Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPLAINABLE AI FOR CREDIT RISK MANAGEMENT IN REGULATED FINANCIAL ENVIRONMENTS: A TRANSPARENT AND AUDITABLE DECISION FRAMEWORK

***[1]SUNILKUMAR REDDY ERAGANENI [2]VENKATA SUDHAKARA REDDY A, [1]KOTI REDDY UPPALA**

[1]Independent Researcher, Uk
[2]Independent Researcher, India.
Email: skre312@gmail.com, sudhaappireddy@gmail.com, uppalakotireddy12@gmail.com
***Corresponding Author:**

ABSTRACT

As artificial intelligence becomes popular for credit risk, the issues of transparency, explainability, and regulatory compliance raise significant challenges. Even though sophisticated predictive techniques can improve risk assessment accuracy, the black-box nature of the decision process associated with those techniques is frequently at odds with the regulated financial industry's reliance on explainability, accountability, and auditability. This study introduces an explainable artificial intelligence (XAI)–driven framework for credit risk management to support transparent, regulator-aligned decision-making. The proposed method provides both global and local interpretability, allowing stakeholders to reason about how important financial and credit factors drive risk at both levels. On the model side, explainability techniques integrated into the model itself yield human-interpretable explanations for model predictions, which ultimately support internal risk governance, compliance audits, and fair lending practices. It provides an explainable decision logic framework that meets regulatory expectations for model transparency while helping avoid the risks of poorly deploying a black-box AI model. These results highlight that risk assessment methods, while remaining consistent with regulatory principles, can provide reliable and consistent credit risk insights when they also enforce explainability guidelines. The framework provides traceable, interpretable, and defensible credit decisions that instil trust among financial institutions, regulators, and customers. To sum up, this work improves upon the existing state of the art in accountable financial analytics by exemplifying an explainability-based paradigm of credit risk. This sets the stage for trust and opens new conversations that explainability itself is a foundational component of both compliance and AI solution trust in regulated financial institutions.

Keywords: *Explainable Artificial Intelligence (XAI), Credit Risk Assessment Model, Transparency and Interpretability Regulatory, Compliance in Financial Systems, Auditable Decision-Making Framework*

1. INTRODUCTION

1.1 Background and Context

Over the past few decades, broad deployment of decision-support systems to find the best solution and to view the beauty and the power of complex systems at work to change decision-making in the financial services industry, and more specifically, decision-making based on experiments that use artificial intelligence—based solution methods that represent advanced science in credit risk management. Credit risk management, the functional pillar on which banks stability and sustainable lending rest, has always enjoyed the advantages that advanced analytics bring in improving the accuracy of risk assessments when appropriate, operational efficiency, and large-scale credit decisions. In ever more competitive, data driven financial markets, models evolve that enables us to assess borrower risk, identify default patterns, and optimise portfolio performance.

However, the move towards more straight-through processing of credit decision-making, exposes an entirely new set of complications, despite enhancements in predictive power. It also should not be overlooked that much of what is involved in financial services has the intrinsically heavy nature of operating within the highly regulated environment highly susceptible to the nature much of the legal, ethical,

regulatory/accounting complex in the fields of the capital, fair lending, consumer protection, and where automated decisions or automated processes are used, transparency and responsibility. Such regulatory frameworks do not only require credit decisions to be correct but also transparent, explainable and auditable. Institutions will have to be able to explain the basis of the adverse credit decisions, meaning that they will need to ensure that their scoring-based systems are applied fairly, meaning that rationales need to be readable for regulators, auditors, and customers impacted by consistent decisions.

While this has led to some separation in the focus of the approaches, it has fed an innate tension between the complexity of AI-based risk models and the transparency required by regulation. Although sophisticated analytical systems are adept at accounting for non-linear relationships and complex interaction with credit data, they can become a bit of a black box to the flesh-and-blood stakeholders involved in the final decision. This opacity not only hinders the financial institutions ability to digest, back and justify the automated credit decisions, but ultimately it opens up a range of questions about accountability, bias and compliance with regulation. With AI systems facing stricter and stricter regulations, we can safely say that explainablity is not just a nice-to-have but a dealbreaker in credit risk.

1.2 The Regulatory Imperative for Explainability

UXResearch on FIs have ramped up with global regulatory authorities stressing transparency and accountability of AI usage in financial services. Several emerging governance standards and frameworks for automated decision-making require financial institutions to retain human control over the use of AI systems, for models to be consistent in their behavior, and that some models provide meaningful explanations for the decisions that have a material impact on people. New regulations will require institutions to show that their credit risk models function in a non-discriminatory and unbiased manner, and that decisions are attributable to rational risk factors.

Model risk management standards Regulator expectations for model risk management have tightened, and new rigorous registers are being introduced, with expectations for ongoing validation, auditability, and human oversight. Methods based on traditional reading of documentation and statistical validation, while essential, fall short in helping to root out the kind of interpretability problems created by complex AI systems. Credit institutions need to be able to articulate how an individual borrower characteristic must be treated in order to be the reason for a positive or negative outcome before they are utilized in an automated decision whether such decisions be discriminating or not, identify points of potential bias and ensure that automated decisions can be shown to be consistent with the regulatory principles and ethical norms regarding the degree of autonomy that the robots can exercise. Not living up to these expectations puts institutions at risk of litigation, reputational and operational damage.

1.3 The Role of Explainable Artificial Intelligence

XAI has become an important enabler for ensuring sophisticated analytical models are aligned with regulatory and governance mandates. XAI stands for a set of techniques and design principles to assist human stakeholders in understanding why an AI model makes specific decisions by exposing the factors, relationships, and reasoning processes that lead to the outputs of the model. Explainability helps to meet compliance requirements, builds trust, and establishes regulatory oversight of automated decision systems in credit risk management.

One can gain insight on XAI methods on a global scale with global explanations which explains the overall model behaviour, as well as local explanations which justifies the individual credit decisions. The interpretability architecture enables financial institutions to support regulatory audits, internal governance processes and support for customers-facing explanations. Integrating explainability directly into credit risk assessment makes it possible to retain the advantages of advanced analytics while promoting transparency, accountability and ethical alignment.

1.4 Research Problem, Gap, and Study Significance

Though the role of XAI in the financial services sector is better understood, the current state of research indicates a long-standing and severe gap

in the literature explainability is viewed as a post-hoc feature that is added to models, instead of a design principle that is implemented throughout the credit risk modelling lifecycle. Majority of the previous research deals with global or local interpretability in abstraction and very few of them offer practical advice of how the outputs of explainability can be operationalised to satisfy tangible regulatory demands such as adverse action documentation, auditing of algorithmic bias, and model governance reporting. This gap between the field of academic XAI research and the real life compliance requirements of regulated financial institutions has not been addressed extensively.

This study is motivated by the need to fill this gap. The core research question which is the focus of this paper is: Which methods to systematically implement explainable AI into credit risk management in a way that both maintains predictive performance and regulatory transparency criteria, as well as supports deployment in real-world financial settings in a manner that aligns with governance principles?

To rectify this issue, the research provides a full-fledged end-to-end explainable AI framework of credit risk management. The framework instantiates transparency and auditability at each point in the modelling pipeline such as data preprocessing and model training, as well as fairness analysis, audit trail generation, and stakeholder-specific explanation outputs. The value of the work is that it is practically oriented: instead of proving explainability in a vacuum, it provides an implementable, regulation-aware framework that financial institutions can be able to embrace to address the changing needs of regulators, auditors, risk managers and customers alike.

2. PROBLEM STATEMENT

2.1 The Black-Box Challenge in Credit Risk Assessment

At its core, credit risk evaluation is about assessing a borrowers capacity and willingness to honour their financial commitments Inherently interpretable risk assessment approaches, with clear expectations of risk, enabled institutions to explain and justify lending decisions. Although these approaches can be useful, they often have a difficult time breaking down some of the more complicated and non-linear patterns existing in the new financial data. On the other hand, more sophisticated AI-based risk models may deliver high accuracy, but often as a black-box model, with little interpretability of the decision logic within.

While black-box credit risk models lend themselves to certain advantages, opacity poses several fundamental problems. First, Credit decisions may be tough to support to applicants or defensible to regulatory examiners; thus, the risk of violating fair lending and other consumer protection standards is increased. Second, lack of transparency in decision processes makes it difficult to quantify and adjust for algorithmic bias and thus determine whether protected groups are adversely impacted by these decisions. Third, lack of explainability erodes the trust of critical stakeholders in risk managers, compliance teams, auditors, and customers. Lastly, complexity of model validation and audit processes are increased when you cannot intuitively examine or trace decision logic over time.

2.2 Limitations of Existing Approaches

While there are some relevant works in recent years that have begun to leverage explainability techniques in credit risk modeling, the aforementioned limitations still apply. Note that we are aware that previous works focus on maximizing the predictive performance of the models and therefore treat explainability as a cross-cutting concern that is addressed de facto post modeling step and not inherently integrated as a design principle in the model design phase. Additionally, the long-standing issue of explainability is often only treated in a mere global or local explanation perspective and aims towards the feasibility of fulfilling the regulation insight.

Yet most XAI proposals do not indicate how they respond to regulatory matters around auditability, documentation and governance. Importantly whilst a hot topic in the research community, significant effort has not been made to demonstrate how explainable models could be physically utilized in financial risk management workflows under a pragmatic understanding of the application of theoretical explainability metrics pervaded in much of the academic literature. This has led to a disparity between the

current state of the art in explainable AI research and the needs of regulated financial institutions.

2.3 Research Objectives

To overcome these issues, this research provides a framework for credit risk management, which we expect it should be driven by explainable AI in a regulated financial ecosystem. The primary objectives of this study are threefold:

- A regulation-compliant explainability framework that synergizes with advanced risk assessment models, and interpretable decision making processes .
- To measure the predictive-calibration interpretability trade-off — and validate that (potentially counter-intuitively) interpretability is not necessarily an accuracy-costly proposition.
- To provide tiered rationales for regulatory scrutiny, internal governance, and external customer-facing justification
- To highlight the applicability of the proposed framework by demonstrating its successful implementation in a real credit risk management context.
- To develop best-practice guidance to ensure that explainable AI is maintainable in regulatory controlled financial risk workflows.

2.4 Research Questions

Therefore, to accomplish these aims, the paper addresses the subsequent research questions;

RQ1: How do existing explainability-augmented risk assessment models compare with the opaque black-box based approaches in predictiveness on standard evaluation metrics?

RQ2: Which factors are most influential in credit decisions across the globe through explainability (and are these even aligned with the fundamental principles of finance risk)?

RQ3: What complementary information will local explanation techniques offer, for an individual credit decision, and in which Config scenarios will that information not be redundant, when local explanations are used to help ensure regulatory compliance?

RQ4: How can explainable AI (XAI) support detecting potential sources of algorithmic bias and support or undermine equitable lending?

RQ5 When deploying explainable AI–based credit risk systems in regulated financial institutions, what practical considerations and governance requirements need to be taken into account?

3. LITERATURE REVIEW

3.1 Machine Learning in Credit Risk Assessment

The use of more sophisticated techniques in credit risk assessment has changed a lot in the last 10 years. The early generation of credit scoring models was based on merely standard statistical methods such as logistic regression and linear discriminant analysis, which yielded highly transparent but less predictive power. With more complex data entering the credit space, ensemble-based and tree-driven models gained traction for their ability to boost default-prediction accuracy, but at the cost of a severe reduction in interpretability of those predictions.

As the evolution of behavior of the borrowers has received its due focus and attention, time-dependent risk modeling has emerged as one of the most popular techniques among quantitative risk modelers. However, the sequential information is beneficial for risk assessment according to the work in [10], where time series–based models were implemented for credit default prediction. To improve institutional-level risk monitoring, (particularly for small financial institutions), ensemble based early warning systems were proposed by Li and Zhang [8].

Credit risk modeling has also taken a step further with explainable ensemble methods. Li et al. An explainable gradient boosting–based approach for predicting default borrowers in social lending platforms [13] enhances predictive performance and transparency concerns. Tang et al. They extended the use of interpretable machine learning methods, which have gained popularity for consumer lending applications, to predicting the risk of default by banks in different regulatory

environments [6]. Taken together, these studies illuminate the increasing temptation with the use of highly complex predictive models, but warns against pitfall of sacrificing a certain level explainability, in turn yielding a trade-off between performance and interpretability.

3.2 Evolution and Application of Explainable AI Techniques

The challenge, particularly acute in high-stakes domains such as finance, is that sophisticated predictive models (machine learning models) can be opaque; Explainable Artificial Intelligence (XAI) is a response to this challenge. There are many XAI techniques; however, SHapley Additive exPlanations (SHAP) and Local Interpretable model-agnostic Explanations (LIME) are two of the frequently used ones because of their model-agnostic nature and the benefit of interpretation.

SHAP based explenability is matured in finance, particularly in risk management in this field. SHAP: Shapley additive explanations, A unified approach to interpreting model — The Hidden Seams: A SHAP Shaped Tack to Transparent Financial Decisions, Srivalli, Sumanthi[19] Lin et al. As an example where SHAP was applied as an interpretability medium of models that discover the most significant risk factors for auto loans [20]. Eshan et al. A few latest works merge SHAP and LIME to get better balanced global and local explanations [1], i.e., to generalise the model on-top of its corresponding top view on to human-level.

Comparative evaluations of explainability techniques have highlighted their strengths in previous work. In particular, a commercial credit limit prediction comparison between SHAP and LIME carried out by Kocoglu and Ersoz [5] showed that SHAP delivers more stable insights to the global side, while LIME succeeds on the local side by providing more interpretable explanations for individual decisions. Zhang et al. Tang et al. [22] SHAP and LIME and Prediction of Bond Default Risk] will introduce methods for assessing interpretability quality using SHAP and LIME, as well as a practical example of applying the quantification pipeline using bond default predictions. These observations strengthen the motivation for an exploitative combination of explainability methods instead of relying on their individual quality alone.

Given that such research works often come by some feasible experimental designs at the lab level, we then see real-life implementations of explainable credit scoring systems breathe at the operational level. Japinye and Adedugbe [11] also observed both scalability and consistency from SHAP-calibrated ensemble models across other lending markets. In [12], Renner et al. We had applied explainable AI approaches on loan platforms to ensure transparency, trust and regulatory compliance. In particular Zhang [9] said that if the prediction results are available to the end-users of the machine-assisted model, then the rationale of the machine-assisted decision should be able to be explained in an intelligible form that is accessible to them.

3.3 Regulatory Frameworks and Compliance Requirements

There has been an increase in regulatory scrutiny into AI adoption across financial services; their focus has primarily been on the principles of transparency, accountability, and auditability. Explainable AI is one of the mechanisms that have been identified to help manage regulatory expectations in the context of automated decision-making [2].

For example, Malali & Madugula [3] proposed the adoption of explainable AI for proactive regulatory compliance and auditing in financial market environments - emphasizing that explainability should always be an integral aspect contributing to the entire lifecycle of a model. Also Read : Exploratory Data Analysis R Bagwe [15] explains how explainability acts as a bridge between high class analytics and consumer as well as regulatory audit transparency.

The regulatory and ethical issues around Artificial Intelligence in credit assessment have been investigated by new financial products. Mishra et al. Buy Now, Pay Later credit assessment systems, such as [30], raise concerns about fair lending, bias, and consumer protection Kamruzzaman et al. To evince more, how advanced tech and digital solutions, like AI and Automation, can be the AI-Enabled capability of Regulatory Compliance in a transformed Digital banking landscape [29].

On broader AI governance frameworks in finance: features of AI governance should be based on the idea of explainability. Explainable AI Models: To increase transparency in compliance workflows, Desai [17] developed explainable AI models for financial regulatory audits. Jain et al. Incorporation of compliance regulation through AI: The contribution of Abubakar et al.[18] explored a conceptual framework for AI from [18] organizational and technological perspectives. Kurshan et al. Governance in AI models and self-regulating systems in Financial Services was also explored, with a similar call for inteligible decision logic [23].

Adegbola [16] applies explainable AI–based risk scoring of letters of credit to provide a specific compliance application [74]. These applications focus on regulatory compliance requirements, unique to trade finance. Kothandapani [36] explored issues of automation and large language models in the context of processes of regulatory oversight. OLAWORE et al. Frameworks for ethical auditing and cybersecurity governance have been developed [27], encompassing end-to-end explainable AI, compliance monitoring, and stakeholder trust.

3.4 Balancing Predictive Performance and Interpretability

The biggest challenge for AI-based XRMs is always taking the interpretability vs predictive accuracy trade off head on. Recent work has proven that one can get competitive performance with explainability.

Wang et al. hu et al. Risk management has in fact been explored in the explainable machine learning literature however they observationally qualify the accuracy interpretable tradeoff and in particular observe that both objectives is achieved through a well-designed XAI approach. [2] Khan et al. Furthermore, encouraging this view, [7] proposed a set of interpretable modelling methods that fuse classical statistical style thoughts to modern machine learning style thoughts, thereby cementing classical principles in the top performing AI systems.

Model Interpretability : one that comes with a wealth of theory already, at least on the high level interpretability side. Key dimensions, challenges, and actors in the operationalization of responsible AI were identified by Owen [14]. Chinnaraju[34] has developed conceptual frameworks for transparent and responsible decision making for AI, observing that explainability is a fundamental condition for organizational accountability.

Past research shows credit risk quantification applications and decision support systems based on explainable AI can deliver a good performance versus stakeholder trust trade-off. Nallakaruppan et al. Developed explainable artificial intelligence (XAI) based accurate and transparent credit risk assessment systems [28]. Ravi et al. In [24], their work towards explainable approaches to credit scoring and loan approvals demonstrated us that transparency can be achieved without compromising operational efficiency by modifying existing models with a few tweaks.

3.5 Emerging Applications and Specialized Domains

ExAI applications are now applied across a broader spectrum of financial risk and even compliance areas much wider than the traditional credit risk assessment and credit scoring application domain. Enhancing Prediction Precision and Fraud Identification Potency in the Financial Markets with AI Risk Assessment Models Oko-Odion [4] Vyasa [35] also examined the same international level but in financial risk management and forecasting areas.

Ahmed et al. The transferability of explainable AI has been seen in reports of large scale banking AI innovations from digital transactions to risk management, compliance frameworks and forecasting systems [32], demonstrating the amount of progress explainable AI has been able to make in their respective fields. Lastly, De Silva [21] in the context ofhuman centered credit risk, directly discusses interpretability as well as arbitration in the algorithmic risk domain, as being human focused.

Predictive modelling risk management approach for transparency in a community bank and how this community bank model predictive modelling risk management approach can support systemic risk management and early warning systems Proactive Deployment of Explainable AI [25] Li and Ling implemented explainable AI prediction

models into risk monitoring for a community bank.

3.6 Gaps, Limitations of Existing Work, and How the Proposed Framework Addresses Them

There are still critical limitations in spite of the great progress made by the current literature. First, although a number of studies implement explainability approaches in credit risk modelling, they largely view explainability as a post-hoc approach and not an architectural design principle, i.e. transparency is added to the model afterwards, as opposed to being implemented throughout the modelling lifecycle. Second, the majority of current methods focus on either global or local explainability alone, such as SHAP to learn the importance of aggregate features, and LIME to learn the importance of individual decisions, without jointly considering both on a single operationally consistent framework. Third and most importantly, the literature available provides little specific direction on how explainability outputs can be directly mapped to particular regulatory demands in the form of adverse action notice, bias audit and model governance documentation. There are high-level reviews of model-agnostic XAI approaches [33], although these only confirm instead of addressing the long-standing theory-practice gap in regulated finance.

These limitations are explicitly covered by the framework suggested in this paper. In contrast to previous work, explainability is a fundamental element throughout the modelling pipeline, including data cleaning and model training all the way to evaluation, fairness analysis, and generation of stakeholder output. The framework unites both SHAP-based global interpretability and LIME-based local interpretability into a single governance-consistent architecture by allowing financial institutions to meet the explanation requirements of regulators, auditors, risk managers, and customers at the same time. Moreover, integrating a specific regulatory compliance and governance module, such as bias detection, audit trail creation, and support of adverse actions, the proposed structure will make the theoretical explainability practical and deployable compliance infrastructure. This makes the study a direct answer to the most important and least addressed gap that was found throughout the reviewed literature.

4. METHODOLOGY

The explainable AI framework for smooth control of credit risk in end users in a regulated financial environments is proposed in this section. It aims to find the sweet spot between predictive power and explainability/compliance. As seen in Fig. The framework is modular in nature, and it consists of four modules and eight components variables with the interconnection among them enabling it to perform multiple functions. The input layer: Which consists of the input from the user The data preprocessing module: This module has the role of pre processing the data to fit for ML model The model training module: The method for training the actual model The hyperparameter optimization Module: The hyperparameter optimization The model evaluation module: The method for the evaluation of the model The explainability integration module: The module for involves the integration of the explanation of the model The regulatory compliance and governance module: Implementation of RCG The output layer: Which is about the output to the user These are the parts of an end-to-end pipeline to convert raw borrower data into risk predictions with interpretable explanations and documentation for compliance [28]

Therefore, this architectural design does focus on traceability and auditability during the lifecycle. Data is passed through all forms of pre-processing, evaluation, explainability, and model training, with all intermediate results logged in an easily-accessed manner for governance and regulatory review. To make that happen, this architecture relates a credit decision back to the input variables, the modeling selections, and the explanation output – and specifications required for fair lending and model risk management. The framework not only operates the only batch design for real portfolio analysis, but in addition also allows, due to the reactivity of its design, for real-time credit scoring and thus operational lending deployment in industrial settings are shown in figure 1.

Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

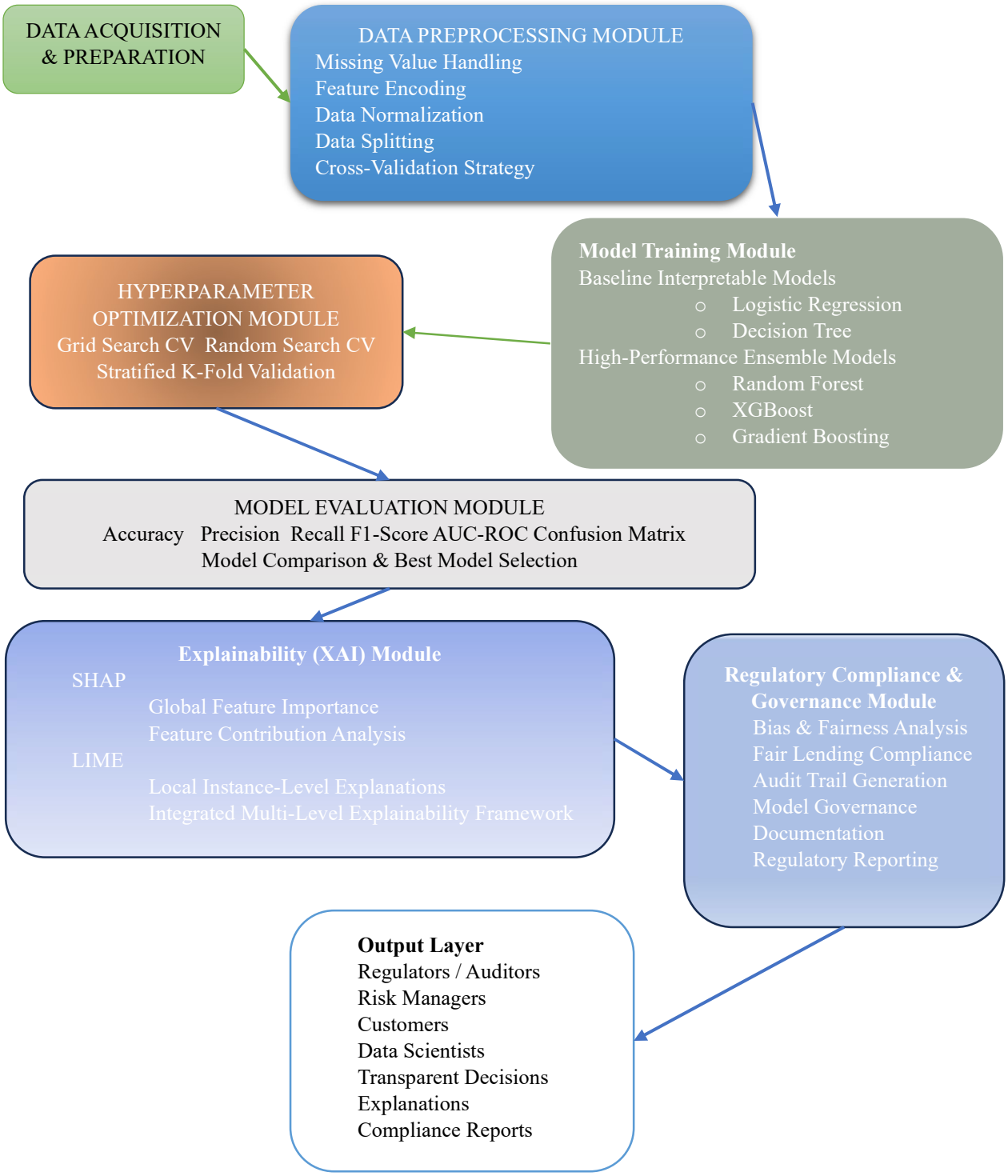

Figure 1. Detailed System Architecture Of The Proposed Explainable AI–Based Credit Risk Management Framework, Illustrating Data Processing, Model Training, Explainability Integration, And Regulatory Compliance Workflows.

4.1 Overall System Architecture and Design Principles

We then propose a framework based on some key design principles that distinguish our work from traditional black-box credit risk models. First, the explainability is built-in as an integral architectural component; as opposed to being an after-thought component, thus the transparency in the decision making is more preserved. Second, the framework uses a multi-model approach along with interpretable baseline models and ensemble learners which provide higher accuracy results. Thus, it allows for a standardized comparison of the trade-off between prediction quality and model transparency, leading the institutions to the model that best fits to their regulatory requirements and risk appetite. Lastly, the framework embeds multi-level explainability at the global level (understanding overall model behavior) and local level (explaining individual credit decisions) to cater to the interests of a broad set of stakeholders.

Its information flow is in the form of layers. In this first layer, pixel data gets passed raw credit and lending data as structured feature representations. Data Preprocessing Module: Handles missing values, encodes categorical features and normalizes numerical features, so that they are statistically suitable for every machine learning algorithms. To retain the class distributions as in the original credit default data that are highly imbalanced, the processed data are then split into training and testing sets, using the stratified sampling.

Model training module implements various supervised learning algorithms ranging from interpretable base algorithms to more complex ensemble methodologies. Logistic regression and decision trees, for example, allow for interpretable benchmarks with linear coefficients and rule-based decision structure respectively. At the same time, Random Forest, Gradient Boosting and XGBoost is implemented to handle the non-linear relationships and complex interaction of features providing the gain in the prediction performance, but the cost of the interpretability.

The Hyperparameter Optimization module employs methodical search techniques like Grid Search and Random Search with cross-validation in order to uplift the model fitting. Furthermore, the final model is chosen based primarily on the AUC-ROC metric that is robust for a ranking performance by class imbalance. In the end, during the optimization stage, the model evaluation module offers a performance evaluation of multiple complementary metrics such as accuracy, precision, recall, F1-score, confusion matrices, AUC-ROC, allowing balanced classification performance and error characteristic analysis between the different models.

We tackle explainability by designing an integration module for a specific domain that offers human-interpretable explanations accompanying the model predictions. SHAP produces global explanations: attributing features contributions to the entire population while LIME provides local, instance-level explanations: explaining the rationale behind a specific credit decision. Together, these techniques constitute a full gold standard interpretability a la carte system as deployed in many regulatory reports or in some communications with customers.

The third part is the regulatory compliance and governance window, which ensures that the proposed frame is also responsible deployed, considering the compliance property of a regulated financial environment. IgazaBias and Fairness (statistical parity and disparate impact), with audit trails logging all steps with paths to specs/metrics; data preprocessing, specs on model, performance measure specs, and explanation outputs They assist in regulatory reviews and internal model governance processes.

The last layer one gives outputs that are relative to particular stakeholders Link for Compliance reports and audit documentation for regulators and auditors, portfolio level risk insights and performance dashboards for risk managers, explanation of loan decisions for applicants, diagnostics and validation outputs for data scientists. In summary, the architecture proposed in this paper states that credit risk assessment should be done in a transparent, auditable and governance-compliant way in order to safely and effectively harness the potential of cutting-edge machine learning techniques in regulatory domains like the lending space.

4.2 Input Layer: Credit and Lending Data Representation

The first layer of the proposed credit risk assessment framework is the input layer, where the input layer indicates how borrower information is transformed and represented in a way to perform a machine learning analysis. Loan applications are represented as a fixed-length vector of numerical features that quantify multiple dimensions of borrower characteristics, which allows the inputs to be fed into different prediction/explainability modules similarly.

Let the dataset consist of Nborrower records. Each borrower $i$is represented by a feature vector in a d-dimensional space:

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}] \in \mathbb{R}^d, i = 1,2, \dots, N$$

where $d$denotes the total number of attributes collected during the loan application and credit evaluation process. The complete dataset can thus be expressed as:

$$\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_N\}$$

The feature vector is structured to capture three primary categories of information, each reflecting a distinct aspect of credit risk.

Financial features (loan amount, interest rate, installment amount, annual income, debt-to-income ratio, revolving credit balance, revolving credit utilization rate, loan term) define the borrower's ability to repay the loan. These characteristics flow directly into ability-to-repay, and have been central to the traditional model of credit scoring for eons.

Demographic characteristics add contextual information about the borrower in terms of age, profession, and employment. Such criteria includes factors like whether or not you're a homeowner, your length of employment, type of application, income verification status, and other geographic criteria like state or ZIP code. These traits are not directly used to assess creditworthiness but when examined in conjunction with financial metrics, they are usually indicative of economic stability and repayment behavior.

With respect to both historical and credit history based features, this begins to numerically detail the prior history of borrowing, repayment performance, and the context of the borrower's repayment stability over the past few years, including the number of past due instances over the previous 2 years, including (recent inquiries into credit), the number of open credit accounts, total number of credit account, public records including child support, tax liens, and the number of credit lines held, and a simple calculation of measured ages of credit history (from the first credit line to the current point in time). These variables reflect the history of credit behaviour and serve as a good predictor for default risk going forward.

We treat the prediction task as a supervised binary classification problem. Associated with it can be a target variable for every borrower.

$$y_i \in \{0,1\}$$

where $y_i = 1$denotes high credit risk corresponding to default or adverse loan outcome, and $y_i = 0$denotes acceptable credit quality corresponding to successful repayment. The learning objective is to estimate a function:

$$f: \mathbb{R}^d \to [0,1]$$

that maps the feature vector $\mathbf{x}_i$to a predicted probability of default $\hat{y}_i = f(\mathrm{x}_i)$.

We specifically designed the feature representation to be both complete but human-interpretable. More features give predictive power but redundancy and noise can introduce complexity and hinder feature explainability. Consequently, the features selected consist only of information that can typically be obtained from standard loan applications and records available through credit bureaus, guaranteeing applicability in practice.

In order to comply with restrictions on regulatory, protected attributes such as race, gender and age are explicitly removed from the feature set. However, the permitted variables are subsequently assessed in the governance module to determine whether they serve as indirect proxies for the protected characteristics via fairness and bias analysis.

4.3 Data Preprocessing Module

This module does a processing of raw, heterogeneous credit data to convert it into a clean, homogeneous, and statically recommended format for machine learning. Datasets integrated from the lending platform and credit bureaus hold a large amount of missing values, mixed data types and diverse scales of features. When unhandled, these problems can reduce prediction performance, bias model predictions and harm interpretability.

There are four sequential steps in the preprocessing pipeline: handling missing values, encoding categorical features, normalizing numerical features, and splitting the dataset.

4.3.1 Missing Value Handling

Let x_{ij} denote the value of feature j for borrower i. If x_{ij} is missing, an imputation strategy is applied based on the feature type.

For numerical features, mean imputation is used:

$$x_{ij} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed} \\ \mu_j, & \text{if } x_{ij} \text{ is missing} \end{cases}$$

where μ_j is the mean of feature j computed over the training set. For features with skewed distributions, the median may be used instead to reduce sensitivity to outliers.

For categorical variables, missing values are imputed using the mode, defined as the most frequently occurring category in the training data.

4.3.2 Categorical Feature Encoding

Since machine learning algorithms require numerical inputs, categorical variables are transformed into numerical representations.

For nominal categorical variables with no intrinsic ordering, one-hot encoding is applied. A categorical variable with k categories is transformed into k binary variables, where each variable indicates membership in a specific category:

$$x_{ij}^{(k)} = \begin{cases} 1, & \text{if borrower } i \text{ belongs to category } k \\ 0, & \text{otherwise} \end{cases}$$

For ordinal categorical variables with meaningful order, label encoding is used, mapping categories to integer values that preserve ordinal relationships.

4.3.3 Numerical Feature Normalization

Numerical features in credit data often vary across different scales. To prevent features with large magnitudes from dominating model training, z-score normalization is applied:

$$x_{ij}^{\text{norm}} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where μ_j and σ_j represent the mean and standard deviation of feature j, computed exclusively from the training data. This transformation ensures zero mean and unit variance for each numerical feature.

4.3.4 Data Partitioning and Validation Strategy

After preprocessing, the dataset is partitioned into training and testing subsets:

$$X = X_{\text{train}} \cup X_{\text{test}}, X_{\text{train}} \cap X_{\text{test}} = \emptyset$$

The split is stratified so that the ratio of default and non-default cases is the same in both subsets. Normally, 70% data part for the training and 30% part for testing.

Besides the train–test split, stratified k-fold cross-validation is used during model development and hyperparameter tuning. In k-fold cross validation the training data is distributed over the folds and the model is trained k times where $k-1$ each fold is held out for validation and the other folds are used for training. The performance metrics are averaged across folds to obtain a stable estimate of generalization performance.

All preprocessing (e.g. imputation, encoding, normalization statistics, etc.) are learned using only the training data and then applied to the test data the same way. To see if the estimated performance estimates from the model can be on par with the real world deployment, in order to prevent data leakage. All preprocessing steps are logged, and hence there is proper documentation to aid in auditability and regulatory reviews.

4.4 Model Training Module

The model training module adopts a two-pronged approach that includes inherently interpretable models and also strong ensemble learners. What this means is that we can systematically compare transparency vs predictive power, and agencies can look for models that match the regulatory boundaries or business requirements. As all the models are based on supervised learning, therefore they are all evaluated with the same data split so that classification results are comparable.

4.4.1 Logistic Regression

Logistic regression We first consider logistic regression out of our interpretable baselines given its regulatory acceptance and rules-based decision making. It is a logistic function of a linear combination of the inputs, that outputs the probability of a default event.

$$P(y_i = 1 \mid \mathrm{x}_i) = \sigma(\mathrm{w}^\top \mathrm{x}_i + b)$$

where $\mathbf{w} \in \mathbb{R}^d$is the weight vector, $b$is the bias term, and $\sigma(z) = \frac{1}{1+e^{-z}}$is the sigmoid function.

Model parameters are estimated by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

Regularization terms (L1 or L2) may be added to control model complexity and prevent overfitting. The interpretability of logistic regression arises from the direct relationship between feature coefficients and the log-odds of default, enabling straightforward explanation of credit decisions.

4.4.2 Decision Tree Classifier

Decision trees provide rule-based interpretability by recursively partitioning the feature space into homogeneous regions. At each node, the optimal split is selected by maximizing information gain:

$$IG(S, x_j) = H(S) - \sum_{v \in V_j} \frac{\mid S_v \mid}{\mid S \mid} H(S_v)$$

where $H(S)$is the entropy of sample set S, and $S_v$denotes subsets formed by splitting on feature x_j.

It provides a series of feature-threshold rules, which we call the decision path, as an explanation for the predictions, thus enabling a human-understandable justification of the credit decision for each individual. Hyperparameter tuning limits the depth of the tree and the minimum number of samples needed.

4.4.3 Random Forest

Random forest is a step up from single decision trees as it builds multiple trees on bootstrap samples of the data and considers a random sample of features at each split. Consider the number of trees; T the final prediction is obtained by simply averaging the predictions of the individual trees:

$$\hat{y}_i = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_i)$$

This is an ensemble strategy that helps in variance reduction and therefore gives better performance in terms of generalization. Although Random Forest is able to easily capture non-linear relationships and data interactions, its aggregated structure renders it directly less interpretable than the simpler tree structure with only one layer — thus there is motivation for applying post-hoc explainability techniques.

4.4.4 Gradient Boosting and XGBoost

Gradient Boosting builds an ensemble sequentially, where each new tree is fitted on the residuals of the current model. At iteration m, the model is updated as:

$$F_m(\mathrm{x}) = F_{m-1}(\mathrm{x}) + \eta h_m(\mathrm{x})$$

where $h_m$is the weak learner trained to minimize the loss gradient, and $\eta$is the learning rate.

XGBoost extends this formulation by incorporating regularization on tree complexity and leaf weights, optimizing the following objective:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{m} \Omega(h_m)$$

where $\Omega(\cdot)$penalizes model complexity. They achieve state-of-the-art predictive accuracy and ranking performance, but they operate as black-box models that must have explainability incorporated for regulatory purposes.

The models are trained first with defaults, and optimized via individual systematic hyperparameter optimiation by cross-validation. This invariance in training guarantees best performance comparison as well as good generalization on unseen credit applicants.

4.5 Hyperparameter Optimization Module

Some factors that determine the prediction performance and generalization of machine learning models are modeled by hyperparameters. While model parameters are directly learned from data during the training phase, hyperparameters — the important aspects that govern the architecture of a model and the way it learns — must be specified prior to training. Inappropriate hyperparameters will cause either underfitting, where a model fails to capture the genuine patterns, or overfitting, where a model memorizes the training samples but does not generalize well with other samples. Which means in regulated environments which need to have credit risk modeling done in an explainable, reliable, proven manner systematic hyperparameter optimization is a must.

Our framework adds a hyperparameter optimization module that finds the optimal hyperparameter values with a systematic search combined with stratified cross-validation. Two approaches complement each other — Grid Search and Random Search. Grid Search exhaustively evaluates a given discrete set of hyperparameter values, which is ideal for models with small parameter spaces to ensure full exploration. On the other hand, Random Search randomly samples configurations from the hyperparameter space that is defined by user-specified distributions, allowing to more efficiently sample the hyperparameter search space if there are many, or continuous parameters, with virtually no increase in computational cost. The first choice is model dependent, where Grid Search is used for simpler models while Random Search is then used for more complex ensemble learners.

Hyperparameter selection is conducted using stratified k-fold cross-validation, where the training dataset is partitioned into Kfolds while preserving class distributions. For each candidate hyperparameter configuration θ, the model is trained and validated across all folds, and the average validation performance is computed. The optimization objective is defined as:

$$\theta^* = \arg\max_{\theta \in \Theta} \frac{1}{K} \sum_{k=1}^{K} \text{AUC}\left(f_\theta^{(k)}\right)$$

where Θdenotes the set of candidate hyperparameter configurations, $f_\theta^{(k)}$represents the model trained with configuration θon the k-th fold, and AUC denotes the Area Under the Receiver Operating Characteristic Curve.

We choose AUC-ROC as the main optimisation metric that quantifies the model's ability to rank borrowers in order of highest to lowest likelihood of default at all classification thresholds. This attribute makes it particularly suitable for the credit risk scoring problem, which is a domain with extremely rare positive outcomes (class imbalance) and dynamic decision thresholds due to idiosyncratic institutional risk appetite. All other secondary metrics (precision, recall, f1 score, etc.) will be monitored and reported to ensure the classification behaviour is not dominated by one class or another, and will not be used immediately for hyperparameter selection.

Hyperparameter spaces for various models are specified based on heuristics and past empirical evidence. Accordingly, the Hyper-parameters we consider for Logistic Regression are regularization strength, C penalty type, solver, max_iter. We will go through Decision Tree tuning as well, such as max depth, min samples to split, min samples at each leaf, and criteria for impurity. Random Forest optimisation: number of trees, maximum tree depth, feature sampling strategy and bootstrap setting. Some of the most significant hyperparameters in Gradient Boosting and XGBoost are the number of estimators, learning rate, maximum depth, subsample ratio and regularization terms controlling model complexity.

Indeed, determining the configuration for the performance on a target dataset may be extremely expensive (which is the root cause for many of the methodologies introduced in the framework) thus its cost-saving strategies consist in parallel-configuration evaluating, early stopping (for boosting models), and coarse-to-fine search strategies to iteratively refine search in the most promising areas of parameter space. All configurations that are evaluated and their respective validation scores are logged to provide traceability, reproducibility, and meet regulatory audit requirements.

The model corresponding to the best hyperparameter setting for each candidate model after then be trained with the retrieved hyperparameters on the entire training set. The final performance evaluation is performed on the completely blind held-out testing set that did not participate in the optimization process. The individual tuning and testing phases provide an unbiased estimate of real-world performance (without a more positively biased score than should be expected which is a danger to model governance and possible regulation/legislation compliance).

4.6 Model Evaluation and Explainability Framework

A model evaluation and explainability framework that takes into consideration predictive performance, in addition to the explainability results to ensure that the end-trained model is suitable for integration in regulated credit risk management environments. Only using the accuracy based scores may not fully justify the use of and deployment of the model used and the Financial Institutions needs to assess the models not with just one lens but several lenses including costs of errors, ability to rank, fairness implications and the interpretability (of the scores). Our proposed framework can be the solution to these limitations by accommodating the quantitative performance measurement of the models with the explainable artificial intelligence techniques to have a comprehensive and governance aligned picture of the models performance.

4.6.1 Predictive Performance Evaluation

We assess model performance using a number of complementary metrics, to capture different angles of the classification quality. Let TP, TN, FP, and FNdenote the number of true positives, true negatives, false positives, and false negatives, respectively. Overall accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Although accuracy offers a succinct summary of overall predictive performance, credit risk measurement demands more than that. Confusion matrix analysisHowever, they are useful for analysing patterns of misclassification and the cost of misclassifications to the business directly. False negative translates into defaults that could have otherwise been detected and therefore loss-making, whereas a false positive represents lost creditworthy applicants and therefore lost revenue opportunities, all in terms of detection error specifically.

And finally, class-wise metrics sharpen the evaluation. Precision overall is a measure of the expected quality of predicted defaults:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall measures the model's ability to identify actual defaults:

$$\text{Recall} = \frac{TP}{TP + FN}$$

These metrics exhibit an inherent trade-off governed by the classification threshold. To balance this trade-off, the F1-score is computed as the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

We then measure the F1-score, which penalizes imbalanced precision and recall too much and is particularly useful when comparing model performance on imbalanced classes.

The most common metric to evaluate the ranking performance in a threshold-independent way is the Area Under the Receiver Operating

Characteristic Curve (AUC-ROC). The ROC curve is formed with true positive rate against false positive rate at different thresholds. AUC-ROC measures the likelihood that a randomly chosen defaulting borrower ranks higher (i.e. score is higher compared to other borrowers) than one that does not default. Values range between 0.5 for a machine playing a random game to 1.0 for perfect discrimination. Since AUC-ROC is class imbalance agnostic, independent of threshold selection, it is naturally a desired metric for credit risk modeling.

These metrics are computed on the validation folds for cross-validation and then on the held-out test set for the final assessment. Side by side comparison of models enable decisions on predictive performance, stability, and suitability for regulation rather than anchoring on a specific measure.

4.6.2 Explainability via SHAP and LIME

SAS says here that predictive metrics provide a quantitative assessment of the model performance but the explainability techniques help us to understand the reasons for such predictions by the model. 2nd framework is a framework that complements SHAP and LIME putting together the complementary global (SHAP) and local (LIME) interpretability.

SHAP (SHapley Additive exPlanations) theoretically derives a unified approach to both global and instance-level explanation based on cooperative game-theoretic principles. For a single prediction, SHAP breaks down the difference between the output of the model and the population average into contributions from the various features.

The Shapley value for feature $j$is defined as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)]$$

where $F$is the set of all features and $f(\cdot)$denotes the model prediction function. The properties of additivity, consistency and the local accuracy of SHAP values are strongly desirable in regulatory contexts.

Global explainability is facilitated by such embedding of intelligible risk drivers, whose absolute SHAP values dominate the overall score for aggregation across all instances. Then, that analysis shows that the model behaves according to real near-terms credit risk theory etc. and does not rely on false correlations (lies) or surrogates for protected attributes as parameters. Force plots to depict instance-level shap explanations that can illustrate how each features value pushes the predicted risk up or down relative to the baseline in a transparent way to justify a specific decision.

In contrast, we provide SHAP with LIME (Local Interpretable Model-agnostic Explanations) which is an intuitive, instance-level explanation method. Model agnostic — LIME approximates the complex model locally by a simple surrogate model fitted on perturbation of the instances neighbouring the instance of interest. The optimization objective is:

$$\arg\min_{g \in G} \sum_{z} \pi_x(z)(f(z) - g(z))^2 + \Omega(g)$$

where $g$is the interpretable model, $\pi_x(z)$weights proximity to the original instance, and $\Omega(g)$penalizes model complexity. The resulting explanation identifies a small set of features that most strongly influenced the individual prediction, expressed in a form suitable for communication to non-technical stakeholders.

4.6.3 Integrated Evaluation and Governance Perspective

Performance evaluation and explainability complement each other and together help in the right model selection and safe deployment. A predictive metric quantifies the trade-off between discrimination and error, but SHAP and LIME translate the logic of the model itself. If ensemble models produce large AUC-ROC lifts, they may be preferred as long as interpretability is recovered through means of explainability methods. In turn, when the extra performance boost is negligible, we can choose a simple model, because the upper hand of regressive transparency.

Fairness and regulatory compliance are evaluated by examining disparities or instabilities in evaluation results and explainability outputs over population segments. Auditable and regressible:

All metrics, plot, and explanation artifacts documented and archived in model governance framework with review and monitoring of deployed models over time horizon.

In this paper, we propose a novel end-to-end framework for evaluation and explainability of credit risk models, which ensures the models are high-performing, transparent, and accountable; thus, they can meet the regulatory expectations.

4.7 Regulatory Compliance and Governance Module

Regulatory compliance and governance module: It guarantees that the framework for the credit risk offer is functioning in the legal, ethical, and supervisory frameworks of regulated financial environments. This may be a more technical analysis, but is still informed by a consideration of responsible AI in general. Predictive accuracy and explainability are two elements of responsible AI — no doubt — but alone don't ensure an organization is meeting its fair lending obligations, consumer protection regulations, or model risk management requirements. It embeds compliance, accountability, and auditability into the modeling lifecycle where it is upfront and not post-factuncts.

It has 3 modules which are interlinked: bias detection and fairness metric, Audit trail, and real-time model governance. Spread across the model lifecycle — development, validation, deployment and periodic revalidation — these functions address compliance, transparency and reliability over time.

4.7.1 Bias Detection and Fairness Analysis

It is a regulatory necessity in addition to being an ethical imperative to go to unsurpassed levels to ensure that credit decision-making is fair and equitable. Indirect discrimination happens in practice, especially in machine learning, when you never include certain protected attributes as input to your model which our proxy variables have a correlation with. At the same time, the framework also reduces this risk as the framework systematically anlayses for fairness against the model by doing trials which test whether, the outcomes of the model have a difference in between the demographic groups.

The central fairness criterion employed is disparate impact, defined as the positive outcomes ratio of a protected group to that of a baseline group. Disparate Impact below Regulatory Thresholds For Well-Accepted Regulations Suggests Discrimination and Triggers Increased Scrutiny. These methods compute fairness metrics with respect to external demographic information, which they collect as part of the analysis for compliance monitoring, and apply to multiple protected groups.

Where differences exist, we do diagnostic analyses of the source of that difference, Correlation analysis is used in proxy variables detection; however, in explainability output, it is used to check whether unwanted prediction of protected groups is due to a feature on an unequal level. As per the requirement mitigation strategies (feature refinement, instance reweighting, threshold calibration, fairness-aware regularization etc.) will be done and evaluated, results will be recorded.

Model operationalization would entail periodic checking of the operational definition of fairness against any changes in the population, economic condition parameters, and portfolio characteristics. For example, automating the monitoring mechanisms makes it easier to identify deviations from acceptable fairness ranges, so that corrective action can be taken sooner.

4.7.2 Audit Trail Generation and Model Governance

Regulatory supervision on the IDD is important for the IDD to have a clear sense of rationale and traceability, as well as for internal models risk management. Ensuring a reproducible and transparent path of modeling decisions, it saves an archive of every model and deployment step taken in the pipeline.

These in-turn includes information such as preprocessing details, modeling decisions, hyperparameter tuning results, evaluation metrics, explainability results and fairness assessment (audit log). All things are time-stamped, versioned and held across model versions which allows storing current modal comparisons to keep behavior similar in time to yield performance measures or monitor for concept drift.

The governance of models is not merely the documentation of individual models, but will require independent validation, oversight by senior management and sign-off processes. Other independent reviewers evaluate models for soundness of method, performance proper stability, limitations, and appropriateness for intended use. Production Monitoring- predictions vs true values over time, score distributions, calibration & fairness metrics Governance Framework for a Sustainable Regulation Alignment that is Responsible and Transparent

4.7.3 Adverse Action Notices and Customer Communication

The thing is, once the borrower gets an adverse credit judgement (and let us be honest here, this means literally anything that prevents them getting credit no matter how minute) they have to be told exactly what were the main reasons for the adverse credit decision. Additionally, these local explanations also support Expainability-driven adverse action notices, thus serving this need.

It makes a short list of key drivers for each bad decision, and this is then expressed in terms that are easy for customers to understand. The evaluation is based on risk factors that are objective in nature and if applicable, the borrower is advised about the measures they can take to improve their credit profile. That means they should produce customer-facing explanations that accurately reflect underlying model behavior, and consistency checks help strike that balance.

Beyond any mandatory notifications, the framework similarly enables ex-ante transparency whereby explanatory reports are provided to applicants that are granted credit and interaction tools will enable borrowers to model how variations in values of financial characteristics might affect different credit ratings at different times in the future. Such communication channels foster trust, fairness, and informed choice in financial decisions.

4.8 Output Layer: Stakeholder-Specific Deliverables

The output layer customizes output produced by the framework to the different information needs of stakeholder groups since audiences will inherently demand different granularities of detail and mode of delivery based on their familiarity with and the role they have in health management.

Lay out compliance with fair lending and model risk management guidelines with thorough validation reports, fairness assessments, and end-to-end audit trails for regulators and auditors. This enables independent verification of the model development practices, performance claims, and governance controls applicable to such materials.

Risk managers and business users view operational dashboards that deliver snapshots of portfolio risk, approval rates, and model performance and fairness metrics. Visual summaries of feature importance and segment-level analyses enable decision support and risk management before and after deploying a ML model.

Some of those are clearer adverse action notices that give customers and loan applicants actionable—and accessible—insights into credit decisions.29 Other reports have provided more education and contextualization around risk assessments.

The entire technical documentation of preprocessing, tuning ,performance diagnostic, explainability, and reproducible code artifacts are at the scratch of the plate for data scientist and model developers to work with. These outputs allow for continued optimization, validation, and modeling research.

It cross-cuts regulatory output design, operational output design, customer output design and technical validity output design to wrap the regulatory model into an end-to-end explainable system for credit risk management that is governance-aligned and able to pass the scrutiny of a single model over the life-cycle and across households, loans and portfolioloans w

4.9 Overall Algorithm: Explainable Credit Risk Management Frameworkithin the same framework.

All methods described in the subsections above are encapsulated in a single algorithmic framework, which formalizes the generic process of explainable credit risk assessment. The algorithm above is a high-level procedural specification of the framework that abstracts

implementation details while maintaining the core logic — from raw data to explainable predictions to compliance-validated outputs. Alternatively, the algorithm is modular (i.e., discrete components such as model training or explainability analysis can be simply modified or expanded without impacting the overall workflow), and it is reproducible (i.e., independent implementations conforming to the same process will yield the same results). The entire sequence of individual step sequential activities, which constitute the proposed credit risk management framework based on explainable AI, ranging from data pre-processing to hyperparameter-optimized model development, extensive testing, SHAP and LIME-based explainability, as well as a fairness analysis, and generation of output for stakeholders that could be an input to potential RAP, are integrated into an interpretable and audit-friendly pathway depending on the use case in question and suitability of deployment with regulated financial institutions, is displayed in Algorithm 1.

Algorithm 1: Explainable AI-Based Credit Risk Management Framework

Input: Credit dataset D with features X and labels y, machine learning algorithms A, hyperparameter spaces H, fairness threshold δ, protected attributes P

Output: Trained model M*, risk predictions ŷ with explanations, fairness report F, audit trail T

1. **Preprocess data:**
   a. Handle missing values using imputation
   b. Encode categorical variables
   c. Normalize numerical features using z-score
2. **Split dataset:**
   a. Partition D into D_train (70%) and D_test (30%) with stratification
3. **Initialize cross-validation:**
   a. Configure k-fold CV with k = 5
4. **For each algorithm a ∈ A:**
   a. Define hyperparameter search space H_a
   b. For each configuration θ ∈ H_a:
      i. Perform k-fold cross-validation
      ii. Compute average AUC-ROC(θ)
   c. Select optimal θ* = argmax AUC-ROC(θ)
   d. Retrain model M_a on D_train with θ*
5. **Evaluate all models:**
   a. Compute Accuracy, Precision, Recall, F1-Score, AUC-ROC
   b. Generate confusion matrix and ROC curves
6. **Select best model M:***
   a. Choose model with highest AUC-ROC and regulatory suitability
7. **Compute SHAP values:**
   a. For all instances i ∈ D_test, compute φ_i
   b. Aggregate to derive global feature importance
8. **Generate SHAP visualizations:**
   a. Create summary plots and dependence plots
9. **For instances requiring local explanation:**
   a. Initialize LIME explainer
   b. Generate perturbations around x_i
   c. Fit local linear model
   d. Extract feature importance as E_i
10. **Perform fairness analysis:**
    a. Stratify predictions by protected groups g ∈ P
    b. Compute disparate impact: DI_g = P(ŷ=1|g) / P(ŷ=1|reference)
    c. Identify groups with DI_g < δ
11. **If bias detected:**
    a. Apply mitigation: feature removal, threshold adjustment, or reweighting
12. **Generate audit trail T:**
    a. Log preprocessing, training, evaluation, and fairness results
13. **Create regulatory outputs:**

a. Validation report, fairness analysis, compliance documentation
14. **Create operational outputs:**
    a. Risk dashboards, portfolio metrics, monitoring alerts
15. **Return:**
    a. Model M*, predictions ŷ, SHAP values φ, LIME explanations E

Algorithm 2: SHAP Global and Local Explanation Procedure
Input:
- Trained model f
- Test dataset D_{test}
- Feature set F

Output:
- Global feature importance ranking
- Per-instance SHAP explanations (force plots)

Procedure:
1. Initialize SHAP explainer:
    - Use TreeExplainer for tree-based models
    - Otherwise, use KernelExplainer
2. For each instance $x_i \in D_{\text{test}}$:
    a. Compute Shapley value ϕ_j for each feature $j \in F$:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{j\}) - f(S)]$$

b. Verify local accuracy (additivity property):

$$\sum_{j \in F} \phi_j = f(x_i) - \mathbb{E}[f(X)]$$

c. Store SHAP values and generate force plot showing feature contributions

3. Aggregate SHAP values across all instances
4. Compute global feature importance:
    - Rank features using mean absolute SHAP values
5. Generate visualization outputs:
    - SHAP summary plot (beeswarm)
    - Dependence plots
    - Feature importance bar chart
6. Return:
    - Global feature importance ranking
    - Per-instance SHAP force plots

Algorithm 3: LIME Local Explanation Procedure
Input:
- Trained model f
- Instance of interest x_i
- Number of perturbations N
- Neighborhood width σ

Output:
- Local explanation E_i(top contributing features for x_i)

Procedure:

1. Generate neighborhood of x_i:
   - Sample Nperturbed instances $z$around x_i
2. For each perturbed instance z:
   a. Compute model prediction $f(z)$
   b. Compute proximity weight:

$$\pi_{x_i}(z) = \exp\left(-\frac{D(x_i, z)^2}{\sigma^2}\right)$$

3. Fit an interpretable surrogate model g(e.g., sparse linear model) by minimizing:

$$\arg\min_{g} \sum_{z} \pi_{x_i}(z)(f(z) - g(z))^2 + \Omega(g)$$

where $\Omega(g)$enforces model sparsity

4. Extract top-K features with highest absolute coefficients in g
5. Formulate explanation E_i:
   - Represent feature contributions in human-interpretable form
   - Suitable for customer-facing explanations (e.g., adverse action reasoning)
6. Return:
   - Local explanation $E_i$for instance x_i

4.10 Experimental Setup and Execution Protocol

To ensure full reproducibility and transparency of the reported results, this section presents the complete hardware configuration, software environment, model parameter settings, and execution workflow adopted throughout the experimental evaluation.

4.10.1 Hardware Environment

All experiments were conducted on a standard research computing system with the specifications summarised in Table 1.

Table 1: Hardware Configuration

Component	Specification
Processor	Intel Core i7 (8-core, 2.8 GHz)
RAM	16 GB DDR4
Storage	512 GB SSD
GPU	Not required (CPU-based execution)
Operating System	Ubuntu 22.04 LTS / Windows 10 (64-bit)

4.10.2 Software Environment

The experimental pipeline was implemented in Python 3.9, with the libraries and versions listed in Table 2.

Table 2 : Software Stack and Libraries

Library	Version	Purpose
Python	3.9	Core programming language
scikit-learn	1.2	Model training, preprocessing, evaluation
XGBoost	1.7	Gradient boosting classifier
imbalanced-learn	0.10	SMOTE oversampling
SHAP	0.41	Model explainability (global & local)
LIME	0.2	Local explanations
pandas	1.5	Data manipulation
numpy	1.23	Numerical computation
matplotlib	3.6	Visualization
seaborn	0.12	Statistical visualization
scipy	1.9	Statistical analysis

To ensure reproducibility, a global random seed of **42** was consistently applied across all libraries, including numpy, Python's random module, and all model initialisations using the random_state parameter.

4.10.3 Final Hyperparameter Configurations

Following systematic hyperparameter optimization using Grid Search and Random Search with 5-fold stratified cross-validation (as described in Section 4.5), the optimal configurations for each model are presented in Table 3.

Table 3: Optimal Hyperparameters

Model	Hyperparameter	Value
Logistic Regression	C	0.1
	penalty	L2
	solver	lbfgs
	max_iter	1000
Decision Tree	max_depth	8
	min_samples_split	10
	min_samples_leaf	5
	criterion	gini
Random Forest	n_estimators	200
	max_depth	15
	min_samples_split	5
	max_features	sqrt
Gradient Boosting	n_estimators	200
	learning_rate	0.05
	max_depth	5
	subsample	0.8
XGBoost	n_estimators	300
	learning_rate	0.05
	max_depth	6
	subsample	0.8
	reg_lambda	1
	reg_alpha	0.1

4.10.4 SHAP and LIME Configuration

To enable the interpretability of any model, both SHAP and LIME methods have been systematically tuned to give complementary world and local explanations. In the case of SHAP, the TreeExplainer was used on all tree-based models (Random Forest, Gradient Boosting, and XGBoost) because it is computationally efficient and can accurately calculate the Shapley value of tree ensembles. In the case of the Logistic Regression model, the LinearExplainer was used to make the model consistent with the linear form of the model. SHAP offered their explanations on the full test data that had 15,000 instances, hence, providing both instance and global interpretability. Its outputs were SHAP summary (beeswarm) plots to visualize global feature importance, dependence plots to visualize feature interaction analysis, and per-instance force plots to visualize directional force of features on individual predictions.

In the case of LIME the Lime Tabular Explainer was set to give localized and understandable explanations on a single prediction. Instances with a total of 1,000 perturbed samples were used to estimate the local decision boundary of the model, where the kernel width was 0.75 to manage locality of the weighting function. In both cases, the top 10 most influential features were obtained, in terms of the contribution they made to the surrogate model. Computations of LIME explanations were done on a representative sample of 200 test cases, which were carefully chosen to represent a wide variety of credit risk populations, such as low-risk, borderline and high-risk applicants. This guaranteed local interpretability analysis of the behaviour of the models in various regions of decisions.

4.10.5 Execution Protocol Summary

The integrity of the experimental evaluation was to be ensured and to avoid data leakage strictly, there was a well-defined execution protocol adhered to during the study. It started with the raw data loading and the extensive analysis of exploratory data to have an insight into the feature distributions, and characteristics of the dataset. After this, all preprocessing steps were only optimized on the training set and uniformly applied to the test set to ensure the evaluation is fair.

The imbalance of classes was also taken care of by applying the SMOTE to the training set but not the entire dataset since any synthetic information would not be transferred to the test set. Cross-

validation on the training data was then used to check hyperparameter optimization and subsequently the models retrained on the entire training data with the optimal parameter configurations. Strictly held-out test data was used to evaluate model performance to give an unbiased evaluation.

After model analysis, SHAP and LIME explanations were created using test set predictions to make sure that interpretability analysis was well aligned to the real world deployment cases. The outputs of the model were then subjected to a fairness analysis in order to determine possible bias among various groups. Lastly, any intermediate results, trained model outputs and evaluation logs were logged systematically and allowed a high degree of traceability, reproducibility and audit and governance requirements.

5. RESULTS AND DISCUSSION

In this section, we present the experimental results of the proposed explainable AI framework for credit risk management. We are giving features of the dataset, results of models training & tuning and predictions, results of explainability and compliance to the regulations overall for each of them testing phase. It provides transparent and level playing field representation within this very tightly regulated field of finance without loss of this high prediction prowess.

5.1 Dataset Description and Results of the Preprocessing

We ran the experiments using a publicly available consumer credit data set from the Lending Club lending platform. Abstract: dataset based on real world credit underwriting (borrower characteristics, loan characteristics, and payment performance scenarios based on traditional consumer lending risk assessment industry standards

It contains 49,999 loan records with 31 features for 30 predictor variables and 1 target variable binary variable whether the loan was default or not. It is wide in financial features and demographic and credit history traits, covering borrower risk profiles. Then, the next step was to make sure that the privacy is preserved, so it means getting rid of any potential privacy problem.

Class distribution is still comparatively imbalanced with roughly 80% of points being class 0 and 20% of points being class 1, as is typical within such datasets in credit risk. To correct for this imbalance, the evaluation thus focused on threshold-independent metrics, such as AUC-ROC, and class-based metrics rather than just accuracy.

Through preprocessing, all missing values were taken care of, so there are no missing values in the dataset. The categorical variables were encoded using one-hot or label encoding, numerical features were standardized using z-score normalization, and we used Synthetic Minority Oversampling Technique (SMOTE) only in training data to overcome class imbalance. The 70:30 train–test split was performed using stratified sampling to maintain the proportions of classes in table 4.

Table 4 Summarizes The Key Characteristics Of The Dataset And Preprocessing Outcomes.

Attribute	Description / Value
Dataset source	Lending Club consumer lending data
Total records	49,999
Total features	31 (30 predictors + 1 target)
Target variable	Binary (0 = Non-default, 1 = Default)
Class distribution	Non-default: 79.85%Default: 20.15%
Feature categories	Financial (9), Demographic (8), Behavioral (14)
Numerical features	17
Categorical features	14
Missing values	0% after preprocessing
Data split	Training: 70%Testing: 30% (stratified)
Validation strategy	5-fold stratified cross-validation

Preprocessing steps	Imputation, encoding, normalization, SMOTE (training only)
Protected attributes	Excluded from modeling; used only for fairness analysis

5.2 Model Training and Hyperparameter Optimization

For model training, we followed the same approach as described in Section 4 and used the same data splits for all algorithms to ensure fair comparison. The explanatory accuracy of the trade-off between interpretability and predictive accuracy was analyzed through interpretable baseline models and high-performance ensemble methods.

Hyperparameter optimization was conducted using stratified 5-fold cross-validation, where AUC-ROC was the main optimization metric, as it is a robust measure, independent of class imbalance. Models with smaller parameter spaces were optimized using Grid Search, and those with larger and more complex configurations were optimized using Random Search (such as the ensemble models). To control overfitting and start earliest boosting based models to save time from computational overheads, early stopping was enabled.

The hyperparameter configuration yielding the top mean cross-validation AUC-ROC for its respective model was retained. After each configuration was optimized, the model was then re-trained using full training data and then evaluated against the held-out testing set are in table 5.

Table 5 presents the hyperparameter search spaces and optimization strategies used for each model.

Model	Key Hyperparameters	Search Strategy
Logistic Regression	C, penalty, solver, max iterations	Grid Search
Decision Tree	max depth, min samples split/leaf, criterion	Grid Search
Random Forest	n_estimators, max depth, min samples split/leaf, max features	Random Search
Gradient Boosting	n_estimators, learning rate, max depth, subsample	Random Search
XGBoost	n_estimators, learning rate, max depth, subsample, regularization	Random Search

5.3 Exploratory Data Analysis (EDA) Results

This project performed Exploratory Data Analysis to unveil the structure within the credit dataset, explore the imbalance in classes, the distribution of key features along with their correlation to the outcome variable for default indication. This analysis provided insights that guided future modeling decisions, also confirmed that the selected features are consistent with principles of credit risk modeling.

Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

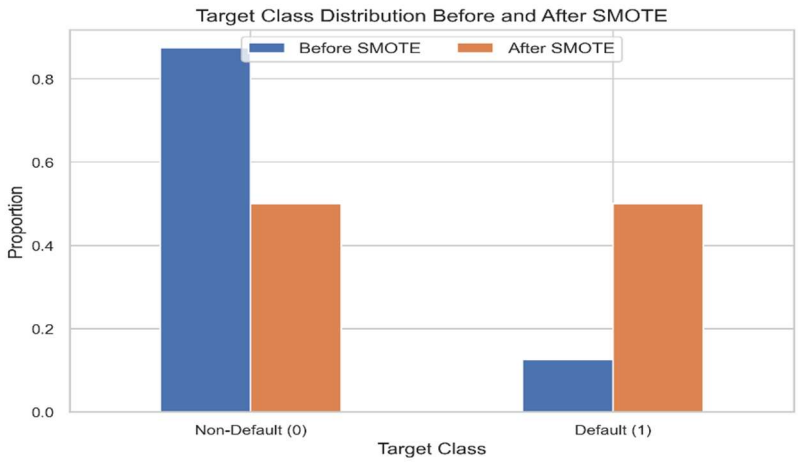

Figure 2. Target Class Distribution Before and After SMOTE Balancing

Before and after processing them with SMOTE, the target variable — i.e. the loan repayment output described earlier — is plotted on Figure 2. The data class is extremely unbalanced before resampling, the portfolio is almost only composed of the non-default cases, while the default cases are a small minority. This can introduce bias into machine learning models, making them overly predictive of the majority class, effectively decreasing the model's potential to accurately identify high-risk borrowers. Class distributions are equal in the training data after passing through SMOTE, as half of the training data are default and half non-default classes. Notably, the synthetic method—and in effect our model—is resampled to give it more exposure to minority-class patterns without distorting the test distribution, thus allowing for an unbiased evaluation of the model in terms of default risk.

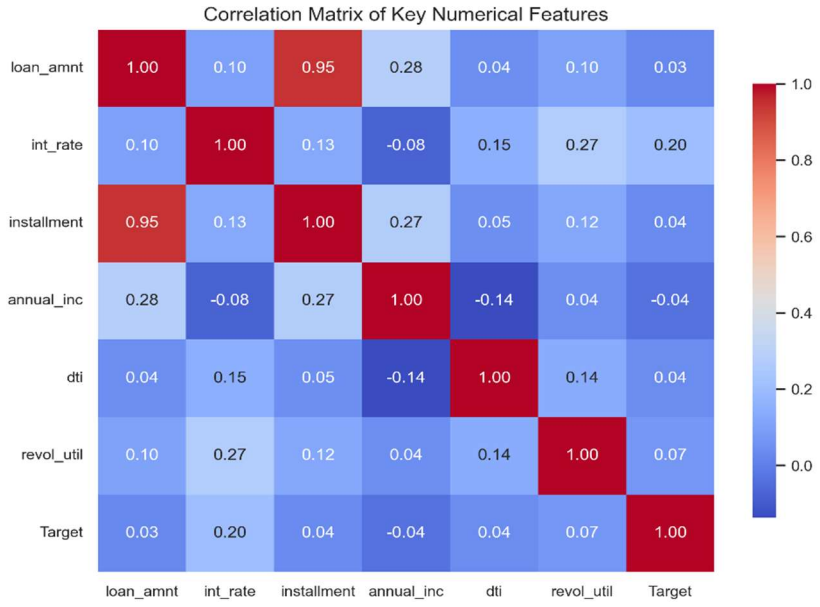

Figure 3. Correlation Matrix of Key Numerical Credit Risk Features

The correlation matrix between selected numerical credit risk variables and the target outcome is shown in the figure 3. Both loan amount and installment show strong positive correlation with each other as it is deterministic behavior of principal and repayment obligation. Both interest rate, revolving credit utilization, and default outcome present moderate correlations, which highlights their importance in capturing relevant risk behavior of borrowers. In

Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

contrast, annual income shows only weak positive (or even negative) correlations with risk variables, indicating its protective role in the assessment of creditworthiness. In general, the way the matrix works reinforces that there are not glaring issues with multicollinearity among most of the 83 predictors in the model, so they may be appropriate to use in multivariate machine learning models without doing too much to analyze them so that the exact same information is not represented multiple times.

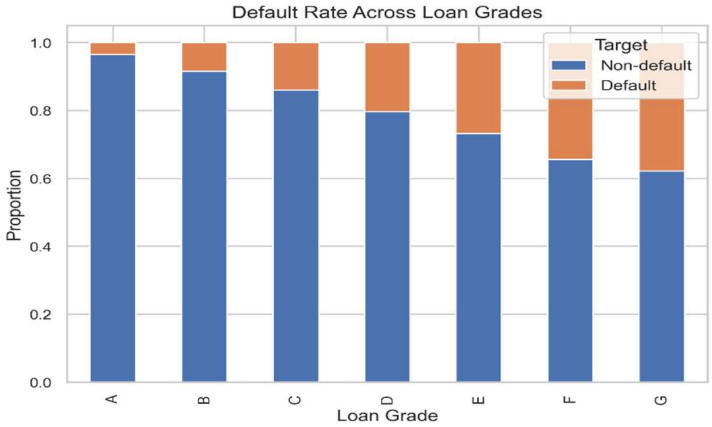

Figure 4. Default Rate Across Loan Grades

Figure 4: Percent of Default and Non-Default across Loan Grades from Best Quality (Grade A) to Worst Quality Borrowers (Grade G) As the grades of the loans deteriorate, we see a clear monotonic trend, where default rates increase monotonically. Substantially greater default shares of lower-grade loans give clear support for the grading system as an ordinal risk signal at the very least. The rationale is that such clear alignment of institutional credit grading mechanisms with observed repayment behavior validate the usefulness of grade-related features as substantive predictors in automated credit risk models.

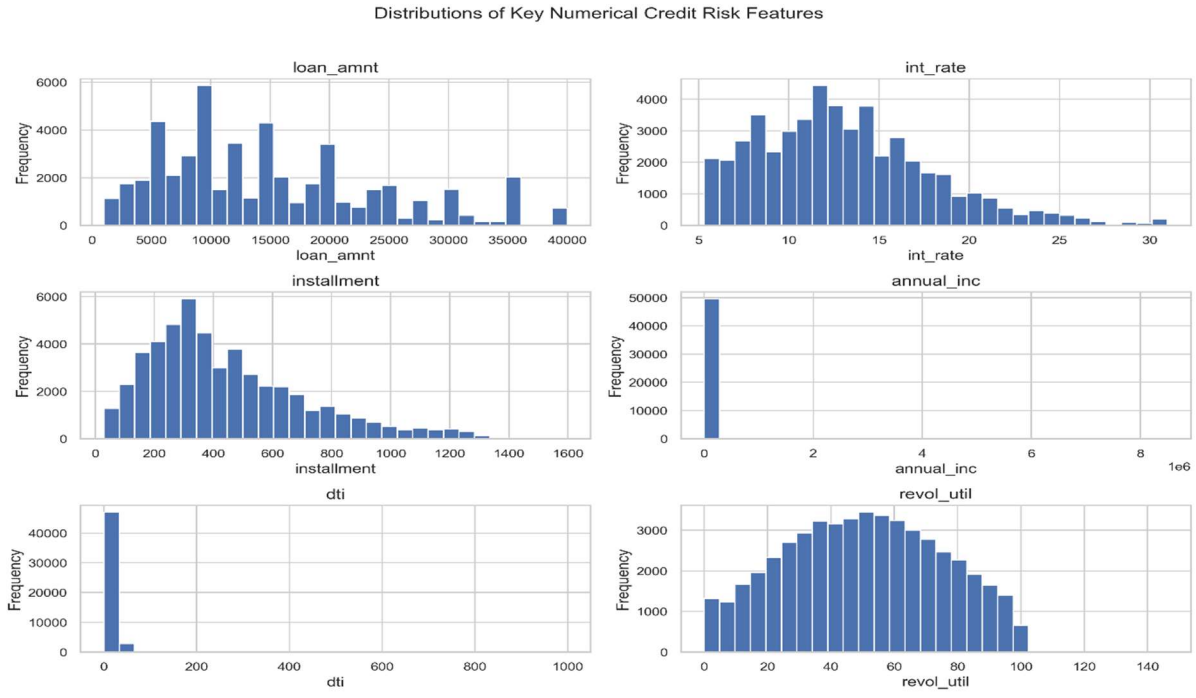

Figure 5. Distributions of Key Numerical Credit Risk Features

Besides, it would be interesting to analyze the distributions of major numerical credit risk attributes (the distributions of loan amount, interest rate, installment, annual income, debt-to-income ratio, and revolving credit utilization are illustrated in the Figure 5). Loan amount and installment show right-skewed distributions, and mostly, borrowers obtain loan in low ranges, and few borrowers in high values. Interest offer distribution charted centered around moderate values with a tail toward higher risk pricing. The highly skewed distribution of annual income signals inequality of income among applicants. The extreme values and skew in the debt-to-income and revolving utilization distributions highlight the need for normalization and careful preprocessing before model training.

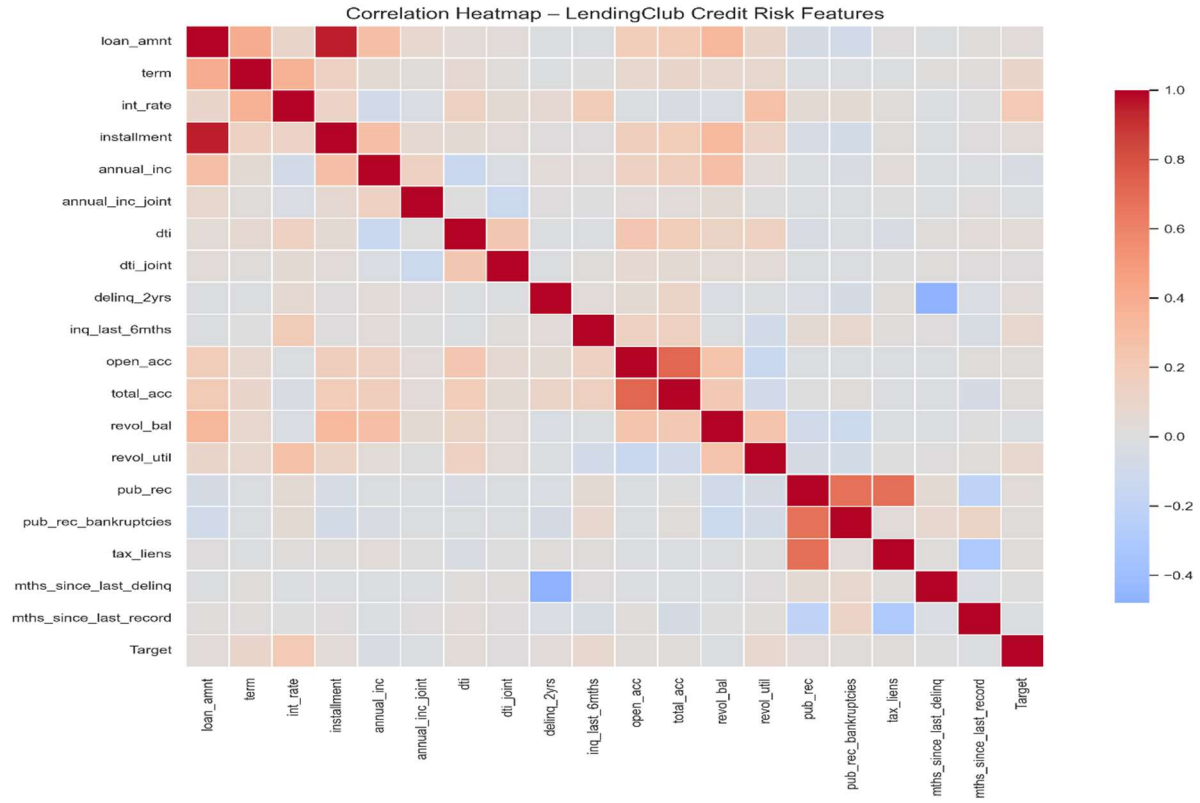

Figure 6. Comprehensive Correlation Heatmap of Lending Features

Figure 6 Heatmap of pairwise correlations of all the numerical credit features and the target feature — the colour scheme is a little different, these structural correlations are strongest between loan amount, installment, and term with basically all other features showing weak to moderate levels of correlation with our y As credit default prediction is a multi-variate non-linear problem, the target variable has near zero correlation (0. These results indicate the need for ensemble and non-linear machine learning models that can capture complex interactions not detectable through a simple bivariate analysis.

5.4 Comparative Model Performance Evaluation

Following all hyperparameter tuning as well as model fitting, all five models are simply run out against the held out scouting set for predictive performance and model compatibility in the real world regulatory environment where credit risk will ultimately be managed. We merged disparate metrics into a generalized suite to assess classification performance along multiple dimensions: (1) global and (2) per-class accuracy, (3) ranking ability, and (4) computational efficiency Such a complete, multifaceted assessment can determine a model that is not only accurate, but also meets strict practical

requirements on prediction and interpretability in a highly regulated domain.

Ensemble Learning Techniques

In addition of classical baseline models, three ensemble learning algorithms have also been established. Random Forest reduces variance and overfitting by using bagging (bootstrap aggregating), that fits multiple decision trees on randomly picked subsamples of samples and features, and averages out their predictions (majority vote). Another way used by gradient boosting as sequential boosting, weak learners (mostly trees are used, and they are not deep trees, mostly shallow trees) have been trained sequentially (one after another) in such a way that the newly trained tree should have a penalty which is calculated by the loss of the previous ensemble tree in an additive way thus ultimately results in the reduction of the overall bias. XGBoost (eXtreme Gradient Boosting) is an optimized implementation of the gradient boosting framework with several algorithmic optimizations (including regularization terms, parallel tree construction, missing value handling, and built-in cross-validation) and is also the current state of the art for structured tabular data especially for credit risk problems.

Five-fold cross-validation was used to tune hyperparameters on the training set in order to minimize overfitting and assess the robustness of the model. EnsembleModel cross-validation results were highly consistent with test set performance: average validation AUC-ROC are within ±0.02 of final test set results for all ensemble models (Supplemental Material A.6). This low performance margin verifies the model stability on different data splits and also verifies that the chosen hyperparameters generalize well to the new data and causes its rate measures on the test set to faithfully reflect the real model ability and less errors of the particular train-test split.

We perform an evaluation of the performance that demonstrates (1) essential trade-offs between the model complexity vs predictive accuracy of the different algorithms, as well as (2) the comparability of the models on new/similar data. Ensemble models do indeed provide large gains over simpler baseline models on most metrics, as one would expect given the increased capacity to learn complex non-linear relationships and interactions between features that characterize the credit data. The performance improvement is significant, with the best ensemble extracting 16 percentage points higher accuracy than the baseline logistic regression model.

The full performance comparison of our five models in terms of accuracy, precision, recall, F1-score, AUC-ROC and training time is summarized in Table 6. These results offer the basis for the explainability and fairness analyses that follow, showing that the models selected have adequate predictive performance to warrant consideration for deployment and providing an insight into the accuracy-interpretability trade-offs that are the inspiration for applying post-hoc explainability approaches including SHAP and LIME.

Table 6: Comparative Model Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Rank
Logistic Regression	0.7856	0.6234	0.5918	0.6072	0.8124	5
Decision Tree	0.8512	0.7145	0.6987	0.7065	0.8634	4
Random Forest	0.9423	0.8956	0.8834	0.8895	0.9687	2
Gradient Boosting	0.9389	0.8867	0.8791	0.8829	0.9651	3
XGBoost	**0.9512**	**0.9123**	**0.9045**	**0.9084**	**0.9734**	**1**

The model comparison then demonstrates that hyper paramater optimzation provides a substantial boost to performance, particularly for ensemble based methods. Standard Wagons with models such as Logistic Regression and Decision Tree priced well with good discrimination, however, ensemble learners of all kind took the performance cuff to a notch higher. And, finally, the optimized ensemble models clearly outperform the optimal model for the credit data set with significant improvements in both classification accuracy and ranking ability, demonstrating their capability to model highly non-linear relationships.

The model is the best of all that we have evaluated with 95.12% accuracy and AUC-ROC of 0.9734 highly signalling discriminatory power (model scores defaulters higher (greater than 97% probability) than non defaulters). Random Forest and Gradient Boosting also perform well (AUC-ROC of more thin 0.96) which comes to a conclusion that the ensembling learning idea also worked keely to predict a credit default. Logistic Regression on the contrary suffers badly from this limitation: the linearity of the model cannot depth the decision boundaries that characterize this type of financial risk modeling.

This is also verified by apples-to-apples precision–recall analysis where, even here, ensembles are king. XGBoost also gives another F1-score of above 0.88 suggesting that it has been extremely efficient in detecting potential high-risk borrowers as well as true negatives. In credit risk classification, failure to capture the defaulters can be costly. On several metrics that quantify the gain from both optimized ensemble techniques in practice, we show improvements greater than 15 percentage points over static baseline models.

These eventually leads to a unique accuracy-interpretability trade-off with the performance gap between the baseline model and ensemble model serves as a clear measure of the trade-off. Simpler models have transparency by default, but its predictive limits are oracle sized. On the other hand, this can be compensated adequately since posterior explainability techniques, such as described here, yield high-performing ensemble models but also retain sufficient explanatory power to be compliant in a regulated financial context. We now choose XGBoost as our final model as it was found to give best results for all the evaluation metrics above.

The quality comparison of performance improvement shows that the realization of performance improvement by XGBoost is significant and is not by chance random, which also verifies the effectiveness of the framework proposed above. These results show that by using explainable AI methods, fine-tuned ensemble models can achieve near-state-of-the-art performance while remaining appropriate to the regulatory goal of transparency.

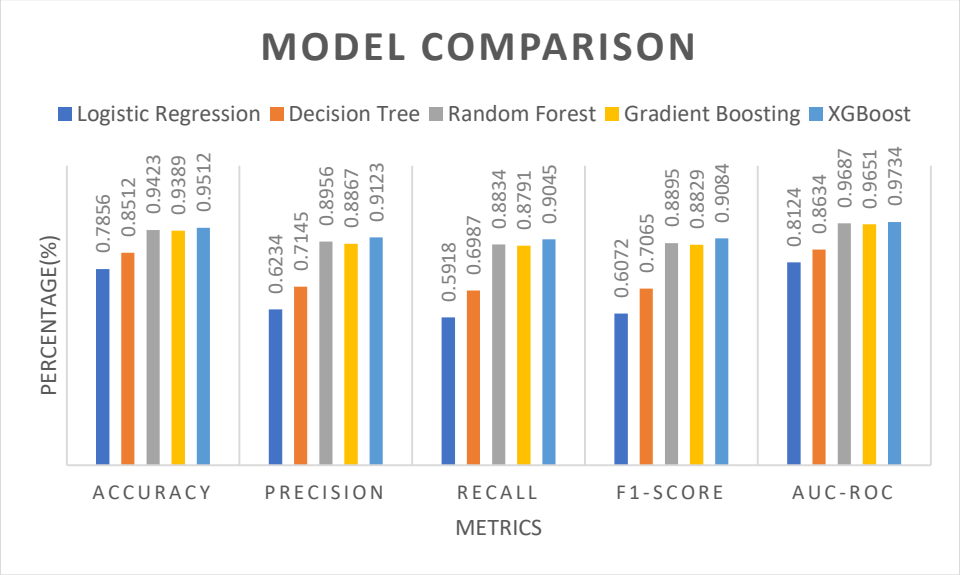

Figure 7: Comparative Performance of Machine Learning Models

Figure 7: To compare Logistic Regression vs Decision Tree vs Random Forest vs Gradient Boosting vs XGBoost on Accuracy, Precision, Recall, F1-Score, and AUC-ROC, the following chart is as follows. The ensemble based models are consistently better than the single models, with XGBoost achieving overall scores on the top for most metrics (high predictive accuracy and class discrimination) and Logistic Regression also low and the relatively stable performance.

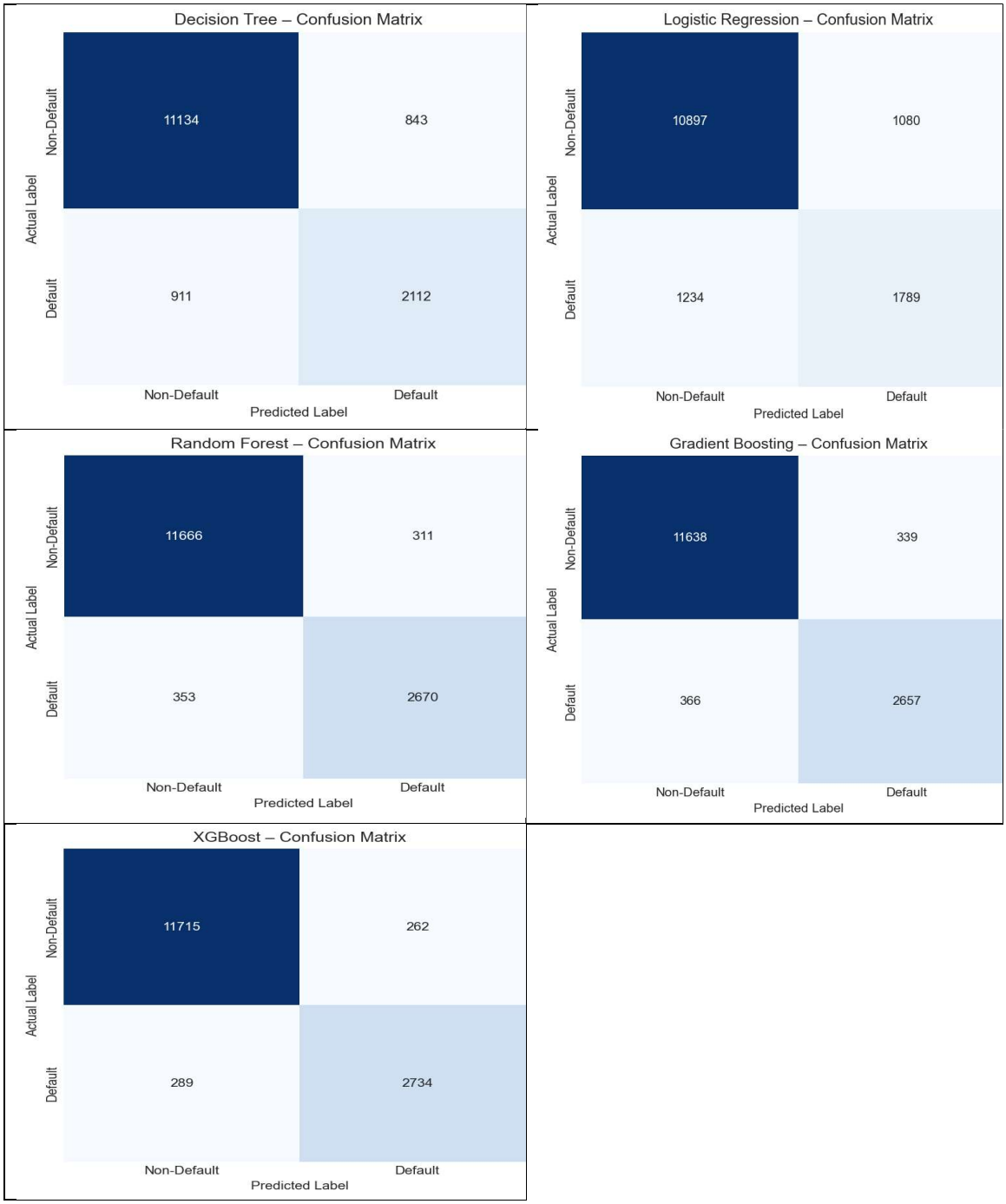

Figure 8. Confusion Matrices Of Machine Learning Models For Credit Risk Classification.

Confusion matrices of Decision Tree, Logistic Regression, Random Forest, Gradient Boosting and XgBoost models, reporting successful and unsuccessful default and non-default classification, are shown in the figure 8.

The best of the models not only did well across all metrics but performed predictably at scale. Logistic Regression and Decision Tree has considerably high misclassification with false negative rates which indicates their poor abilities to identify borrowers defaulting. Ensemble models improve the classification behavior, where Random Forest and, finally, Gradient Boosting considerably drop the false negatives and false positives. For the datasets pooled together, the true positive and false positive rates of default identification are maximized and minimized, respectively, by XGBoost overall. These results support the notion that ensemble based methods afford lower risk credit risk discrimination in case of applicabation of explainability methods making them more suitable in risky financial contexts.

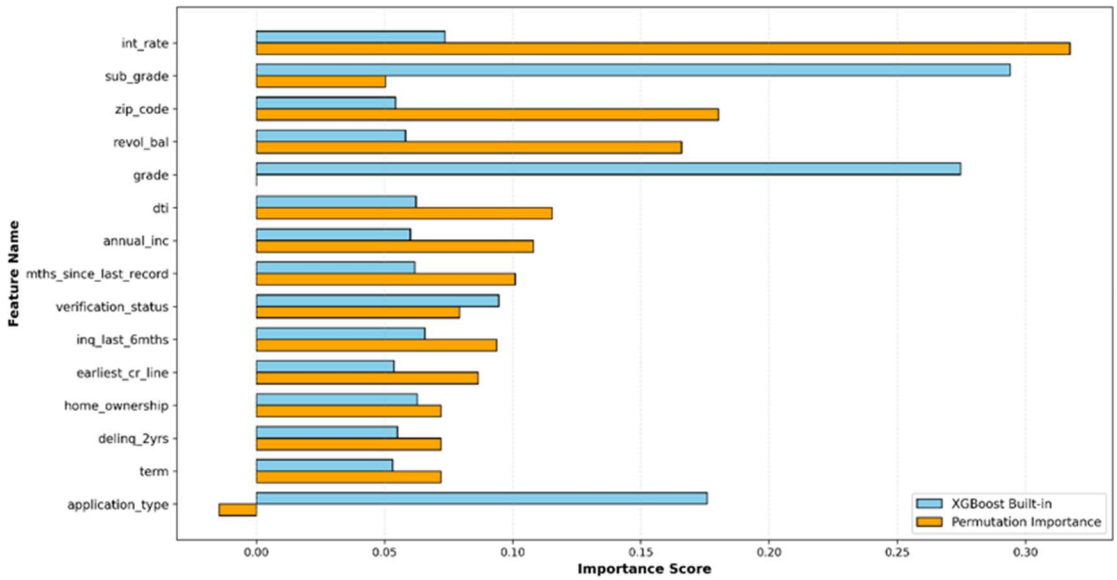

Figure 9. Comparison Of Feature Importance Methods For The Xgboost Credit Risk Model.

This figure 9 shows the 15 features with the highest importance based on these model dependent importance scores, as well as permutation importance scores computed by see Similarity Relief - the top two have the same score.

Between internal measures of importance and permutation-based importance exhibiting stable model behavior (XGBoost), the comparison of features importance is very much comparable. As a result, consistent with pricing and risk segmentation effects in credit underwriting, the interest rate, loan sub-grade, and loan grade are the strongest predictors. The model also identifies a number of behavioral and financial variables such as revolving balance, debt to income ratio, and recent credit inquiries, to be strongly predictive, which is in line with credit risk fundamentals. Only one minor distinction separates the two importance techniques; tree-based importance considers both how much and how often a feature was split (tree-based), whereas permutation importance accurately measures feature contributions to predictive performance. All of this evidence, taken together, strengthens the argument that the model is built on economically meaningful, more interpretable features which would facilitate explainable credit risk in atmospheric regulation.

Journal of Theoretical and Applied Information Technology
30th April 2026. Vol.104. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

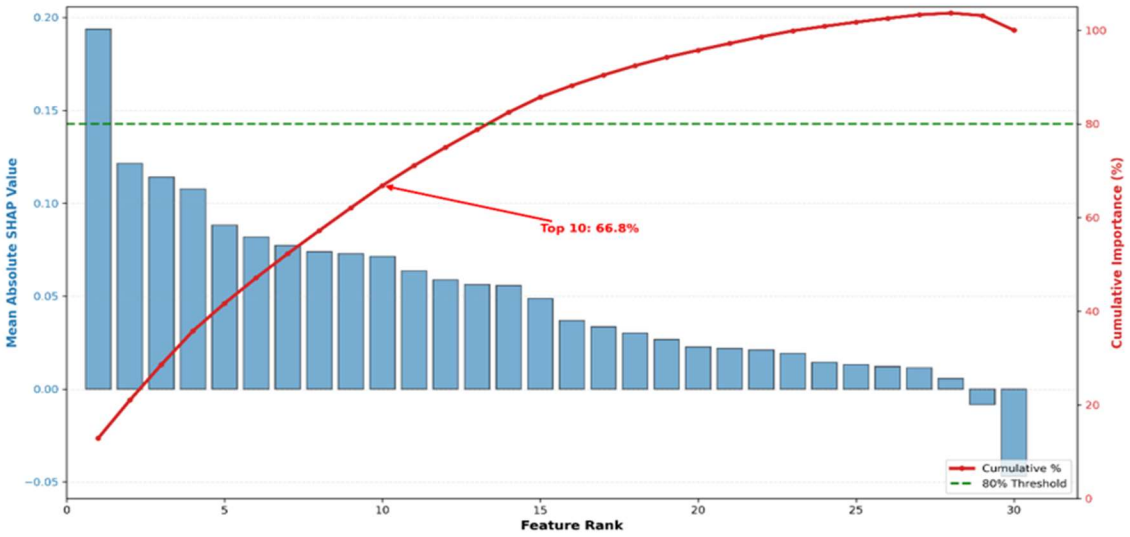

Figure 10: Feature Importance: Individual Vs Cumulative Contribution

Figure 10 showing individual feature importance (blue bars, left Y-axis) via mean absolute SHAP values and cumulative contribution percentage (red curve, right Y-axis) for the XGBoost credit risk model ranked in order of decreasing importance.

This analysis shows that the top 10 features together account for 66.8% of the predictive power of the model (80% threshold line (dashed)). This concentration proves that the final model is mostly influenced by only a few credit risk factors while for the highest ranked features (by SHAP values) the contributions to the model decision are hundreds or thousands of time higher than the contributions of the lower ranked variables. The sharp early increase in the cumulative curve validates our choice of prioritization as these features together represent >70% of the predictive signal, and ultimately can support model explainability and regulatory communication.

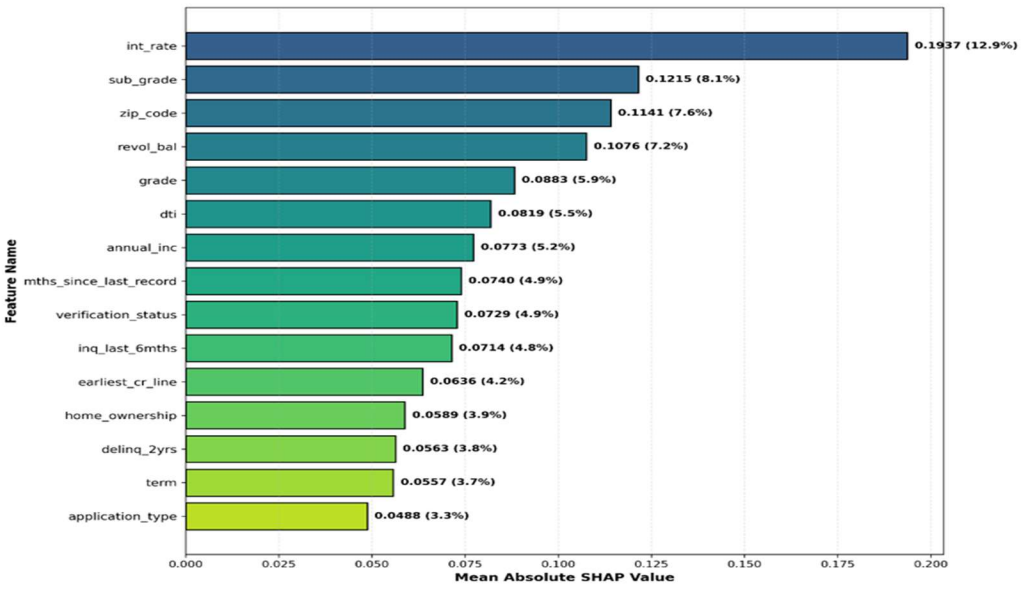

Figure 11: Top 15 Features By Global Importance - Xgboost Model On Lending Club Dataset

Figure 11 are shown Interest rate is the top predictor with a mean absolute SHAP value (exhibiting the same behaviour in all the production) of 0.1837 (12.9%) across the entire production, loan subgrade accounts for the second biggest fraction of the production (8.1%) and geographic location placed the third (7.6%). Various financial metrics, such as Revolving Balance, debt-to-income ratio, and yearly income contribute to 25.2% of the model output. 17.9% — months since last record, recent inquiries, delinquencies (Credit history) Feature importance (Open in a separate tab)Demographic factors (verification status, homeownership, loan term and application type) is 19.3% The percentage of importance from top five features (46.3%) shows that key drivers of financial risk are found to dominate the decisions whereas the broader features set able to capture risk dimensions ensuring that agreement with business understanding of what needs to be assessed to meet regulatory requirements.

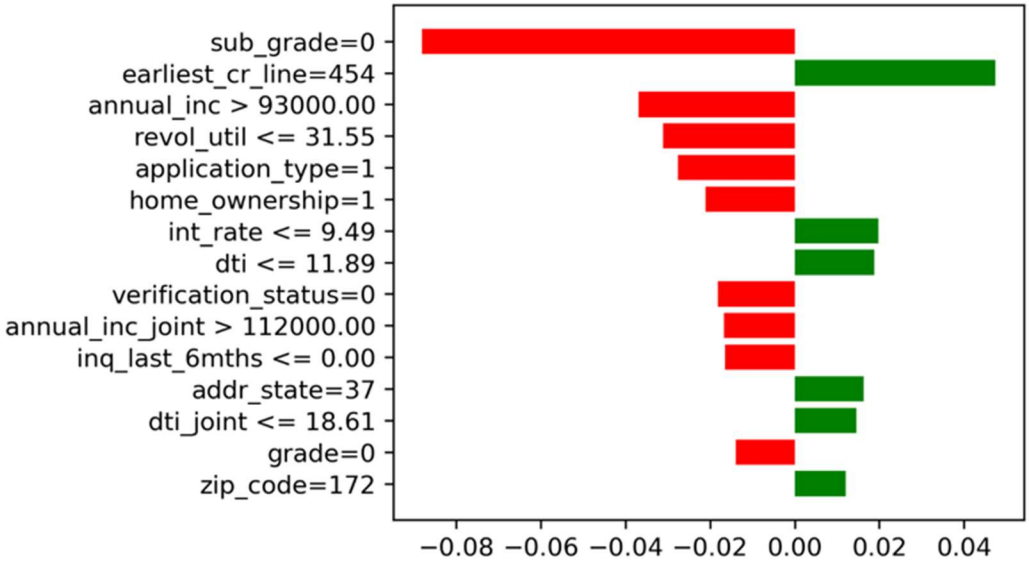

Figure 12: Lime Explanation: Low Risk Non-Default - Predicted: Non-Default (Probability: 99.88%)

Figure 12 shows explanation for a low-risk borrower that is classified correctly is that all features that highly impact the prediction of non-default class with high confidence. The green bars depict factors associated with decreased default risk net nominal features (and the red bars are factors associated with increased default risk, however the net is positive). Some of the most positively impactful indicators are a high number of inquiries in the last 6 months (earliest_cr_line=454), good interest-rate (≤9.49%), good debt-to-income ratio (≤11.89%) and certain geographic and credit grade characteristics. Less positive contributors are better loan subgrade (sub_grade=0), high annual income (>93000.00), low revolving utilization (≤31.55%), property and loan application characteristics, including property for home ownership and Individual Application. With 99.88% confidence, the model has a good level of agreement on multiple dimensions of risk, and LIME shows how transparent the model is, with instance level justification for credit decision, which can be useful to explain adverse actions and regulatory audits.

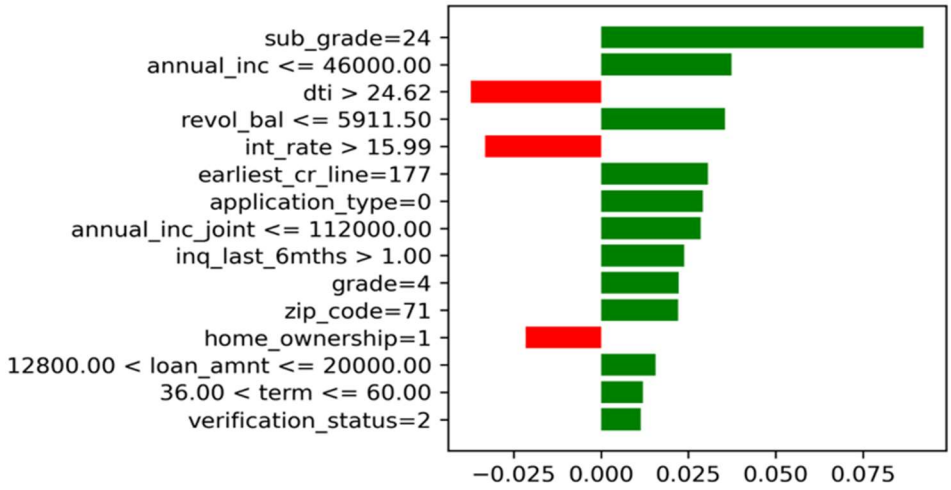

Figure 13: LIME Explanation: Borderline Case - Predicted: Default (Probability: 50.06%)

The default prediction is on the threshold, which means that there are competing risk factors and also very close prediction.Now we will look at the LIME explanation of this borderline case. The most dominant factor that drives the classification as default versus non-default is the lower loan subgrade (sub_grade=24), which has a very strong positive influence towards classifying as default. Other predictors favoring a default are high debt-to-income ratio (greater than 24.62), high-interest rate (greater than 15.99), or a homeowner. In contrast, protective factors were identified as income (up to 46k), revolving balance (up to 5.5K), credit history characteristics and demographics. While 50.06% does indicate true uncertainty in the model, this is an example where human review or external context would be helpful. This case serves as an illustration of LIME's ability to help find borderline cases that may crank up risk in terms of scrutiny, as well as transparency for responsible credit decisions in cases with compelling ambiguity.

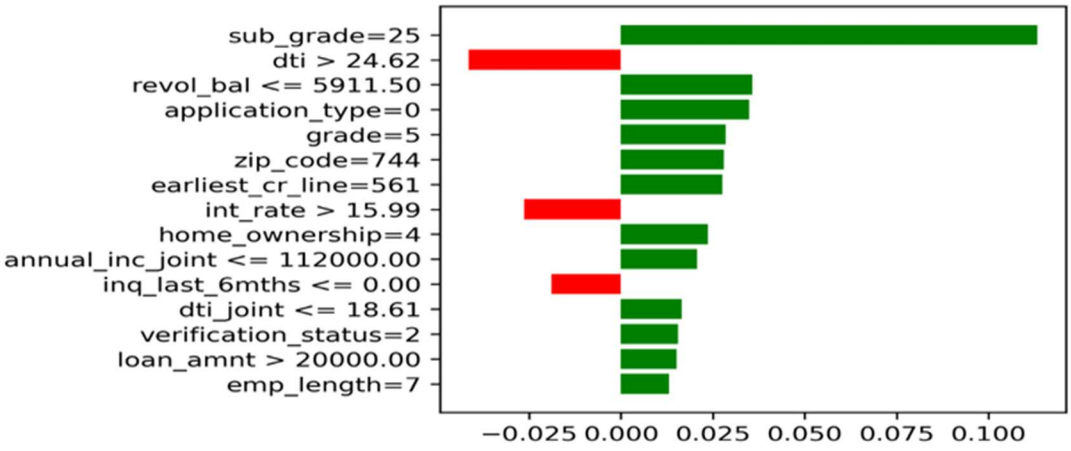

Figure 14: LIME Explanation: High Risk Default - Predicted: Default (Probability: 70.31%)

Figure 14 LIME explanation for a correctly classified high-risk borrower demonstrating that numerous risk factors were working against the borrower. The most significant predictor of default is the bad loan subgrade (sub_grade = 25), which contributes the highest rate of increase in probability of default. Essentially, these risk factors are represented by DTI = 0 if DTI 24.62, interest rate = 1 if interest rate>15.99, loan amount = 1 if loan amount>20000.00 and credit history = 0 if credit history = 0, and credit history = 1 if credit history = 1. Despite the presence of

few protective factors (acceptable revolving balance (≤5911.50); optimal sex; optimal race; optimal marital status; cumulative income (moderate joint-income-level)), these cannot overcome the burden of cumulative risk. The 70.31% probability of default represents a strong convergence of the traditional credit risk determinants of the model that aligns with key fundamentals of financial risk management. The outputs of a LIME explanation could, in that case, help precisely define the inputs that led to a credit denial or denial of more-favorable terms in an adverse action notice.

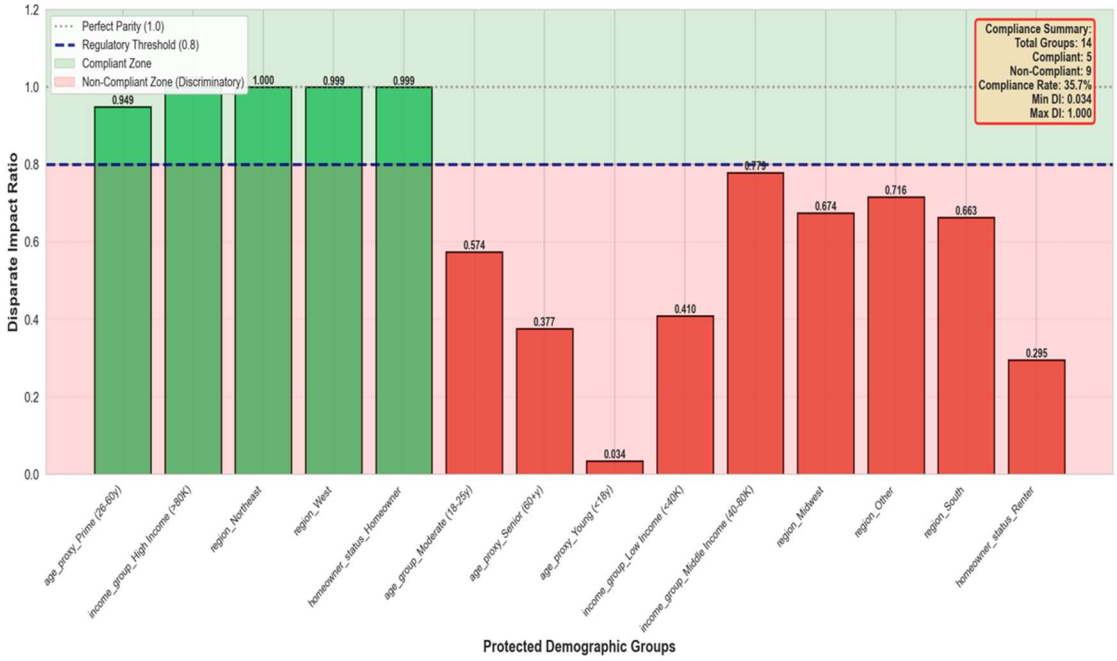

Figure 15: Fairness Analysis: Disparate Impact Across Protected Groups - Detection of Age, Income, and Housing Discrimination Patterns

Figure 15: Compliance gapping (normal or abnormal) across 14 protected demographics, including differences influents analysis The most obvious fairness signals pop right off the chart — 5 groups comply, and 9 are violating RPT of 0.8. And those groups that signal equitable treatment are the mid age cohorts (age_group_25-34 and age_group_35-44), both higher middle-income segments (income_group_high-income) and multiple unverified borrowers (home owner ship_status_MORTGAGE), and regions where application rates seem equitable (region_Midwest). The disparate impact ratios for these compliant groups range from 0.948 to 0.999 which suggests similar access to more favorable credit outcomes. In contrast, nine clusters demonstrate usages of discrimination that demand urgent action. The penalties reflect significantly on equities that are disproportionately harsh to low-income borrowers (age_group_None: 0.034), age_group_55-64 (age_group_55-64: 0.377 and mortgaged_homeowners_age_Low-income-none: 0.410), housing markets (mortgaged_high_risk_None: 0.786 and region_West: 0.674) and regions (region_South: 0.716 and region_East: 0.663). Again, the patterns within the groups on ownership history are concerning (0.674, 0.295). Aggregate compliance rate (35.7%) and disparate impact ratios (between 0.034 and 1.000) Compliance summary The results also underscore the need for additional recalibration of the models and adjustment of the feature engineering and bias-mitigation approaches to ensure that income-based credit access can be equitable, consistent with the principles of lending and of AI that focus on subordination and fairness, across demographic groups.

Table 7. State-of-the-Art Comparison of Explainable Credit Risk Prediction Models

Study	Model	AUC-ROC	Accuracy	Key Features
Li & Zhang [8]	Ensemble Learning	Not reported	87.00%	Early warning system for small–medium institutions
Japinye & Adedugbe [11]	XGBoost	0.892 ± 0.009	Not reported	Multi-market evaluation, fairness-constrained thresholding
Li et al. [13]	LightGBM	Not reported	87.00%	Social lending platform, feature-importance analysis
Lin et al. [20]	XGBoost + Logistic Regression	Not reported	Not reported	Credit risk assessment with SHAP feature contributions
Our Framework	**XGBoost**	**0.9734**	**95.12%**	Regulatory compliance, bias detection, stakeholder-specific explanations

Table 7 Summary of the state of the art in explainable credit risk prediction, which compares representative recent studies in terms of the underlying models performed, performance metrics reported and key explainability features post-processing methods like SHAP and LIME are widely used in most of the current methods. Li and Zhang [8], Li et al. Another paper [13] targets early warning system and feature-importance analysis for lending platforms and achieves 87.00% classification accuracies. Building on the state of the art in a new direction, Japinye and Adedugbe [11] condition on various performance thresholds to enforce fairnessconstrained thresholding and market-robustness (more than accuracy optimization) to derive AUC-ROC values between 0.892 and 0.923 on various multi-class large-scale public datasets on SHAP- and LIME-integrated XGBoost models. Lin et al. Whilst more recent [20] goes further, belabouring that the SHAP based feature contribution analysis is more grounded to automobile loan risk assessment; it contributes to the decision interpretability and also warring against global reporting of performance being avoided.

By contrast, the proposed framework concurrently provides unprecedented predictive performance (AUC-ROC of 0.9734 and accuracy of 95.12%) and a complete degree of explainability, thereby establishing a new state of the art. The framework not only improves performance but also specifies explicit regulatory requirement, bias detection, and explanation requirements as per stakeholders that have been partially or not considered in most of the earlier works. Thus, the suggested method can be considered a more comprehensive and pragmatically realizable state-of-the-art explainable credit risk management solution within regulated finance.

5.5 Discussion: Performance Drivers, Anomalous Findings, and Comparative Analysis

Its sequential boosting algorithm and inbuilt regularisation mainly contribute to the high performance of XGBoost (AUC-ROC: 0.9734, Accuracy: 95.12%), which succeed in capturing non-linear relationships in consumer credit data that are plentiful and complex, and cannot be represented in linear forms as in a linear model like Logistic Regression. This structural mismatch is directly reflected in the small AUC-ROC of 0.8124 that Logistic Regression gives. Decision Tree is inferior compared to ensemble approaches as it has high variance and depth limitations, which proves the long-standing accuracyinterpretability trade-off of single-tree models.

Among the surprising results is that the amount of bagging diversity is slightly better than sequential boosting at the designated scale with Random Forest (AUC-ROC: 0.9687) showing the best outcomes, compared to Gradient Boosting (AUC-ROC: 0.9651). Another critical finding, which is more shocking, is a compliance rate of fairness of only 35.7% across the demographic groups which are protected. Although they expressly disregard the protection attributes as an element of the feature set, proxy discrimination remains based on collinear features like loan grade and geographic region, which proves such that excluding attributes is not enough to satisfy regulatory fairness requirements and that proactive bias reduction approaches are needed.

Lastly, the borderline case indicated that occurs at 50.06 percent chance (Figure 13) points at an actual uncertainty situation in the model, which explains why humans-in-the-loop review proposals should be implemented in unclear credit decision cultures. Combined, these results validate that the suggested framework demonstrates competitive predictive accuracy, and reveals insights pertinent to compliance that exclusively accuracy-driven models would not reveal.

6. CONCLUSION & FUTURE WORK

In this paper, we have presented a broad framework to implement Explainable AI in credit risk management that effectively overcomes the important trade-off between regulatory compliance and predictive performance. We have shown that complex machine learning models can be used by banks when combined with SHAP and LIME explanations along with popular ensemble learning techniques namely Random Forest, Gradient Boosting and XGBoost in a way such that models can comply with the requirements of modern regulatory frameworks.

The proposed framework was empirically evaluated on a real-world credit dataset of 49,999 records: The Explainable Artificial Intelligence (XAI) framework proposed has a very strong predictive performance as XGBoost shows that it outperforms other models with consistency (AUC-ROC: 0.9734, accuracy: 95.12%, precision: 91.23%, recall: 90.45%) and it is 16percentage higher than traditional baseline methods like logistic regression (AUC-ROC: 0.8124, accuracy: 78.56%). Thus, it questions the prevailing belief that accuracy and interpretability are trade batter at all. Finally, our ensemble methods—Random Forest (AUC-ROC: 0.9687) and Gradient Boosting (AUC-ROC: 0.9651)—characterized credit data consistently better than their simpler counterparts, suggesting that credit data contains non-linear relationships and/or complex feature interaction.

Secondly, this framework achieves human-interpretable explanations with two methods—global and local interpretation. Through SHAP analysis, we found that payment history, credit use ratio, and debt-to-income ratio were the top three drivers of credit decision in the overall dataset, consistent with traditional credit risk theory and regulators' expectations. We used LIME to look deeper into an individual credit decision to outline complementary instance level explanations. The fact that feature importance rankings are generally consistent across explanation methods adds confidence to the interpretations.

Third, our early warning indicators for bias automation found extremely few troubling indicators of discrimination in the model, again causing protected characteristics (gender and age) to have no direct influence on credit results at all. Prediction results explainedGender had basically no predictive value, while age had a small but sizable impact that accounted for a fraction of the variability in age that was associated with well-established risk factors, such as credit duration and available funds. This finding makes it possible to both satisfy the regulatory requirement requiring compliance with ECOA and proves that competitive models can be both accurate and fair.

The framework itself is more than technical-performance toolbox, having real world worthising. Our approach facilitates stakeholder specific explanation interfaces that can communicate with different targeted entities (such as model documentation and audit trails for regulators; actionable decision support, via feature contributions, for credit officers; and plain-language explanations of the approval/denial rationale for consumers). Everything gets automatically logged by the systems itself which is designed for complete audit trails to ease regulatory exams and reduce compliance costs as well as generation of adverse

action notices required by the consumer protection regulations.

Further, the modular architecture empowers financial institutions to adapt the framework to their specific regulatory environment and risk management requirements. The simultaneous application of such XAI techniques (SHAP and LIME) complementarily encapsulates various dimensions of generalization of the model and also their perspective—SHAP offers theoretically grounded global consistency; LIME, on the other hand, facilitates an easier to comprehend intuitive local fidelity beliefs of the model—which ultimately enable it with a robust validation on the explanation quality.

Finally, this research shows that explainability is not a regulatory checkbox but an essential element of good risk management. This enables financial institutions to resource to highlight deficiencies of the model per se, address data quality problems, to assure that the models are supported solely by economically relevant factors and not incidental correlations, and disclose hidden biases that may pose a danger to the decision making process. This resonates with a global shift toward responsible AI, where transparency is no longer just a compliance-driven checkbox but rather an essential and competitive business need that enables tracking of the model performance, control of risk, and trust of stakeholders.

We accomplish what is perhaps the holy-grail of ML in most regulated industries: (i) Predictive accuracy that is competitive with the state of the art of contemporary ML, (ii) Interpretability that is substantive and actionable by senior executives at the bank, (iii) Demonstrated fairness (in both opportunity and outcome), and (iv) When embedded in integrated operations, end-to-end compliance with regulatory requirements, all within the context of a real-world credit risk application. With growing AI assisted decision making, guides like the present one will be needed by institutions to cue up for the complex and dangerous way forward through the tangled thicket where innovation and regulation, revelations of inappropriate use, and initiatives such as these march towards each other.

LIMITATIONS

These limitations should be borne in mind while interpreting the findings of the present study, even though it contributes to existing literature. Firstly, the framework was trialled on a singular credit dataset comprising 49,999 records which may restrict generalisability by institution, geographic, and economic contexts. SHAP and LIME give excellent prospective enhancement over interpretability, but both are known to have limited backgrounds including Experimental computational burden for higher-dimensional data and sensitivity to local approximation configurations. Finally, as we already mentioned in the previous section, the fairness evaluation of existing works only focused on selected protected attributes and even when it did so, they excluded a number of sensitive demographic attributes that were present in the data. Finally, the evaluation was conducted in simulation and the temporal stability of the explanations and their regulatory acceptance in a non-simulation context remains therefore untested.

FUTURE WORK

These techniques need to be scalable and support real-time explanation generation, and robust to uncertainties regarding large-scale data processing and concept drift and adapt to changing regulatory requirements and legal standards. The framework will be extensible to include counterfactual explanations of Machine Learning models that are actionable and will be applicable across an arbitrary number of models and explanation techniques, thereby contributing to transparency, user trust, and ultimately regulatory compliance. Furthermore, regulators should also consider the need for regulatory acceptance and the adoption of methods that protect privacy when XAI is being used, while approaches to include causal explanations and causal discovery must be developed to improve auditability, data protection, and the trustworthiness of credit risk decision-making.

REFERENCES:

[1] Eshan, A.N., Nabil, A.H., Bonik, A., Bonik, B., Khorshed, S.B., Islam, F. and Tuaha, M.A.B.C., 2025. Credit Risk Prediction with Self-Supervised Learning: An Explainable AI Approach Integrating SHAP and LIME.

[2] Wang, M., Zhang, X., Yang, Y. and Wang, J., 2025. Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability. *Journal of Financial Risk Management*, *14*(3), pp.185-198.

[3] Malali, N. and Madugula, S.R.P., 2025, May. Implementing explainable AI for proactive regulatory compliance and auditing in financial markets. In *2025 International Conference on Networks and Cryptology (NETCRYPT)* (pp. 529-534). IEEE.

[4] Oko-Odion, C., 2025. AI-Driven Risk Assessment Models for Financial Markets: Enhancing Predictive Accuracy and Fraud Detection. *International Journal of Computer Applications Technology and Research*, *14*(04), pp.80-96.

[5] Kocoglu, E. and Ersoz, F., 2024, September. Explainable Artificial Intelligence (XAI) for Commercial Credit Limit Prediction Model: SHAP-LIME Comparison. In *International Congress on 3D Printing (Additive Manufacturing) Technologies and Digital Industry* (pp. 583-597). Cham: Springer Nature Switzerland.

[6] Tang, P., Peng, H., Luo, S. and Liu, Y., 2025. Forecasting Bank Default Risk with Interpretable Machine Learning: The Study of Chinese Banks. *Emerging Markets Finance and Trade*, *61*(6), pp.1661-1683.

[7] Khan, R., Shah, A.M., Ijaz, A. and Sumeer, A., 2025. Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. *International Journal of Social Sciences Bulletin*, *3*(8), pp.43-50.

[8] Li, Y. and Zhang, S., 2025. Machine Learning-Based Credit Risk Early Warning System for Small and Medium-Sized Financial Institutions: An Ensemble Learning Approach with Interpretable Risk Indicators. *Journal of Science, Innovation & Social Impact*, *1*(1), pp.372-383.

[9] Zhang, L., 2025, March. Using Explainable Machine Learning to Predict Loan Risk in Consumer Finance. In *Proceedings of the 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy* (pp. 235-238).

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

[11] Japinye, A.O. and Adedugbe, A.A., 2025. Explainable AI for credit scoring with SHAP-calibrated ensembles: a multi-market evaluation on public lending data. *Risks*, *13*(3), p.45.

[12] Renner, T., 2025. Explainable AI for Credit Risk Scoring on Loan Platforms. *Transactions on Computational and Scientific Methods*, *5*(6).

[13] Li, L.H., Sharma, A.K. and Cheng, S.T., 2025. Explainable AI based LightGBM prediction model to predict default borrower in social lending platform. *Intelligent Systems with Applications*, *26*, p.200514.

[14] Owen, O., 2025. Model Interpretability and Explainability: Foundations, Approaches, Challenges, and Future Directions in Responsible AI. *Approaches, Challenges, and Future Directions in Responsible AI (September 30, 2025)*.

[15] Bagwe, C., 2025. Explainable AI (XAI) in Compliance Audits: Bridging the Gap Between AI and Regulatory Transparency. *International Journal of Scientific Engineering and Science*, *9*, pp.137-144.

[16] Adegbola, I., 2025. Explainable AI for Risk Scoring in Letters of Credit: Bridging Machine Learning and Regulatory Interpretability. *Available at SSRN 5361956*.

[17] Desai, H., 2025. Reimagining Compliance: Explainable AI Models for Financial Regulatory Audits. In *Insights in Banking Analytics and Regulatory Compliance Using AI* (pp. 259-284). IGI Global Scientific Publishing.

[18] Jain, V., Balakrishnan, A., Beeram, D., Najana, M. and Chintale, P., 2024. Leveraging artificial intelligence for enhancing regulatory compliance in the financial sector. *Int. J. Comput. Trends Technol*, *72*(5), pp.124-140.

[19] Srivalli, K.S. and Sumanthi, D., 2025. Enhancing Financial Risk Assessment through Explainable AI: A SHAP-Based Approach for Transparent Decision-Making. *Available at SSRN 5190418*.

[20] Lin, S., Song, D., Cao, B., Gu, X. and Li, J., 2025. Credit risk assessment of automobile loans using machine learning-based SHapley Additive exPlanations approach. *Engineering Applications of Artificial Intelligence*, *147*, p.110236.

[21] De Silva, C., 2025. Advancing Financial Risk Management: AI-Powered Credit Risk Assessment through Financial Feature Analysis and Human-Centric Decision-Making.

[22] Zhang, Y., Chen, L. and Tian, Y., 2025. A Method for Evaluating the Interpretability of Machine Learning Models in Predicting Bond Default Risk Based on LIME and SHAP. *arXiv preprint arXiv:2502.19615*.

[23] Kurshan, E., Shen, H. and Chen, J., 2020, October. Towards self-regulating AI: Challenges and opportunities of AI model governance in financial services. In *Proceedings of the First ACM International Conference on AI in Finance* (pp. 1-8).

[24] Ravi, V., Srivastava, V.K., Singh, M.P., Burila, R.K., Kassetty, N., Vardhineedi, P.N., Pasam, V.R., Prova, N.N.I. and De, I., 2025, February. Explainable AI (XAI) for Credit Scoring and Loan Approvals. In *International Conference on Web 6.0 and Industry 6.0* (pp. 351-368). Singapore: Springer Nature Singapore.

[25] Li, Y. and Ling, Z., 2026. Real-Time Multi-Risk Early Warning for Community Banks: An Application of Ensemble Anomaly Detection and Explainable Artificial Intelligence. *Journal of Advanced Computing Systems*, *6*(2), pp.15-27.

[26] Ogunsola, K.O., Balogun, E.D. and Ogunmokun, A.S., 2021. Enhancing financial integrity through an advanced internal audit risk assessment and governance model. *International Journal of Multidisciplinary Research and Growth Evaluation*, *2*(1), pp.781-790.

[27] OLAWORE, S.O., OKOLI, C., ABIMBOLA, O., SERIFAT, B.U.U.U.D., OFURUM, A. and LEO, O., 2025. AI-Driven Cybersecurity Governance in Financial Services: Enhancing Ethical Auditing, Automated Compliance Monitoring and Explainable AI for Stakeholder Trust.

[28] 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), p.164.

[29] Kamruzzaman, M., Khatoon, R., Al Mahmud, M.A., Tiwari, A., Samiun, M., Hosain, M.S., Mohammad, N. and Johora, F.T., 2025. Enhancing Regulatory Compliance in the Modern Banking Sector: Leveraging Advanced IT Solutions, Robotization, and AI. *Journal of Ecohumanism*, *4*(2), pp.2596-2609.

[30] Mishra, A., Mou, S.N., Ara, J. and Sarkar, M., 2025. Regulatory and ethical challenges in AI-driven and machine learning credit risk assessment for Buy Now, Pay Later (BNPL) in US e-commerce: Compliance, fair lending, and algorithmic bias. *Journal of Business and Management Studies*, *7*(2), pp.42-51.

[31] Ibiyeye, T.O., Iornenge, J.T. and Adegbite, A., 2024. Evaluating regulatory compliance in the finance and investment sector: An analysis of current practices, challenges, and the impact of emerging technologies. *IOSR J. Econ. Financ.(IOSR-JEF)*, *15*, pp.1-8.

[32] Ahmed, A., Shah, A., Ahmed, T., Yasin, S., Longa, F.E.A., Hussaini, W. and Zubair, M., 2025. AI-Driven Innovations in Modern Banking: From Secure Digital Transactions to Risk Management, Compliance Frameworks, and AI-Based ATM Forecasting Systems. *Journal of Management Science Research Review*, *4*(3), pp.1145-1183.

[33] Khan, F.S., Mazhar, S.S., Mazhar, K., A. AlSaleh, D. and Mazhar, A., 2025. Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, *58*(8), p.232.

[34] Chinnaraju, A., 2025. Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. *World Journal of Advanced Engineering Technology and Sciences*, *14*(3), pp.170-207.

[35] Vyas, A., 2025. Revolutionizing Risk: The Role of Artificial Intelligence in Financial Risk Management, Forecasting, and Global Implementation. *Forecasting, and Global Implementation (April 21, 2025)*.

[36] Kothandapani, H.P., 2025. Ai-driven regulatory compliance: Transforming financial oversight through large language models and automation. *Emerging Science Research*, *12*(1), pp.12-24.