu, 2023 note LightGBM vs. XGBoost differences were minor). They picked LightGBM for its speed and known strong performance. So, while they may not have exhausted every algorithm, they doubt any would clearly beat LightGBM in this context by a large margin.

8. Conclusion

This study presented a comprehensive modeling framework for SME loan default prediction that bridges the gap between predictive performance and interpretability. By leveraging nearly 899,000 SBA loans, an unusually large dataset by academic standards, authors were able to rigorously evaluate a state-of-the-art ensemble model (LightGBM) against interpretable and traditional models. The results are compelling: LightGBM achieved a ROC-AUC of 0.969 with over 91% recall of defaulting loans, substantially outperforming logistic regression and even a neural network. Meanwhile, the Explainable Boosting Machine (EBM) delivered almost matching performance (ROC-AUC 0.963) while providing full transparency into its decision-making. When calibrated and combined, this champion–challenger pair offers a powerful and practical solution for lenders: high-accuracy risk predictions that can be explained to stakeholders and regulators.

Authors also demonstrated how to derive meaningful risk tiers from the model's probabilistic output, essentially constructing a three-tier credit risk rating system aligned with IFRS 9 stages. Loans were sorted into Low, Medium, and High risk with observed default rates of ~2.5%, ~49%, and ~90%, respectively, validating the model's ability to separate the portfolio into distinct risk bands. This has direct applicability in credit risk management, enabling targeted interventions (e.g., heightened monitoring for the medium group, denial or special handling for the high group). The fact that the entire modeling pipeline can be executed on commodity hardware in under a minute (for LightGBM) also underscores the practicality of this approach for widespread adoption.

In conclusion, our work offers a blueprint for modernizing SME credit scoring: use gradient boosting for maximal predictive power and calibrate its outputs for risk segmentation; concurrently, maintain an interpretable model like EBM to ensure transparency and compliance. This approach marries the strengths of AI with the trust required in finance. Authors believe such frameworks can

drive the next generation of credit risk analytics, where "glass-box" performance (models that are both accurate and interpretable) becomes the norm. The LightGBM–EBM stack proposed here is a step in that direction, showing that lenders need not trade off accuracy for explainability. As the financial industry increasingly embraces machine learning, studies like ours help chart a path toward models that are not only powerful and efficient but also fair, accountable, and aligned with regulatory principles.

9. Future Work

The promising results of this study open several avenues for future research:

- **Fairness and Bias Mitigation:** A natural next step is to ensure the model's decisions are unbiased. Future work could incorporate fairness metrics or constraints (e.g., equal opportunity, disparate impact analysis) into model training. Techniques like synthetic minority oversampling or adversarial debiasing could be tested to see if performance can be maintained while reducing any unwanted bias. Additionally, proxy features (like ZIP code or demographic indicators) could be introduced to explicitly audit the model's fairness and apply mitigation strategies (such as reject option-based classification).
- **Adversarial Robustness:** As with any predictive model, especially one used in lending, adversaries might try to game the system (e.g., by manipulating input features to appear at low risk). Future research could evaluate the robustness of LightGBM and EBM to adversarial perturbations in data input. Methods to improve robustness, such as imposing monotonicity constraints (to ensure logically consistent behavior) or adversarial training (training on slightly perturbed data), could be investigated. Ensuring the model is not easily fooled is important for deployment.
- **Macro-Economic Integration:** The current model uses loan-specific features and inherently handles some macro conditions due to the time span of the data, but explicit integration of macroeconomic indicators could be valuable. Future models could incorporate variables such as GDP growth, unemployment rates, or other economic indices at the time of loan origination or during loan life. This would enable dynamic stress-testing: assessing how portfolio default risk might change under different economic scenarios (similar to stress test frameworks in banking). Time-series models or scenario analysis could be layered on top of the PD model to simulate performance under recession versus expansion.
- **Extension to Loss Given Default (LGD) Modeling:** As mentioned in the limitations, predicting default is only one part of credit risk. Future research could pair the PD model with an LGD model for a more complete risk assessment. EBM or LightGBM could be used to predict LGD given default (using historical recovery data). Combining these, one could estimate expected loss for each loan, which is ultimately what banks need for capital allocation. Another angle is multi-task learning, where a single model predicts both probability of default and expected loss, though this can be complex.
- **Real-Time Decision Support:** Deploying this model in an interactive decision tool for loan officers is another potential direction. For example, an interface could take applicant data and return not just a score and tier but also an explanation ("Debt-to-Income is high, which contributes X% to risk, consider requiring a co-signer or reducing loan amount"). Usability studies could evaluate whether human decision-makers improve decision quality or consistency using these explanations.
- **Comparative Studies with Other Explainable ML:** While EBM was our choice for interpretability, other methods exist, such as Explainable Neural Networks (XNNs), GA2M (Generalized Additive Models with interactions), or even simpler rule-based classifiers. Future work could benchmark these on the same problem to see if any lighter-weight interpretable method can match EBM's performance. Additionally, exploring SHAP or LIME not just for explanation but to create simplified surrogate models might be interesting, though surrogate models typically lose fidelity.
- **Cross-Validation with Other Data:** Testing the framework on different datasets (e.g., LendingClub data, mortgage datasets, or non-U.S. SME loans) would ensure robustness. Broad validation could confirm that the LightGBM–EBM approach generalizes across credit contexts, or indicate what adjustments (e.g., hyperparameter tuning, recalibration) are needed.
- **Automated Machine Learning (AutoML):** An AutoML approach could be applied to this problem, where the system searches over model architectures (including pre-processing and feature engineering) to see if any combination can exceed our manual approach. AutoML might discover interactions or transformations authors did not explicitly code. However, maintaining interpretability would be a challenge if the best model found is highly complex. Balancing AutoML with interpretability constraints could itself be a research topic.

By addressing these future directions, researchers and practitioners can further enhance the reliability, fairness, and utility of machine learning models in credit risk. The goal is a robust, transparent credit scoring system that stakeholders trust and that demonstrably improves financial outcomes. Our study lays a strong foundation, and authors anticipate continued advancements built upon this work in the quest for better credit risk modeling.

Acknowledgement

The authors sincerely thank the professors of Liverpool John Moores University, Liverpool, England, for their invaluable support, encouragement, and for providing the necessary facilities and opportunities to carry out this research.

Authorship Contribution

Minh Nguyen Hoang: Conceived of the presented idea, developed the theory and performed the computations, and wrote the manuscript with support from other authors.
Thota Sai Karthikeya: Verified the analytical methods, contributed to the final version of the manuscript and provided critical feedback and helped shape the research, analysis and manuscript.
Thota Sree Mallikharjuna Rao: Supervised the project and findings of this work, discussed the results and contributed to the final manuscript**.**

Ethical Approval

This research did not involve studies with human participants or animals performed by any of the authors. Therefore, ethical approval was not required.

Funding

This research received no specific grant from any funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

The authors declare no specific declarations. Both authors have read and approved the final version of the manuscript and agree to be accountable for the integrity and accuracy of the work.

Conflict of Interest

The authors declare no conflict of interest related to this work.

References

Agarwal, R., Frosst, N., Zhang, X., Caruana, R., & Hinton, G. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems (NeurIPS)*.

Arik, S. Ö., & Pfister, T. (2021). Explainable Neural Networks (XNNs): A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627– 635.

Baesens, B., Roesch, D., & Scheule, H. (2023). *Machine learning and AI for credit risk*. Wiley & Sons.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting, 29*(4), 563–574.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Proceedings of CHI 2018*. https://doi.org/10.1145/3173574.3173951

Bone-Winkel, G. F., & Reichenbach, F. (2024). Improving credit risk assessment in P2P lending with explainable machine learning survival analysis. *Digital Finance*. https://doi.org/10.1007/s42521-024-00114-3

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A: 1010933404324

Carmody, E., & Ajani, T. (2023). Predicting credit default with machine learning: An Australian case study. *Journal of Risk Modeling, 7*(1), 45–62.

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Journal of Financial Innovation, 15*(2), 100–124.

Dessain, J., Bentaleb, N., & Vinas, F. (2023). Cost of explainability in AI: An example with credit scoring models. In *Advances in Financial Machine Learning* (pp. 498–516). Springer. https://doi.org/ 10.1007/978-3-031-44064-9_26

Do, T. T., Babaei, G., & Pagnottoni, P. (2024). Explainable machine learning for credit risk management when features are dependent. *Measurement: Interdisciplinary Research and Perspectives, 22*(4), 315–340.

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://doi.org/10.4236/ajibm. 2024.1412085

Hjelkrem, L. O., & Lange, P. E. (2023). Explaining deep learning models for credit scoring with SHAP. *Journal of Risk and Financial Management, 16*(4), 221. https://doi.org/10.3390/jrfm16040221

Jakubik, P., & Teleu, S. (2025). Improving credit risk assessment in uncertain times: Insights from IFRS 9. *Risks, 13*(2), 38. https://doi.org/10.3390/risks13020038

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

Li, H., & Wu, W. (2023). Loan default predictability with explainable machine learning. *Finance Research Letters, 52*, 104993.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of KDD 2012* (pp. 150–158). ACM.

Nallakaruppan, M. K., Balusamy, B., Shri, M. L. A., & Kannan, K. (2023). Transforming credit risk assessment: A systematic review of AI and machine learning approaches. *Journal of Banking and Financial Technology, 27*(3), 234–256.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Noriega, J. R., Rivera, L., & Herrera, A. (2023). Ensemble-based machine learning algorithm for loan default risk prediction. *Journal of Computational Finance and Economics, 2*(1), 33–47.

Sinap, V. (2024). A comparative study of loan approval prediction using machine learning methods. *Gazi* Üniversitesi *Fen Bilimleri Dergisi (GUJSC)*. https://doi.org/10.29109/gujsc.1455978

Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021). Prediction of modernized loan approval system based on machine learning approach. *Proceedings of CONIT 2021 (IEEE)*. https://doi.org/10.1109/CONIT51480.2021.9498475

Tavakoli, M., Chandra, R., Tian, F., & Bravo, C. (2023). Multi-modal deep learning for credit rating prediction. arXiv:2301.01234.

Uddin, N., Ahamed, M. K., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). Ensemble machine learning-based bank loan approval prediction. *International Journal of Cognitive Computing in Engineering, 4*, 327–339. https://doi.org/10.1016/j.ijcce.2023.09.001

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of KDD 2002* (pp. 694–699).