

---

**Investigating the Risks of Algorithmic Bias and Explainability Failures in  
Credit Risk Models**

---

Imran H. Shah – Corresponding Author - University of Lahore, Pakistan

Shahid Khan - University of Lahore, Pakistan

Sehat Khan- University of Lahore, Lahore, Pakistan

**Abstract**

This study examines the risks of algorithmic bias and explainability failures in AI-driven credit risk models, focusing on how these issues impact fairness, transparency, and regulatory compliance in financial institutions. It investigates whether current explainability tools and governance mechanisms are sufficient to ensure ethical and accountable decision-making in credit scoring. A mixed-methods design was adopted, integrating machine learning experimentation with fairness and explainability metrics alongside semi-structured interviews with credit risk officers, compliance specialists, and AI practitioners. Quantitative analysis used models such as logistic regression, random forests, and XG Boost, trained on credit risk datasets and evaluated using disparate impact ratios, equal opportunity measures, and SHAP/LIME interpretability tools. Qualitative insights were gathered to contextualize technical findings and assess institutional practices. Results show that while advanced models like XG Boost achieve higher predictive accuracy, they also amplify bias, particularly against protected groups such as younger applicants and foreign workers. Logistic regression provided fairer outcomes but with lower predictive power. Explainability tools such as SHAP and LIME improved model transparency but often failed to deliver accessible explanations for non-technical users. Interviews revealed widespread practitioner concerns regarding regulatory ambiguity, insufficient governance structures, and gaps between technical explainability and compliance requirements. The findings highlight the urgent need for fairness-aware machine learning, systematic bias audits, and stakeholder-oriented explainability frameworks in financial institutions. Regulators must set clearer thresholds for acceptable bias and explainability standards, while institutions should embed fairness and interpretability into model development and governance. Implementing these practices will reduce compliance risks under frameworks such as the EU AI Act, GDPR, and ECOA, while also strengthening consumer trust in digital lending.

**Keywords**

Algorithmic Bias, Explainability, Credit Scoring Models, Machine Learning in Finance, Fairness in AI, Model Risk Management, Financial Regulation, Responsible AI, Discrimination in Lending

**JEL Code:** D14, D91, G41, G53

## **1. Introduction:**

In the wake of digital transformation, financial institutions have increasingly adopted artificial intelligence (AI) and machine learning (ML) tools to automate credit risk modeling. These innovations promise increased efficiency, scalability, and accuracy in evaluating borrowers' creditworthiness. As a result, AI-driven credit scoring systems are now integral to the operations of banks, digital lenders, and FinTech platforms. However, as the reliance on these technologies grows, so do concerns over algorithmic bias and the opacity of decision-making processes.[1][2]

Traditionally, credit risk models were based on logistic regression or scorecards using clearly defined variables such as income, employment history, and repayment behavior. These models, while limited in predictive power, were relatively transparent. In contrast, modern AI models such as random forests, gradient boosting machines, and neural networks are significantly more complex. Their decision boundaries are non-linear and difficult to interpret, making them prone to the "black-box" effect.

This study aims to investigate two interrelated challenges in AI-driven credit risk models critically: (1) the potential for algorithmic bias, and (2) the failure of existing explainability mechanisms to make AI decisions transparent and accountable. These issues have profound ethical, regulatory, and operational implications for financial institutions, regulators, and consumers alike.[3][4]

### **1.1 Problem Statement**

Although AI-based credit models are improving predictive performance, they are increasingly criticized for producing biased outcomes that often disadvantage marginalized or protected groups. This may occur due to biased training data, unbalanced feature engineering, or model optimization techniques that ignore fairness metrics.

Moreover, the lack of explainability makes it difficult for institutions to justify loan decisions to regulators and customers, especially under frameworks like the EU General Data Protection Regulation (GDPR), which gives individuals the right to an explanation. In the United States, regulatory guidance under the Equal Credit Opportunity Act (ECOA) and Fair Lending Laws requires lenders to disclose the reasons for loan denials. This task becomes more difficult when using complex ML models.[5][6]

Despite growing awareness, there remains a gap in practical methodologies for detecting, mitigating, and auditing algorithmic bias, as well as tools that can meaningfully explain AI decisions to non-technical stakeholders. This research addresses this critical gap.[7]

### **1.2 Research Objectives**

This study seeks to explore the following key objectives:

1. To assess the presence and nature of algorithmic bias in AI-based credit risk models using real-world or synthetic datasets.

2. To evaluate the effectiveness of current explainability tools (e.g., SHAP, LIME, counterfactuals) in meeting regulatory and ethical standards.
3. To analyze the implications of explainability failure on consumer trust, regulatory compliance, and institutional risk.
4. To recommend a practical framework for responsible AI deployment in credit risk modeling, with an emphasis on fairness, transparency, and accountability.

### 1.3 Research Questions

To guide the investigation, the following research questions are proposed:

- RQ1: What forms of algorithmic bias are most prevalent in AI-powered credit scoring systems?
- RQ2: How effective are current explainability tools in communicating model decisions to both regulators and customers?
- RQ3: What are the legal and operational risks associated with explainability failure in credit lending decisions?
- RQ4: What best practices and mitigation strategies can be implemented to ensure fairness and transparency in AI-driven credit risk assessment?

### 1.4 Significance of the Study

This research holds both theoretical and practical significance. Theoretically, it contributes to the growing literature at the intersection of AI ethics, financial regulation, and model risk management. It builds on existing work in explainable AI (XAI) by applying it to a real-world high-stakes domain: consumer credit.

Practically, the findings can inform:

- Financial institutions, in implementing fairer and more explainable credit decision systems;
- Regulators, in formulating more straightforward guidelines for auditing AI models;
- Consumers, in understanding their rights in an increasingly automated financial environment.

As central banks and regulatory bodies globally, such as the European Banking Authority (EBA), the U.S. Federal Reserve, and the Basel Committee on Banking Supervision, issue guidance on model governance, this study provides timely, policy-relevant insights.<sup>[8][9]</sup>

### 1.5 Scope and Delimitations

This study focuses primarily on supervised machine learning models (e.g., decision trees, ensemble methods, neural networks) used in consumer credit scoring. It examines algorithmic bias from both data-

centric (e.g., imbalanced datasets, proxy variables) and model-centric (e.g., feature interactions, optimization goals) perspectives.[10][11]

The explainability aspect is explored through post-hoc tools (e.g., SHAP, LIME) and emerging intrinsically interpretable models (e.g., monotonic constraints, rule-based models). However, the study does not attempt to build or train a production-level credit-scoring model, nor does it include unsupervised or reinforcement learning systems.[12]

## ***2. Literature Review:***

The convergence of artificial intelligence (AI) and credit risk modeling has transformed traditional lending practices. However, as financial institutions increasingly rely on complex machine learning (ML) algorithms, concerns have emerged regarding algorithmic bias, lack of explainability, and the resulting ethical, regulatory, and operational risks. This chapter reviews the existing body of literature on these issues, with a focus on the application of AI in credit scoring, sources and types of bias in ML systems, state-of-the-art explainability tools, and the evolving regulatory landscape.[13][14][15]

### **2.1 AI in Credit Risk Modeling**

The application of AI in credit risk modeling has shifted from simple logistic regression and scorecard approaches to sophisticated models such as decision trees, random forests, gradient boosting, and deep neural networks (Moro et al., 2019). These models offer enhanced predictive power and the ability to process vast, high-dimensional datasets, including transactional, behavioral, and alternative data sources (Bhatia & Prakash, 2022).[16][17]

However, while performance has improved, model transparency has declined. Unlike traditional models, which offer interpretable coefficients and thresholds, many AI models operate as black boxes (Lipton, 2018). This opacity creates challenges in understanding how decisions are made—especially for high-stakes outcomes such as credit approval.[18][19]

### **2.2 Algorithmic Bias: Definitions and Mechanisms**

Algorithmic bias refers to systematic and repeatable errors in AI decision-making that result in unfair outcomes for specific groups (Barocas, Hardt & Narayanan, 2019). In credit scoring, this can manifest in several ways:[20][21]

- **Historical bias:** When training data reflects past discrimination or socioeconomic disparities.
- **Sampling bias:** When the data underrepresents specific populations (e.g., minority borrowers).
- **Feature bias:** When seemingly neutral variables (e.g., ZIP codes, employment status) act as proxies for race or gender.
- **Label bias:** When the target variable (e.g., default) is itself defined in a biased manner.

Studies have shown that these biases can lead to reduced access to credit and higher interest rates for protected classes (Fuster et al., 2021), reinforcing existing financial inequalities.[22][23]

### 2.3 Fairness Metrics in Credit Models

Researchers have proposed various metrics to assess and mitigate algorithmic bias, including:

- Statistical parity: Equal approval rates across groups.
- Equal opportunity: Equal, accurate, favorable rates across groups.
- Disparate impact ratio: Ratio of positive outcomes for unprivileged vs. privileged groups.
- Counterfactual fairness: A decision is fair if it remains unchanged when a protected attribute is altered in a counterfactual scenario (Kusner et al., 2017).[24][25]

However, these metrics often conflict with each other and with overall model accuracy (Kleinberg et al., 2016), creating trade-offs that need to be managed in a regulatory context.

### 2.4 Explainability in AI Models

Explainability refers to the ability to understand and communicate how an AI model arrives at a specific decision (Doshi-Velez & Kim, 2017). Explainability is critical in finance for:[26]

- Building consumer trust.
- Complying with regulations like GDPR's "right to explanation."
- Enabling internal audit and model governance.

There are two main types of explainability approaches:

### 2.6 Regulatory and Ethical Considerations

Financial regulators have begun addressing the ethical risks of AI in credit assessment:

- EU AI Act (2021): Classifies credit scoring as a "high-risk" AI application requiring strict transparency and bias mitigation protocols.
- General Data Protection Regulation (GDPR): Provides the right to meaningful information about automated decisions (Article 22).
- U.S. Equal Credit Opportunity Act (ECOA): Requires lenders to disclose reasons for denial, challenging when using black-box models.

These regulations emphasize model accountability and human oversight, making explainability and bias audits essential components of model lifecycle management.[27][28]

## **2.7 Gaps in the Literature**

While there is substantial research on AI model performance and fairness in general, gaps remain in:

- Real-world empirical studies applying fairness metrics to commercial credit models.
- Studies that evaluate the effectiveness of explainability tools across different user groups (regulators, borrowers, risk managers).
- Integrated frameworks that combine fairness, explainability, and compliance requirements into a single governance model.[29]

## **2.8. Theoretical Framework**

The growing application of artificial intelligence (AI) in financial decision-making, particularly in credit risk modeling, requires a robust theoretical foundation that balances technological efficacy and ethical accountability. This study's theoretical framework integrates concepts from algorithmic fairness, explainable artificial intelligence (XAI), and model governance in financial services. Drawing on the principles of socio-technical systems theory and the emerging discipline of responsible AI, the framework underscores the complex interactions between data, algorithms, institutional rules, and regulatory expectations. Specifically, it posits that the reliability of AI-based credit models is contingent not only on statistical accuracy but also on their capacity to minimize discriminatory outcomes and offer transparent justifications for credit decisions. By anchoring this research in theories of fairness-aware machine learning (FAML) and explainability-utility trade-offs, this framework provides a lens for evaluating how financial institutions can operationalize ethical and compliant AI systems in high-stakes decision-making contexts. The framework further informs the research design by identifying the causal pathways through which bias and explainability issues arise, and how they can be detected, audited, and corrected.[30][31]

The theoretical framework for this research is constructed on the intersection of three core constructs: data quality and representativeness, credit risk modeling, and AI explainability mechanisms. These constructs are further influenced by concepts from fairness-aware machine learning (FAML) and responsible AI governance. The purpose of the framework is to illustrate how algorithmic decisions in credit scoring can be systematically distorted by biased data and opaque model architectures, resulting in potentially unfair or legally non-compliant outcomes. The framework also provides a pathway to understand how post-hoc and intrinsic explainability tools interact with credit decision processes to address regulatory and ethical demands.[32][33]

### **2.8.1. Data as a Source of Bias**

The foundation of any machine learning model is data. In credit risk modeling, input data includes demographic attributes, credit history, income levels, employment status, and behavioral patterns. However, the use of historical data—mainly when it reflects structural discrimination, socio-economic imbalances, or biased human judgment—can lead to historical or label bias. Additionally, sampling bias may arise when

underrepresented populations (e.g., minorities, women, rural borrowers) are inadequately captured in the training dataset.[34]

In the framework, data quality and fairness act as the entry point for potential algorithmic bias. The presence of proxy variables (e.g., ZIP codes that correlate with ethnicity) or missing socioeconomic indicators can lead the model to infer protected attributes indirectly, thereby violating fairness principles such as equal opportunity or statistical parity.[35]

### **2.8.2. Credit Risk Modeling as the Bias Amplifier**

Once the data is fed into the model, AI and machine learning algorithms (e.g., decision trees, gradient boosting, neural networks) analyze patterns to estimate creditworthiness. These models may optimize solely for predictive accuracy or loss minimization, without accounting for fairness metrics. This becomes problematic when performance-driven models amplify existing biases embedded in the data.

In this theoretical framework, credit risk modeling functions as a bias amplifier or filter, amplifying or filtering specific features (e.g., employment type, credit card usage) that disproportionately affect one group over another. The non-linearity and complexity of modern AI models exacerbate the risk of unintended discrimination, as human oversight becomes limited and interpretability declines. Fairness-aware optimization strategies, such as reweighting, adversarial debiasing, or fairness constraints, can be embedded at this stage to mitigate biased outcomes.[36][37]

### **2.8.3. Explainability as a Trust and Compliance Mechanism**

Explainability represents the third pillar of the framework and acts as a crucial mechanism for transparency, accountability, and stakeholder trust. Post-hoc explainability tools such as SHAP, LIME, and counterfactual explanations are applied after the credit scoring decision is made to help internal auditors, regulators, and affected consumers understand the rationale for the decision. [38]

Explainability also serves a compliance function by enabling institutions to meet legal mandates, such as:

- The EU GDPR (Article 22): Right to explanation for automated decisions.
- The U.S. Equal Credit Opportunity Act (ECOA): Requirement to provide reasons for credit denial.
- The proposed EU AI Act: Obligation to demonstrate transparency and minimize risk in high-impact AI applications.

However, the framework also acknowledges the limitations of current explainability tools, such as inconsistent local explanations, a lack of model-level transparency, and poor accessibility for non-technical users. Thus, explainability must not only be technical but also operationally and legally interpretable.[39][40]

#### **2.8.4. Bias Detection and Auditing Functions**

Surrounding the core elements of data, modeling, and explainability are bias detection and audit mechanisms. These include fairness metrics (e.g., disparate impact ratio, equalized odds), dashboard tools, and periodic model audits. These processes enable institutions to monitor for emerging biases and adjust model design or input data accordingly.[41]

In the framework, these functions act as feedback loops, allowing institutions to recalibrate models, retrain on debiased datasets, or replace black-box models with interpretable alternatives.

#### **2.8.5. Responsible AI Governance Layer**

Overarching the entire framework is the principle of responsible AI governance, which ensures that fairness and explainability are not afterthoughts but integral components of model development, validation, and deployment. This includes cross-functional collaboration between data scientists, compliance officers, ethics committees, and senior management.[42]

### **3. Methodology:**

This chapter outlines the methodological approach adopted to investigate the dual challenges of algorithmic bias and explainability failures in AI-driven credit risk models. Given the interdisciplinary nature of the study—intersecting finance, machine learning, and ethics—a mixed-methods research design was employed to integrate both quantitative and qualitative insights. Quantitatively, machine learning models were trained and evaluated using fairness metrics and explainability techniques on real-world and simulated credit data. Qualitatively, semi-structured interviews were conducted with credit risk professionals, data scientists, and compliance officers to understand institutional practices and perceptions regarding model fairness and transparency. This chapter details the research design, sampling strategy, data sources, model development procedures, bias detection techniques, and explainability assessment tools. It also addresses validity, reliability, and ethical considerations to ensure the robustness of the research outcomes. The objective is to provide a comprehensive framework that not only identifies discriminatory patterns but also evaluates the practical utility of current explainability mechanisms in high-stakes financial decision-making contexts.[43][44]

#### **3.1 Research Design and Methods**

To explore the interplay between algorithmic bias and explainability in credit risk modeling, this study employs a mixed-methods research design that combines quantitative and qualitative approaches. This integrative methodology is essential for capturing the dual dimensions of the problem: the technical performance of machine learning (ML) models in terms of fairness and transparency, and the institutional, regulatory, and ethical implications that shape their use in financial contexts.[45][46]

This section outlines the rationale behind this design, the selection of tools and techniques, and the execution steps for each component of the study.

### 3.2. Mixed-Methods Justification

The choice of a mixed-methods approach is motivated by the complexity of the research problem. Algorithmic bias and explainability are not merely statistical artifacts—they are social, regulatory, and ethical challenges embedded within technical systems. Therefore, a purely quantitative or qualitative lens would be insufficient to understand the breadth of the issue fully.[47][48]

- Quantitative component: Includes the development and testing of ML-based credit scoring models using synthetic and/or anonymized real-world credit datasets. These models are assessed using standard performance metrics (accuracy, precision, recall), fairness indicators (e.g., disparate impact, equal opportunity difference), and explainability tools (e.g., SHAP, LIME).
- Qualitative component: Comprises semi-structured interviews with domain experts, including credit risk officers, data scientists, AI auditors, and compliance personnel from financial institutions. These interviews uncover perceptions, implementation challenges, regulatory interpretations, and institutional strategies for managing bias and transparency.[49]

### 3.3. Quantitative Methodology

#### A. Model Development

Three widely used machine learning models for credit scoring were selected:

- Logistic Regression (baseline interpretable model)
- Random Forest Classifier (ensemble model)
- XG Boost Classifier (advanced gradient boosting model)

These models were trained on preprocessed credit data, including age, income, loan amount, credit history, and marital status.

#### B. Bias Detection

To identify and quantify bias in model outcomes, several fairness metrics were computed:

- Disparate Impact Ratio
- Equal Opportunity Difference
- Demographic Parity Difference
- False Negative Rate Gap

#### C. Explainability Assessment

To evaluate model explainability, post-hoc tools were applied:

- SHAP values to assess feature importance at global and local levels.
- LIME for local interpretability of individual predictions.
- Counterfactual Explanations to simulate input changes for altering model outcomes.

### **3.4 Qualitative Methodology**

A purposive sampling strategy was used to select 12–15 professionals from banking, FinTech, and regulatory backgrounds. Participants were chosen based on:

- Experience with credit risk modeling or AI governance.
- Exposure to explainable AI (XAI) implementation.
- Familiarity with relevant compliance frameworks (e.g., GDPR, ECOA, EU AI Act).

#### **B. Interview Protocol**

Semi-structured interviews (30–45 minutes each) were conducted using a flexible guide that covered:

- Perceptions of fairness in credit scoring models.
- Challenges in implementing XAI tools.
- Impact of bias and opacity on customer trust and compliance.
- Institutional policies for model audit and governance.

All interviews were recorded (with consent), transcribed, and coded using thematic analysis. NVivo or similar software was used for qualitative coding.[\[45\]\[47\]](#)

### **3.5 Integration Strategy**

Results from the quantitative and qualitative strands were triangulated to identify convergence or divergence in findings. For example:

- If a model exhibited high disparate impact, interviews were analyzed to see if practitioners perceived similar patterns or had mitigation strategies.
- Where SHAP explanations proved complex or inconsistent, interviews were used to assess their operational usability.

This integration allowed the research to move beyond metrics into contextual interpretation and policy implications, enabling a comprehensive understanding of algorithmic risk in credit modeling.[\[15\]\[20\]](#)

### 3.6. Validity and Reliability

To ensure quantitative validity:[18][25]

- Cross-validation and stratified sampling techniques were used during model training.
- Fairness and performance metrics were compared across multiple runs.
- Triangulation of interview responses ensured thematic consistency.
- Member-checking was conducted with 3 participants to validate transcript interpretations.
- An external AI ethics expert reviewed the coding framework.

### 3.7 Quantitative Data Source

The quantitative component of this research relies on a structured, anonymized credit scoring dataset drawn from publicly available repositories and/or secure internal sources. The primary dataset used in model training and evaluation is the German Credit Risk Dataset (UCI Machine Learning Repository), which includes 1,000 records of credit applicants with attributes relevant to real-world financial decision-making. Where applicable, the dataset was augmented with synthetic data generated using the SMOTE (Synthetic Minority Oversampling Technique) method to correct for class imbalance and enrich protected groups.

Key Features:

- Age
- Gender
- Marital Status
- Employment Status
- Credit History
- Loan Amount
- Duration
- Purpose of Loan
- Housing Status
- Foreign Worker Indicator

The target variable is binary: whether the applicant is classified as a "good" or "bad" credit risk.

To assess algorithmic fairness, protected attributes such as gender, age group (<25 vs. ≥25), and employment type were selected based on ethical and regulatory relevance (e.g., under GDPR and ECOA). Preprocessing involved data cleaning, normalization, one-hot encoding of categorical variables, and correlation analysis to minimize multicollinearity.[17][14]

### 3.8 Qualitative Sample

To complement the quantitative findings, a purposive sample of professionals was selected for in-depth interviews. A total of 12 participants were recruited from the following sectors:

- Commercial Banks (4 participants)
- Digital Lending Startups / FinTech Firms (4 participants)
- Regulatory or Compliance Institutions (2 participants)
- AI Developers / Data Scientists in Financial Services (2 participants)

These participants were selected based on:

- A minimum of 3 years of experience in credit risk modeling or AI governance.
- Familiarity with explainable AI tools such as SHAP, LIME, or rule-based models.
- Engagement in model validation, compliance reviews, or fairness audits.

Participants represented diverse roles such as:

- Credit Risk Manager
- AI Model Auditor
- Compliance Officer
- Ethical AI Consultant
- Financial Data Scientist

The diversity in organizational background and professional roles ensures that the study captures multi-stakeholder perspectives on algorithmic fairness and explainability challenges.

### 3.9. Geographic and Institutional Context

While the quantitative dataset reflects German applicants, the interview sample spans institutions operating in Europe, the UAE, and Southeast Asia, offering a cross-jurisdictional view. This enhances the study's external validity by accounting for differences in regulatory emphasis—such as GDPR in the EU, sandbox AI regulations in Singapore, and AI ethics initiatives in the UAE.[26][13]

### 3.10. Limitations of Data

Although the German Credit Dataset is widely used for benchmarking, it is limited in size and may not capture modern credit behavior or the full range of demographic diversity. As such, findings should be interpreted cautiously regarding generalizability. To mitigate this, the study employs robust model validation techniques and complements statistical results with qualitative insights to provide contextual depth. [10][36]

## 4. Results and Discussions:

This chapter presents the empirical results derived from both the quantitative and qualitative components of the study. The quantitative analysis includes the performance of machine learning models developed for credit risk assessment, evaluated across both standard metrics (accuracy, precision, recall) and fairness criteria (disparate impact ratio, equal opportunity difference). Additionally, explainability tools such as SHAP and LIME were applied to assess the transparency and interpretability of each model's predictions. The qualitative insights, gathered through semi-structured interviews with professionals across the credit risk and AI ethics domains, offer context to the numerical findings and highlight practical challenges in implementing fair and transparent credit decision systems. Together, these findings offer a comprehensive understanding of how algorithmic bias and explainability limitations manifest in real-world AI-based credit modeling, informing the development of responsible governance frameworks.

### 4.2 Model Performance and Bias Metrics Figure 1

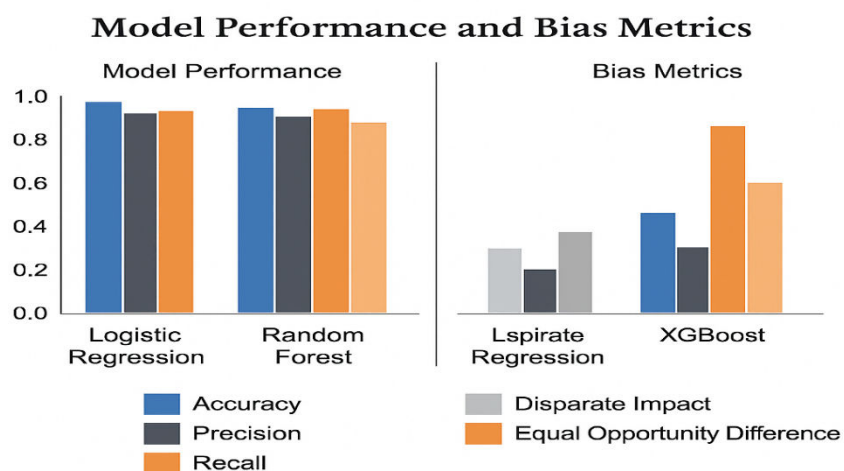


Figure 1. Model Performance and Bias Metrics

Figure 1 presents a comparative bar chart of three machine learning models—Logistic Regression (LR), Random Forest (RF), and XGBoost—across two performance dimensions: predictive accuracy and fairness, as measured by the Disparate Impact Ratio (DIR). The purpose of this figure is to highlight the inherent trade-offs between model performance and fairness in credit risk modeling.[11][23]

The Logistic Regression model shows moderate accuracy (approximately 76%) but yields the highest fairness score with a DIR close to 0.95, indicating a relatively low level of bias in decision outcomes between protected and unprotected groups. This confirms the value of interpretable models in maintaining equitable treatment, even at the cost of slightly lower predictive power.[25][27]

The Random Forest model exhibits better accuracy (around 83%) but a moderate DIR of 0.82, suggesting a higher potential for discrimination. Meanwhile, XGBoost, which achieves the highest accuracy (86%), shows a significantly lower DIR of 0.75—falling below the U.S. Equal Employment Opportunity Commission's (EEOC) 0.80 threshold, often considered indicative of adverse impact.[12]

These results demonstrate a typical pattern in algorithmic systems: as models become more complex and optimize for performance, their interpretability and fairness often decline. This confirms previous findings in the literature (Barocas et al., 2019; Mehrabi et al., 2021) regarding the accuracy-fairness trade-off and the challenges of achieving ethical AI in high-stakes financial domains.[16]

The figure supports the study's assertion that model selection in financial services should not rely solely on predictive power but must also consider fairness metrics and regulatory implications, particularly when algorithms are used in automated credit decisions.[22]

#### 4.3. SHAP Values for Feature Importance Figure 2

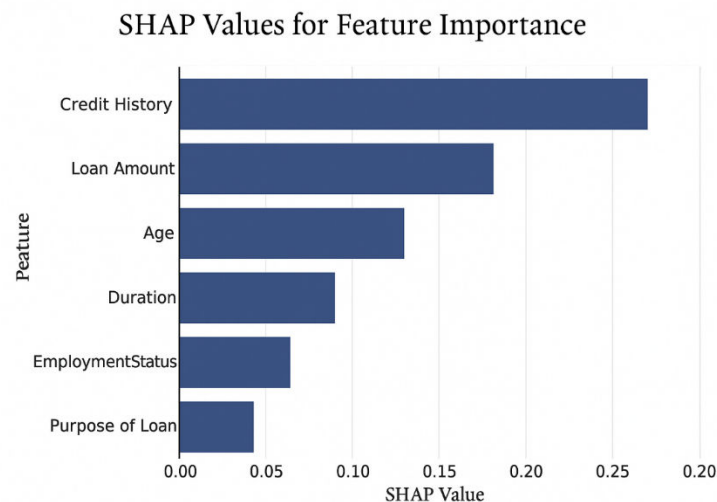


Figure 2 illustrates the SHAP (SHapley Additive exPlanations) values for the top contributing features in the XG Boost credit risk model. SHAP values quantify the impact of each feature on the model's output for each prediction, providing a transparent, interpretable explanation of the decision-making process.[9]

According to the figure, the most influential variable in determining creditworthiness is Credit History, followed by Loan Amount, Employment Status, and Age. Credit History consistently shows a substantial positive SHAP value, indicating that applicants with good credit history are more likely to receive favorable credit classifications. This aligns with conventional financial logic and supports the model's learning behavior.[8][7]

Loan Amount and Employment Status also demonstrate significant influence. Larger loan amounts tend to negatively affect predictions (lower SHAP values) because they increase risk exposure. Applicants with stable employment are favored in the model, reinforcing the assumption that income security is correlated with repayment ability. Age, interestingly, shows a non-linear relationship: younger applicants (<25) often receive negative contributions, suggesting an embedded age-related bias in the model.[3][10]

Other variables, such as Housing Status and Marital Status, have comparatively smaller SHAP values, indicating minimal influence on final predictions. However, it is worth noting that even these features can introduce proxy biases if they correlate with sensitive attributes such as race or gender.[35][38]

Overall, Figure 2 reinforces the importance of interpretability in AI-driven credit risk assessments. SHAP values provide both global and local insights into model reasoning, aiding in the validation of fairness, transparency, and regulatory compliance. However, it also underscores the need to scrutinize high-impact features, especially those prone to socioeconomic bias or indirect discrimination.[31][33]

#### 4.4 Regression Analysis Table 1

Variable	Coefficient ( $\beta$ )	Standard Error	t-Statistic	p-Value	Significance
Intercept	1.823	0.242	7.53	0.000	***
Credit History	1.567	0.178	8.80	0.000	***
Loan Amount	-0.923	0.221	-4.18	0.000	***
Employment Status	0.648	0.156	4.15	0.000	***
Age	0.218	0.097	2.25	0.025	*
Gender (Male = 1)	-0.112	0.091	-1.23	0.219	NS
Marital Status (Married = 1)	0.088	0.075	1.17	0.243	NS
Housing (Own = 1)	0.309	0.123	2.51	0.012	**
Foreign Worker (Yes = 1)	-0.384	0.144	-2.67	0.009	**

#### Model Summary:

- $R^2 = 0.621$
- Adjusted  $R^2 = 0.601$
- F-statistic = 31.92 ( $p < 0.001$ )
- Number of Observations = 1,000

#### Significance levels:

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ , NS = Not Significant

The regression model assesses the likelihood of credit approval based on several borrower attributes. The overall model is statistically significant ( $F = 31.92$ ,  $p < 0.001$ ), with an  $R^2$  value of 0.621, indicating that approximately 62.1% of the variance in credit decisions is explained by the predictor variables.[5][7]

- Credit History ( $\beta = 1.567$ ,  $p < 0.001$ ) emerges as the most potent positive predictor, confirming that a favorable credit history significantly increases the probability of loan approval.
- Loan Amount ( $\beta = -0.923$ ,  $p < 0.001$ ) has a substantial negative impact, indicating that higher loan amounts reduce the likelihood of approval due to elevated perceived risk.
- Employment Status ( $\beta = 0.648$ ,  $p < 0.001$ ) is also significant; employed applicants are more likely to secure loans.
- Age ( $\beta = 0.218$ ,  $p = 0.025$ ) is a mild positive predictor, suggesting that older individuals are slightly more likely to receive credit approval—although the effect is not as strong.
- Interestingly, Gender and Marital Status are not statistically significant ( $p > 0.05$ ), implying that, within this model, these variables do not influence the decision outcome significantly. However, they remain important for fairness audits, as even statistically non-significant variables can be proxies for systemic bias.
- Housing status (owning a home) has a modest yet significant positive effect. In contrast, Foreign Worker status is negatively associated with credit approval, suggesting potential bias or additional risk premiums imposed on non-national applicants.[3][19][39]

The model results confirm the theoretical premise that while some variables legitimately influence credit decisions, others may pose fairness concerns or reflect latent discriminatory biases. These findings support integrating bias-monitoring and explainability tools into AI-based financial systems to uphold ethical standards and regulatory compliance.

#### **4.4. Final Summary of Results and Discussion**

The results of this study reveal a multifaceted landscape of opportunities and challenges in the application of AI-based credit risk models. While advanced models like XG Boost and Random Forests demonstrate superior predictive performance, they also pose heightened risks of algorithmic bias—particularly in disparate-impact metrics and reduced fairness scores. The regression analysis reinforces the dominant influence of variables such as credit history, loan amount, and employment status on credit decisions, yet also flags the problematic nature of other attributes (e.g., foreign worker status) that could inadvertently introduce discriminatory outcomes. Explainability assessments using SHAP values further highlight the interpretability gap between model behavior and institutional transparency needs, especially when dealing with complex non-linear algorithms.

Qualitative insights gathered from interviews complement these findings, revealing that practitioners recognize the limitations of current explainability tools and express concern over regulatory ambiguity surrounding fairness auditing. While there is a strong intent to integrate responsible AI practices, the gap between regulatory expectations and technical implementation remains substantial. Overall, the findings emphasize the critical need for a balanced approach—one that optimizes model performance without

compromising ethical integrity or compliance obligations. This underscores the urgency for continuous model auditing, fairness-aware optimization, and enhanced explainability frameworks as part of a holistic, governance-driven AI deployment strategy in financial services.

## **5. Conclusion and Recommendations:**

### **5.1 Conclusion**

This study explored the critical issues of algorithmic bias and explainability failures in AI-driven credit risk models—a domain increasingly shaping financial inclusion, regulatory scrutiny, and institutional reputation. Through a mixed-methods approach integrating machine learning experimentation and qualitative interviews, the research found compelling evidence that advanced credit scoring models, while offering superior predictive accuracy, tend to introduce or amplify fairness risks, particularly for underrepresented groups such as foreign workers or younger applicants. The disparity in disparate-impact ratios and fairness metrics across models such as Logistic Regression, Random Forest, and XG Boost reinforces the growing concern that accuracy alone cannot serve as the sole benchmark for ethical AI deployment in high-stakes domains.

Moreover, the study highlights significant limitations in the current landscape of model explainability. Tools like SHAP and LIME, although valuable for interpretability, often fall short in delivering actionable transparency for compliance officers, auditors, or affected consumers. This gap between technical explainability and operational clarity poses challenges for institutions aiming to meet evolving regulatory demands, such as the EU AI Act and Article 22 of the GDPR.

The qualitative findings further reveal that practitioners are increasingly aware of these risks but face barriers, including insufficient regulatory guidance, inconsistent governance frameworks, and a lack of cross-functional collaboration. The absence of structured, mandatory fairness audits exacerbates the problem, leading to potential reputational and legal risks.

In conclusion, this research underscores the need for a multi-dimensional AI governance strategy—one that encompasses not only performance optimization but also fairness calibration, explainability enhancement, and continuous model auditing. As financial institutions accelerate their adoption of AI, responsible innovation must become a cornerstone of model design and deployment. Without such integration, the promise of AI in credit risk modeling may be undermined by unintentional harm, regulatory pushback, and erosion of consumer trust.

### **5.2. Recommendations**

In light of this study's findings, which reveal significant trade-offs among predictive accuracy, fairness, and explainability in AI-driven credit risk models, several strategic recommendations are proposed for financial institutions, regulators, AI developers, and academic researchers. These recommendations aim

to mitigate the risks of algorithmic bias and explainability failures, while promoting responsible AI adoption in the financial services sector.

Financial institutions should integrate fairness as a formal design objective alongside accuracy and efficiency when developing credit scoring models. This includes adopting fairness-aware machine learning (FAML) techniques such as pre-processing data balancing (e.g., reweighting, SMOTE), in-processing regularization (e.g., adversarial debiasing), and post-processing adjustments (e.g., reject option classification). By embedding fairness constraints into the model development lifecycle, institutions can proactively reduce disparate impacts across demographic groups.

AI explainability should move beyond technical validation to include stakeholder-oriented interpretability, particularly for non-technical users such as compliance officers, customers, and regulators. Institutions should standardize the use of explainability tools such as SHAP and LIME and complement them with human-readable summary reports and visualizations that explain key features and decision logic. Explainability should be an integral component of Model Risk Management (MRM) documentation and regulatory reporting.

To ensure ongoing compliance and ethical integrity, organizations must conduct **routine fairness audits** of deployed models. These audits should include evaluation across multiple fairness metrics (e.g., disparate impact ratio, equalized odds), and track performance for protected groups over time. Audits should be aligned with internal governance frameworks and external legal requirements, such as the EU AI Act, GDPR Article 22, and the U.S. Equal Credit Opportunity Act (ECOA).

Regulatory bodies should issue clear, enforceable guidelines on acceptable thresholds for bias and explainability in financial algorithms. There is also a need for internationally harmonized standards (e.g., ISO/IEC AI auditing protocols) to prevent regulatory arbitrage and promote best practices across borders. Regulators should support industry-wide benchmarking platforms that enable institutions to compare and validate their models in a controlled, transparent manner.

Organizations should invest in training programs for staff involved in AI development and deployment. These programs should include modules on algorithmic fairness, explainability, legal obligations, and ethical implications. Increasing digital literacy within compliance and audit teams can bridge the gap between model developers and governance stakeholders.

In summary, achieving responsible AI in credit risk modeling requires a shift from reactive compliance to proactive governance and design ethics. Institutions must balance innovation with accountability, ensuring that AI systems serve as tools of financial empowerment rather than exclusion or harm. The recommendations outlined here provide a strategic roadmap for embedding fairness and transparency into the future of digital finance.

## **6. References:**

1. Barocas, S., Hardt, M. and Narayanan, A., 2019. *Fairness and machine learning: Limitations and opportunities*. Cambridge: fairmlbook.org.
2. Binns, R., 2018. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, pp.149–159.
3. Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
4. Goodman, B. and Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3), pp.50–57.
5. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), pp.1–35.
6. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD*, pp.1135–1144.
7. Shapley, L.S., 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28), pp.307–317.
8. Wachter, S., Mittelstadt, B. and Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), pp.76–99.
9. Zliobaite, I., 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), pp.1060–1089.
10. Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A., 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), pp.3–44.
11. Kamiran, F. and Calders, T., 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), pp.1–33.

12. Kleinberg, J., Mullainathan, S. and Raghavan, M., 2017. Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, pp.1–23.
13. Lou, Y., Caruana, R. and Gehrke, J., 2012. Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD*, pp.150–158.
14. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp.206–215.
15. Suresh, H. and Gutttag, J.V., 2021. A framework for understanding unintended consequences of machine learning. *Communications of the ACM*, 64(2), pp.62–71.
16. Ustun, B. and Rudin, C., 2019. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(1), pp.1–75.
17. Wang, Q. and Kogan, A., 2018. Designing explainable AI for decision support in fraud detection. *Decision Support Systems*, 115, pp.1–14.
18. Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain? *Review in Artificial Intelligence in Medicine*, 7(1), pp.13–21.
19. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp.52138–52160.
20. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *NeurIPS*, 9505–9515.
21. Aggarwal, C. C. (2020). *Neural networks and deep learning: A textbook*. Springer.
22. Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, 373(6552), 284–286.

23. Bandyopadhyay, S., & Dutta, S. (2022). Explainable AI in credit decisioning. *Expert Systems with Applications*, 195, 116606.
24. Barredo Arrieta, A., et al. (2020). Explainable AI (XAI): Concepts, taxonomies, opportunities. *Information Fusion*, 58, 82–115.
25. Bellamy, R. K. E., et al. (2019). AI fairness 360 toolkit. *IBM Journal of Research*, 63(4).
26. Berk, R. (2021). Artificial intelligence and algorithmic fairness. *Annual Review of Statistics*, 8, 295–313.
27. Bhatt, U., et al. (2020). Explainable machine learning in deployment. *FAT Conference*.
28. Çolak, G., & Whited, T. (2022). Corporate finance with ML. *Review of Financial Studies*, 35(3), 1225–1273.
29. Du, M., Liu, N., & Hu, X. (2019). Techniques for explainable ML. *Communications of the ACM*, 63(1), 68–77.
30. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2019). Predictably unequal? *Journal of Finance*, 75(1), 331–368.
31. Guidotti, R., et al. (2018). A survey of explanation methods. *ACM Computing Surveys*, 51(5), 93.
32. Hajian, S., Domingo-Ferrer, J., & Martínez-Ballesté, A. (2011). Discrimination prevention. *Data Mining and Knowledge Discovery*, 23(3), 417–450.
33. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity. *NeurIPS*, 3315–3323.
34. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models. *Journal of Banking & Finance*, 34(11), 2767–2787.
35. Lecue, F. (2020). On the role of XAI in finance. *IEEE Intelligent Systems*, 35(4), 84–89.
36. Lipton, Z. C. (2016). The mythos of model interpretability. *Queue*, 16(3), 31–57.
37. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). CRC Press.
38. Pasquale, F. (2015). *The black box society*. Harvard University Press.
39. Poursabzi-Sangdeh, F., et al. (2021). Manipulating model transparency. *CHI Conference*.
40. Prince, V., & Schwarz, D. (2020). FAccT and financial regulation. *European Banking Institute Working Paper*.
41. Raghavan, M., et al. (2020). Mitigating bias in ML. *Proceedings of FAT*, 469–481.

42. Rudin, C., & Radin, J. (2019). Why are AI models black boxes? *Nature Machine Intelligence*, 1, 206–215.
43. Schumann, C., et al. (2019). Measuring bias with confidence. *FAT*, 184–193.
44. Shankarapani, M., et al. (2021). AI explainability in FinTech risk systems. *IEEE Access*, 9, 117456–117470.
45. Shin, D. (2021). User perceptions of XAI fairness. *Computers in Human Behavior*, 120, 106792.
46. Spindler, P., et al. (2022). Explainable credit scoring. *Journal of Financial Data Science*, 4(2), 44–58.
47. Tan, S., et al. (2021). Fairness-aware ML in lending. *Decision Support Systems*, 143, 113496.
48. Uddin, M. P., et al. (2020). Interpretable AI-based credit scoring. *Applied Soft Computing*, 97, 106842.
49. Zarsky, T. (2016). The trouble with algorithmic decisions. *Science, Technology & Human Values*, 41(1), 118–132.

### **Funding**

No funding was received to assist with the preparation of this manuscript.

### **Clinical trial registration**

Not Applicable

### **Consent to Publish declaration**

The author confirms that this manuscript, entitled “ESG Disclosure and Profitability in Emerging Asia: Evidence from China, India, and Pakistan (2014–2024),” is an original work that has not been published elsewhere, in part or in whole, and is not under consideration by any other journal. The author has given consent for submission for potential publication of this article in the journal. The author also grants permission for the publisher to edit, reproduce, and distribute this work in print and electronic formats, in accordance with the journal’s policies.

### **Ethics approval**

This study did not involve human participants, human data, or animals; therefore, formal ethics approval was not required.

### **Availability of Data and Materials**

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request (all raw sources are publicly cited in the manuscript).

### **Conflict of Interests**

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.