DOI: https://doi.org/10.4038/sajbi.v5i4.84

An Efficient Loan Default Risk Assessment Framework Using TPE-Optimized CatBoost Model

South Asian Journal of Business Insights
2026, Special Issue, 5-20
ISSN 2773-7012 (print)
ISSN 2773-6997(online)
© 2026, Faculty of Management and Finance University of Ruhuna, Sri Lanka



Baidyanath Sou*
Dept. of Computer Science, Jagannath Kishore College, India

Abstract

Reliable estimation of loan default risk plays a vital role in ensuring financial system resilience and enabling sound credit decision-making in today's lending landscape. Although conventional statistical techniques offer transparency, they frequently struggle to model the intricate nonlinear patterns embedded in high-dimensional financial datasets. To address this limitation, the present study introduces a robust framework for default risk evaluation that employs a CatBoost model optimized through a Tree-Structured Parzen Estimator (TPE). This methodology combines Bayesian-driven hyperparameter tuning with gradient boosting techniques to improve both predictive accuracy and the calibration of probability estimates. The performance of the proposed model is validated using three widely recognized benchmark datasets: German Credit Risk, Polish Companies Bankruptcy, and Taiwan Credit Card Default. The proposed approach is evaluated in comparison with several conventional machine learning models, namely Logistic Regression, Random Forest, and Support Vector Machine. The assessment is carried out using a diverse set of performance indicators, including ROC–AUC, recall, F1-score, Brier score, and the Kolmogorov–Smirnov (KS) measure. Empirical analysis across multiple datasets reveals that the CatBoost model enhanced through TPE-based optimization consistently delivers superior results relative to the benchmark methods, particularly in terms of classification strength, probability estimation accuracy, and effective differentiation of risk levels. Notably, the improvement is more pronounced in datasets characterized by class imbalance, emphasizing the advantage of the adopted optimization technique. The results indicate that combining advanced boosting methods with systematic hyperparameter tuning yields a dependable and scalable framework for credit risk evaluation, offering meaningful benefits to financial institutions through more precise and data-informed lending decisions.

Keywords: CatBoost, Credit Risk Assessment, Gradient Boosting, Hyperparameter Optimization, Loan Default Prediction, Tree-Structured Parzen Estimator (TPE)

* Corresponding Author – sou.baidyanath@gmail.com

Introduction

Credit risk assessment, particularly the analysis of loan default risk, is crucial to guaranteeing the stability and viability of financial institutions. Inaccurate estimates of loan failure probability can expose banks and lending institutions to significant financial losses, lower profitability, and increased capital requirements, thereby spreading instability throughout the financial ecosystem (Chang et al., 2024). Beyond institutional consequences, loan default prediction has a systemic impact on capital allocation, liquidity distribution, and overall market stability.

The complexity of financial risk dynamics has increased significantly due to the rapid expansion of digital lending platforms, such as peer-to-peer platforms, fintech-driven microcredit, and unsecured personal loans. The shortcomings of conventional loan default prediction methods are especially noticeable during times of economic hardship when default rates increase across countries. A paradigm change from conventional to dynamic, data-driven decision-making frameworks has been sparked by the integration of high-frequency data, behavioral indicators, and alternative credit signals (Ahelegbey & Giudici, 2023). As a result, rather than being a one-time examination, loan default assessment is now an ongoing and flexible procedure.

Conventional loan default forecasting systems, which rely heavily on linear statistical models and historical credit bureau data, are becoming unsuitable for capturing modern borrowers' complex behavioral patterns. The increasing availability of big data, real-time risk monitoring technologies, and alternative information sources has prompted financial institutions to implement complex analytical frameworks capable of handling large, multidimensional datasets. These techniques help to identify intricate non-linear risk correlations that standard methods often fail to capture (Hlongwane et al., 2024).

Globally, the difficulty of managing rising loan default rates has grown because of economic volatility, inflationary pressures, and systemic disruptions like the COVID-19 pandemic. The foregoing factors have increased default risk for both individual and corporate debtors (Chang et al., 2024). In advanced economies, rising default rates in consumer loan markets illustrate the limits of traditional models based on inflexible assumptions and limited feature sets. At the same time, the democratization of lending via digital and mobile platforms has allowed borrowers with weak or informal credit histories to enter the financial system.

This transition has driven research into the use of other data sources, such as phone records, e-commerce transactions, and psychometric characteristics, to improve forecast accuracy and increase financial inclusion (Lee et al., 2024). In this circumstance, data-driven decision-making has shifted from a competitive advantage to a strategic imperative. Machine learning (ML) systems can handle large datasets, detect hidden patterns, and adjust as new information becomes available. Compared to rule-based or manual techniques, ML-driven models improve borrower classification precision and flexibility, allowing for more responsive and risk-sensitive lending decisions (Malempati, 2025).

The evolution of loan default prediction modelling reflects broader technological trends in financial analytics toward automation, scalability, and intelligence. Historically, default modelling relied on statistical techniques such as discriminant analysis and logistic regression (LR), which estimate default probabilities using linear relationships between borrower attributes

and repayment outcomes. While these models offer transparency and ease of interpretation, they often struggle to capture the complex, non-linear interactions characteristic of modern financial data (Bücker et al., 2022). ML techniques address these limitations by learning directly from multidimensional and heterogeneous datasets. Ensemble methods, including Random Forests (RF) and Gradient Boosting Machines, have demonstrated superior predictive performance in identifying potential defaulters. Gradient boosting frameworks have emerged as benchmarks in credit risk prediction due to their robustness in handling missing values, feature interactions, and non-linearities (Alvi et al., 2024).

Despite recent advances in ML, several challenges remain in loan default prediction. Available models often show bias toward minority borrower groups, depend on static features, and perform poorly on imbalanced datasets (Hlongwane et al., 2024). Also, they struggle to effectively incorporate emerging behavioural and contextual data, such as mobile transactions, social network activity, and psychometric information, which increasingly influence financial behaviour. Although ensemble learning methods have substantially improved predictive performance, many operate as "black boxes," limiting transparency, interpretability, and regulatory acceptance (Bücker et al., 2022). Furthermore, their performance is highly dependent on hyperparameter selection, making efficient optimization essential for achieving reliable predictive outcomes.

Recent studies have shown the effectiveness of gradient boosting algorithms in credit risk prediction, while Bayesian optimization techniques have emerged as efficient approaches for hyperparameter tuning. These developments have largely been investigated independently. Despite CatBoost's proven ability to handle categorical variables and reduce overfitting in structured financial datasets, there is still limited systematic research on integrating the Tree-Structured Parzen Estimator (TPE) with CatBoost for loan default prediction across diverse credit datasets. Consequently, empirical evidence regarding the effectiveness and generalizability of this integrated framework remains limited.

This unresolved gap gives rise to an important research problem: whether integrating the TPE with CatBoost can consistently improve predictive performance, probability calibration, and risk discrimination across diverse loan default datasets compared with conventional ML approaches. Addressing this problem is important because CatBoost's predictive performance depends heavily on appropriate hyperparameter optimization, while existing optimization strategies often fail to efficiently explore complex search spaces. Accordingly, this study aims to develop and evaluate a TPE-optimized CatBoost framework for loan default risk assessment. To achieve this objective, the study aims to: (i) investigate whether TPE-based hyperparameter optimization enhances CatBoost's predictive performance, (ii) compare the proposed framework with widely used baseline models, including LR, RF, and Support Vector Machines (SVMs), and (iii) evaluate its effectiveness using risk-relevant performance measures, including ROC-AUC, Recall, F1-score, Brier score, and the Kolmogorov–Smirnov (KS) statistic. Through this comprehensive evaluation, the study seeks to provide a robust, scalable, and practically applicable framework for data-driven credit risk assessment and financial decision-making.

The remainder of this paper is organized as follows. Section two reviews the relevant literature on loan default prediction modelling and hyperparameter optimization. Section three outlines the proposed methodology. Section four describes the experimental setup, while Section five

presents the results and discussion. Section 6 highlights the challenges and future research directions, and finally, Section 7 concludes the paper.

Literature Review

For several decades, loan default prediction has relied primarily on classical statistical approaches, particularly LR and LDA, to estimate borrowers' probability of default. Among these, LR has remained the dominant technique because of its simplicity, interpretability, and regulatory acceptance. By modelling a linear relationship between borrower characteristics, such as income, debt-to-income ratio, and credit history, and the probability of default, LR provides transparent and auditable decision rules. Similarly, LDA separates borrowers into default and non-default groups by projecting data onto discriminant axes. Despite these advantages, both methods assume relatively simple linear relationships and often fail to capture the complex nonlinear interactions present in modern financial datasets, limiting their effectiveness in heterogeneous credit environments (Levy & Baha, 2021).

Traditional lending decisions have also relied heavily on rule-based expert systems and credit bureau information, where predefined heuristics guide borrower evaluation (Soni et al., 2024). Although these approaches remain fundamental to conventional credit scoring, they primarily reflect historical repayment behaviour and frequently exclude under-banked or informal borrowers, thereby reinforcing structural barriers to financial inclusion. Furthermore, their dependence on manual feature selection and heuristic threshold setting introduces subjectivity and reduces adaptability to changing borrower behaviour and economic conditions (Jemai & Zarrad, 2023). Empirical evidence further indicates that the predictive performance of LR deteriorates substantially in high-dimensional, heterogeneous, and class-imbalanced datasets. Collectively, these studies suggest that while traditional statistical models remain valuable because of their transparency and regulatory acceptance, their limited ability to model complex financial relationships has driven the transition toward data-driven ML approaches capable of exploiting large-scale and multidimensional financial data (Roy & Vasa, 2025).

The emergence of ML techniques has significantly transformed loan default prediction by enabling automatic feature learning, nonlinear interaction modelling, and scalable analysis of large financial datasets. Algorithms such as RF, SVMs, Gradient Boosting methods, and Artificial Neural Networks (ANNs) have consistently demonstrated superior predictive performance compared with traditional statistical models (Soni et al., 2024). Decision Trees (DTs) provide interpretable classification rules but are sensitive to data perturbations, leading to the development of ensemble learning approaches. RF reduce this instability through bootstrap aggregation, improving robustness and predictive accuracy, while SVMs effectively address class imbalance by maximizing the decision margin between borrower classes, although their performance remains highly dependent on kernel selection and regularization (Wang, 2024). Similarly, boosting-based methods sequentially minimize prediction errors and improve the detection of minority-class instances, particularly when integrated with resampling strategies or hybrid meta-learning frameworks (Roe et al., 2020). Overall, the literature consistently demonstrates that ensemble learning methods outperform conventional statistical models in terms of predictive accuracy and robustness. However, comparative studies also reveal that model performance varies considerably across datasets, indicating that prediction accuracy

depends not only on the learning algorithm but also on effective model configuration and hyperparameter optimization.

Among ensemble learning methods, CatBoost has emerged as a competitive gradient boosting algorithm because of its ability to process categorical variables efficiently while reducing overfitting through ordered boosting (Prokhorenkova et al., 2017). Its predictive performance depends mainly on appropriate hyperparameter selection, which remains a computationally demanding task. Conventional optimization methods, including grid search and random search, often struggle to efficiently explore complex and high-dimensional hyperparameter spaces. Bayesian optimization techniques, particularly the TPE, provide a more efficient and scalable alternative by guiding the search process toward promising hyperparameter configurations (Watanabe, 2023). Although previous studies have independently demonstrated the effectiveness of CatBoost and Bayesian optimization, the integration of these two techniques has received comparatively limited studies in the context of loan default prediction. Existing studies primarily emphasize predictive accuracy, while comparatively less attention has been given to probability calibration, risk discrimination, and the robustness of optimized CatBoost models across datasets with different characteristics. The comparative advantages of integrating TPE with CatBoost over widely adopted baseline models such as LR, RF, and SVM remain insufficiently investigated. These observations indicate that further research is required to determine whether efficient Bayesian hyperparameter optimization can consistently enhance CatBoost's predictive capability across diverse credit risk scenarios.

Motivated by these research gaps, the present study advances the existing literature by systematically integrating TPE-based Bayesian optimization with CatBoost for loan default prediction and evaluating its performance across multiple benchmark credit datasets. Unlike previous studies that primarily focus on predictive accuracy, the proposed framework provides a comprehensive evaluation using discrimination, calibration, and risk-separation measures, including ROC-AUC, Recall, F1-score, Brier score, and the KS statistic. Through this comprehensive assessment, the study contributes to literature by providing stronger empirical evidence regarding the effectiveness and generalizability of TPE-optimized CatBoost models for credit risk assessment, thereby supporting more reliable and data-driven financial decision-making.

Methodology

Loan default prediction is formulated as a binary classification problem, where each borrower is labelled as either a defaulter (1) or a non-defaulter (0). The objective is to build a predictive model that determines the likelihood of default using financial and behavioral attributes of borrowers. Achieving high classification accuracy is essential, as the costs of misclassification are unequal; in particular, failing to correctly identify defaulters can lead to substantial financial losses.

Catboost Model

CatBoost enhances predictive capability while overcoming several shortcomings associated with traditional boosting approaches, especially in the treatment of categorical variables. In comparison to conventional gradient boosting methods that rely on preprocessing steps such as

one-hot or ordinal encoding, CatBoost processes categorical features through ordered target-based statistics, which helps retain information and limits unnecessary expansion of feature space. One of its distinguishing features is the ordered boosting technique, which reduces issues such as prediction shift and target leakage by training models over sequential permutations of the dataset. This design contributes to better generalization and minimizes overfitting, an essential requirement in credit risk analysis (Prokhorenkova et al., 2017).

CatBoost builds its model as a sequential ensemble of decision trees, with each new tree being trained to optimize the loss function to repair the residual mistakes of previous ones. Symmetric tree architectures provide effective prediction and increase processing performance. The algorithm is especially well-suited for real-world financial data, which is frequently incomplete and diversified. It also integrates adaptive learning rate techniques and can naturally handle missing values. These features make CatBoost a scalable and dependable method for identifying intricate nonlinear patterns in loan default prediction issues.

Tpe-based hyperparameter optimization

TPE is a Bayesian optimization technique that effectively identifies appropriate hyperparameter settings for models such as CatBoost. It was selected in this study because it efficiently explores high-dimensional hyperparameter search spaces and is well suited for optimizing computationally expensive ensemble models such as CatBoost. Unlike exhaustive search strategies such as grid search and random search, TPE adaptively samples promising hyperparameter configurations based on information obtained from previous evaluations, thus improving search efficiency. Many Bayesian optimization methods model the objective function directly. On the other hand, TPE models the distribution of hyperparameter settings using the observed objective values. It uses a threshold derived from the quantiles of recorded loss values to divide configurations into two groups: those producing better results and those producing worse outcomes. Kernel density estimation is then employed to model these two distributions, enabling TPE to capture complex and conditional interactions within the hyperparameter search space (Watanabe, 2023).

At each iteration, the method recommends new hyperparameter combinations by maximizing the predicted improvement criterion, effectively selecting configurations that are more likely to belong to the "good" group than the "bad" one. This probabilistic search strategy allows TPE to focus on promising regions of the search space while leveraging knowledge gained from previous evaluations. Consequently, it requires fewer evaluations than conventional search methods, making it particularly suitable for optimizing computationally expensive models such as CatBoost in loan default prediction applications.

Proposed Framework

The proposed framework, illustrated in Figure 1, uses a standardized and unified ML pipeline that ensures consistency, comparability, and robustness across all models. The method is separated into five steps: data loading, preprocessing, model training, hyperparameter optimization, and performance evaluation.

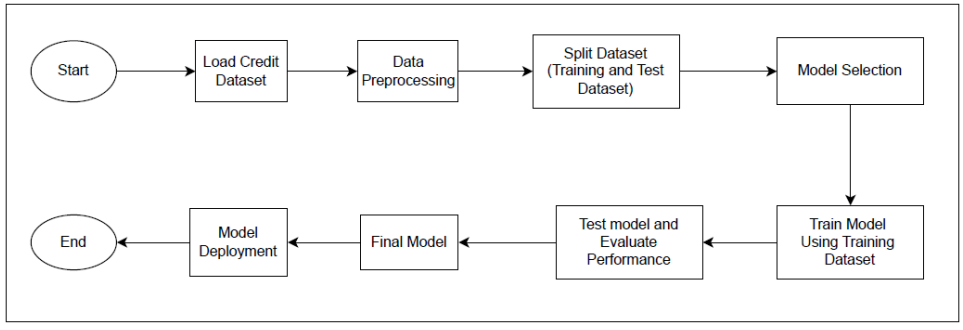


Figure 01. The framework of proposed model

Initially, the dataset is imported and prepared by separating the input features from the target variable representing loan default status. To keep class distribution and avoid sample bias, a 70:30 stratified train-test split is used. To prevent data leakage, all preprocessing procedures are applied to the training data first, followed by the test data.

To ensure numerical stability and homogeneous feature contribution, data preprocessing involves median imputation for missing values and min-max scaling to normalize feature values. The altered training and testing datasets are then applied consistently across all models, allowing for a fair comparison.

The selected benchmark models (LR, RF, and SVMs) are trained on the same dataset. Class-balanced weighting is used to address class imbalance and find minority class occurrences. Each model generates class predictions and probability estimates, which are then used for evaluation.

To improve predictive performance, a CatBoost model is optimized using the TPE technique, which is applied in Optuna. The optimization method explores a preset hyperparameter search space that includes learning rate, tree depth, regularization strength, number of iterations, and sampling parameters. To analyze various configurations in the training data, a stratified k-fold cross-validation procedure is utilized with the purpose of optimizing the ROC-AUC value.

Once the optimum hyperparameters have been identified, the CatBoost model is retrained on the whole training dataset and tested on a new set. The final performance of all models is then evaluated using a consistent evaluation methodology, resulting in an objective assessment of the suggested approach in comparison to traditional machine learning techniques.

Model Evaluation

A held-out test set is used to test model performance in order to provide an objective evaluation. Numerous evaluation criteria are employed to evaluate classification performance and probability dependability due to the financial significance of loan default prediction. The model's capacity to differentiate between defaulters and non-defaulters at different levels is evaluated using ROC-AUC (Fawcett, 2006). Recall evaluates the model's ability to accurately identify

defaulters, which is essential for reducing financial risk. While the Brier score assesses the calibration quality of predicted probabilities, the F1 score strikes a balance between precision and recall (Glenn, 1950; Saito & Rehmsmeier, 2015). In credit risk modelling, KS statistic is frequently employed to assess the degree of divergence between default and non-default distributions (Hand & Henley, 1997).

Experimental Setup

Dataset Description

To assess the effectiveness and generalizability of the proposed framework, experiments are conducted on three widely used benchmark datasets: the German Credit Risk dataset, the Polish Companies Bankruptcy dataset, and the Taiwan Credit Card Default dataset. These datasets are obtained from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.html) and a summary of their key characteristics is presented in Table 1. The datasets were selected because they are widely recognized benchmark datasets in credit risk research and collectively represent diverse prediction scenarios with different sample sizes, feature dimensions, and levels of class imbalance. Their diversity enables a comprehensive assessment of the robustness and generalizability of the proposed framework under varying credit risk conditions.

The German Credit Risk dataset consists of borrower-level financial and demographic attributes, such as credit history, loan purpose, and employment status, along with a binary target variable indicating credit risk (good or bad). The dataset contains approximately 1,000 instances and 20 features, with moderate class imbalance.

The Polish Companies Bankruptcy dataset includes financial ratio indicators derived from corporate financial statements and is designed to predict bankruptcy outcomes. It is characterized by a high-dimensional feature space and significant class imbalance, reflecting real-world rare-event prediction scenarios.

The Taiwan Credit Card Default dataset contains detailed information on credit card clients, including repayment history, bill amounts, and demographic characteristics. It comprises 30,000 observations with 23 features and exhibits class imbalance, making it suitable for evaluating model robustness under skewed class distributions.

These datasets enable a comprehensive evaluation of the proposed framework across diverse credit risk scenarios characterized by differences in dataset size, feature dimensionality, and class imbalance.

Table 01. Summary of the used datasets

Dataset	Number of Samples	Number of features	Class Distribution
German Credit Risk	1000	20	Moderate
Polish Companies Bankruptcy	7027	64	High
Taiwan Credit Card Default	30000	23	Moderate

Data Preprocessing

Data preparation ensures consistency and reliability across datasets. Missing numerical values are handled via median imputation, which preserves the underlying distribution while minimizing the impact of outliers. All features are then scaled using min-max normalization to guarantee that variables of varying magnitudes contribute consistently during model training. Categorical variables were transformed into numerical representations using appropriate encoding techniques to ensure compatibility with ML algorithms.

To prevent data leaking and assure unbiased evaluation, each dataset is divided into training and testing subsets using a stratified 70:30 split, while retaining the original class distribution. Preprocessing transformations are first fitted to the training data and then applied to the test data.

Model Configuration

A single experimental framework is used to provide fair comparisons between models. The baseline models selected for comparison are LR, RF, and SVMs, as they represent widely used and well-established approaches for credit risk assessment. LR serves as the traditional statistical baseline owing to its interpretability and regulatory acceptance, whereas RF and SVM represent widely adopted ML techniques with demonstrated effectiveness in modelling nonlinear relationships and handling class-imbalanced credit datasets. Their inclusion enables a comprehensive comparison between conventional statistical methods, ensemble learning, and the proposed optimized boosting framework. LR is implemented with class-balanced weighting and optimized with the L-BFGS solver. To allow for probabilistic evaluation, the SVM model uses a radial basis function kernel with probability estimation enabled.

The proposed model employs CatBoost, a gradient boosting algorithm optimized using the TPE within the Optuna framework. The hyperparameter search space consists of learning rate, tree depth, L2 regularization, number of boosting iterations, bagging temperature, and border count.

To achieve robust hyperparameter selection, the TPE optimization process is carried out over 50 trials, with stratified k-fold cross-validation ($k = 5$) applied to the training dataset. The objective function is specified as maximizing the ROC-AUC score. To improve model generalization, the optimized CatBoost model is trained using an internal validation split with early stopping to reduce overfitting and enhance prediction stability. Class imbalance is addressed through class-weight adjustments in the CatBoost model.

All models are trained and evaluated under identical experimental conditions, ensuring a fair and consistent comparison of performance across different learning approaches.

Results and Discussion

The comparative performance of the proposed TPE-optimized CatBoost model and the baseline models is evaluated using the performance metrics described in the Model Evaluation section. Table 2 summarizes the experimental results across the German Credit Risk, Polish Companies Bankruptcy, and Taiwan Credit Card Default datasets. The following discussion analyzes the

comparative performance of the models and highlights the strengths of the proposed framework under different credit risk scenarios.

In the German Credit Risk dataset, the TPE-optimized CatBoost model achieves the highest overall performance, with a ROC–AUC of 0.7644, Recall of 0.9476, and F1 score of 0.8379, outperforming all baseline models. It also yields the lowest Brier score (0.1720), indicating superior probability calibration. Although Logistic Regression exhibits a slightly higher KS statistic (0.4667), CatBoost demonstrates a better balance between discrimination and recall, making it more suitable for risk-sensitive applications. This performance can be attributed to CatBoost's ability to capture complex nonlinear relationships among borrower attributes, while TPE efficiently identifies suitable hyperparameter configurations. These findings are consistent with previous studies highlighting the effectiveness of gradient boosting and Bayesian optimization for structured financial datasets (Prokhorenkova et al., 2017; Watanabe, 2023).

In the Polish Companies Bankruptcy dataset, the effectiveness of the proposed framework is more pronounced. With a ROC–AUC of 0.9623, the TPE-optimized CatBoost model outperforms Random Forest (0.8939) and other baseline models. It also achieves the highest KS statistic (0.7908) and F1 score (0.5739), indicating a strong ability to distinguish between bankrupt and non-bankrupt firms. Furthermore, it records the lowest Brier score (0.0190), reflecting superior probability calibration. These results highlight the effectiveness of the proposed TPE-optimized CatBoost framework in handling highly imbalanced and high-dimensional datasets. CatBoost's ordered boosting mechanism, combined with TPE-based hyperparameter optimization, enables better adaptation to complex feature interactions, consistent with previous findings (Prokhorenkova et al., 2017; Watanabe, 2023).

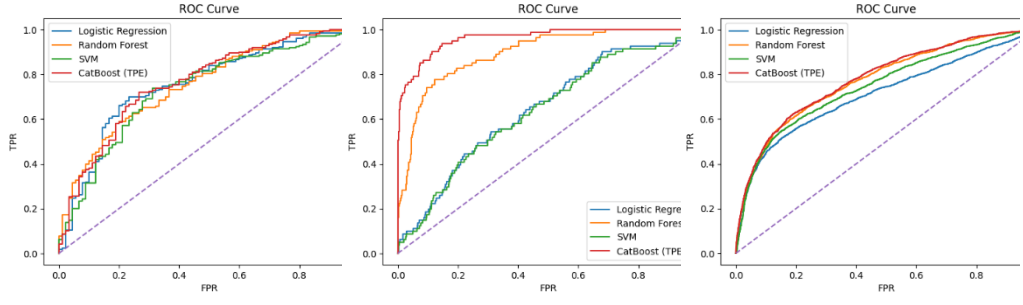
For the Taiwan Credit Card Default dataset, the TPE-optimized CatBoost model achieves the highest ROC–AUC (0.7814) and KS statistic (0.4324), indicating strong discriminatory performance. It also records the lowest Brier score (0.1350), suggesting improved calibration compared to baseline models. However, Random Forest slightly outperforms CatBoost in terms of Recall (0.5585 vs. 0.3551), indicating a trade-off between recall and overall model calibration. Despite the lower Recall, CatBoost achieves better overall discrimination and probability calibration than the baseline models. This trade-off suggests that the proposed framework provides more reliable risk estimates for practical credit risk assessment, where balanced predictive performance is essential.

Table 02. Comparative performance of models across three datasets

	Model	ROC-AUC	Recall	F1 Score	Brier Score	KS
German Credit Risk	LR	0.7549	0.6857	0.7680	0.1971	0.4667
	RF	0.7561	0.8810	0.8241	0.1740	0.3921
	SVM	0.7342	0.6810	0.7546	0.1790	0.4270
	CatBoost + TPE	0.7644	0.9476	0.8379	0.1720	0.4524

Polish Companies Bankruptcy	LR	0.6446	0.5556	0.1064	0.2295	0.2335
	RF	0.8939	0.2346	0.3167	0.0352	0.6545
	SVM	0.6372	0.5556	0.1079	0.0371	0.2206
	CatBoost + TPE	0.9623	0.4074	0.5739	0.0190	0.7908
Taiwan Credit Card Default	LR	0.7123	0.6213	0.4683	0.2085	0.3625
	RF	0.7740	0.5585	0.5415	0.1691	0.4186
	SVM	0.7470	0.5625	0.5198	0.1417	0.3961
	CatBoost + TPE	0.7814	0.3551	0.4606	0.1350	0.4324

Figure 2 illustrates the comparative performance of the evaluated models using ROC, Precision–Recall, and calibration curves across all three datasets. The ROC curves indicate that the CatBoost (TPE) model consistently achieves better separation between defaulters and non-defaulters compared to baseline models. This reflects its superior ranking capability and robustness across diverse financial datasets.



(a) (b) (c)

Figure 02. ROC curves for (a) German Credit Risk, (b) Polish Companies Bankruptcy, and (c) Taiwan Credit Card Default.

The Precision–Recall curves in Figure 3 further highlight the effectiveness of CatBoost in handling class imbalance. In particular, for the Polish Companies Bankruptcy dataset, where class imbalance is severe, CatBoost maintains substantially higher precision across a wide range of recall levels. This demonstrates its enhanced ability to detect rare default events, which is critical in real-world financial risk assessment scenarios.

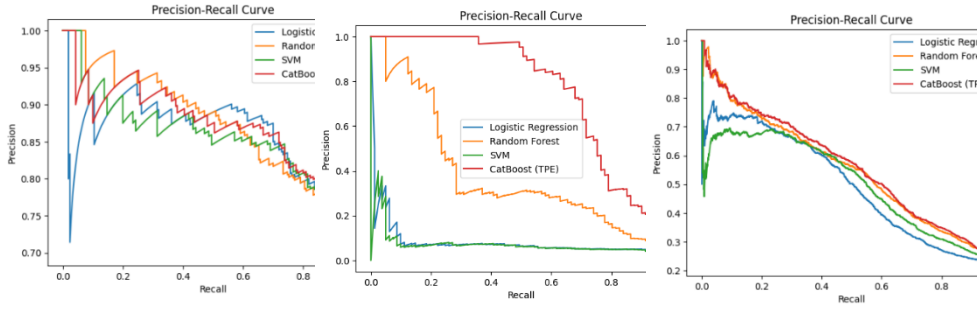


Figure 03. Precision–Recall curves for (a) German Credit Risk, (b) Polish Companies Bankruptcy, and (c) Taiwan Credit Card Default.

The calibration curves for the CatBoost model in Figure 4 indicate that the predicted probabilities closely align with the observed default rates. This observation is supported by consistently lower Brier scores across all datasets, confirming improved probability calibration. Well-calibrated probabilities are essential in financial applications, as they directly influence decision-making processes such as credit approval, pricing, and risk provisioning. Consequently, the improved calibration of the proposed model enhances its practical applicability in real-world financial environments.

The superior performance of the proposed TPE-optimized CatBoost framework can be attributed to the combination of CatBoost's ability to model complex nonlinear relationships and TPE's efficient hyperparameter optimization. These findings support previous studies demonstrating that integrating advanced boosting algorithms with Bayesian optimization can improve both predictive accuracy and probability calibration in credit risk modelling

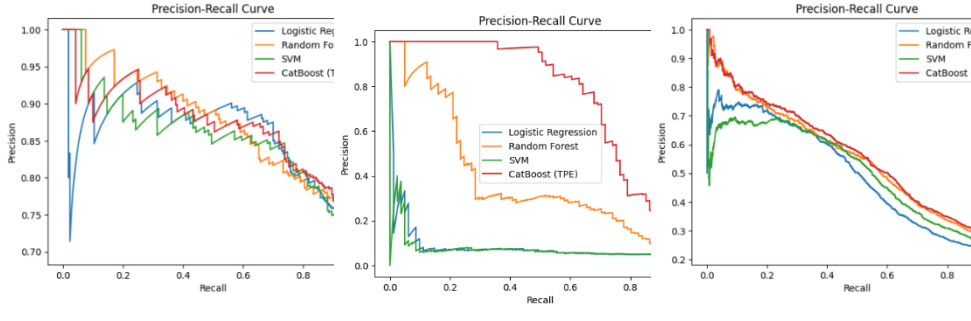


Figure 04. Calibration curves for (a) German Credit Risk, (b) Polish Companies Bankruptcy, and (c) Taiwan Credit Card Default.

The combined analysis of ROC, Precision–Recall, and calibration curves demonstrates that the proposed framework not only achieves strong predictive accuracy but also ensures reliable probability estimation. This balance between discrimination and calibration is particularly important in loan default prediction modelling, where both accurate ranking of borrowers and trustworthy probability estimates are required for informed financial decision-making.

Challenges and Future Directions

The integration of ML into risk assessment has significantly improved computational efficiency and predictive capability in loan default modelling. In particular, the proposed TPE-optimized CatBoost framework demonstrates strong performance across multiple datasets in terms of discrimination, calibration, and risk separation. However, despite these advancements, several challenges remain that may affect the practical deployment and generalization of such models. These challenges include data integrity, privacy protection, model transparency, algorithmic fairness, and governance. The increasing use of advanced AI techniques further amplifies ethical and regulatory concerns, highlighting the need for default prediction systems that are transparent, compliant, and sustainable over time.

Conventional loan default prediction models rely on large and diverse datasets encompassing financial, behavioral, and demographic information. In operational settings, access to such data is often limited by privacy legislation, data residency requirements, and organizational data silos, leading to incomplete or biased representations (Crisanto et al., 2024). Regulatory regimes such as the EU General Data Protection Regulation impose strict controls on data reuse and cross-border data transfers, necessitating the use of anonymization, or synthetic data generation. While these techniques support regulatory compliance, excessive data masking can degrade feature fidelity and adversely affect model performance.

Privacy-enhancing learning paradigms, including federated learning and differential privacy, have emerged as viable alternatives that enable decentralized model development without direct data sharing (De Oliveira & Basso, 2025). These methods introduce additional computational complexity, governance challenges, and increased reliance on distributed infrastructures. Future research is likely to focus on hybrid solutions that combine controlled anonymization with secure data-sharing mechanisms to balance privacy preservation and predictive performance.

Ensuring algorithmic fairness remains a central concern in ML–driven loan default prediction, as historical datasets may reflect entrenched social and demographic inequalities. Variables such as geographic location or employment category can unintentionally serve as proxies for protected characteristics, increasing the likelihood of discriminatory outcomes (Barocas et al., 2019). Bias mitigation strategies are commonly implemented through pre-processing, in-processing, or post-processing interventions designed to reduce unfairness at different stages of the modelling pipeline (Kozodoi et al., 2022).

Despite progress in fairness-aware modelling, achieving transparency remains difficult for complex algorithms, including ensemble and boosting-based approaches such as CatBoost. Although explainable AI methods, including SHAP and LIME, offer partial solutions by elucidating feature contributions to model predictions, they may not fully resolve regulatory concerns regarding model accountability. Ethical AI governance must extend beyond

interpretability tools to encompass accountability mechanisms, audit trails, and effective human oversight throughout the model lifecycle.

While the proposed TPE-optimized CatBoost framework demonstrates strong predictive performance, certain limitations remain, particularly in handling extreme class imbalance and maintaining high recall across all datasets. For example, the observed variation in recall performance across datasets suggests that model sensitivity to minority class detection can vary depending on data characteristics. Future research may therefore explore cost-sensitive learning strategies, adaptive thresholding, and advanced resampling techniques to further improve default detection performance.

Further advancements in hyperparameter optimization, including hybrid Bayesian and meta-learning approaches, may enhance model efficiency and robustness. Research may also focus on improving model interpretability within boosting frameworks to better align predictive performance with regulatory requirements. Ultimately, the long-term success of AI-enabled credit scoring systems depends on the development of models that are not only accurate but also transparent, fair, and adaptable to evolving financial environments.

Conclusion

This study presents an efficient loan default risk assessment framework based on a TPE-optimized CatBoost model for improving predictive accuracy and reliability in credit risk modelling. By integrating Bayesian hyperparameter optimization with CatBoost, the proposed framework effectively captures nonlinear relationships and demonstrates strong generalization across the German Credit Risk, Polish Companies Bankruptcy, and Taiwan Credit Card Default datasets. The experimental results show that the proposed model consistently achieves superior performance in terms of ROC–AUC, F1 score, Brier score, and KS statistic while maintaining competitive recall, highlighting its effectiveness for credit risk prediction.

Beyond improving predictive accuracy, the proposed framework advances current research on credit risk modelling by systematically integrating TPE-based Bayesian hyperparameter optimization with CatBoost and evaluating its performance using discrimination, calibration, and risk-separation metrics across multiple benchmark datasets. Unlike many previous studies that primarily emphasize classification accuracy, this study demonstrates the importance of combining efficient hyperparameter optimization with reliable probability estimation for practical credit risk assessment. Consequently, the proposed framework contributes to the development of more robust and generalizable ML approaches for data-driven financial decision-making.

From a practical perspective, the proposed framework can support financial institutions in improving borrower risk segmentation, credit approval, and risk management through more accurate and well-calibrated default risk estimation. Although the proposed framework demonstrates strong predictive performance, its effectiveness may vary depending on dataset characteristics. Future research should validate the framework using larger real-world credit datasets, investigate advanced optimization and cost-sensitive learning techniques, and integrate explainable AI methods to further enhance transparency, robustness, and practical applicability in financial decision-making.

References

Ahelegbey, D. F., & Giudici, P. (2023). Credit Scoring for Peer-to-Peer Lending. *Risks*, *11*(7), 123. https://doi.org/10.3390/risks11070123

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), e39770. https://doi.org/10.1016/j.heliyon.2024.e39770

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*.

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

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), 174. https://doi.org/10.3390/risks12110174

Crisanto, J. C., Leuterio, C. B., Prenio, J., & Yong, J. (2024). *Regulating AI in the Financial Sector: Recentdevelopments and Main Challenges*. Bank for International Settlements, Financial Stability Institute.

De Oliveira, N. A., & Basso, L. F. C. (2025). Advancing Credit Rating Prediction: The Role of Machine Learning in Corporate Credit Rating Assessment. *Risks*, *13*(6), 116. https://doi.org/10.3390/risks13060116

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Glenn, W. B. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

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

Jemai, J., & Zarrad, A. (2023). Feature Selection Engineering for Credit Risk Assessment in Retail Banking. *Information*, *14*(3), 200. https://doi.org/10.3390/info14030200

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, *297*(3), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023

Lee, J. Y., Yang, J., & Anderson, E. (2024). *Who Benefits from Alternative Data for Credit Scoring? Evidence from Peru*. SSRN. https://doi.org/10.2139/ssrn.4852032

Levy, A., & Baha, R. (2021). Credit risk assessment: A comparison of the performances of the linear discriminant analysis and the logistic regression. *International Journal of Entrepreneurship and Small Business*, *42*(1/2), 169. https://doi.org/10.1504/IJESB.2021.112265

Malempati, M. (2025). The evolution of credit data and the role of machine learning in modern credit scoring. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5330882

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: Unbiased boosting with categorical features* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1706.09516

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., Shpitser, I., & Taylor, C. O. (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. *PLOS ONE*, *15*(4), e0231300. https://doi.org/10.1371/journal.pone.0231300

Roy, J. K., & Vasa, L. (2025). Transforming credit risk assessment: A systematic review of AI and machine learning applications. *Journal of Infrastructure Policy and Development*, *9*(1), 9652. https://doi.org/10.24294/jipd9652

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Soni, U., Jethava, G., & Ganatra, A. (2024). Latest Advancements in Credit Risk Assessment with Machine Learning and Deep Learning Techniques. *Cybernetics and Information Technologies*, *24*(4), 22–44. https://doi.org/10.2478/cait-2024-0034

Wang, Z. (2024). Artificial Intelligence and Machine Learning in Credit Risk Assessment: Enhancing Accuracy and Ensuring Fairness. *Open Journal of Social Sciences*, *12*(11), 19–34. https://doi.org/10.4236/jss.2024.1211002

Watanabe, S. (2023). *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2304.11127