Application of Explainable AI and Uncertainty Quantification in Credit Risk Assessment

Mulavhelesi Rambauli , Thakhani Ravele * , Caston Sigauke

Posted Date: 30 April 2026

doi: 10.20944/preprints202604.2092.v1

Keywords: credit risk; interpretability; LIME; machine learning; SHAP; uncertainty quantification

 doi:10.20944/preprints202604.2092.v1

Article

Application of Explainable AI and Uncertainty Quantification in Credit Risk Assessment

Mulavhelesi Rambauli, Thakhani Ravele * and Caston Sigauke

Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou 0950, Limpopo, South Africa
* Correspondence: thakhani.ravele@univen.ac.za

Abstract

Credit risk modelling is essential for assessing the likelihood of borrower default and supporting informed lending decisions. Despite advances in predictive algorithms, challenges remain in ensuring model transparency, reliability, and robustness to uncertain inputs. This study investigates integrating explainable AI (XAI) and uncertainty quantification (UQ) to enhance interpretability and confidence in credit risk predictions. Three modelling approaches, Logistic Regression, Random Forest, and XGBoost, were evaluated using the Home Equity (HMEQ) dataset, with performance assessed on predictive accuracy, probability calibration, interpretability, and uncertainty handling. Ensemble methods achieved superior predictive performance, exceeding 98% accuracy and yielding near-perfect AUC scores above 0.999, whereas Logistic Regression exhibited substantially lower performance. Calibration analysis revealed a discrepancy between accuracy and probabilistic reliability: Random Forest, despite high accuracy, produced less well-calibrated predictions (ECE = 0.0475), while XGBoost achieved both strong predictive performance and reliable confidence estimates (ECE = 0.0117). Entropy-based uncertainty quantification identified instances where the model's predictions were highly uncertain, effectively highlighting challenging cases. SHAP and LIME consistently identified DELINQ, DEROG, and DEBTINC as primary drivers of default risk, aligning with established financial risk logic. By combining SHAP, LIME, and entropy-based UQ, this study proposes a unified framework that enhances interpretability, supports regulatory compliance, and increases trust in automated lending systems, emphasising the importance of reliable confidence alongside predictive accuracy.

Keywords: credit risk; interpretability; LIME; machine learning; SHAP; uncertainty quantification

1. Introduction

1.1. Overview

The global financial ecosystem has undergone a profound paradigm shift over the past few decades, driven by rapid technological advances and the growing availability of large-scale data sources. These developments have fundamentally transformed traditional banking practices, reshaping how financial institutions operate and compete within an increasingly digital environment [2,3]. Artificial intelligence (AI), particularly machine learning (ML), is accelerating this shift, reshaping standard frameworks for risk evaluation, client support, fraud detection, and market analysis. Thus, finance sectors have undergone significant gains in how smoothly things run, better choices made, and more precise outcomes in interaction with customers throughout all stages of the value process [10,11].

Credit risk models using machine learning often achieve high accuracy but lack transparency and reliable measures of uncertainty, which can affect trust and regulatory compliance in lending decisions. Existing research largely treats prediction, explainability, and uncertainty separately. This study addresses this gap by investigating a unified approach that integrates explainable AI and uncertainty quantification to improve the transparency, calibration, and reliability of credit risk predictions.

 doi:10.20944/preprints202604.2092.v1

1.2. Literature Review

1.2.1. The Evolution of Credit Risk Assessment Models

The modelling of credit risk - the possibility of losses when borrowers do not repay loans - has changed significantly in recent decades. In earlier times, evaluating this risk relied on human judgment and fixed rules. As data availability and computational power increased, statistical methods gained prominence. Logistic regression (LR), for example, became common early on because it was straightforward and easy to understand [7,23]. Although useful for small data sets and standard financial indicators, these models struggled with complex variable relationships or curved trends.

Subsequent advances introduced discriminant analysis and probit models, which offered incremental improvements but maintained restrictive parametric assumptions. Although limited, these approaches formed the foundation for credit scoring systems that became standard in financial institutions during the 1980s and 1990s. One major drawback of such approaches was their reliance on handpicked variables and fixed distributional assumptions; this often led to skewed or less effective forecasts in real-world applications [6].

These limitations paved the way for more flexible, data-driven techniques in the early 2000s. Pioneering applications of Support Vector Machines (SVMs) demonstrated their potential to capture complex, non-linear relationships in credit data. [53] showed that hybrid SVM approaches with genetic algorithm-based feature selection could match or exceed neural network performance on benchmark credit datasets, establishing SVMs as viable alternatives to traditional methods. Similarly, ensemble methods gained prominence as researchers demonstrated that combining models could improve predictive performance. [54] conducted comparative assessments and found that techniques such as bagging, boosting and stacking significantly enhanced the performance of individual learners, with bagging often outperforming boosting in credit-scoring applications. Neural network ensembles also proved effective; [55] showed that the Random Subspace approach combined with multilayer perceptrons consistently outperformed single classifiers across Australian, German and Japanese credit datasets.

However, these early studies revealed a critical insight: no single algorithm dominated universally. [56] tested multiple algorithms across five different datasets and found that model rankings changed substantially with dataset characteristics, advocating for cross-dataset evaluation rather than single-benchmark claims. reinforced this finding citebrown2012experimental, who investigated classifier performance under progressive class imbalance and discovered that RF and XGboost maintained robustness under severe imbalance, while traditional classifiers (C4.5, QDA, kNN) degraded substantially. The most comprehensive benchmark came from [9], who evaluated state-of-the-art classifiers across multiple real-world credit datasets and concluded that algorithmic superiority is dataset- and context-dependent, establishing the need for rigorous, multi-dataset evaluation protocols that remain standard practice today.

These foundational studies highlighted three important insights for credit risk modelling: (1) ensemble methods typically deliver better performance than single classifiers, particularly in situations with imbalanced classes; (2) the effectiveness of a model depends heavily on its specific context, necessitating validation for each institution; and (3) increasing model complexity intensifies the trade-off between predictive accuracy and interpretability.

1.2.2. Machine Learning in Credit Risk Assessment

Machine learning has substantially improved predictions and decision-making in credit risk evaluation. While LR remains widely used for its interpretability, methods like RF and XGBoost bring distinct advantages alongside important trade-offs. Many contemporary studies favour ensemble approaches, which often deliver stronger results under diverse testing conditions [39,58].

XGBoost handles complex datasets effectively by handling missing values and categorical variables through specialised procedures, and often achieves higher precision and F1 Scores, particularly in imbalanced credit risk datasets when combined with techniques such as SMOTE [40]. However,

 doi:10.20944/preprints202604.2092.v1

its limited transparency complicates model validation and regulatory compliance. RF also demonstrates strong predictive performance, resists overfitting, and provides interpretability through feature importance measures [48,49]. However, a critical weakness persists: as [25] note, these performance gains come at the cost of interpretability, creating a "black box" problem that concerns regulators requiring transparency and fairness. LR, while often underperforming ensemble methods, offers clear, interpretable coefficients that support regulatory compliance and improve stakeholder trust [49,59]. Evidence from large-scale studies further shows that ensemble approaches remain robust across regions and datasets, whereas LR continues to serve as a transparent and reliable baseline [60].

Despite the widespread adoption of XGBoost, RF, and LR, their comparative performance reveals important trade-offs that challenge the notion of a universally superior method. Large-scale benchmarking indicates that no single classifier consistently dominates across datasets, although ensemble methods—particularly gradient boosting and RF, tend to perform more reliably [9]. Under conditions of severe class imbalance, which are common in credit risk modelling, ensemble techniques maintain stability while traditional classifiers deteriorate. A key tension therefore emerges: XGBoost and RF provide stronger predictive accuracy but at the expense of transparency and regulatory interpretability, whereas LR supports compliance through its explainable structure despite weaker predictive performance. This trade-off is not merely technical but institutional: banks must balance predictive performance against regulatory requirements for explainability.

1.2.3. Empirical Evidence Across International Contexts

The performance of ML methods varies substantially across different institutional, geographic, and economic contexts, highlighting the importance of context-specific validation rather than relying solely on benchmark datasets.

Research from China, North Africa, the Middle East and Nordic banking systems demonstrates that while advanced methods such as support vector machines, neural networks and gradient-boosting models often improve classification accuracy, their success is highly dependent on the availability of relevant financial variables and the characteristics of institutional datasets [61–64]. Many studies rely on single-bank data and in-sample evaluations, which may produce optimistic results that do not generalise across institutions or regions. Furthermore, findings indicate that feature engineering and domain-specific variables can be as important as algorithm selection in improving predictive performance.

The increasing adoption of advanced ML in finance has led to what [12] called the "black box issue", where systems perform well yet their internal logic stays unclear. When assessing credit risk, missing clarity creates serious issues. Banks now use complex methods such as deep learning and ensemble models to score creditworthiness; these boost predictive quality; however, [22] challenges this position as overly restrictive. They note that modern ML models substantially outperform traditional methods in credit risk prediction, and that rejecting them entirely sacrifices predictive accuracy that could benefit both lenders and borrowers through better risk assessment and potentially more inclusive lending.

They suggest that robust post-hoc explanation methods, combined with rigorous validation, can provide sufficient transparency for regulatory compliance without abandoning high-performance models. Compliance matters greatly: regulations, including the U.S. Equal Credit Opportunity Act and the General Data Protection Regulation (GDPR), require transparency in algorithmic decisions. In particular, GDPR's Article 22 grants people the right to understand automated decisions [13], highlighting why transparent AI tools are urgently needed across financial services.

1.2.4. Explainable AI in Credit Risk

XAI helps improve AI-based financial tools by making their choices clearer, reducing confusion caused by opaque models [4]. Using ML in tightly supervised areas brings difficulties - these systems rarely show how they arrive at outcomes [1]. Without clear insight, users struggle to grasp results, meeting legal rules becomes harder, and confidence in automation may drop.

 doi:10.20944/preprints202604.2092.v1

To address these problems, XAI tools are gaining more attention in evaluating credit risk. While local post hoc techniques like LIME [5] and SHAP [8] are widely used, their roles vary across contexts. For instance, SHAP values help clarify neural network outputs on mortgage defaults [1]; at the same time, game-theoretic models support the interpretation of consumer scores [18]. However, challenges persist despite this progress. Some XAI approaches give reasons without measuring prediction uncertainty - or even uncertainty in the explanations [19,20]. Also, similar cases may be interpreted differently, weakening reliability. Few XAI methods in credit scoring align well with principles such as the "right to explanation," often delivering results that lack clarity or depth for users to truly grasp [21]. Still, making models both accurate and easy to understand remains difficult - some prioritise simplicity over power, others focus on precision but hide how decisions are made.

[19] and [20] identify a critical weakness: similar cases may receive different explanations, undermining reliability. If two borrowers with nearly identical profiles receive different explanations for the same decision, the explanations fail to support consistent, fair treatment—a core regulatory requirement.

1.2.5. Uncertainty Quantification in Credit Decisions

Beyond just explaining results, measuring uncertainty addresses a key flaw in credit scoring: standard models do not show how confident they are in their forecasts. Instead of fixed guesses, these methods separate clear-cut cases from unclear ones, helping banks prepare for unexpected risks [14]. Studies suggest that incorporating uncertainty into loan portfolio management improves performance, especially during market declines. When predictions carry high doubt, institutions may choose stricter actions - for example, human evaluation - to manage possible defaults [15]. For this to work, the model's probability outputs must match real-world outcomes; without calibration, uncertainty signals lose meaning.

UQ is now key in machine learning systems, especially where risks matter - like credit evaluation. Different approaches exist, each with strengths and clear drawbacks. Bayesian Neural Networks offer a solid foundation for handling uncertainty due to limited knowledge; yet they demand heavy computation, making them hard to use in large-scale financial settings.To address this issue, lighter techniques such as Monte Carlo Dropout [16] have gained popularity. While faster and easier to run, these workarounds tend to produce overly confident predictions when data changes substantially, reducing their trustworthiness over time as lending conditions shift.

A more data-driven option comes from Deep Ensembles, which merge several separately trained networks to reflect diverse prediction patterns [17]. Rather than relying on single models, this approach provides better reliability and uncertainty estimates than Bayesian Neural Networks or Monte Carlo techniques. Because of that, it is increasingly used in practical applications. Similar shifts have happened in credit risk analysis, where group-based strategies like RF or boosted trees are preferred; they handle messy, uneven data well while scaling efficiently.

A major benefit of ensemble methods is their capacity to produce probability-based results, allowing a clear assessment of uncertainty. While predictive entropy is a common measure, it assesses uncertainty by analysing the output probability distribution. Research such as [45] highlights its usefulness for detecting financial fraud: entropy levels often match confidence and accuracy trends: lower values suggest more confident predictions, while higher values signal doubt. Such findings reinforce the idea that entropy can reliably reflect uncertainty.

Yet findings agree that entropy makes sense mainly if the model is properly calibrated. If not adjusted, systems tend to produce overly confident predictions, resulting in misleadingly low entropy that masks real uncertainty. Research points out this flaw: [50] finds that Bayesian approaches may yield inaccurate confidence ranges without direct calibration steps. According to [45], accurate uncertainty assessment depends on prior calibration and the use of ECE as an evaluation tool. In parallel, [51] reports that methods such as temperature tuning or label smoothing lower miscalibration and boost prediction trustworthiness; meanwhile, [52] claims that calibration and uncertainty measurement are deeply linked, forming a core requirement for dependable outcomes.

 doi:10.20944/preprints202604.2092.v1

Though entropy-based UQ with calibration and ensembles works well for credit scoring, some concerns appear in studies. One, these ensemble techniques demand more computing power and storage than single models, creating hurdles for banks with limited resources. Instead, entropy captures overall uncertainty but fails to separate knowledge-driven from randomness-driven errors, making analysis harder than with structured Bayesian methods. Also, even if calibrated outputs yield better probability estimates, they do not automatically make models easier to explain - a key need under strict lending regulations.

1.3. Summary of Literature Review

The literature on XAI and UQ in credit risk assessment indicates that ensemble methods, such as RF and XGBoost, generally achieve higher predictive accuracy than traditional statistical models, particularly under class imbalance. However, performance depends on dataset characteristics and context. No single model consistently performs best, and effectiveness varies with institutional, dataset, and economic factors, necessitating context-specific validation. While complex models pose interpretability challenges, XAI methods such as SHAP and LIME provide post-hoc explanations, mitigating the accuracy–interpretability trade-off, though they have methodological limitations. Uncertainty quantification can be achieved through ensemble methods and calibration, enabling probabilistic predictions for improved risk management.

Despite these advances, current research shows several limitations. XAI methods are prone to misleading feature attributions, perturbation artefacts, and instability, with no standardised evaluation for regulatory compliance. Empirical validation is often restricted to single-institution or small datasets, and issues of fairness, bias, and demographic disparities remain underexplored. XAI and UQ are rarely integrated, leaving a gap in methods that jointly provide prediction uncertainty and explanation confidence. Calibration is frequently neglected, undermining the reliability of uncertainty estimates. Overall, research remains fragmented and methodologically fragile, emphasising the need for robust, integrated frameworks that deliver accurate predictions, calibrated uncertainty estimates, and trustworthy explanations across diverse contexts.

This study addresses the identified research gaps through three main contributions. First, it implements and compares multiple XAI methods on a common loan default dataset, explicitly evaluating explanation stability and handling of feature dependencies. Second, it develops an integrated framework that combines ensemble-based predictions with uncertainty quantification and explainability, enabling both reliable predictions and interpretable insights. Third, the study validates the framework's performance in terms of predictive accuracy and calibration quality, providing evidence of its robustness and practical applicability in credit risk assessment.

The literature review is summarised in Table 1. The progression from judgment-based techniques to ML approaches is highlighted, with an emphasis on ensembles which perform better than individual classifiers. Major limitations identified include nonlinearity, opacity of black-box models, non-generalizable data, inconsistency across XAI techniques, and the absence of UQ integration in XAI.

 doi:10.20944/preprints202604.2092.v1

Table 1. Summary of literature review.

Theme	Key Findings	Methodologies	Gaps & Limitations	Cited Examples
1. Evolution of Credit Risk Models	Shift from judgment to ML. No single algorithm dominates. Ensembles outperform single classifiers.	LR, SVMs, NNs, Bagging, Boosting, RF, XGBoost	Early models struggle with non-linearity. Rankings change across datasets.	Altman (1968); Huang (2007); Lessmann (2015)
2. ML in Credit Risk	LR: interpretable but weaker. RF/XGBoost: accurate but black box. Institutional trade-off.	LR, RF, XGBoost, SMOTE	Transparency loss complicates compliance.	Li (2020); Blessing (2024); Trinh (2024)
3. International Evidence	Performance varies by context. Feature engineering key. Post-hoc + validation may suffice.	SVMs, NNs, Gradient Boosting, SHAP, LIME	Single-bank data non-generalizable. GDPR requires transparency.	Chen (2009); Bracke (2019); Goodman (2017)
4. XAI in Credit Risk	Improves clarity, but inconsistent explanations across similar cases undermine fairness.	LIME, SHAP, game-theoretic models	No uncertainty measurement. Poor alignment with right-to-explanation.	Nallakaruppan (2024); Alkhyeli (2023)
5. UQ in Credit Decisions	Separates ambiguous cases. Deep Ensembles best. Entropy needs calibration.	BNNs, MC Dropout, Deep Ensembles, ECE	BNNs are heavy. Entropy cannot separate uncertainty types.	Lakshminarayanan (2017); Habibpour (2023)
6. Research Gap	Prediction, XAI, UQ treated separately. This study integrates them for transparent, reliable credit risk predictions.	Unified: XAI + UQ + modelling	No existing integrated framework.	Adadi (2018); Gomber (2018)

1.4. Contributions and Research Highlights

The current study shows that there is much more to creating a good AI system for credit scoring than simply building a prediction model, since it must deliver calibrated probabilities, explanations and uncertainty estimates (e.g., based on entropy). The strongest ground for developing a good AI system appears to be the XGBoost algorithm.

The research highlights of this study are:

- Ensemble models significantly exceed the performance of logistic regression. The ROC AUC of both Random Forest (0.9992) and XGBoost (0.9991) surpasses 98% accuracy and an F1-score of >0.985, while logistic regression attains merely 77.4% accuracy and generates 388 misclassifications (against <60 by ensembles).
- The most informative features for the default class include DELINQ, DEROG, and DEBTINC. On average, defaulters exhibit 1.23 instances of delinquency, 0.71 derogatory credit items, and a DEBTINC ratio of 39.39%, compared with 0.25, 0.13, and 33.25% for non-defaulters, respectively.
- XGBoost provides the best calibration (ECE = 0.0117) compared with Random Forest (ECE = 0.0475, overconfidence) and logistic regression (ECE = 0.0330, despite low accuracy).
- Predictive entropy highlights that correct predictions tend to lie close to zero entropy, whereas misclassifications happen at higher entropy values (e.g., 0.4-0.7 in the case of logistic regression

The rest of the paper is as follows. Section 2 will cover the methodology approach and the prediction models. Empirical findings are presented in Section 3, while Section 4 covers the discussion of he results. The conclusion is covered in Section 5.

 doi:10.20944/preprints202604.2092.v1

2. Methodology

This section outlines the methodological framework used to address the study's core objective: examining how explainable AI and uncertainty quantification can enhance the transparency and reliability of credit risk prediction. These choices were informed by a gap in the literature, where predictive performance is typically prioritised while interpretability and uncertainty are treated as secondary concerns. Three models were selected to allow comparison across varying levels of complexity. Logistic Regression served as an interpretable baseline in line with regulatory expectations, while Random Forest and XGBoost were chosen for their ability to handle nonlinear relationships and improve predictive accuracy. This combination enables the study to assess whether increased complexity comes at the cost of interpretability and reliability. SHAP and LIME were selected as complementary explainability techniques, offering both global and local insights into model behaviour. Entropy and Expected Calibration Error were used to quantify predictive uncertainty and evaluate the accuracy of probability estimates.

Together, these choices provide a structured framework for jointly evaluating predictive performance, interpretability, and uncertainty in credit risk modelling.

2.1. Data Source

The study uses a panel dataset of 50,000 U.S. home loan applicants, tracked over 60 periods, including origination and repayment outcomes. The data are publicly available at https://www.listendata.com/2019/08/datasets-for-credit-risk-modeling.html?m=1 (accessed on 13 February 2025).

2.2. Data Preprocessing

Before model training, the dataset was prepared using domain-specific preprocessing steps suitable for credit risk analysis. Numerical features with missing values — such as MORTDUE, VALUE, YOJ, DEROG, DELINQ, CLAGE, NINQ, CLNO, and DEBTINC — were imputed using mean values, while categorical features such as REASON and JOB were imputed using the most frequent category. To address the class imbalance between default and non-default cases in the BAD target variable, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training set, generating synthetic examples of the minority class and producing a balanced dataset for model learning. Categorical variables such as REASON (e.g., HomeImp) and JOB (e.g., Other, Office, Sales, Mgr) were transformed using one-hot encoding to convert them into a suitable numerical format for the classification models. These preprocessing steps ensured that the dataset was complete, balanced, and properly formatted, supporting reproducibility and reliable predictive performance in credit risk assessment.

2.3. Models

The following section presents the three classification models employed in this study for predicting loan default. Each model is described in detail, including its underlying principles, operational mechanisms, and how it is applied to the dataset to distinguish between default and non-default borrowers.

2.3.1. Random Forests

The development of RF originated from earlier work on decision tree methodology. [27] first introduced decision trees as a statistical classification tool, which [26] later advanced by proposing the RF ensemble learning technique. The RF method integrates two foundational approaches: Breiman's bootstrap aggregating (bagging) and Ho's random subspace method, and its construction involves three sequential stages [30,31].

Bagging produces m distinct training samples, denoted as S_j for $j = 1, 2, \ldots, m$, each drawn with replacement from the original dataset and containing the same number of observations as the full dataset. The random subspace method is then applied to construct m decision trees from these samples, represented as $\{f(Z, \Phi_j), j = 1, 2, \ldots, m\}$, where Z is the feature vector for classification, and Φ_j is an

 doi:10.20944/preprints202604.2092.v1

independent random variable introducing diversity into each tree. The overall prediction is obtained by aggregating the outputs of all m trees through majority voting, expressed as:

$$F(z) = \arg\max_{C} \sum_{j=1}^{m} I(f_j(z) = C),$$

where $F(z)$ denotes the ensemble's final prediction for input z, $f_j(z)$ represents the prediction of the j-th individual tree, C indicates a candidate class label (e.g., default or non-default), and $I(\cdot)$ stands for an indicator function - its value is 1 if the condition holds, 0 in any other case. The expression picks the class C that gets most votes across the m decision trees.

In credit risk analysis, every decision tree learns from a randomly picked group of borrowers along with a random selection of financial factors. Because of this variability, the chance of overfitting drops while the system better picks up varied customer patterns and complex links between money-related traits [9]. Thanks to its combined structure, RF achieves strong prediction accuracy and resists errors caused by messy data - useful when handling uneven, mixed datasets common in predicting loan defaults [39].

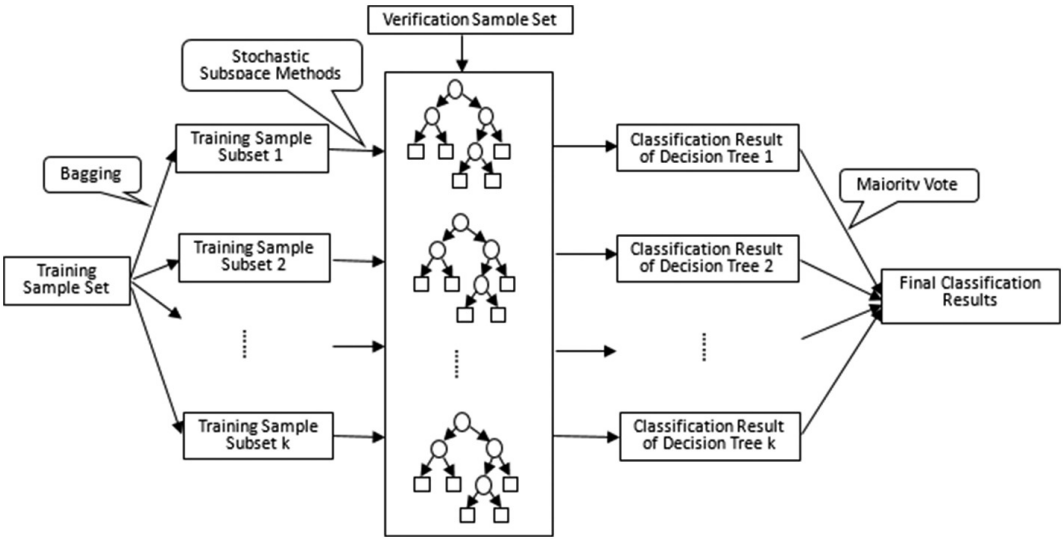


Figure 1. Primary operational framework of Random Forest in credit risk assessment. (Source: [32]).

Beyond prediction, RF includes ways to assess feature relevance - helping spot main drivers of default risk, improving clarity [24]. Such assessments matter to bankers and supervisors because they show which aspects - like debt usage, repayment patterns, or borrowing period - affect default likelihood the most. Yet, while powerful, RF can demand high computing resources on big credit records; also, it's less clear than simpler methods like LR [38]. Still, due to its mix of precision and consistency, RF remains widely used in modern credit scoring systems.

2.3.2. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) improves standard gradient boosting by combining gradient descent with second-order Taylor expansions, increasing speed and accuracy [33,34]. Instead of working independently, trees are added step-by-step - each one targeting mistakes made earlier. Within this setup, new models focus on leftover errors ignored so far. Because updates happen sequentially and adaptively, the method performs well in classification, forecasting, and ordering problems. Unlike RF - which uses vote-based outcomes - this system combines outputs through weighted tree contributions (illustrated in Figure 2):

$$\hat{y}_i = \sum_{m=1}^{M} g_m(w_i), \quad g_m \in \mathcal{G},$$

 doi:10.20944/preprints202604.2092.v1

where $\mathcal{G}$ is the set of tree-like functions, g_m stands for one specific tree, $g_m(w_i)$ gives the result from the i-th tree, while $\hat{y}_i$ shows the forecasted value for input w_i.

XGBoost is a flexible tree-based system built to handle big machine learning tasks. Because it works well in many real-world cases, data scientists have widely adopted it [35]. Instead of just fitting data tightly, the method uses a penalised goal that blends prediction error with complexity control - this helps avoid overlearning. It builds trees step by step through gradient boosting, using quadratic approximation for faster, more precise updates [19,34]. The objective function that XGBoost seeks to minimise is expressed as:

$$\mathcal{O}(\psi) = \mathcal{C}(\psi) + \mathcal{P}(\psi),$$

where $\mathcal{C}(\psi) = \sum_{n=1}^{N} L(a_n, \hat{r}_n)$ constitutes the cost function evaluating the difference between the predicted outcome $\hat{r}_n$ and the observed value a_n, whereas $\mathcal{P}(\psi) = \sum_{j=1}^{J} \mathcal{P}(f_j)$ represents a regularisation component that limits structural complexity. The learning algorithm operates through successive augmentation. Defining $\hat{r}_n^{(j)}$ as the forecast for the n-th sample at the j-th stage, the iterative update is given by:

$$\hat{r}_n^{(j)} = \hat{r}_n^{(j-1)} + f_j(z_n).$$

XGBoost was used here because it works well for credit scoring, while also balancing prediction accuracy with simplicity [36]. Its gradient boosting structure helps detect small trends in borrowing habits; at the same time, built-in regularisation improves reliability on new data [37].

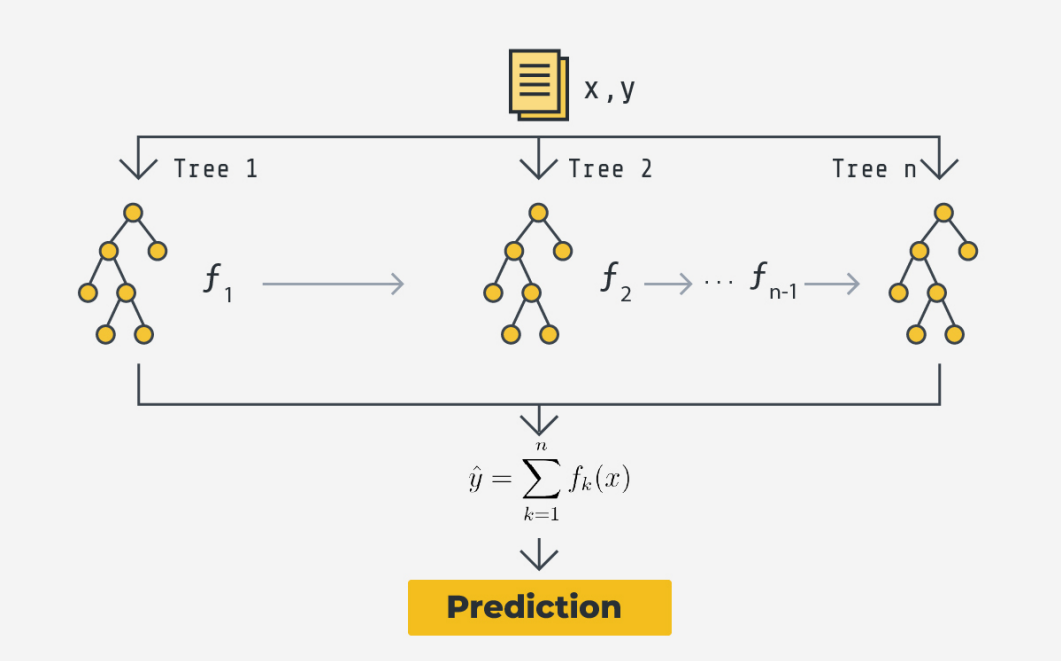


Figure 2. Primary operational framework of XGBoost.

2.3.3. Logistic Regression

Logistic regression (LR) serves as the baseline method because it works well when outcomes fall into two categories [41]. Instead of raw sums, LR uses the sigmoid function to convert weighted inputs into values between 0 and 1 - these represent probabilities. Parameters get estimated via maximum likelihood, a technique that picks values most likely to reproduce the observed data while maintaining statistical reliability. Unlike complex models, LR offers clarity: its coefficients show how strongly - and in what direction - each variable affects the result [42]. This transparency makes LR particularly useful in applications requiring explainable decision-making, which is essential in regulated domains. Its strength, consistency plus ease of use mean logistic regression works well as a baseline when compared to advanced techniques. The usual steps for using it are shown in Figure 3.

 doi:10.20944/preprints202604.2092.v1

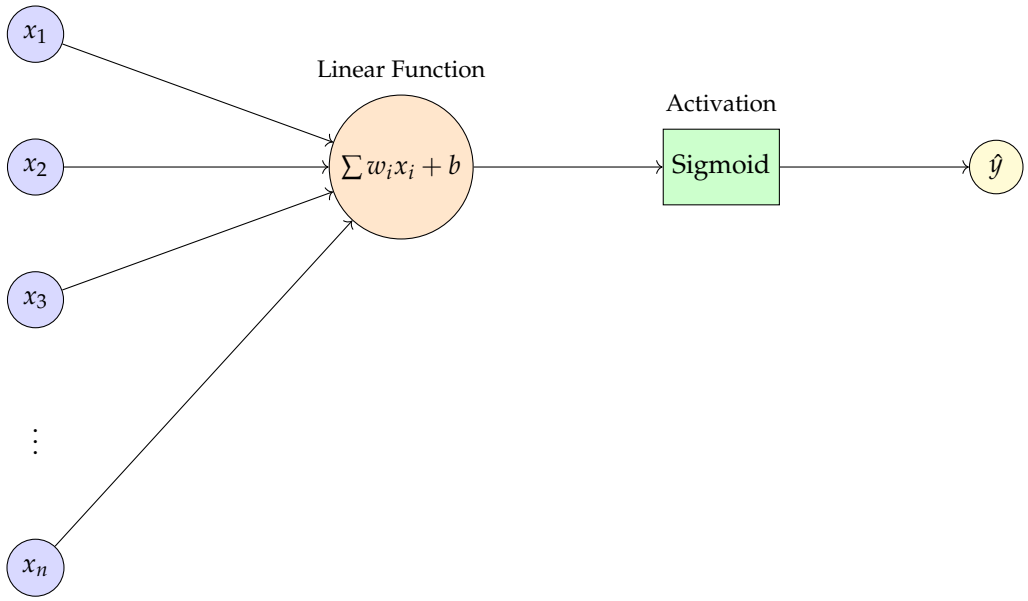


Figure 3. Primary operational framework of Logistic Regression.

2.4. Model Development and Implementation

The study employs supervised machine learning, where algorithms learn predictive models from datasets containing both explanatory variables and corresponding labels. Following feature selection, the dataset was split 80/20 for training and testing. Three algorithms—XGBoost, RF, and LR were trained on the training set. Hyperparameters were optimised via 5-fold cross-validated grid search, selecting the configuration that maximised ROC-AUC performance.

Model Selection: The three models were compared using ROC-AUC, F1-score, ECE, computational efficiency, and interpretability. This framework ensured that the selected model achieved high predictive accuracy while providing reliable probability estimates and transparent reasoning, supporting practical decision-making in credit risk assessment.

2.5. Explainable AI Approaches

2.5.1. Local Interpretable Model-Agnostic Explanations

LIME helps explain how ML models make decisions - making their outputs easier to understand. Introduced by [5], it tackles the issue of opaque models that act like "black boxes." Rather than analysing the full system, LIME focuses on one prediction at a time. For each case, it builds a simpler model near that specific input point. By tweaking inputs slightly, it gathers predictions from the original model on these modified examples. Then, using those results - with more weight given to similar cases - it trains a transparent approximation [43]. The surrogate model subsequently identifies the features that had the greatest impact on the prediction. For a given instance x, the LIME explanation is defined as: The LIME explanation for an instance x is defined as

$$\phi(x) = \arg\min_{g \in \mathcal{G}} \mathcal{L}(M, g, w_x) + \mathcal{C}(g), \tag{1}$$

as proposed by Gramegna and Giudici [43]. where:

- $M : \mathbb{R}^p \to \mathbb{R}$ is the black-box model,
- $\mathcal{G}$ is the set of interpretable surrogate models,
- $g \in \mathcal{G}$ is a candidate surrogate,
- $w_x(z)$ is a proximity kernel giving higher weight to points z near x,
- $\mathcal{L}(M, g, w_x)$ is the locality-weighted loss, measuring how well g approximates M locally,
- $\mathcal{C}(g)$ is a complexity penalty to ensure interpretability.

LIME explanations were generated for the same instances using the `lime_tabular` module with default parameters to compare local fidelity

 doi:10.20944/preprints202604.2092.v1

2.5.2. Shapley Additive Explanations

SHAP is a unified framework for interpreting machine learning model predictions based on game theory. Introduced by [8], it connects several existing explanation methods and provides a solid theoretical foundation using Shapley values from cooperative game theory. SHAP assigns each feature a Shapley value, representing its average marginal contribution to a particular prediction across all possible feature combinations. In essence, SHAP quantifies how much each feature pushes the prediction away from the base value (the average prediction), providing a fair distribution of the prediction among features and delivering both local and global explanations [44]. Mathematically, the SHAP value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)], \tag{2}$$

as proposed by Gramegna and Giudici [43].

where S is a subset of features excluding i, F is the set of all features, and f is the model prediction function. The key properties and advantages of SHAP values are:

- Local accuracy: the sum of all SHAP values equals the difference between the model's prediction and the average prediction.
- Missingness: variables absent from the model get a SHAP score of zero.
- Consistency: when a model shifts so that one feature affects the outcome more, its SHAP value either stays the same or goes up - never down.

Model-agnostic explanations were generated for the test set. SHAP values were calculated using the shap. KernelExplainer for global interpretation and individual case analysis.

2.6. Quantifying Predictive Uncertainty

UQ means checking how reliable a model's predictions can be. Building prediction systems - especially in high-stakes cases like credit scoring - requires more than just outcomes; confidence in those results matters equally. Instead, understanding the model's doubt helps decision-makers weigh possible errors. This awareness reduces blind trust, improving both robustness and acceptance of AI-based tools [45].

2.6.1. Entropy for Uncertainty Quantification

Entropy measures how uncertain a probability distribution is. It was first developed by Shannon in information theory [46] to capture randomness in data. In machine learning, predictive entropy (PE) helps assess how sure a model is about its output [45]. For two-class problems ($j = 2$), PE can be written as:

$$PE = -\sum_{j=1}^{2} p_j \log p_j, \tag{3}$$

where p_j stands for the estimated likelihood of class j (either default or non-default). Smaller PE scores suggest more certainty, whereas larger ones point to less predictability. Uncertainty drops if a single result becomes much more likely, hitting zero once the forecast is completely sure. Predictive entropy was computed for each test set prediction.

In credit risk analysis, entropy helps by measuring how reliable default probability estimates are [47]. Because it highlights uncertain predictions, experts can target ambiguous cases - asking for extra data or modifying evaluation criteria. Instead of relying solely on model outputs, teams may use alternate methods when uncertainty is high. Quantifying this unpredictability leads to better choices, lowering errors in classifying risky applicants while strengthening the accuracy of scoring systems.

 doi:10.20944/preprints202604.2092.v1

2.6.2. Model Calibration

Model calibration means tweaking prediction models so their forecasts match real-world results [45]. In credit risk, accurate PD estimates must mirror true default rates - this supports better loan choices. However, even if a model sorts risky from safe borrowers well, its probability scores might still be off, leading to flawed judgments. Here, we assess calibration via ECE, a metric capturing the gap between predictions and actuals within grouped probability ranges [65].

$$ECE = \sum_{k=1}^{K} \frac{|S_k|}{T} |\bar{y}_k - \bar{p}_k|, \tag{4}$$

In this formula, S_k stands for the group of forecasts in bin k, whereas $|S_k|$ counts how many cases are in that group. Here, T gives the overall count of predictions made. Calibration was assessed by grouping predictions into 10 equal-width bins and calculating the ECE. The value $\bar{y}_k$ shows the actual proportion of positive results in bin k, although $\bar{p}_k$ reflects the average predicted likelihood for the same bin. Models that are well calibrated yield probabilities which can be clearly understood and trusted, aiding better decision-making under risk. Also, calibration works alongside entropy measures of uncertainty by making sure the model's confidence in default probability estimates matches real outcomes.

2.7. Performance Evaluation

2.7.1. ROC and AUC

The ROC curve assesses binary classifiers, displaying TPR versus FPR at various threshold levels. While it illustrates how sensitivity trades off against specificity, the AUC condenses this into one metric - where 1.0 means flawless prediction, whereas 0.5 suggests chance-level results. In mathematical terms:

$$\text{TPR (Recall)} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$
$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}$$

ROC AUC works well even when classes are uneven, since it checks performance across every threshold rather than one fixed point. Different cutoffs can skew results, so looking at the full range gives a clearer picture

2.7.2. F1-Score

The F1-Score serves as a standard measure in binary classification, especially when class distribution is skewed. This value combines Precision and Recall using their harmonic average, thus addressing both Type I and Type II errors.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$
$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

A value of 1 means high accuracy in both recall and precision, whereas a result near zero shows weak outcomes. This measure becomes relevant when incorrect predictions - either positive or negative - affect results in distinct ways.

 doi:10.20944/preprints202604.2092.v1

3. Empirical Results and Discussion

3.1. Dataset Description

The HMEQ dataset was used here, containing data from 5,960 requests for home equity loans. Such a loan relies on property value as security against repayment. Information covers applicant traits along with loan conditions - delinquency record; how much money borrowed; earnings level; time at job; past credit behavior; whether default happened. An example appears in Table 2(a); further variable explanations are given in Table 2(b).

Table 2. (a) Credit Risk Dataset – Sample Records.

bad	loan	mortdue	value	reason	job	yoj	derog	delinq	clage	ninq	clno	debtinc
0	1700	97800	112000	HomeImp	Office	3	0	0	93.33	0	14	NaN
1	1700	30548	40320	HomeImp	Other	9	0	0	101.47	1	8	37.11
1	1800	48649	57037	HomeImp	Other	5	3	2	77.10	1	17	NaN

(b) Description of Variables.

Variable	Description
Bad	Loan default status: 1 = defaulted or seriously delinquent; 0 = otherwise
Loan	Loan amount requested
Mortdue	Balance on existing mortgage
Value	Current property value
Reason	Purpose of the loan (DebtCon = debt consolidation; HomeImp = home improvement)
Job	Occupation type
Yoj	Years in current job
Derog	Number of major derogatory credit reports
Delinq	Number of delinquent credit lines
Clage	Age of oldest credit line
Ninq	Number of recent credit inquiries
Clno	Total number of credit lines
Debtinc	Debt-to-income ratio

Software and Packages

This research uses Python because it is adaptable, simple to work with, or well-supported for tasks in data science and ML. Analyses run on Python 3.x, which works alongside common stats and ML tools. For cleaning and handling data, Pandas along with NumPy handle the operations. Visuals are created through Matplotlib paired with Seaborn. Building models depends heavily on scikit-learn - especially when applying LR, RF, and XGBoost. To improve clarity, LIME along with SHAP are used as XAI methods; at the same time, uncertainty measures are included. Together, they form a solid approach for dependable and understandable credit risk evaluation.

3.2. Exploratory Data Analysis

In this part, early findings from the data are shared. Because the dataset is large, only key features get close attention. Attention shifts instead toward outlining overall traits of the information. The study looks at the outcome variable separately to clarify how it spreads out or acts. Then, clear trends and key points seen during review are highlighted.

A full check of how complete the data were processed before any analysis began. In an initial check, for numerical data, Table 3 displays summary stats across the full sample - alongside separate Definitions for borrowers who did not default (ND) and those who defaulted (D). Instead of simple averages, it includes mean values plus standard deviation measures. Differences between ND and D groups are captured using the Kolmogorov–Smirnov (KS) metric shown in one column. This value reflects how distinct the distributions are; larger numbers imply better distinguishing power regarding loan risk. When defaults occur, individuals typically show more late payments, worse credit records,

 doi:10.20944/preprints202604.2092.v1

elevated debt relative to income, while also having less time spent building credit history. For defaults, the mean number of late payments is 1.23; for others, it's just 0.25 - similarly, negative records stand at 0.71 versus 0.13. Debt relative to income reaches 39.39% in default cases but only 33.25% otherwise. Credit history length averages 150 months when defaulted, yet spans 187 months if not. Each of these factors shows high KS scores, meaning they clearly distinguish both groups.

Table 3. Descriptive Statistics and KS Values.

Var	Mean(All)	SD(All)	Mean(ND)	SD(ND)	Mean(D)	SD(D)	KS
LOAN	18607.97	11207.48	19028.11	11115.76	16922.12	11418.46	0.1386
MORTDUE	73760.82	44457.61	74829.25	43584.99	69460.45	47588.19	0.0955
VALUE	101776.05	57385.78	102595.92	52748.39	98172.85	74339.82	0.1028
YOJ	8.92	7.57	9.15	7.68	8.03	7.10	0.0885
DEROG	0.25	0.85	0.13	0.51	0.71	1.47	0.2249
DELINQ	0.45	1.13	0.25	0.67	1.23	1.90	0.3216
CLAGE	179.77	85.81	187.00	84.47	150.19	84.95	0.2192
NINQ	1.19	1.73	1.03	1.53	1.78	2.25	0.1591
CLNO	21.30	10.14	21.32	9.68	21.21	11.81	0.0735
DEBTINC	33.78	8.60	33.25	6.95	39.39	17.72	0.2648

Non-defaulting individuals tend to have more stable finances, which aligns with the idea that timely payments, manageable debt levels, and established credit records reduce default risk. On the other hand, factors like borrowed sum, home worth, employment duration, total accounts, or outstanding mortgage loans display weaker distinctions and modest KS scores - indicating limited power in predicting defaults.

As seen in Table 4, default levels differ by REASON and also by JOB type. To examine links between these categories and the outcome, we ran a Chi-square test instead of assuming independence. People taking loans for home upgrades (HomeImp) defaulted more often - 22.25% - compared to individuals paying off existing balances (DebtCon), where it was 18.97%. Among occupations, defaults are highest for Sales (34.86%) and Self-employed (30.05%), while Office and ProfExe roles have lower default rates (13.19% and 16.61%). This confirms that both loan purpose and occupation are significant predictors of default risk.

Table 4. Categorical Variables and Default Rates.

Category	Relative Frequency	Default Rate	Chi2 p-value
REASON: DebtCon	0.6591	0.1897	0.004576
REASON: HomeImp	0.2987	0.2225	0.004576
JOB: Mgr	0.1287	0.2334	3.3067e-16
JOB: Office	0.1591	0.1319	3.3067e-16
JOB: Other	0.4007	0.2320	3.3067e-16
JOB: ProfExe	0.2141	0.1661	3.3067e-16
JOB: Sales	0.0183	0.3486	3.3067e-16
JOB: Self	0.0324	0.3005	3.3067e-16

Figure 4 presents both the proportion and count of missing values across several variables, excluding those with complete observations. Supervised learning methods, for example RF, are capable of managing missing values, while NNs generally require fully observed data for reliable performance [29]. The variable DEBTINC shows the highest proportion of missing data (21.3%), followed by DEROG (11.9%) and DELINQ (9.7%).

 doi:10.20944/preprints202604.2092.v1

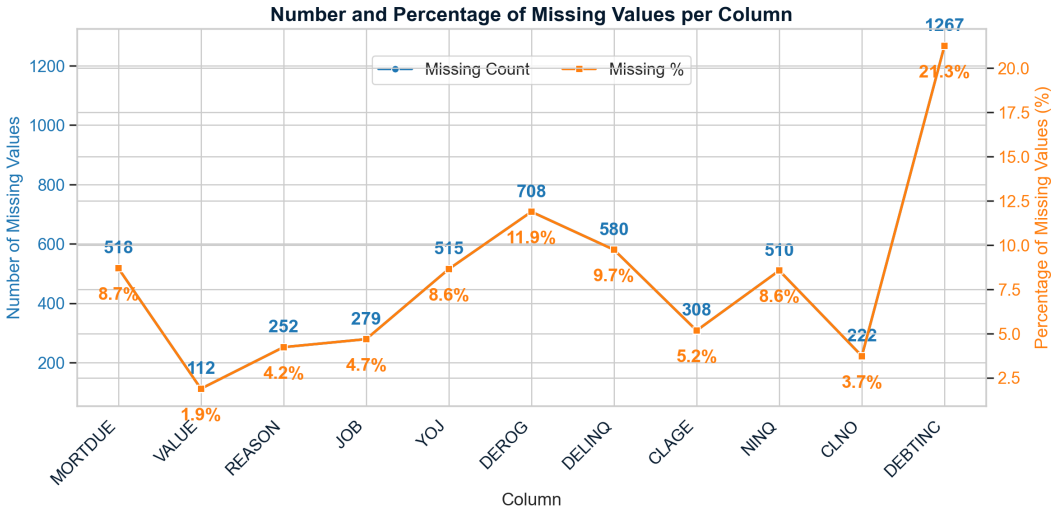


Figure 4. Missing values.

As a result, appropriate imputation techniques were applied to ensure data completeness. The detailed approach to handling missing values is described in Section 3.3.1.

The target variable, symbolised by B, indicates whether a borrower has defaulted on a loan. It is defined as:

$$B = \begin{cases} 0, & \text{for non-default cases,} \\ 1, & \text{for default cases.} \end{cases}$$

The outcome variable takes only two possible values, indicating that the task at hand is a binary classification problem. The dataset comprises 5 960 loan records, of which 1 189 (19.9%) represent defaults and 4 771 (80.1%) represent non-defaults, as represented in Figure 5 and Figure 6, which illustrate the class distribution using a pie chart and a histogram, respectively.

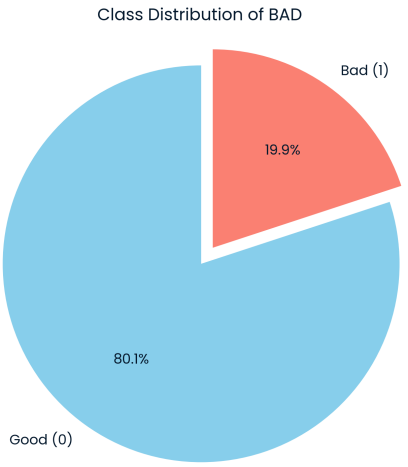


Figure 5. Pie chart showing the proportion of Good and Bad loans.

 doi:10.20944/preprints202604.2092.v1

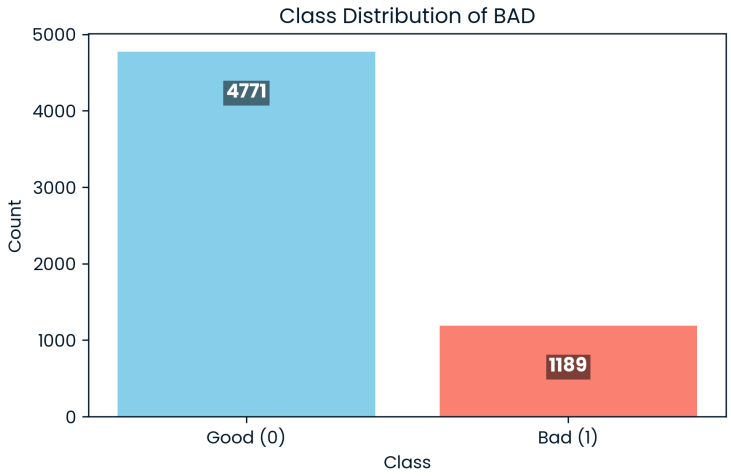


Figure 6. Histogram illustrating the class distribution of the dataset.

This reflects a considerable class imbalance, which may bias model training towards the majority class. The imbalance suggests that standard classifiers may perform poorly in identifying the minority (default) cases. To mitigate this, the SMOTE was applied to the training data, as detailed in Section 3.3.2.

3.3. Data Preprocessing

3.3.1. Handling Missing Values

In real-life data, gaps often appear - these may harm how well algorithms work. To keep results trustworthy while maintaining quality, incomplete entries were handled differently: numbers versus categories got distinct treatments. For numeric columns, gaps were filled with the average value from each respective column. By doing so, the overall center of the data stays intact - no records need to be dropped. In particular, if an entry y_i in a numeric variable y was missing, it got substituted by:

$$y_i = \frac{1}{k} \sum_{j=1}^{k} y_j,$$

where k is the number of non-missing values in the column. For categorical variables, gaps were filled by picking the mode per variable. Because this approach uses existing levels only, it stops artificial groups from appearing - keeping the data's natural spread intact.

After filling in missing data, categorical variables got turned into numbers through one-hot coding. For every group, this method creates a yes-or-no flag - though it leaves out the initial category per variable to reduce overlap issues. The result keeps all original details intact while making them usable for algorithmic models.

3.3.2. Handling Class Imbalance Using Synthetic Minority Over-Sampling Technique

The outcome being studied showed clear unevenness, where one group made up only a tiny fraction of all cases. Fitting classifiers straight onto skewed datasets usually shifts focus toward the larger group - this weakens recognition of rare outcomes such as defaults. To mitigate this issue, the SMOTE proposed by [28] was employed to artificially balance the training data. SMOTE operates by generating synthetic minority instances through linear interpolation between existing minority samples and their nearest neighbours in the feature space. For each minority sample x_i, one of its k nearest neighbours x_{zi} is randomly selected, and a new synthetic observation is created according to the following formula:

$$x_{\text{new}} = x_i + \delta(x_{zi} - x_i). \tag{5}$$

 doi:10.20944/preprints202604.2092.v1

Here, $\delta \sim U(0,1)$, indicating a random variable drawn from a uniform distribution over $[0,1]$. SMOTE generates synthetic minority-class samples without duplicating existing data, reducing overfitting while preserving the original distribution. In this study, SMOTE was applied using Python's `imblearn` library with $k = 5$ nearest neighbors, only on training folds during cross-validation to avoid data leakage. Model training and evaluation were then performed on these resampled training sets to ensure balanced learning and unbiased assessment.

3.4. Variable Selection

Figure 7 illustrates the feature selection results obtained via RFECV using a RF classifier as the estimator. The analysis indicates that all features contributed positively to model performance except for `Job_Self`, which was identified as non-informative and subsequently excluded. The remaining predictors were retained and employed to fit the final RF model, alongside all other classifiers evaluated in the study.

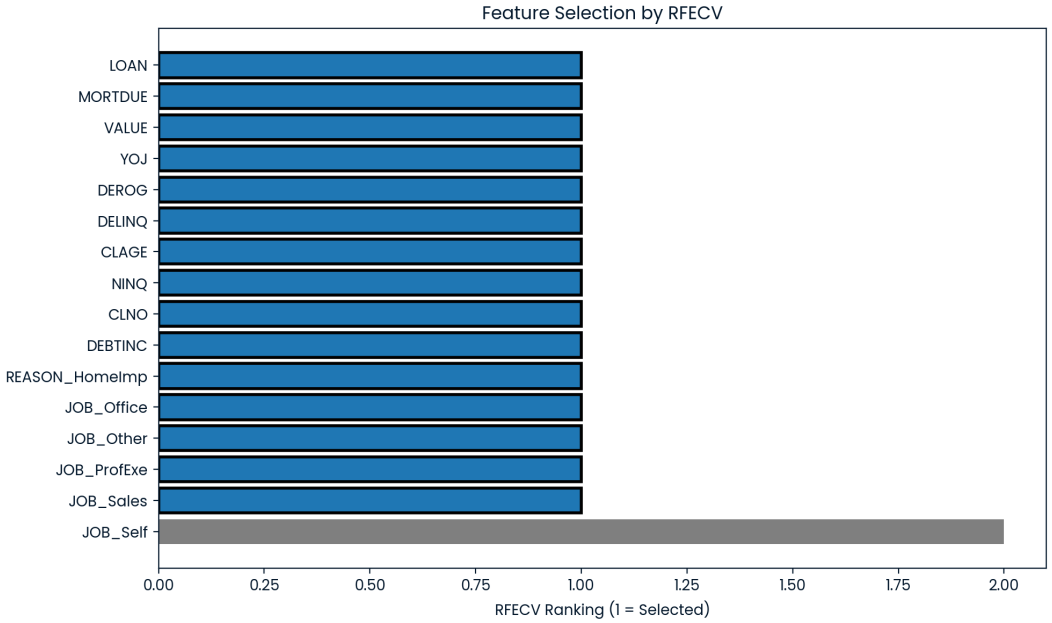


Figure 7. Feature selections.

3.5. Supervised Learning Models

The current investigation employs supervised machine learning frameworks. Within the supervised learning paradigm, algorithms derive mathematical models through training on datasets that encompass both explanatory variables and their corresponding ground-truth labels. After variable selection, the data was split into an 80/20 ratio for training and testing. Three supervised learning algorithms namely XGBoost, RF, and LR were trained using the training dataset. Hyperparameters for each model were optimised using grid search with 5-fold cross-validation, selecting the combination that maximised the ROC AUC score.

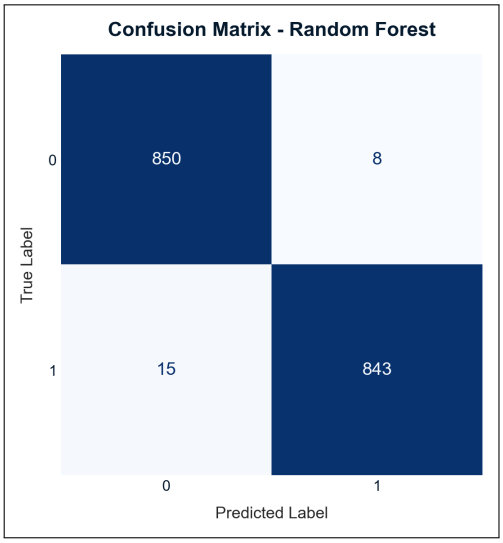
3.5.1. Best Model Parameters

For the XGBoost Classifier, the hyperparameters selected through grid search were a learning rate of 0.2, a maximum depth of 7, 200 estimators, and a subsample ratio of 0.8. For the RF Classifier, the optimised parameters were `bootstrap = False`, `max_depth = None`, `max_features = 'sqrt'`, `min_samples_leaf = 1`, `min_samples_split = 2`, and 200 estimators. For LR, the chosen hyperparameters were `C = 1`, `penalty = 'l1'`, `solver = 'liblinear'`, and `l1_ratio = None`. These configurations were employed to train the final models and were subsequently used for evaluation on the test set, ensuring that each model achieved optimal predictive performance while maintaining interpretability and robustness.

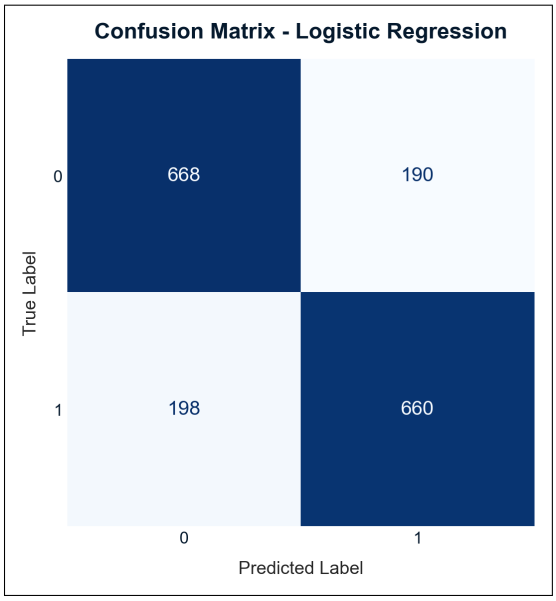
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2026 doi:10.20944/preprints202604.2092.v1

3.5.2. Model Performance Evaluation

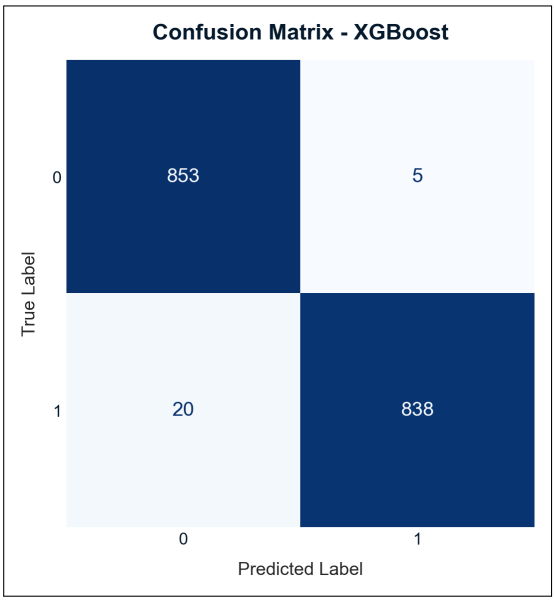
The performance of the three classification models RF, LR, and XGBoost was assessed using confusion matrices (Figure 8), ROC curves (Figure 8), and standard classification metrics summarised in Table 5. Together, these results offer a thorough insight into each model's performance and predictive reliability on the test set.



(a) Random Forest.



(b) Logistic Regression.



(c) XGBoost.

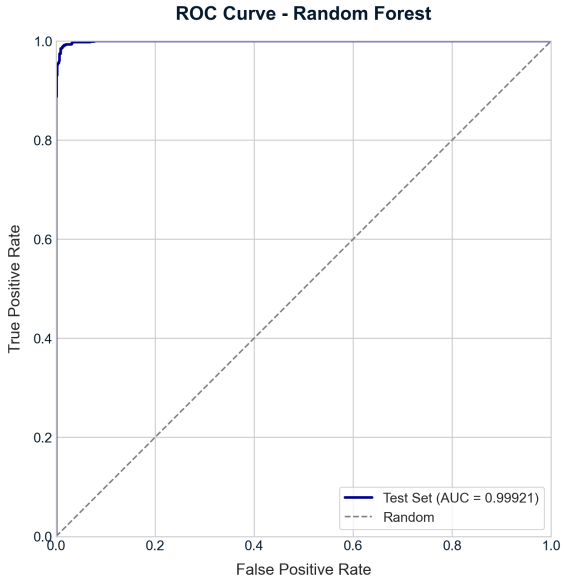
Figure 8. Confusion matrices for three models evaluated on the test set.

The confusion matrices in Figure 8 show distinct model performances. While RF performs well, XGBoost also achieves high accuracy, with few misclassifications. RF correctly identified 850 true negatives and 843 true positives; XGBoost reached similar results, with 853 correct negatives and 838 positives. These findings indicate that ensemble methods reliably detect both defaults and non-default borrowers, reflecting their capacity to model complex, non-linear relationships in credit data.

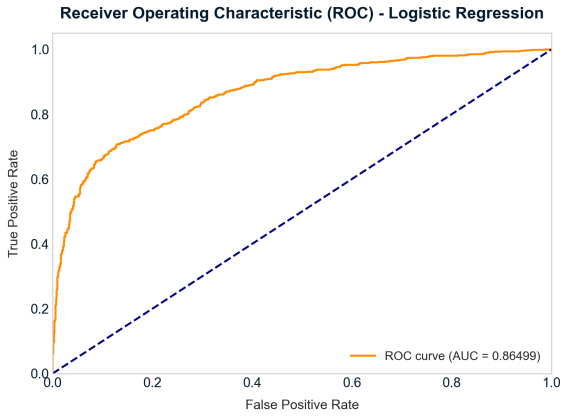
In contrast, Logistic Regression (LR) struggles to distinguish between groups. The model produced 190 false positives and 198 false negatives, highlighting limitations in capturing intricate patterns. Its weaker results across all measures account for the lower overall accuracy. The ROC curves in

 doi:10.20944/preprints202604.2092.v1

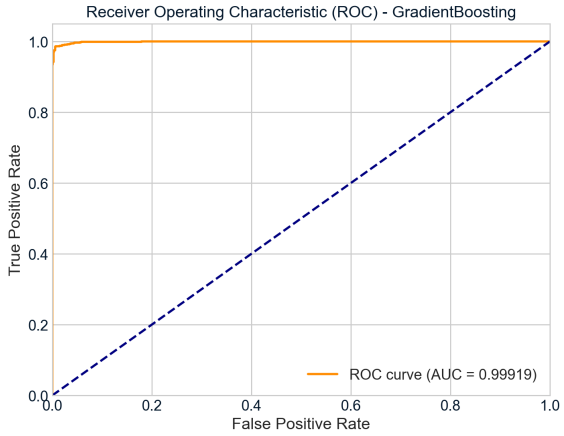
Figure 9 illustrate these differences. RF achieved an AUC of 0.9992, with XGBoost slightly behind at 0.9991, demonstrating near-perfect discrimination between defaulted and repaid loans. LR recorded 0.8649, indicating weaker separation and a reduced ability to handle non-linear borrower behaviours. Although LR's performance may suffice for simple linear approaches, it falls short compared to tree-based ensembles in predicting defaults.



(a) Random Forest.



(b) Logistic Regression.



(c) XGBoost.

Figure 9. ROC curves for three models evaluated on the test set.

Table 5 summarises model performance. LR underperforms, while RF and XGBoost achieve over 98% accuracy and F1-scores above 0.98. XGBoost attains the highest precision (0.9941), minimising false positives—critical in credit risk, where approving high-risk borrowers can result in significant financial loss. LR's lower accuracy (0.7739) and F1-score (0.7728) reflect its inability to capture non-linear relationships in the data. These results emphasise that ensemble models not only improve predictive performance but also enhance the reliability of credit risk assessments, supporting informed lending decisions.

Table 5. Performance metrics for different classification models.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.9866	0.9906	0.9825	0.9865
Logistic Regression	0.7739	0.7765	0.7692	0.7728
XGBoost	0.9854	0.9941	0.9767	0.9853

 doi:10.20944/preprints202604.2092.v1

3.6. Explainable AI for Credit Risk Assessment

A thorough evaluation of model performance must be complemented by clear insight into why the model arrives at specific predictions, ensuring transparency and trust in credit-risk decisions. This section applies XAI techniques specifically LIME and SHAP to deliver both global and local interpretability for the trained tree-based models.

3.6.1. Model Explanation for Tree Ensembles Using LIME

Following the establishment of the procedure for generating explanations with the LIME framework, a technical constraint was identified when applying it to the XGBoost and Random Forest classifiers, as LIME is not compatible with models wrapped in a GridSearchCV object. To resolve this issue, both classifiers were retrained using the best-performing hyperparameters obtained from the tuning stage, allowing them to be stored as native models compatible with LIME.

3.6.2. Insights from LIME Explanations

This section presents interpretative insights derived from LIME through three representative case studies. Figure 10 shows a loan predicted as "Fully Paid" (BAD = 0). LIME highlights the top nine factors affecting the outcome along with their individual impact scores. Key positive influences include zero delinquencies ("DELINQ = 0"), clean credit records ("DEROG = 0"), and office employment ("JOB_Office = 1"). These patterns are consistent with prior research showing that past payment behaviour and credit cleanliness are strong predictors of creditworthiness [66].

For risky cases (Figure 11), LIME identifies major negative contributors such as negative legal records, recent credit enquiries, and unstable employment. This aligns with literature showing that frequent credit applications, prior adverse records, and job instability are associated with higher default probability [68,70]. In the XGBoost example (Figure 12), high debt-to-income ("DEBTINC") and prior delinquencies contribute strongly toward default prediction. These results are supported by previous studies using ML and XAI, which consistently find debt ratios and delinquency counts as top predictors of default risk [67]. Overall, the LIME explanations provide transparent insights into how each feature influences predictions, confirming patterns widely documented in the credit scoring literature.

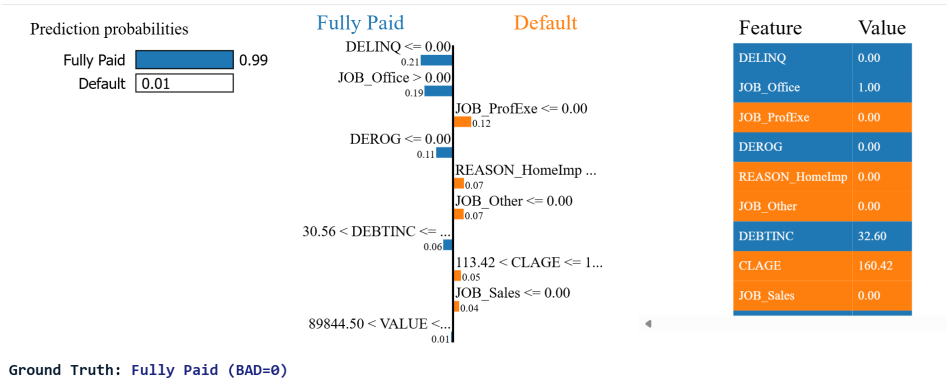


Figure 10. LIME explanation for a customer classified as a "Fully Paid" loan by the Random Forest model.

 doi:10.20944/preprints202604.2092.v1

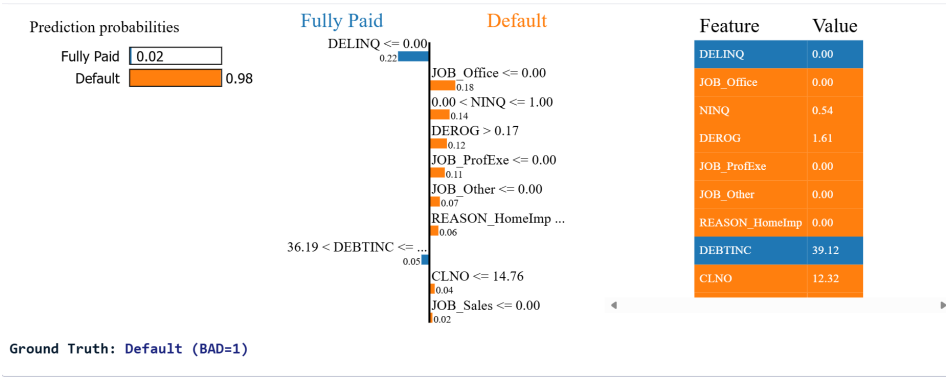


Figure 11. LIME explanation for a customer classified as a "Default" loan by the Random Forest model.

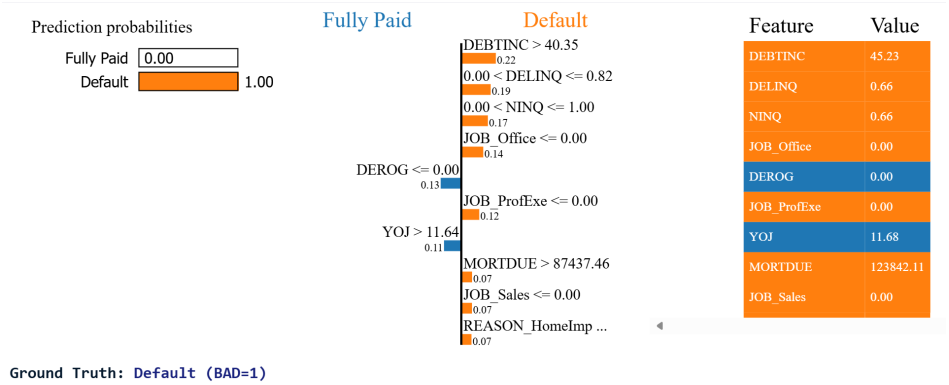


Figure 12. LIME explanation for a customer classified as a "Default" loan by the XGBoost model.

3.6.3. Global and Local Model Explanation Using SHAP

SHAP was used to interpret the tree-based models at both global and local levels. Unlike LIME, which approximates local behaviour, SHAP provides exact contributions of each feature using game-theoretic principles, allowing for both directional and magnitude interpretation across the dataset. Definitions 13 and 14 illustrate these effects.

Globally (Figure 13), "DELINQ", "DEROG", and "DEBTINC" are the top predictors of default. This pattern is consistent with empirical credit scoring research, where past delinquencies, derogatory events, and high debt ratios are widely reported as strong risk indicators [66].

The SHAP summary plot (Figure 14) further shows directional effects. High values of "DELINQ" and "DEBTINC" (red points) shift predictions toward default, whereas low values push predictions toward repayment. Employment features display more heterogeneous effects, reflecting variation in income stability across borrower segments, a phenomenon also reported in recent studies on ML credit scoring [69]. Overall, SHAP results confirm and complement the local insights provided by LIME, demonstrating that XAI methods reliably highlight established financial risk factors, providing both transparency and validation of model behaviour.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2026 doi:10.20944/preprints202604.2092.v1

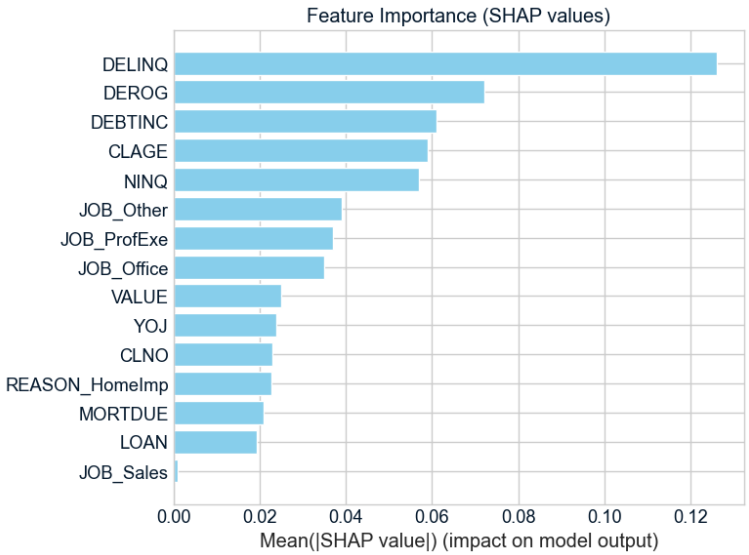


Figure 13. Global feature importance derived from SHAP values.

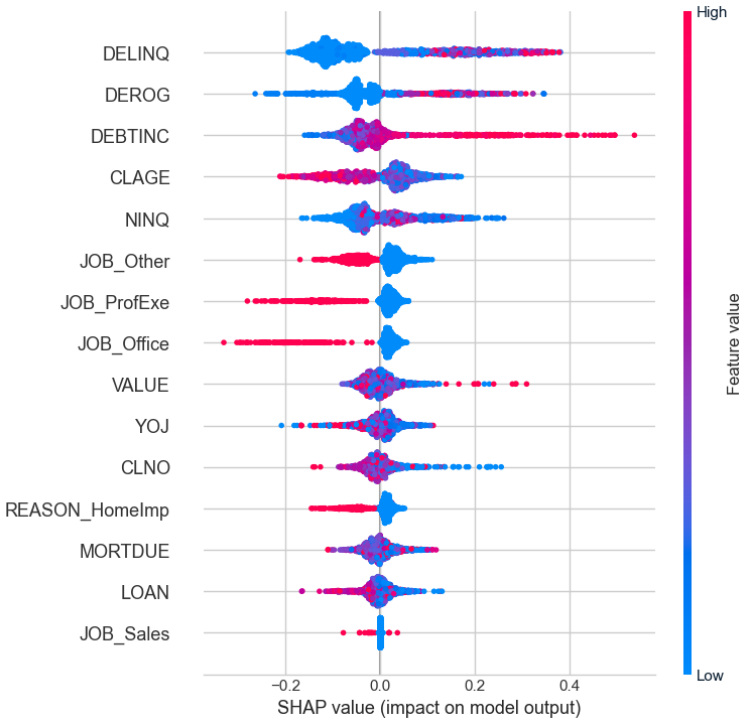


Figure 14. SHAP summary plot showing feature impact direction and magnitude.

3.7. Quantifying Predictive Uncertainty

We assess model confidence in default risk forecasts using predictive entropy, which captures variability in outcome probabilities, and ECE, which compares average predictions to real-world outcomes. While RF generates inherent variation via bootstrap aggregation, XGBoost and LR lack this built-in stochasticity. Therefore, bootstrap resampling was applied to XGBoost and LR to introduce controlled randomness, ensuring fair and consistent uncertainty assessment across all models.

Across all three models, predictive entropy and model errors exhibit a consistent pattern: higher uncertainty is associated with a greater likelihood of misclassification. For LR, misclassified instances concentrate in the higher-entropy region (approximately 0.4-0.7), indicating that although the linear model struggles to separate the underlying data structure fully, reflected in its overall accuracy of 77%, it nevertheless assigns higher uncertainty to cases it finds difficult. In contrast, XGBoost demonstrates

 doi:10.20944/preprints202604.2092.v1

substantially stronger pattern-recognition ability. As shown in Figure 16, the majority of its correct predictions (approximately 1,400 samples) lie in the near-zero entropy range, with only a small cluster of errors occurring where entropy is high, consistent with its accuracy of 98.54%. A similar behaviour is observed for RF (Figure 15): when tree agreement is high (entropy close to zero), the model achieves near-perfect accuracy, and the limited number of misclassifications arises only in regions where the ensemble displays maximal disagreement. These results show that all three models capture meaningful structure in the data and can clearly separate confident predictions from ambiguous ones. The accuracy differences mainly reflect how well each model shifts samples into the low-entropy "well-understood" region: LR is limited by its linear boundary, while XGBoost and RF better model the underlying non-linear patterns. Across all models, misclassifications occur almost exclusively at high entropy, confirming predictive entropy as a reliable measure of uncertainty and a strong indicator of when predictions should be treated with caution.

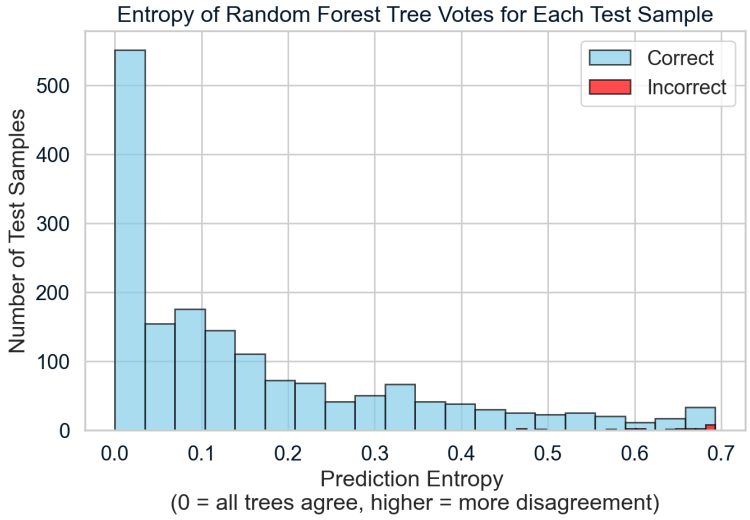


Figure 15. Random Forest Predictive Entropy.

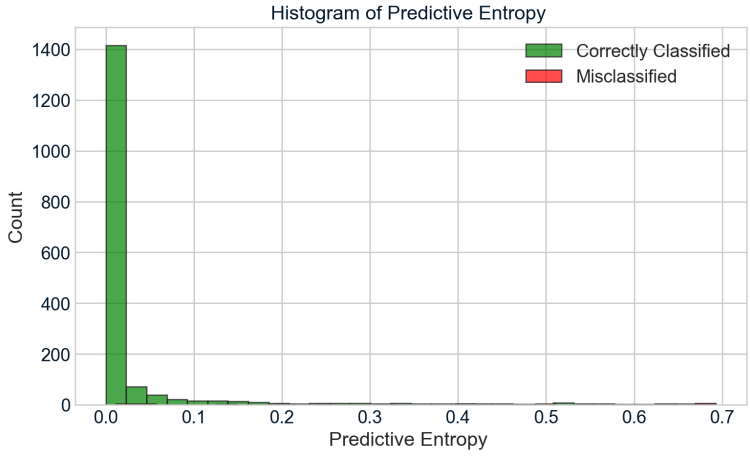


Figure 16. XGBoost Predictive Entropy.

 doi:10.20944/preprints202604.2092.v1

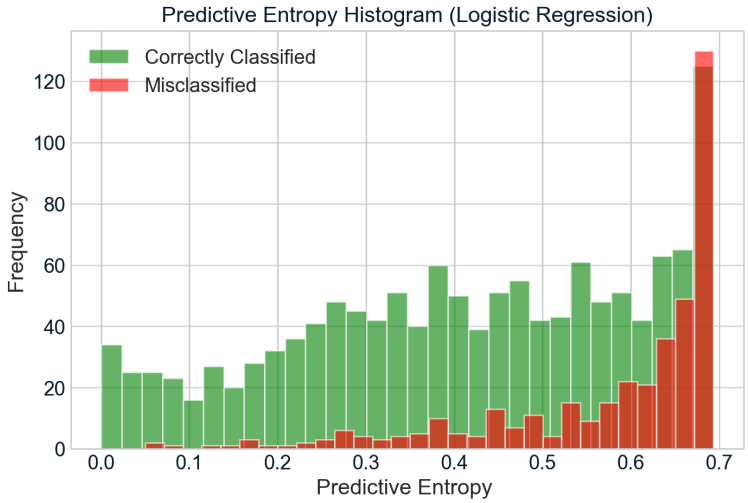


Figure 17. Logistic Regression Predictive Entropy.

Although RF shows slightly stronger predictive metrics than XGBoost (see Table 5), the difference between the models is minimal. Their calibration behaviour, however, differs substantially. As shown in the calibration diagrams in Definitions 18 – 20, RF exhibits clear miscalibration (ECE = 0.0475), reflecting overconfident PD estimates.

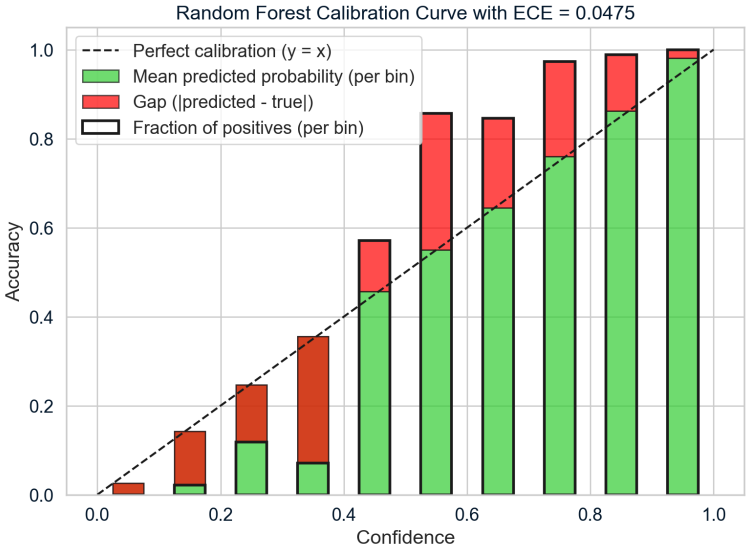


Figure 18. Random Forest ECE.

 doi:10.20944/preprints202604.2092.v1

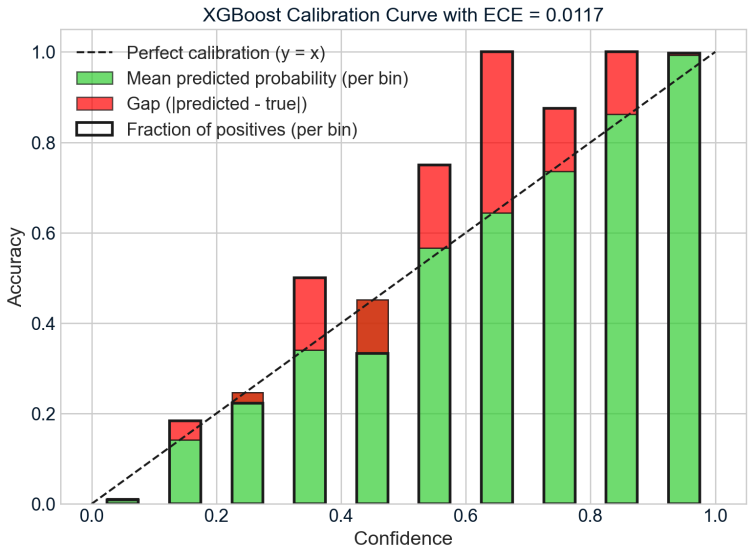


Figure 19. XGBoost ECE.

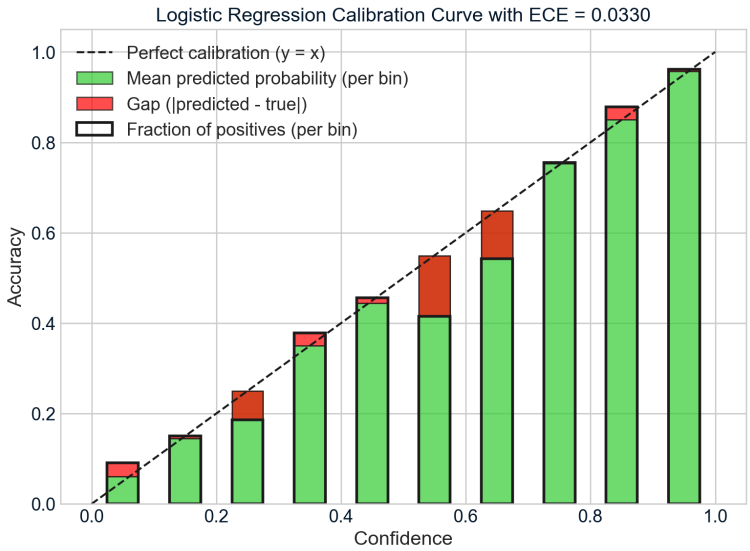


Figure 20. Logistic Regression ECE.

XGBoost achieves markedly better calibration (ECE = 0.0117), with its predicted PD values aligning more closely with observed default frequencies. LR provides an interesting contrast: although its accuracy is noticeably lower than the ensemble models, its calibration performance (ECE = 0.0330) is better than RF's, though still inferior to XGBoost's. In credit risk modelling, well-calibrated PD estimates are more valuable than marginal differences in predictive accuracy, as miscalibration can distort lending decisions, pricing strategies, and capital calculations. Therefore, even with comparable predictive accuracy, XGBoost provides the most reliable basis for probability-of-default estimation within an uncertainty-informed framework.

3.7.1. Interpretation of Predictive Entropy and Calibration for Credit Risk Decision-Making

Predictive entropy and calibration extend beyond accuracy by indicating when model outputs require caution in decision contexts. Higher entropy is consistently associated with unstable class probabilities and increased misclassification risk, identifying cases where automated predictions should be treated conservatively or subjected to additional review. In lending environments, such uncertainty signals borrowers whose profiles do not closely match learned patterns, supporting risk-sensitive actions such as stricter approval thresholds, further verification, or adjusted pricing.

 doi:10.20944/preprints202604.2092.v1

Calibration complements this by assessing whether predicted probabilities align with observed default rates. Reliable calibration is essential for probability of default estimation, portfolio risk assessment, and expected credit loss calculations. Poor calibration, particularly overconfidence, can distort provisioning, pricing, and capital allocation, increasing model risk.

Together, entropy and ECE provide a dual evaluation framework: entropy highlights instance-level uncertainty, while calibration evaluates the reliability of probability estimates at the portfolio level. Their joint use strengthens uncertainty-aware modelling and supports defensible, risk-aligned credit decisions.

4. Discussion of Results

An examination of the HMEQ data set, comprising 5,960 home equity loan applications with a default rate of 19.9%, reveals numerous insights into predicting credit risk, interpreting models, and assessing uncertainties. Ensemble algorithms significantly outperform logistic regression at recognising loan defaults. Both Random Forest and XGBoost achieve almost perfect discrimination, with ROC AUC values of 0.9992 and 0.9991, respectively, and accuracies above 98%, along with F1 scores above 0.985. Logistic regression achieves only 77.4% accuracy and an AUC of 0.8649, making 388 incorrect predictions, compared with fewer than 60 by each of the ensemble algorithms. XGBoost offers the highest precision (0.9941), thereby avoiding costly false positives.

As per the existing literature on credit scoring, an exploratory analysis reveals that the discriminating variables are DELINQ, DEROG, and DEBTINC, with p-values of 0.3216, 0.2249, and 0.2648, respectively. Individuals who have defaulted have poor credit ratings: Average delinquencies of 1.23 against 0.25, derogatory credit of 0.71 against 0.13, and debt income ratio of 39.39% against 33.25%. Categorical analyses indicate a higher default probability for home improvements than for debt consolidations (22.25% vs 18.97%) and for sales and self-employed professions (34.86% and 30.05% respectively).

Although both techniques have equal prediction accuracy, calibration behaviour varies widely. Calibration is superior in XGBoost (ECE = 0.0117) when compared to the random forest technique (ECE = 0.0475). The random forest technique demonstrates overconfidence. Logistic regression has moderate calibration performance (ECE = 0.0330) but poor prediction accuracy. Predictive entropy results indicate consistent performance across all cases; uncertainty is positively correlated with classification error. Predictions with accurate results occur in low-entropy areas (close to zero), whereas errors occur mostly in high-entropy situations (around 0.4-0.7 in logistic regression).

About the use of default probability estimates to make credit risk decisions, calibrated estimates provide greater value than marginal increases in accuracy, as they prevent distortions in pricing, provisioning, and capital computations. When used together, predictive entropy and expected calibration error enable uncertainty-informed lending: high entropy is treated by making conservative lending decisions, whereas low calibration indicates model risk at the portfolio level. XGBoost is the most solid foundation for default probability estimates.

5. Conclusions

This study demonstrates that dependable AI for credit scoring requires more than predictive accuracy. Models must provide calibrated probabilities alongside interpretable explanations to support stable, fair, and transparent decisions. XGBoost stood out among the tested methods, offering strong predictions, reliable confidence estimates, and compatibility with SHAP and LIME for explainability. Entropy-based uncertainty successfully identified high-risk or ambiguous cases, reinforcing the value of uncertainty quantification in real-world lending decisions.

By explicitly linking model logic to awareness of uncertainty, this study provides evidence supporting the informed deployment of AI in credit risk assessment. Recommendations for banks and future researchers are derived directly from these findings, highlighting practical and ethical implica-

 doi:10.20944/preprints202604.2092.v1

tions, while addressing identified limitations. This work thus adds to the field by demonstrating a clear methodology to balance predictive power, reliability, and interpretability in credit risk modelling.

Author Contributions: Conceptualization, M.R., T.R. and C.S.; methodology, M.R.; software, M.R.; validation, M.R., T.R. and C.S.; formal analysis, M.R.; investigation, M.R., T.R. and C.S.; data curation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, M.R., T.R. and C.S.; visualization, M.R.; supervision, T.R. and C.S.; project administration, T.R. and C.S.; funding acquisition, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP): http://www.escience.ac.za/, accessed on 15 January 2024.

Data Availability Statement: The study uses a panel dataset of 50,000 U.S. home loan applicants, tracked over 60 periods, including origination and repayment outcomes. The data are publicly available at https://www.listendata.com/2019/08/datasets-for-credit-risk-modeling.html?m=1 (accessed on 13 February 2025).

Acknowledgments: The support of the DST-CSIR National e-Science Postgraduate Teaching and Training Platform (NEPTTP) provided for this research is acknowledged. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NEPTTP. In addition, the authors thank the anonymous reviewers for their helpful comments on this paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUC	Area Under the Curve
CRM	Credit Risk Model
DR	Default Risk
ECE	Expected Calibration Error
GB	Gradient Boosting
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
HMEQ	Home Equity
ML	Machine Learning
NNs	Neural Networks
PD	Probability of Default
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross-Validation
ROC	Receiver Operating Characteristic Curve
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
ECE	Expected Calibration Error
UQ	Uncertainty Quantification
XAI	Explainable Artificial Intelligence
XGBOOST	Extreme Gradient Boosting
IFRS	International Financial Reporting Standards

References

1. Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216. Available online: [CrossRef]
2. Noriega, J.P., Rivera, L.A. and Herrera, J.A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, 8(11), 169. Available online: [CrossRef]

 doi:10.20944/preprints202604.2092.v1

3. Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138–52160. Available online: [CrossRef]
4. Nallakaruppan, M.K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V.P. and Hameed, I.A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks*, 12(10), 164. Available online: [CrossRef]
5. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. Available online: [CrossRef]
6. Hand, D.J. and 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. Available online: [CrossRef]
7. Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589–609. Available online: [CrossRef]
8. Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. Available online: [CrossRef]
9. Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. Available online: [CrossRef]
10. Jagtiani, J. and John, K. (2018). Fintech: the impact on consumers and regulatory responses. *Journal of Economics and Business*, 100, 1–6. Available online: [CrossRef]
11. Gomber, P., Kauffman, R.J., Parker, C. and Weber, B.W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of management information systems*, 35(1), 220–265. Available online: [CrossRef]
12. Rudin, C. and Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 1–9. Available online: [CrossRef]
13. Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3), 50–57. Available online: [CrossRef]
14. Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, 1050–1059. Available online: [CrossRef]
15. Chi, Q. and Li, W. (2017). Economic policy uncertainty, credit risks and banks' lending decisions: Evidence from Chinese commercial banks. *China journal of accounting research*, 10(1), 33–50. Available online: [CrossRef]
16. Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30. Available online: [CrossRef]
17. Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30. Available online: [CrossRef]
18. Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41, 647–665. Available online: [CrossRef]
19. Alkhyeli, K. (2023). Explainable AI for Credit Risk Assessment. MSc. Thesis, *Khalifa University*, Abu Dhabi, United Arab Emirates. Available online: [CrossRef]
20. Chiaburu, T., Haußer, F. and Bießmann, F. (2024). Uncertainty in xai: Human perception and modeling approaches. *Machine Learning and Knowledge Extraction*, 6(2), 1170–1192. Available online: [CrossRef]
21. ASSESSMENT, M.R. (2024). EXPLAINABLE AI TECHNIQUES FOR MORTGAGE RISK ASSESSMENT: A FAIR APPROACH. Available online: [CrossRef]
22. Bracke, P., Datta, A., Jung, C. and Sen, S. (2019). Machine learning explainability in finance: an application to default risk analysis. Bank of England Working Paper. Available online: [CrossRef]
23. Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131. Available online: [CrossRef]
24. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6), 627–635. Available online: [CrossRef]
25. Blessing, E., Abill, R., Kaledio, P. and Louis, F. (2024). Explainable AI: Interpreting and Understanding Machine Learning Models. [CrossRef]
26. Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32. Available online: [CrossRef]
27. Breiman, L. and Ihaka, R. (1984). Nonlinear discriminant analysis via scaling and ACE. Department of Statistics, University of California Davis One Shields Avenue. Available online: [CrossRef]

 doi:10.20944/preprints202604.2092.v1

28. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357. Available online: [CrossRef]
29. Berendsen, S. (2019). Transparency in black box models. MSc Thesis, Vrije Universiteit Amsterdam, De Boelelaan, HV Amsterdam. Available online: [CrossRef]
30. Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123–140. Available online: [CrossRef]
31. Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832–844. Available online: [CrossRef]
32. Tang, L., Cai, F. and Ouyang, Y. (2019). Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. *Technological Forecasting and Social Change*, 144, 563–572. Available online: [CrossRef]
33. Qinghe, Z., Wen, X., Boyan, H., Jong, W. and Junlong, F. (2022). Optimised extreme gradient boosting model for short term electric load demand forecasting of regional grid system. *Scientific Reports*, 12(1), 19282. Available online: [CrossRef]
34. Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M. and Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. *Machine Learning and Knowledge Extraction*, 6(1), 53–77. Available online: [CrossRef]
35. Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794. Available online: [CrossRef]
36. Wang, K., Li, M., Cheng, J., Zhou, X. and Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia computer science*, 199, 1128–1135. Available online: [CrossRef]
37. Ali, Z.A., Abduljabbar, Z.H., Tahir, H.A., Sallow, A.B. and Almufti, S.M. (2023). eXtreme gradient boosting algorithm with machine learning: A review. *Academic Journal of Nawroz University*, 12(2), 320–334. Available online: [CrossRef]
38. Nwafor, C.N., Nwafor, O. and Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. *Scientific Reports*, 14(1), 25174. Available online: [CrossRef]
39. Nandipati, V.S.S. and Boddala, L.V. (2024). Credit Card Approval Prediction: A comparative analysis between Logistic Regression, KNN, Decision Trees, Random Forest, XGBoost. Available online: [CrossRef]
40. Garg, K., Gill, K.S., Malhotra, S., Devliyal, S. and Sunil, G. (2024). Implementing the xgboost classifier for bankruptcy detection and smote analysis for balancing its data. *2024 2nd International Conference on Computer, Communication and Control (IC4)*, 1–5. Available online: [CrossRef]
41. Musa, A.B. (2013). Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, 4(1), 13–24. Available online: [CrossRef]
42. Šarlija, N., Bilandžić, A. and Stanic, M. (2017). Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models. *Croatian operational research review*, 631–652. Available online: [CrossRef]
43. Gramegna, A. and Giudici, P. (2021). SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4, 752558. Available online: [CrossRef]
44. Ariza-Garzón, M.J., Arroyo, J., Caparrini, A. and Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, 8, 64873–64890. Available online: [CrossRef]
45. Habibpour, M., Gharoun, H., Mehdipour, M., Tajally, A., Asgharnezhad, H., Shamsi, A., Khosravi, A. and Nahavandi, S. (2023). Uncertainty-aware credit card fraud detection using deep learning. *Engineering Applications of Artificial Intelligence*, 123, 106248. Available online: [CrossRef]
46. Shannon, C.E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423. Available online: [CrossRef]
47. Li, Y. and Chen, W. (2021). Entropy method of constructing a combined model for improving loan default prediction: A case study in China. *Journal of the Operational Research Society*, 72(5), 1099–1109. Available online: [CrossRef]
48. Bhaskar, A., Rani, R., Jaiswal, G., Dev, A., Sharma, A., Bansal, P. and Gupta, U. (2024). Automatic credit card approval prediction system. *AIP Conference Proceedings*, 2919(1), 050007. Available online: [CrossRef]
49. Shi, S., Tse, R., Luo, W., D'Addona, S. and Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, 34(17), 14327–14339. Available online: [CrossRef]

 doi:10.20944/preprints202604.2092.v1

50. Ouattara, K.I. (2024). Quantifying calibration error in modern neural networks through evidence based theory. *arXiv preprint arXiv:2411.00265*. Available online: [CrossRef]
51. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R. and others (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1), 1513–1589. Available online: [CrossRef]
52. Fakour, F., Mosleh, A. and Ramezani, R. (2024). A structured review of literature on uncertainty in machine learning & deep learning. *arXiv preprint arXiv:2406.00332*. Available online: [CrossRef]
53. Huang, C.-L., Chen, M.-C. and Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4), 847–856. Available online: [CrossRef]
54. Wang, G., Hao, J., Ma, J. and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1), 223–230. Available online: [CrossRef]
55. Nanni, L. and Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, 36(2), 3028–3033. Available online: [CrossRef]
56. Zurada, J., Kunene, N. and Guan, J. (2014). The classification performance of multiple methods and datasets: Cases from the loan credit scoring domain. *Journal of International Technology and Information Management*, 23(1), 5. Available online: [CrossRef]
57. Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, 39(3), 3446–3453. Available online: [CrossRef]
58. Li, Y. and Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, 8(10), 1756. Available online: [CrossRef]
59. Aruleba, I. and Sun, Y. (2024). Effective credit risk prediction using ensemble classifiers with model explanation. *IEEE Access*. Available online: [CrossRef]
60. Trinh, L.T. (2024). A comparative analysis of consumer credit risk models in Peer-to-Peer Lending. *Journal of Economics, Finance and Administrative Science*, (ahead-of-print). Available online: [CrossRef]
61. Chen, W., Ma, C. and Ma, L. (2009). Mining the customer credit using hybrid support vector machine technique. *Expert systems with applications*, 36(4), 7611–7616. Available online: [CrossRef]
62. Karaa, A. and Krichene, A. (2012). Credit–risk assessment using support vectors machine and multilayer neural network models: a comparative study case of a Tunisian bank. *Journal of Accounting and Management Information Systems (JAMIS)*, 11(4), 587–620. Available online: [CrossRef]
63. Bekhet, H.A. and Eletter, S.F.K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(1), 20–28. Available online: [CrossRef]
64. De Lange, P.E., Melsom, B., Vennerød, C.B. and Westgaard, S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management*, 15(12), 556. Available online: [CrossRef]
65. Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*. Available online: [CrossRef]
66. Rafi, M.A., Shaboj, S.M.I., Miah, M.K., Rasul, I., Islam, M.R. and Ahmed, A. (2024). Explainable AI for Credit Risk Assessment: A Data-Driven Approach to Transparent Lending Decisions. *Journal of Economics, Finance and Accounting Studies*, 6(1), 108–118. Available online: [CrossRef]
67. Kim, H., Cho, H. and Ryu, D. (2018). An empirical study on credit card loan delinquency. *Economic systems*, 42(3), 437–449. Available online: [CrossRef]
68. Gerardi, K., Herkenhoff, K., Ohanian, L.E. and Willen, P. (2013). Unemployment, negative equity, and strategic default. *Available at SSRN 2293152*. Available online: [CrossRef]
69. Al Maruf, A., Kowsar, M.M., Mohiuddin, M. and Mohna, H.A. (2024). Behavioral Factors in Loan Default Prediction A Literature Review On Psychological And Socioeconomic Risk Indicators. *American Journal of Advanced Technology and Engineering Solutions*, 4(01), 43–70. Available online: [CrossRef]
70. Gerardi, K., Herkenhoff, K.F., Ohanian, L.E. and Willen, P.S. (2018). Can't pay or won't pay? Unemployment, negative equity, and strategic default. *The Review of Financial Studies*, 31(3), 1098–1131. Available online: [CrossRef]