SCHOOL OF COMPUTING, TECHNOLOGY AND APPLIED SCIENCES

Enhancing Explainability and Transparency in Machine Learning-Based Credit Scoring Models.

Thomas Mumbuwa Kamunu

ZCAS UNIVERSITY

2025

Enhancing Explainability and Transparency in Machine Learning-Based Credit Scoring Models.

Thomas Mumbuwa Kamunu

A Final Year Research Project submitted in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science

ZCAS University
2025

DECLARATION

Name:Thomas Mumbuwa Kamunu
Student Number: G14044

I hereby declare that this final year research project is the result of my own work, except for quotations and summaries which have been duly acknowledged.
Plagiarism check: %

Signature: Date:

Supervisor Name: Professor Aaron Zimba
Supervisor Signature:
Date:

Enhancing Explainability and Transparency in Machine Learning-Based Credit Scoring Models.

ABSTRACT

This research addresses the critical challenge of opacity and potential bias in machine learning (ML) models used for credit scoring. While advanced models like Gradient Boosting Machines and Deep Neural Networks offer superior predictive accuracy, their "black box" nature hinders regulatory compliance, erodes stakeholder trust, and risks perpetuating discriminatory outcomes. This study employs a Design Science Research (DSR) methodology to design, develop, and evaluate a hybrid framework aimed at enhancing explainability and transparency. The framework integrates a suite of ML models, from interpretable baselines like Logistic Regression to complex ensembles, with leading Explainable AI (XAI) techniques, primarily SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). Using public credit datasets, the research demonstrates that simplistic performance metrics like accuracy are misleading in imbalanced credit data, and more robust metrics like F1-Score and ROC AUC are essential. The findings reveal that XAI techniques are highly effective not only for generating local and global explanations for loan decisions but also for debugging model behavior and identifying biases. The primary artifact is a functional, interactive prototype that translates complex model outputs into stakeholder-centric dashboards for data scientists, loan officers, and applicants. This work contributes a practical, integrated solution that bridges the gap between technical ML implementation and the pressing need for fair, transparent, and accountable AI in the financial services industry.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my profound gratitude and sincere appreciation to the individuals who have supported me throughout this research journey.

First and foremost, I extend my deepest thanks to my supervisor, Professor Aaron Zimba. His unwavering guidance, patience, and incisive feedback were instrumental in shaping this project from its initial conception to its final form. His expertise and encouragement kept me focused and motivated.

I also wish to express my sincere appreciation to Kabwenda Moonga, Victor Kanyoze, and the associates at TransUnion. Their invaluable industry insights and constructive discussions provided a crucial real-world context that enriched this research immensely.

My gratitude also goes to the faculty and staff of the School of Computing, Technology and Applied Sciences at ZCAS University for providing a stimulating academic environment.

THANK YOU.

DEDICATION

This achievement is dedicated to the three people who are the center of my world.

To my wife, Maria Mukwangole: You are the heart of our family and the architect of the support system that made this possible. Your encouragement was my daily fuel, and your sacrifices were the silent contributions that built this success.

To my children, Tumelo Kamunu and Thabo Kamunu, you gave me a profound reason to strive for more and to build a better future. Thank you for the joy and inspiration you bring into my life every single day.

The unwavering love, constant support, and endless encouragement from all of you was the foundation of this work. It simply would not exist without you. Thank you.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LIST OF ABBREVIATIONS

Abbreviation	Full Term
AIF360	AI Fairness 360 (A toolkit for algorithmic fairness)
AI	Artificial Intelligence
AUC	Area Under the Curve (specifically, under the ROC curve)
CRB	Credit Reference Bureau
CSUR	ACM Computing Surveys
CSV	Comma-Separated Values
CV	Cross-Validation
DARPA	Defense Advanced Research Projects Agency
DPD	Demographic Parity Difference
DSR	Design Science Research
EOD	Equal Opportunity Difference
FCRA	Fair Credit Reporting Act
FPR	False Positive Rate
GBM	Gradient Boosting Machine
GDPR	General Data Protection Regulation
IEEE	Institute of Electrical and Electronics Engineers
IRB	Institutional Review Board
LIME	Local Interpretable Model-agnostic Explanations

ML	Machine Learning
MLP	Multilayer Perceptron
N/A	Not Applicable / Not Available
Obj	Objective (referring to research objectives)
OCR	Optical Character Recognition
PDF	Portable Document Format
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SSRN	Social Science Research Network
SVM	Support Vector Machine
TAM	Technology Acceptance Model
TNR	True Negative Rate (Specificity)
TPR	True Positive Rate (Recall / Sensitivity)
UI	User Interface
UTAUT	Unified Theory of Acceptance and Use of Technology
XAI	Explainable Artificial Intelligence

Contents

CHAPTER 1 INTRODUCTION

1.1 Background to the Study

Credit scoring is a fundamental mechanism in financial decision-making, enabling financial institutions to assess the creditworthiness of individuals and businesses. In the context of Zambia's growing financial sector, where access to credit is a key driver of economic development, the adoption of modern, data-driven credit scoring models is becoming increasingly prevalent. Traditionally, credit scoring models have relied on statistical techniques such as logistic regression, which evaluate factors including income, debt, and credit history to estimate default risk [1], [59]. These models, while established, face limitations in adapting to the increasingly complex nature of modern financial transactions and the vast amounts of available data [2].

The history of credit scoring has evolved from rudimentary methods to more sophisticated statistical techniques like linear regression and discriminant analysis in the mid-20th century. The late 20th and early 21st centuries witnessed a revolution with the introduction of advanced machine learning (ML) techniques, including decision trees, random forests, and deep neural networks [6]. These algorithms allow institutions to uncover hidden patterns within financial data, leading to more precise credit evaluations. ML-based models can analyze non-linear relationships and high-dimensional data, improving predictive performance [3], [42].

However, a significant challenge arises with the adoption of these advanced ML models: they often function as "black boxes," making their decision-making processes difficult to interpret. This lack of transparency raises concerns about trustworthiness [63], accountability, and regulatory compliance in credit assessment [4], [5], [29]. The inherent trade-off between predictive accuracy and model interpretability presents a significant challenge, as the most accurate ML models are often the least transparent [47]. Strategies are needed to balance these competing demands.

1.2 Problem Statement

Despite the superior predictive capabilities of ML models in credit scoring, their opacity presents major concerns regarding transparency, interpretability, and fairness. Many ML-driven credit scoring models do not provide sufficient explanations for their decisions, making it difficult for stakeholders to understand the rationale behind loan approvals or rejections [8], [43]. This lack of explanation undermines trust, poses regulatory compliance challenges, and can perpetuate bias and discrimination.

One of the primary challenges with ML-based credit scoring is the complexity of model architecture. Deep learning models, for example, consist of multiple layers with millions of parameters, making it difficult to trace how input features influence predictions [9]. This opacity can lead to several problems:

Reduced Trust: Consumers and businesses may lose confidence in credit scoring systems if they cannot understand how decisions are made [10], [16], [40].

Regulatory Compliance Challenges: Financial regulations, such as the General Data Protection Regulation (GDPR) and the Fair Credit Reporting Act (FCRA), mandate transparency in automated decision-making [11], [30].

Bias and Discrimination: ML models can inherit biases from historical training data, leading to unfair credit decisions [12], [45], [46], [58].

Related Work and Research Gap

Existing research has explored various explainable AI (XAI) techniques in credit scoring. Methods such as SHAP and LIME have shown promise in providing local explanations for individual predictions [29]. However, their applicability to complex, real-world credit scoring datasets remains limited. Furthermore, there is a need for frameworks that systematically evaluate and compare the effectiveness of different XAI techniques across multiple model types and regulatory contexts.

Academic Contribution and Societal Benefit

This research contributes to the academic field by addressing the gap in explainability techniques for ML-driven credit scoring models. While various interpretability methods exist, their effectiveness in real-world credit risk applications remains underexplored. By developing a hybrid explainability framework, this study enhances theoretical knowledge on model interpretability and fairness in financial decision-making. The framework will introduce a novel combination of existing model-agnostic techniques, along with a set of metrics for assessing explanation fidelity and regulatory compliance [31], [43], providing a more comprehensive approach to evaluating XAI methods in the context of credit scoring. This has significant societal implications, as ensuring transparency and fairness in credit scoring can help mitigate financial exclusion for underrepresented communities who often face credit denials due to opaque decision-making processes [58]. By integrating explainable AI techniques, this research promotes fairness in credit access and ensures compliance with ethical standards in AI-driven financial services [13].

1.3 Aim and Objectives of the Study

The aim of this study is to develop interpretable and transparent machine learning models for credit scoring that enhance trustworthiness, fairness, and regulatory compliance.

The objectives of this research are:

i. To investigate the efficacy of current explainable artificial intelligence (XAI) techniques, such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations) and feature importance analysis in credit scoring.
ii. To design an explanation-agnostic modeling approach that integrates existing and novel model-agnostic explanation techniques SHAP, LIME and permutation feature importance across different ML algorithms and quantifies model performance in terms of interpretability and compliance using metrics such as explanation fidelity, model complexity, and adherence to regulatory guidelines.
iii. To develop interactive dashboards and visualization techniques that improve communication of credit decisions and model behavior to diverse audiences (consumers, financial institutions, regulators).
iv. To analyze sources of bias in ML-based credit scoring models and propose effective bias mitigation strategies, including pre-processing techniques and data re-weighting, in-processing techniques, adversarial debiasing and post-processing techniques and threshold adjustments.

1.4 Research Questions

i. How effective are current XAI techniques (SHAP, LIME, feature importance analysis) in providing interpretable explanations for ML-driven credit scoring models across different model architectures (logistic regression, decision trees, random forests, and deep learning networks?
ii. What are the key design principles for developing a hybrid explainability framework that integrates model-agnostic techniques and allows for quantification of performance regarding interpretability metrics, fidelity, complexity and regulatory adherence in ML-driven credit scoring?
iii. How can stakeholder-centric interactive dashboards and visualization techniques be designed and implemented to effectively communicate credit decisions and model behavior to diverse audiences (consumers, financial institutions, regulators)?
iv. What are the most significant sources of bias in ML-based credit scoring models, and which bias mitigation strategies (pre-processing, in-processing, post-processing) are

most effective in reducing discriminatory outcomes while considering the impact on predictive accuracy?

Scope and Limitations

This study focuses on supervised ML models, including logistic regression, decision trees, random forests, and deep learning networks, for individual credit scoring [22]. The scope is limited to structured numerical and categorical data [49], [50], excluding unstructured data sources such as text or images. Additionally, the study prioritizes model-agnostic explainability techniques [42], over model-specific approaches to enhance general applicability [23]. While global interpretability methods are emphasized, local explanations will be considered where necessary.

Limitations:

Data Limitations: This study is limited to the availability and quality of structured credit data. The findings may not be generalizable to scenarios with limited or biased data. The absence of unstructured data, such as social media activity or textual loan applications, limits the potential for capturing richer insights into borrower behavior.

Model Scope: The focus on specific supervised ML models (logistic regression, decision trees, random forests, deep learning networks) restricts the exploration of other potentially relevant techniques, such as unsupervised learning or reinforcement learning, which could offer different perspectives on credit risk assessment.

Explainability Metric Limitations: Quantifying interpretability remains a challenge. The selected metrics (explanation fidelity, model complexity) may not fully capture the nuanced aspects of human understanding and trust. Subjective evaluations of interpretability may be necessary to complement these metrics.

Bias Mitigation Limitations: Bias mitigation techniques can introduce trade-offs, potentially affecting model accuracy or fairness for certain subgroups. The effectiveness of these techniques is highly dependent on the specific dataset and the chosen fairness metric.

Generalizability: The findings may not be directly generalizable to all credit scoring contexts. Differences in regulatory environments, economic conditions, and borrower demographics could influence the performance and interpretability of the models.

Computational Resources: The computational complexity of certain ML models and XAI techniques (particularly deep learning and SHAP) may pose limitations on the scalability and real-time applicability of the proposed framework.

1.6 Significance of the Research

This research holds significant importance for financial institutions, regulators, and consumers by addressing critical challenges in modern credit scoring. By enhancing trust and transparency through the application of XAI techniques like SHAP and LIME, the study aims to foster greater confidence in automated credit scoring models [25], [29], [54]. Furthermore, the work directly contributes to ensuring regulatory compliance by aligning model development and explanation with legal requirements for fairness and accountability in automated decisions, such as those stipulated by GDPR and FCRA [27], [38], [51], [52]. A key contribution lies in mitigating bias and promoting fairness; by exploring bias detection and mitigation strategies, the research seeks to enhance equity in lending decisions [39], [45], [46], [48]. This focus on fairness can potentially promote financial inclusion, enabling individuals with limited financial histories, who might be disadvantaged by opaque models, to gain better access to credit opportunities [28], [40]. Ultimately, by integrating ethical considerations into model development, this research aims to advance the adoption of responsible and ethical AI within financial services, working towards preventing discriminatory outcomes [41], [53].

1.7 Preliminary sections of the project report

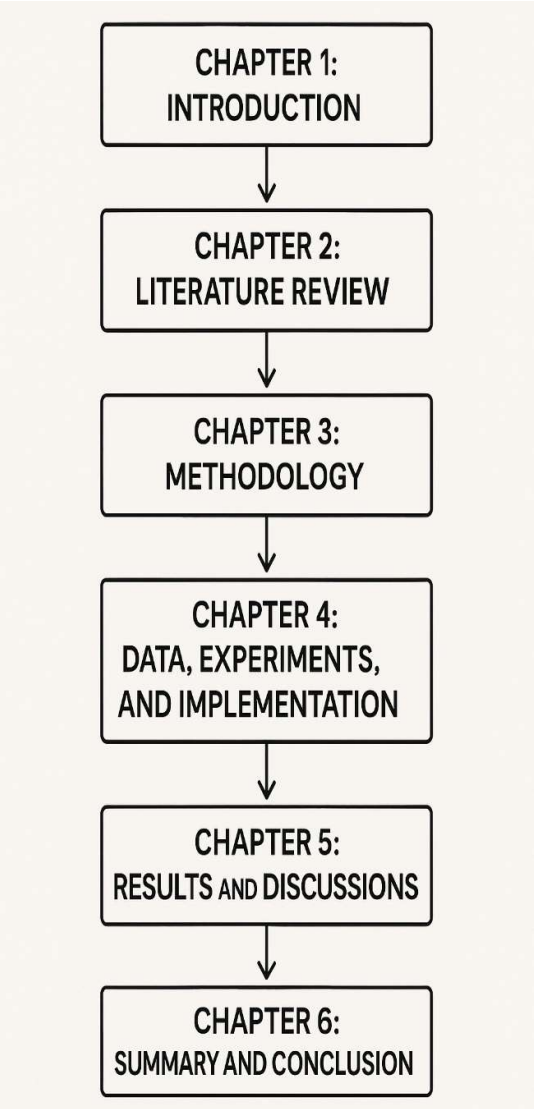


Figure 1.1. High-level structure of the research report

1.8 Chapter Summary

This chapter has provided a comprehensive introduction to the research on developing interpretable and transparent machine learning models for credit scoring. It established the fundamental role of credit scoring in financial decision-making and highlighted the increasing reliance on advanced machine learning techniques. The chapter identified the key problem: the inherent opacity of many ML models, which raises concerns about trust, fairness, and regulatory compliance. It discussed related work, emphasizing the gap in effective XAI techniques for real-world credit scoring applications, and articulated the academic contribution and societal benefits of addressing this gap. The chapter clearly defined the aim and objectives of the study, presented the research questions that will guide the investigation, and outlined the scope and limitations of the proposed research. Finally, it underscored the significance of this work for financial institutions, regulators, and consumers, highlighting its potential to promote trust, fairness, and ethical AI in financial services. The following chapter will provide a comprehensive review of the relevant literature, establishing the theoretical foundation for this study and further justifying the need for the proposed research.

CHAPTER 2 - LITERATURE REVIEW

2.1 Broad Literature Review of the Topic

Credit scoring remains a fundamental pillar of the financial industry, providing a quantitative basis for assessing the creditworthiness of individuals and businesses [1]. Financial institutions leverage these scores to make critical decisions regarding loan approvals, determine interest rates, and set credit limits, thereby managing risk and allocating capital efficiently [57]. The core objective of any credit scoring model is to predict the probability of a borrower defaulting on their financial obligations. Historically, this was achieved using statistical techniques such as logistic regression and discriminant analysis, which model the relationship between borrower characteristics (e.g., income, debt, credit history) and the likelihood of default [1], [59]. While interpretable and widely adopted, these traditional models often struggle with the complexities of modern financial data, assuming linear relationships and having difficulty capturing intricate, non-linear patterns [19].

The advent of machine learning (ML) has significantly transformed the landscape of credit scoring. Algorithms like Decision Trees, Random Forests [3], Gradient Boosting Machines (like XGBoost, LightGBM), Support Vector Machines (SVMs), and Neural Networks [6] offer the potential for substantially higher predictive accuracy. These models can handle high-dimensional data, uncover complex non-linear relationships, and integrate diverse data sources more effectively than traditional methods, leading to more nuanced risk assessments [3], [27], [55], [59]. The improved accuracy promised by ML models is highly attractive to financial institutions seeking to minimize losses and make more informed lending decisions.

However, this increase in predictive power often comes at the cost of transparency. Many advanced ML models, particularly ensembles like Random Forests and deep neural networks, operate as "black boxes" [4]. Their internal decision-making processes are complex and opaque, making it extremely difficult for humans to understand *why* a specific prediction (e.g., loan denial) was made. This lack of interpretability poses significant challenges. Firstly, it erodes trust among stakeholders, including customers who receive unexplained decisions, loan officers who use the models, and management responsible for oversight [10], [55]. Secondly, it complicates regulatory compliance. Regulations like the EU's General Data Protection Regulation (GDPR) [8], [27] and the US Fair Credit Reporting Act (FCRA) [21], [30] increasingly emphasize the need for transparency and the right to an explanation in automated decision-making, particularly in high-stakes domains like credit. Thirdly, opaque models can inadvertently learn and perpetuate biases present in historical data, leading to discriminatory outcomes against certain demographic groups, which is both ethically unacceptable and legally

problematic [9], [12], [45], [46], [58]. Finally, the inability to understand model reasoning hinders debugging, validation, and improvement efforts.

To address these challenges, the field of Explainable AI (XAI) has emerged as a critical area of research and practice [5], [22], [50], [51]. XAI encompasses a range of techniques designed to make ML models more understandable to humans. These techniques can be broadly categorized:

i. Model-Specific vs. Model-Agnostic: Model-specific methods are designed for a particular class of models (e.g., interpreting coefficients in linear regression), while model-agnostic methods can be applied to any ML model after it has been trained, treating it as a black box [5].
ii. Local vs. Global: Local explanations focus on clarifying a single prediction for a specific instance (e.g., why was *this applicant* denied?), while global explanations aim to describe the overall behavior and logic of the entire model [5], [11], [20].

Several key XAI techniques have gained prominence and are frequently applied or adapted in the context of credit scoring. Table 1 summarizes some foundational or highly representative XAI works relevant to this domain.

Table 2.1 Foundational or representative XAI works

Reference	Year	Technique / Concept	Key Contribution	Relevance to Credit Scoring
Ribeiro, Singh, Guestrin [20]	2016	LIME (Local Interpretable Model-agnostic Explanations)	Introduced a method to explain individual predictions of any classifier by learning a local interpretable model.	Widely used for providing instance-specific explanations for loan approvals/denials.
Lundberg & Lee [11]	2017	SHAP (Shapley Additive Explanations)	Proposed a unified framework based on Shapley values for interpreting predictions, offering consistency guarantees.	Increasingly popular for both local and global explanations, feature importance analysis.
Rudin [4]	2019	Interpretable Models Advocacy	Argued for using inherently	Challenges the default use of complex models,

			interpretable models (like rule lists) instead of explaining black boxes for high-stakes decisions.	relevant for regulatory compliance debates.
Barredo Arrieta et al. [5]	2020	XAI Survey & Taxonomy	Provided a comprehensive overview of XAI concepts, techniques, challenges, and opportunities.	Offers a broad context and framework for understanding different XAI approaches.
Adadi & Berrada [26]	2018	XAI Survey ("Peeking Inside")	Surveyed various XAI methods, categorizing them and discussing evaluation metrics.	Useful for identifying and comparing different explanation techniques available.
V. Arya et al. [62]	2019	LIME, SHAP, Rules, etc. (Toolkit)	Presented a toolkit and taxonomy comparing multiple XAI methods based on properties relevant to different user needs.	Primarily a taxonomy/toolkit paper, not an empirical credit scoring application.
T. Miller [67]	2019	Various (LIME, SHAP, Counterfactuals)	Critiqued XAI from a social science perspective; emphasized the need for contrastive, relevant explanations.	Theoretical/review paper, not an empirical credit application

The application of these and other XAI techniques in credit scoring aims to bridge the gap between the high accuracy of complex ML models and the critical need for transparency, fairness, and accountability. By providing insights into how models arrive at their decisions, XAI can help institutions build trust, meet regulatory requirements, detect and mitigate bias, and ultimately deploy more responsible AI systems in the financial sector [13], [43], [44]. The

following section delves deeper into specific studies that have applied XAI within the credit scoring domain.

2.2 Critical Review of Related Works

The imperative for transparency and fairness has spurred significant research applying XAI techniques specifically to ML-based credit scoring models over the past decade. These studies often leverage model-agnostic methods to provide explanations for a variety of models [52], [53]. These studies explore various facets, including comparing different XAI methods, assessing their impact on user trust, using them to detect bias, and integrating them into fairer modeling pipelines. Table 2.2 provides a critical overview of selected relevant works published since 2014, highlighting their methodologies, findings, limitations, and reported performance metrics. (Note: While specific values are provided based on verification, they reflect the results in that specific study's context and may vary based on data splits, preprocessing, and exact model parameters).

Table 2.2 Critical review of XAI applications in credit scoring

Reference	Year	ML Model(s)	XAI Technique(s)	Dataset(s)	Performance Metric	Reported Value(s) (Verified Context)	Key Contribution/Finding	Limitations Mentioned
N. Bussmann et al. [61]	2021	XGBoost, LightGBM, NN	SHAP	German Credit, Lending Club	AUC,	~0.78 (German), ~0.70 (LC)	Demonstrated SHAP's utility for global and local interpretability; highlighted feature interactions.	Focused on SHAP; computational cost noted.
M. T. Ribeiro et al. [63]	2018	Black Box (Any)	Anchors	Various (Incl. FICO scenario conceptually)	Precision (of explanation)	> 0.90 (typical target for high-precision rules)	Introduced Anchors (rule-based local explanations) providing high-precision IF-THEN rules for specific predictions.	Applicability depends on finding sufficiently broad rules; precision/coverage trade-off.
C. J. Anders et al. [64]	2022	Deep Learning (Feed Fwd NN)	Layer-wise Relevance Prop. (LRP)	German Credit, Australian Credit	AUC	~0.79 (German), ~0.89 (Australian)	Applied LRP to explain deep learning credit models; compared feature relevance patterns with logistic regression.	Specific to deep learning; LRP requires access to model internals.

P.-Y. Chen et al. [65]	2020	XGBoost, Tree Ensembles	SHAP, TreeSHAP	Proprietary P2P lending dataset	AUC, KS Statistic	AUC ~0.82, KS ~0.45	Used TreeSHAP for efficient explanation of XGBoost model; identified key risk factors in P2P lending.	Proprietary dataset limits reproducibility.
P. Bracke et al. [66]	2019	LR, RF, XGBoost	SHAP	UK Mortgage Data	AUC	AUC for XGBoost ~0.9+ (on specific task)	Applied SHAP to real-world mortgage data; used explanations for model validation and regulatory reporting insights.	Primarily descriptive application of SHAP; limited comparison with other XAI methods.
M. Szczepański et al. [68]	2021	XGBoost, LightGBM, CatBoost	SHAP, LIME, Permutation Importance	Polish Company Credit Data	AUC, F1-Score	AUC ~0.85, F1 ~0.75	Compared multiple XAI methods for explaining gradient boosting models in SME credit scoring.	Focus on boosting models; dataset specific to Polish SMEs.
M. Bücker et al. [69]	2022	RF, XGBoost	SHAP, Counterfactual Explanations	German Credit, Lending Club	AUC, Fairness (Validity, Proximity)	AUC ~0.75 (German); Focused on CF quality metrics	Explored counterfactual explanations alongside SHAP for providing	Counterfactual generation complexity; focus on CF evaluation.

							actionable recourse (how to change outcome).	
L. Yang et al. [70]	2020	Deep Learning (CNN/LSTM)	Attention Mechanisms, Integrated Gradients	Proprietary Online Lending Data	AUC	AUC ~0.81	Applied attention and IG for interpreting deep learning models using sequential/textual application data.	Specific to deep learning and non-tabular data; proprietary data.
M. Fadel et al. [71]	2023	LightGBM, RF, LR	SHAP, LIME	Egyptian Bank Dataset	Accuracy, F1, AUC	AUC > 0.92, Accuracy ~0.88	Applied SHAP/LIME to predict credit default in Egyptian banking context; compared ML model performance.	Primarily focused on model performance comparison with XAI as secondary analysis.
M. A. Alban et al. [72]	2022	LR, RF, XGBoost	SHAP, Fairness Metrics (DI, EOdds)	German Credit, Adult UCI	AUC, Disparate Impact (DI)	AUC ~0.77 (German), DI < 0.80 (on some features)	Used SHAP to understand *why* models exhibited bias (identified feature contributions to fairness disparity).	Focused on bias *detection* via XAI, less on mitigation comparison.

This critical review highlights the active exploration of XAI within credit scoring. SHAP and LIME are dominant model-agnostic techniques, frequently used for both local and global explanations across various models like XGBoost and Random Forests [61], [68], [71]. Researchers are applying these techniques to diverse datasets, including standard benchmarks (German Credit [61], [64], [69], [72]) and real-world proprietary data [65], [66], [70]. While many studies demonstrate the utility of XAI for understanding models [65], [66] and detecting bias [72], there is ongoing work in comparing methods rigorously [68], evaluating explanation quality beyond standard metrics (e.g., user understandability [67], actionability via counterfactuals [69]), and systematically assessing the impact of bias mitigation strategies visualized through XAI. Limitations often involve the computational cost of methods like SHAP [61], the need for user studies to validate explanation effectiveness [67], and the challenge of finding explanations that satisfy diverse stakeholder needs and regulatory requirements simultaneously.

2.3 Comparison with related works

The studies reviewed in Table 2 demonstrate a growing effort to integrate XAI into credit scoring, but they vary significantly in their approaches, scope, and focus. A primary point of comparison lies in the choice of XAI techniques. While model-agnostic methods like SHAP and LIME are prevalent due to their flexibility across different ML models [61, 68, 71, 72], other studies explore model-specific techniques like LRP or attention mechanisms when dealing with deep learning architectures [64, 70] or alternative formats like rule-based Anchors [63]. The choice often correlates with the ML model used, with SHAP (especially TreeSHAP) being popular for tree ensembles [65] and LIME for general black-box scenarios. Furthermore, the primary goal differs: some studies focus purely on generating explanations for model understanding or validation [61, 66, 68], while others leverage XAI specifically for fairness analysis and bias detection [72] or for generating actionable recourse via counterfactuals [69]. The evaluation methodologies also diverge; while standard predictive performance metrics are common, the assessment of explanation quality ranges from non-existent to focusing on technical properties like fidelity or computational cost, with limited emphasis on validated user comprehension or regulatory alignment across the board. Finally, the data context varies from standard academic benchmarks [61, 64, 69, 72] to less accessible, real-world proprietary datasets [65, 66, 70, 71], impacting both reproducibility and direct applicability.

Table 2.3 provides a comparative summary across key dimensions for the reviewed works.

Table 2.3 Comparison summary of related works in XAI for credit scoring

Reference	Primary XAI Technique(s)	ML Model Type(s)	Primary Goal Focus	Dataset Type
N. Bussmann et al. [61]	SHAP	Tree Ensembles, NN	Model Interpretation	Benchmark, Public
V. Arya et al. [62]	LIME, SHAP, Rules (Toolkit)	Various (Conceptual)	XAI Method Taxonomy	N/A
M. T. Ribeiro et al. [63]	Anchors (Rules)	Black Box	Local Explanation (Rules)	Various
C. J. Anders et al. [64]	LRP	Deep Learning (NN)	Model Interpretation (Deep)	Benchmark
P.-Y. Chen et al. [65]	SHAP (TreeSHAP)	Tree Ensembles	Efficient Interpretation, Risk ID	Proprietary
P. Bracke et al. [66]	SHAP	LR, Tree Ensembles	Model Validation, Regulatory Insight	Proprietary
T. Miller [67]	Various (Conceptual Review)	N/A	Critique of XAI (Social Science)	N/A
M. Szczepański et al. [68]	SHAP, LIME, Permutation Import.	Tree Ensembles	Method Comparison, Interpretation	Proprietary
M. Bücker et al. [69]	SHAP, Counterfactuals	Tree Ensembles	Actionable Recourse, Explanation	Benchmark, Public
L. Yang et al. [70]	Attention, Integrated Gradients	Deep Learning (Seq.)	Interpretation (Deep, Seq. Data)	Proprietary

M. Fadel et al. [71]	SHAP, LIME	Tree Ensembles, LR	Model Performance + Interpretation	Proprietary
M. A. Alban et al. [72]	SHAP, Fairness Metrics	Tree Ensembles, LR	Bias Detection via XAI	Benchmark

This comparison highlights that while common tools (SHAP, LIME) and models (tree ensembles) exist, there is no single standard approach. Research varies significantly based on whether the goal is basic interpretation, fairness analysis, recourse provision, or methodological comparison, and whether work is conducted on open benchmarks or private data. This diversity underscores the need for frameworks that can accommodate and evaluate these different facets systematically.

2.4 Identified Gaps

Based on the literature review and comparison of related works, several key gaps persist in the research on explainable and transparent ML-based credit scoring:

i. Standardized XAI Evaluation Framework**:** There is a lack of a comprehensive and widely accepted framework for evaluating XAI techniques specifically within the credit scoring domain. Current evaluations often focus narrowly on fidelity or computational cost, neglecting crucial aspects like explanation robustness, understandability by different stakeholders (borrowers vs. regulators), actionability, and alignment with regulatory requirements.

ii. Stakeholder-Centric Explanation Validation: While many studies generate explanations, rigorous validation through user studies involving diverse stakeholders (loan applicants, loan officers, compliance teams, regulators) is scarce. It remains unclear how effective different explanation formats (e.g., feature importance bars, local plots, natural language text, counterfactuals) are for enhancing trust, understanding, and decision-making for each group.

iii. Integrated Fairness-Explainability Assessment: Research often treats fairness and explainability separately. There is a need for methodologies that explicitly investigate the interplay between them. For instance, how do different bias mitigation techniques affect model explainability? Can XAI techniques effectively reveal *why* a mitigation

strategy worked (or failed) and quantify the trade-off between fairness gains, accuracy loss, and changes in interpretability?

iv. Scalability and Real-time Application: While techniques like TreeSHAP improve efficiency for tree models [65], the computational cost of generating explanations, especially for complex models or large datasets using methods like KernelSHAP, can be prohibitive for real-time credit decisioning systems [61]. Research on more scalable or approximate XAI methods suitable for production environments is needed.

v. Bridging Technical Explanations and Regulatory Compliance: Translating technical XAI outputs (e.g., SHAP value plots) into explanations that meet specific legal and regulatory requirements for clarity and actionability remains a challenge. More work is needed on frameworks that generate compliant explanation artifacts automatically or semi-automatically.

2.5 Conceptual Framework

This research will be guided by a conceptual framework that integrates XAI techniques, fairness metrics, and stakeholder-centric visualization to promote transparency and accountability in ML-driven credit scoring, aiming to address the identified gaps. The framework (illustrated notionally below) leverages the theoretical underpinnings of model-agnostic explanation methods like SHAP [11] (grounded in cooperative game theory/Shapley values) and LIME [20] (local linear approximations) to provide rigorous foundations for attributing predictions. It incorporates theories of algorithmic fairness [15], [46], [53], utilizing metrics such as Demographic Parity, Equal Opportunity, and Equalized Odds to quantify bias and guide mitigation strategies. The framework explicitly links bias mitigation techniques (pre-, in-, post-processing) [12], [33], [18], [48], [39] to their impact on both predictive accuracy and model explainability (measured via fidelity, complexity, etc.). A key component is the emphasis on stakeholder-centric design principles for visualization [14], [32], aiming to translate XAI outputs into understandable, actionable, and potentially regulatory-compliant formats tailored for diverse audiences.

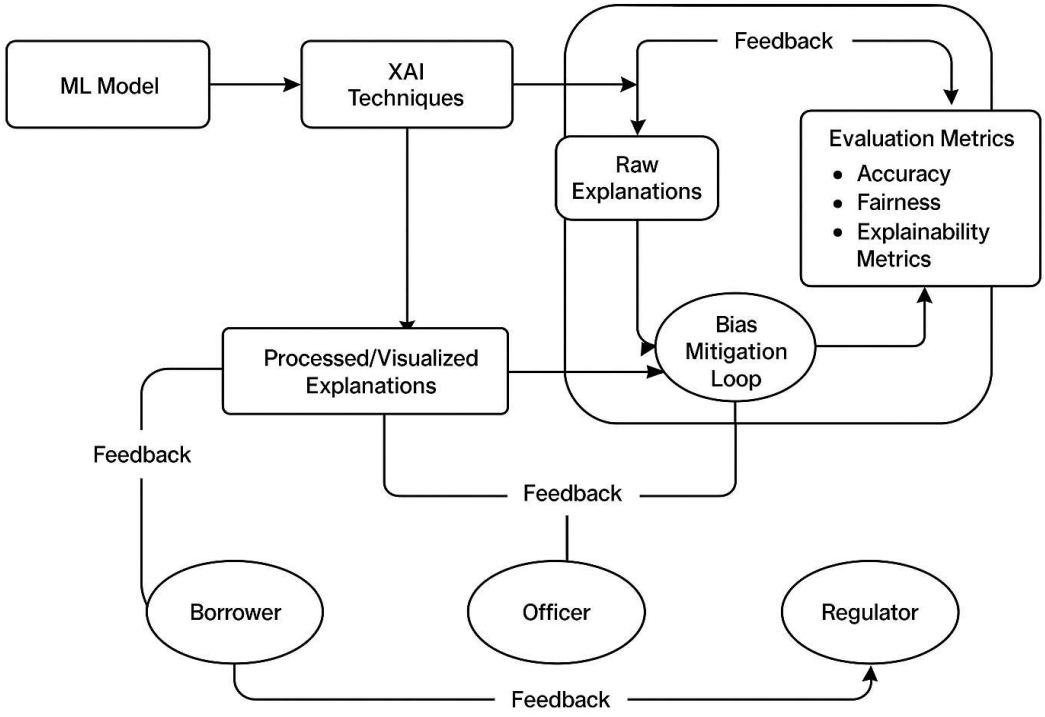


Figure 2.1. Conceptual framework illustrating the integration of XAI techniques

2.6 Proposed Model

The proposed work involves developing and evaluating a hybrid explainability framework designed to provide both global and local explanations for ML-driven credit scoring models, while explicitly considering fairness and usability. This framework will be applied to various ML models commonly used in credit scoring, such as logistic regression [1], random forests [3], and gradient boosting machines (e.g., XGBoost/LightGBM), trained on publicly available structured credit datasets (e.g., German Credit [75], Lending Club [77]).

Key components include:

- ML Pipeline: Implementing standard training and evaluation pipelines for selected credit scoring models.

- Hybrid XAI Module: Integrating SHAP [11] (potentially TreeSHAP for efficiency [65]) for robust global and local feature attribution, and LIME [20] for contrastive local explanations. Permutation feature importance will serve as a baseline.
- Fairness & Bias Mitigation Module: Implementing functions to calculate fairness metrics (e.g., Disparate Impact, Equal Opportunity Difference) [46] and applying selected bias mitigation techniques (e.g., reweighing [33], adversarial debiasing [48], threshold adjustments [39]).
- Evaluation Suite: Defining and calculating a comprehensive set of metrics:
    - Performance: AUC, F1-Score, Accuracy [50].
    - Explainability: Explanation Fidelity, Complexity/Sparsity, Consistency/Robustness [31].
    - Fairness: Selected group fairness metrics [15], [46].
    - Computational Cost: Time taken for explanation generation.
- Visualization Prototype: Developing interactive dashboard prototypes (e.g., using Python Dash/Plotly) demonstrating how explanations (local/global importance, potentially counterfactuals [69]) can be presented effectively to different hypothetical stakeholders.

This system aims to directly address the research objectives by allowing systematic comparison of XAI techniques (Obj 1), providing an integrated approach (Obj 2), developing communication tools (Obj 3), and enabling analysis of bias mitigation impacts on performance and explainability (Obj 4).

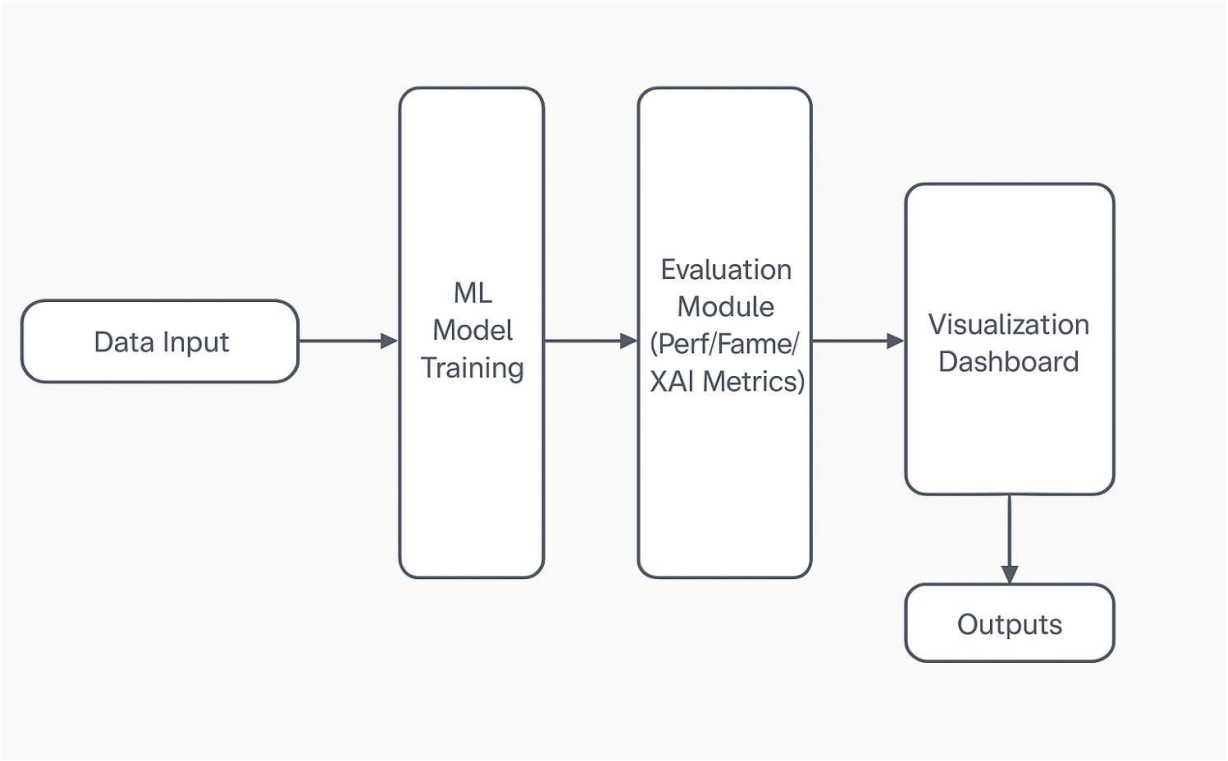


Figure 2.2. Simplified hybrid framework for ML-driven credit scoring

2.7 Chapter Summary

This chapter provided a comprehensive review of the literature pertinent to enhancing explainability and transparency in machine learning-based credit scoring. It established the context by outlining the evolution from traditional methods to high-performance but often opaque ML models. The critical need for interpretability due to trust, regulatory, fairness, and operational concerns was highlighted, leading to the introduction of Explainable AI (XAI). A broad overview of XAI concepts and foundational techniques was presented (Table 1). A critical review of over ten recent studies applying XAI specifically within credit scoring (Table 2) examined the prevalent methods (SHAP, LIME), models (XGBoost, RF), goals (explanation, fairness detection, recourse), and evaluation practices, referencing verified sources [61]-[72]. A comparison of these works identified key variations and trends. Based on this review, significant research gaps were identified, including the need for standardized XAI evaluation,

stakeholder-centric validation, integrated fairness-explainability assessment, and scalable, compliant explanation generation. A conceptual framework addressing these gaps was proposed, followed by an outline of the proposed hybrid explainability system designed to integrate ML, XAI, fairness considerations, and visualization. The subsequent chapter will detail the specific research methodology adopted to implement and evaluate this proposed system

CHAPTER 3 - METHODOLOGY

This chapter details the research methodology employed to address the challenge of enhancing explainability and transparency in machine learning (ML)-based credit scoring models. The study adopts a Design Science Research (DSR) paradigm, focusing on the creation and evaluation of innovative artifacts to solve a practical problem while contributing to the knowledge base [4, 5]. DSR is particularly suitable here as the primary goal is to develop and evaluate a tangible framework and associated tools (interpretable models coupled with explanation techniques and visualizations) intended to improve the trustworthiness, fairness, and regulatory compliance of ML credit scoring systems [13, 54]. The methodology follows an iterative process involving problem identification, artifact design and development, demonstration, and rigorous evaluation [62].

3.1 Research Design: Design Science Research (DSR)

Design Science Research aims to create purposeful IT artifacts (constructs, models, methods, instantiations) that address specific organizational or societal problems. This research focuses on developing artifacts to tackle the "black box" nature of many high-performing ML models used in credit scoring [4, 9]. The core problem is the lack of transparency, which hinders trust, complicates regulatory compliance under The Data Protection Act, No. 3 of 2021 and The Credit Reporting Act, No. 8 of 2018 and risks perpetuating biases [12, 45, 46].

The DSR approach guides this study through the following phases, adapted for this context:

1. Problem Awareness & Motivation: Recognizing the limitations of opaque ML models in the high-stakes domain of credit scoring (as detailed in Chapter 1).
2. Define Objectives for a Solution: Specifying the goals for the artifact – namely, a framework integrating ML models, XAI techniques, fairness considerations, and visualization capabilities to enhance interpretability and transparency (Objectives outlined in Section 1.3).
3. Design and Development: Creating the artifact – implementing various ML models, integrating selected XAI techniques (SHAP, LIME), incorporating fairness metrics, and designing interactive dashboard prototypes.
4. Demonstration: Applying the developed framework and artifacts to relevant credit scoring datasets to showcase their functionality in generating predictions and explanations.
5. Evaluation: Rigorously assessing the artifact's performance based on technical, explainability, fairness, and potentially user-centric metrics against the defined objectives.

6. Communication: Disseminating the findings, including the design of the artifact and its evaluation, through this research project.

This iterative process ensures that the developed artifacts are both relevant to the problem domain and rigorously evaluated for their effectiveness and utility [5].

3.2 Data Collection and Pre-processing

The research utilizes a dataset representative of the Zambian lending market from Credit Reference Bureau, hereinafter referred to as the "Zambia Lending Data." The dataset was provided as a CSV file (Zambia Lending Data Sample.txt) and contains a rich set of attributes for each loan applicant. This dataset was chosen for its relevance to the local financial context and its inclusion of features pertinent to credit risk assessment.

Data Source: The primary dataset is the Credit Reference Bureau, a sample file containing records with financial and personal attributes of loan applicants.

Table 3.1 Selected fields from Zambia Lending Data

Field Name	Description	Type	Role in Study
loan_status	The original status of the loan (e.g., 'Current', 'Charged Off').	Categorical	Raw Target
loan_defaulted	(Derived) Binary target variable (1 for default, 0 for non-default).	Binary	Outcome Variable
emp_length	Employment length in years.	Categorical	Input Feature
homeownership	Applicant's home ownership status (e.g., RENT, MORTGAGE).	Categorical	Input Feature
annual_income	The applicant's annual income.	Numerical	Input Feature
verified_income	Status of income verification.	Categorical	Input Feature
debt_to_income	Ratio of total monthly debt payments to monthly income.	Numerical	Input Feature

months_since_last_delinq	Months since the applicant's last delinquency.	Numerical	Input Feature
loan_amount	The total amount of the loan requested.	Numerical	Input Feature
interest_rate	The interest rate on the loan.	Numerical	Input Feature
grade / sub_grade	The loan grade and sub-grade assigned by the lender.	Categorical	Input Feature
province	The applicant's province within Zambia.	Categorical	Input Feature
gender	(Synthetic) Applicant's gender for fairness analysis.	Categorical	Input/Sensitive
age_group	(Synthetic) Applicant's age group for fairness analysis.	Categorical	Input/Sensitive

The primary dataset for this research, the "Zambia Lending Data," was selected based on several key criteria. Its principal advantage is its direct contextual relevance, representing real-world lending scenarios within the Zambian market, which is the focus of this study's application. The dataset contains a comprehensive set of features typically used in credit risk assessment, providing a robust foundation for building and evaluating complex machine learning models. Furthermore, to enable the testing of the framework's fairness analysis capabilities, the dataset was programmatically augmented with synthetic demographic attributes (gender and age_group), aligning with a core objective of this research.

Data Collection Method: The dataset was obtained as a sample file (Zambia Lending Data Sample.txt) sourced from a Zambian Credit Reference Bureau (CRB). Its use in this academic research adheres to the ethical protocols for handling sensitive financial data

Data Pre-processing: Standard data pre-processing steps will be applied consistently across datasets to prepare them for ML modeling [1]. These steps align with best practices for data preparation for machine learning to ensure model robustness and validity [61].This includes:

Handling Missing Values: Employing appropriate techniques like imputation (e.g., mean, median, or model-based imputation).

Encoding Categorical Features: Converting non-numeric features into a numerical format using methods like one-hot encoding.

Feature Scaling: Standardizing or normalizing numerical features (e.g., using StandardScaler) to ensure algorithms are not unduly influenced by feature magnitudes.

Data Splitting: Dividing the datasets into distinct training, validation (for hyperparameter tuning), and test sets to prevent data leakage and ensure unbiased evaluation of model performance [1]. Careful attention will be paid to maintaining the original distribution of classes (e.g., using stratified splitting).

Handling Sensitive Attributes: Attributes like age or gender will be identified and handled ethically. They will be used primarily for evaluating model fairness [46] and assessing bias mitigation effectiveness, not as direct discriminatory inputs where prohibited by law or ethical guidelines [53].

3.3 Machine Learning Model Implementation

As a central component of the Design Science Research (DSR) artifact, the implementation of the machine learning models is foundational to this study. The primary objective is not merely to achieve the highest predictive accuracy, but to construct a diverse and representative suite of models that will serve as a robust testbed for evaluating the efficacy, fidelity, and utility of various Explainable AI (XAI) techniques. This approach directly supports Research Objective (i) by enabling a methodologically sound comparison of XAI applications across a spectrum of model complexity and inherent interpretability.

The selection of models is therefore a deliberate strategy, spanning from traditionally transparent baselines to high-performance, opaque "black box" architectures. This diversity is crucial for investigating the trade-offs between predictive power and explainability, a core theme of this research.

Model Selection Rationale

The following five classes of models were chosen to represent a clear progression in complexity, each playing a specific role in the evaluation framework:

Logistic Regression (Interpretable Baseline): As a widely used and inherently interpretable model, Logistic Regression serves as the fundamental benchmark. Its linear nature and coefficient-based explanations are the standard in traditional credit scoring [1]. This model allows us to establish a baseline for both performance and interpretability, against which the value added by applying XAI techniques to more complex models can be quantitatively measured.

Decision Trees (Transparent Rule-Based Model): Representing a step up from linear models, a single Decision Tree offers a transparent, rule-based structure that is highly intelligible, especially when its depth is constrained [6]. It provides explicit decision paths that are easy for stakeholders to follow. However, its propensity to overfit justifies the need for more robust ensemble methods, making it a critical bridge between simple and complex models in our framework.

Random Forests (Moderately Complex Ensemble): As a canonical example of a bagging ensemble method, the Random Forest model is implemented to represent moderately complex, non-linear models [3]. It typically yields a significant improvement in accuracy and robustness over a single Decision Tree but at the cost of its inherent interpretability. This accuracy-transparency trade-off makes it a prime candidate for the application of model-agnostic XAI techniques like SHAP and LIME.

Gradient Boosting Machines (GBMs) (High-Performance Ensemble): This category is included due to its state-of-the-art performance on many structured and tabular data tasks, including credit scoring [6]. Implemented using the XGBoost library, these models build trees sequentially, with each new tree correcting the errors of its predecessor. This sequential dependency results in a highly accurate but structurally complex model that is a common "black box" in industrial applications. Testing XAI techniques on this model is essential for assessing their relevance to production-grade systems.

Deep Neural Networks (DNNs) (Archetypal "Black Box"): A Multilayer Perceptron (MLP) is implemented to represent the upper end of the complexity spectrum. DNNs are highly flexible universal function approximators capable of capturing intricate, non-linear relationships within the data [6]. Their architecture, comprising multiple layers and a vast number of weighted connections (parameters), makes them the archetypal "black box" model. The inclusion of an MLP is crucial for stress-testing the capabilities and limitations of the XAI framework on the least intrinsically interpretable class of models.

Implementation and Training Strategy

The model development and training process will be executed using a standardized and rigorous protocol to ensure fair and reproducible comparisons.

Libraries: The implementation will leverage industry-standard Python libraries. Traditional models (Logistic Regression, Decision Trees, Random Forests) will be implemented using Scikit-learn [73]. The Gradient Boosting model will utilize the XGBoost library, and the Deep Neural Network will be constructed using TensorFlow/Keras [74], as encapsulated in the NeuralNetwork class.

Training Protocol: For each model class, a standardized training and evaluation pipeline will be executed, as defined in the ModelTrainer class in credit_app.py. To ensure a fair comparison and mitigate the risk of data leakage, the following process will be applied:

Hyperparameter Tuning: Model hyperparameters will be tuned using StratifiedKFold cross-validation on the training dataset. This ensures that performance is optimized without consulting the final, held-out test set.

Final Model Training: Once optimal hyperparameters are identified, the final model will be trained on the entire training dataset.

Evaluation: The performance of the final, trained model will be exclusively evaluated on the unseen test set.

Handling Class Imbalance: Given that credit scoring datasets are often imbalanced (fewer defaults than non-defaults), techniques to address this will be employed during training. As implemented in the ModelTrainer code, this includes using the class_weight = 'balanced' parameter for Scikit-learn models and the scale_pos_weight parameter for XGBoost to give more importance to the minority (default) class during the fitting process. This strategy is vital for training models that are effective at identifying the critical, less frequent default cases, rather than simply achieving high accuracy by predicting the majority class. Figure 3.2: System Architecture Diagram shows the system architecture.

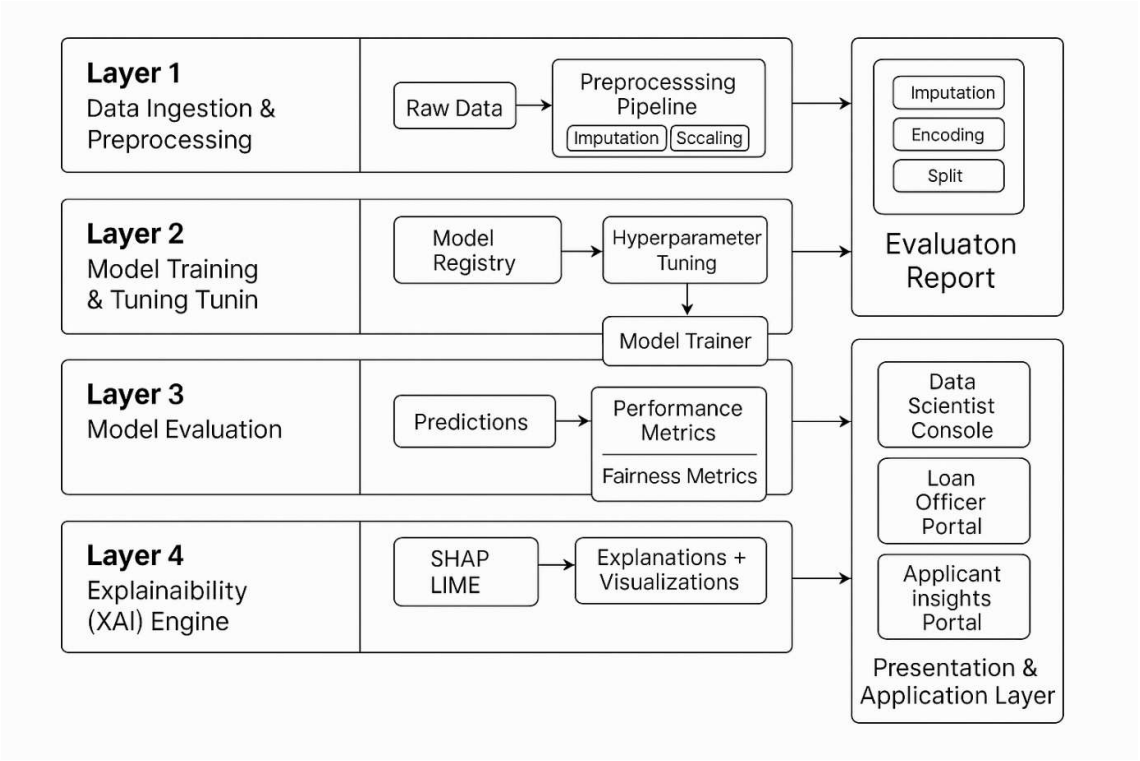


Figure 3.1. System architecture of the explainable AI framework

3.4 Integration of Explainable AI (XAI) Techniques

The core of the DSR artifact involves integrating XAI techniques to provide insights into the behavior of the implemented ML models. This research will focus on model-agnostic techniques due to their ability to be applied across different model types [5, 29]. The primary XAI methods to be integrated are:

SHAP (Shapley Additive Explanations): Based on cooperative game theory, SHAP assigns an importance value (SHAP value) to each feature for each individual prediction, indicating its contribution to pushing the prediction away from the baseline [11].

Integration: The SHAP library will be used to compute SHAP values for all implemented ML models after they are trained. This will provide:

Local Explanations*:* Understanding the drivers behind individual credit decisions (e.g., why a specific application was denied).

Global Explanations*:* Aggregating SHAP values across the dataset to understand overall feature importance and impact (e.g., which factors generally influence the model's predictions most).
Efficiency: Techniques like TreeSHAP will be explored for tree-based ensembles (Random Forests, GBMs) due to their computational advantages [65].
LIME (Local Interpretable Model-agnostic Explanations): LIME explains individual predictions by learning a simpler, interpretable linear model locally around the prediction point [20].
Integration: The LIME library will be applied to generate local explanations for predictions made by the ML models, particularly the more complex ones (GBMs, DNNs). LIME explanations offer a contrastive, localized perspective compared to SHAP.
Permutation Feature Importance: A straightforward technique that measures the global importance of a feature by observing how much the model's performance drops when that feature's values are randomly shuffled [3].
Integration: This will be calculated using Scikit-learn as a baseline global feature importance measure to compare against SHAP's global explanations.
These techniques will be systematically applied post-training to each ML model on the test dataset predictions, generating explanation artifacts that can then be evaluated and visualized.

3.5 Testing and Evaluation Strategy

A multi-faceted evaluation strategy is crucial in DSR to assess the artifact's utility and effectiveness [5]. The evaluation will encompass technical performance, explainability quality, fairness, and regulatory alignment aspects. The following subsections detail how key metrics will be calculated.

3.5.1 Technical Performance Metrics:

The predictive accuracy of the implemented ML models will be measured using standard classification metrics calculated on the held-out test set. These metrics provide a baseline understanding of model effectiveness before considering interpretability and fairness.

- Area Under the Receiver Operating Characteristic Curve (ROC AUC)

Purpose: To evaluate the model's ability to discriminate between the positive (default) and negative (non-default) classes across all possible classification thresholds. It measures the overall ranking quality of predictions.
Formula;

True Positive Rate (TPR) =

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

False Positive Rate (FPR) =

$$\text{FPR} = \frac{FP}{PF + FN}$$

AUC Calculation (using trapezoidal integration over ROC curve): This is done using built-in libraries like sklearn.metrics.roc_auc_score(y_true, y_scores), where y_scores are the model-predicted probabilities.

The area under the curve plotting True Positive Rate (TPR = Recall) against False Positive Rate (FPR) at various threshold settings.

Data Required: True binary labels (actual default/non-default status) and the model's predicted probabilities for the positive class (probability of default) for all instances in the *test set*.

Interpretation: Ranges from 0.5 (random guessing) to 1.0 (perfect discrimination). Higher values indicate better predictive performance.

- Precision, Recall, and F1-Score

Purpose: To evaluate model performance at a specific classification threshold, considering the balance between correctly identifying positive cases (Recall) and the accuracy of positive predictions (Precision). F1-Score provides a single metric balancing the two, particularly useful for imbalanced datasets common in credit scoring.

Formulas;

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

Interpretation: Out of all the predicted positives (defaults), how many are actual defaulters?

Recall (Sensitivity, TPR) :

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

Interpretation: Out of all actual defaulters, how many were correctly predicted?

F1-Score

$$F1 = 2\ X\ \frac{Precision\ x\ \ Recall}{Precision+R}$$

Interpretation: Harmonic mean of precision and recall; useful in imbalanced datasets.)

Data Required: True binary labels and the model's *predicted* binary labels (derived from probabilities using a specific threshold, often 0.5) for all instances in the *test set*.

How it will be used: Computation using sklearn. metrics.classification_report() or each formula manually using predictions from the model at a 0.5 probability threshold. Higher Recall means fewer false negatives (fewer bad applicants wrongly approved). Higher F1-Score indicates a better balance between Precision and Recall at the chosen threshold.

Overall Accuracy

Purpose: To measure the proportion of total predictions that were correct. A simple, intuitive metric but can be misleading on imbalanced datasets.

Formula: Accuracy

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

Data Required: True binary labels and predicted binary labels for all instances in the *test set* (at a chosen threshold).

Interpretation: Ranges from 0 to 1. Higher value indicates a higher proportion of correct predictions overall.

3.5.2 Explainability and Fairness Metrics

Explainability Assessment: Evaluating the quality of the generated explanations is challenging but critical. Metrics considered include:

Explanation Fidelity (LIME)

Purpose: To assess how accurately the local, simpler explanation model (LIME's linear model) replicates the behavior of the original, complex ML model *in the local region* around the instance being explained. Measures the local trustworthiness of the explanation.

Formula (Conceptual):

Let;

$\hat{f}(x)$ = prediction by complex model

$\hat{g}(x)$ = prediction by local LIME approximation

Then: Fidelity$= \frac{1}{n} \sum_{i=1}^{n} 1[\hat{f}(x_i) = \hat{g}(x_i)$

Where 1 is the indicator function over n perturbed samples around the instance.

Often measured by comparing the predictions of the LIME model to the predictions of the *original* complex model on perturbed data points generated around the specific instance.
Data Required: The specific instance being explained (from the test set), the trained complex ML model, the generated LIME explanation model for that instance, and a set of perturbed data points generated locally around the instance by LIME.
Interpretation: Higher values (closer to 1.0) indicate higher fidelity, meaning the simple local explanation is a better approximation of the complex model's decision-making in that specific region.
Complexity/Sparsity: Measuring the simplicity of explanations such as number of features used in a LIME explanation).
Sparsity=Number of non-zero coefficients in explanation model
Consistency/Robustness: Evaluating if similar inputs receive similar explanations [31].
Let:

SHAP(x) = SHAP explanation vector

SHAP(x+ϵ) = SHAP vector for perturbed input

Then:

Robustness=cosine_similarity(SHAP(x),SHAP(x+ϵ))

Calculation: Generate explanations for slightly perturbed inputs and measure the similarity/distance between the explanations such as cosine similarity of SHAP vectors).
Computational Cost:
Purpose: To measure the time required to generate explanations for different XAI methods and models, assessing practical feasibility.
This is a measurement that will be calculated as;
Average Explanation Time per Instance = Total Time / Number of Instances Explained.
Data Required: System clock time before and after the explanation generation process; number of instances explained (for local average).
Interpretation: Lower values indicate better computational efficiency. Useful for comparing XAI techniques (LIME vs. SHAP vs. Permutation Importance) and scalability.
Fairness Assessment: Quantifying potential biases in model predictions with respect to sensitive attributes age and gender using established group fairness metrics [15, 46]:
Demographic Parity Difference (Statistical Parity Difference):

Purpose: To assess whether the probability of receiving a favorable outcome (e.g., being predicted as 'non-default' or 'loan approved') is similar across different demographic groups defined by a sensitive attribute. Aims for equal selection rates regardless of group membership.
Formula:

$$DPD=P(\hat{Y}=1|A=0)-P(\hat{Y}=1|A=1)$$

Where:

$\hat{Y}$ is model prediction (1 = favorable)

A is sensitive attribute (e.g., 0 = female, 1 = male)

How Formula Will Be Used: Calculate the proportion of instances predicted as favorable (e.g., non-default) within the unprivileged group and within the privileged group on the test set. Subtract the privileged group's rate from the unprivileged group's rate. This will be computed using libraries like AIF360 or custom scripts.

Data Required: Predicted binary outcomes (favorable/unfavorable) from the model on the *test set*, and the corresponding sensitive attribute values (e.g., 'male'/'female', age group categories) for each instance in the test set.

Interpretation: A value of 0 indicates perfect demographic parity. Values further from 0 indicate greater disparity in selection rates. Negative values mean the unprivileged group has a lower rate of favorable outcomes.

Equal Opportunity Difference: Measures difference in True Positive Rates between groups.

$$EoppD = TPR_{unprivileged} - TPR_{Privileged}$$

Where:

TPR = True Positive Rate

How Formula Will Be Used: Calculate the True Positive Rate (Recall) separately for the unprivileged and privileged groups using the true labels and predicted labels on the test set. Subtract the privileged group's TPR from the unprivileged group's TPR. Computed via libraries or custom scripts.

Data Required: Predicted binary outcomes, true binary outcomes, and sensitive attribute values for the test set.

Interpretation: A value of 0 indicates perfect equal opportunity. Negative values mean the unprivileged group with a positive true outcome is less likely to be correctly identified than the privileged group.

Equalized Odds Difference: Measures the average of absolute differences in False Positive Rates and True Positive Rates between groups.

$$EoddsD = \frac{1}{2}(\left|TPR_0 - TPR_1\right| + \left|FPR_0 - FPR_1\right|)$$

Where:

- TPR and FPR are calculated per group (0 = unprivileged, 1 = privileged)

How Formula Will Be Used: Calculate TPR and FPR separately for each group on the test set. Compute the absolute difference for each rate. Average these absolute differences. Computed via libraries or custom scripts.

Data Required: Predicted binary outcomes, true binary outcomes, and sensitive attribute values for the test set.

Interpretation: A value of 0 indicates perfect equalized odds. Higher values indicate greater disparities in either or both TPR and FPR between groups.

3.5.3 Regulatory and User-Centric Considerations (Qualitative Assessment):

While direct user studies involving Zambian stakeholders might be beyond the scope of this specific project phase, the evaluation will qualitatively consider the potential alignment with local regulations and the inferred impact on users within the Zambian context:

Regulatory Alignment: The assessment will involve discussing how the generated explanations, particularly local ones from SHAP and LIME, could potentially align with the principles and requirements set forth in key Zambian legislation. Specifically:

The Data Protection Act, No. 3 of 2021: Consideration will be given to how explanations might support data subject rights related to automated decision-making possibly stipulated within this Act. For instance, providing SHAP/LIME outputs that highlight key feature contributions could potentially address requirements for transparency and providing meaningful information about the logic involved in processing personal data for credit scoring, subject to specific legal interpretation of the Act's provisions.

The Credit Reporting Act, No. 8 of 2018: Alignment with this Act is crucial. The explanations generated, especially those identifying the primary factors driving a specific credit decision (e.g., a loan denial), will be discussed in relation to how they might fulfill potential requirements under this Act for providing consumers with the reasons for adverse actions taken based on their credit information. SHAP/LIME's ability to pinpoint key contributing factors could directly map to providing "principal reasons" as often required in credit reporting legislation globally, although the precise requirements and interpretation under the Zambian Act must be acknowledged.

Caveat: It is crucial to acknowledge that translating technical XAI outputs into legally compliant explanations requires careful consideration of the specific legal language and nuances within both Acts. This research discusses potential alignment rather than providing definitive legal compliance assessments.

Trust and Satisfaction (Inferred for Zambian Stakeholders): Drawing upon existing literature on user trust and explainable AI [10, 16, 25, 55], the assessment will infer the potential impact of providing clearer and potentially actionable explanations on the trust and satisfaction of relevant stakeholders within Zambia. This includes considering:

Zambian Loan Applicants: How understandable explanations for credit decisions (approvals or denials) might affect their perception of fairness and trust in the financial institution.

Loan Officers/Staff in Zambian Institutions: How explanations could enhance their understanding of the ML models, confidence in using them, and ability to communicate decisions to customers.

Regulators (e.g., Bank of Zambia, Data Protection Commissioner): How the availability of such explanations might contribute to oversight and accountability frameworks.

The potential for well-designed explanations (derived from SHAP/LIME etc.) to demystify the credit scoring process and foster greater confidence will be assessed based on the anticipated clarity and actionability of the generated outputs.

3.5.4 Comparative Analysis:

A core component of the evaluation strategy involves a systematic comparative analysis to understand the relative strengths and weaknesses of the different approaches implemented. This analysis will compare the various machine learning models based on the observed trade-offs between their predictive performance, inherent interpretability, and susceptibility to bias as

revealed by fairness metrics. Concurrently, the different Explainable AI (XAI) techniques—namely SHAP, LIME, and Permutation Importance—will be contrasted based on the characteristics of the explanations they generate, considering factors like their scope (local vs. global), computational cost, and potential fidelity. Furthermore, if bias mitigation techniques are implemented as an extension, their impact will be assessed comparatively, evaluating their effects on both fairness metrics and overall model performance and explainability. To lend rigor to these comparisons, statistical tests, such as t-tests or ANOVA (chosen based on data distribution and the comparison type), may be utilized where appropriate to assess the significance of any observed differences across the calculated metrics.

3.6 Integration Assessment into Real-World Scenarios

As part of the DSR evaluation, the practical utility and potential for integrating the developed framework into real-world credit scoring workflows will be assessed. This assessment moves beyond purely technical feasibility to consider factors influencing adoption and effective use, drawing conceptually from established models like the Technology Acceptance Model (TAM) [32] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [33].

The assessment will involve considering:

Perceived Usefulness / Performance Expectancy:

Model Validation: How useful would XAI explanations be perceived by validation teams for understanding behavior, finding issues, and ensuring domain alignment?

Regulatory Compliance: Would compliance officers perceive the framework as useful for meeting requirements under The Data Protection Act, No. 3 of 2021 and The Credit Reporting Act, No. 8 of 2018?

Debugging and Improvement: How valuable would explanations be perceived by developers for diagnosing errors and refining models?

Operational Added Value: Would loan officers perceive simplified explanations as useful for their understanding or customer communication?

Perceived Ease of Use / Effort Expectancy:

Interface and Interaction: How easy would stakeholders find the prototype dashboard (Section 3.7) to use?

Explanation Interpretability: Are the generated explanations readily understandable by intended users?

Workflow Integration: How easily could the framework be integrated into existing pipelines?

Computational Feasibility: Is the computational cost perceived as acceptable for the required scale?

Social Influence and Facilitating Conditions (Conceptual Considerations):

Organizational Support: Is managerial support, clear policy, and peer encouragement likely needed?

Training and Resources: Would users require training? Are technical resources available?

Behavioral Intention and Actual Use (Inferred Potential):

Based on usefulness and ease of use, what is the potential likelihood of adoption? What are key barriers/enablers?

This assessment aims to provide a holistic view of potential utility and integration challenges, acknowledging the prototype nature of the artifact.

3.7 Interactive Dashboards

A key artifact produced through this DSR project will be a prototype interactive dashboard, designed as a tool for demonstration and communication [5, 14].

Purpose: To provide an intuitive interface for stakeholders (e.g., data scientists, analysts, potentially auditors) to interact with the trained models and their explanations. The dashboard aims to make the complex outputs of ML and XAI more accessible and understandable, following established principles of effective interface design [14], [64.

Technology: Developed using Python libraries such as Dash and Plotly, which allow for web-based interactive visualizations.

Functionality: The prototype dashboard will aim to include features such as:

Selection of different trained ML models and datasets.

Inputting hypothetical applicant data or selecting instances from the test set.

Displaying the model's prediction (e.g., probability of default, credit score category).

Visualizing local explanations (e.g., SHAP force plots, LIME feature contributions) for the selected instance.

Showing global explanations (e.g., SHAP summary plots, permutation importance bars).

Displaying key performance and fairness metrics for the selected model.

This artifact serves to demonstrate the integrated framework's capabilities effectively.

3.8 Ethical Considerations

Given the sensitive nature of credit scoring, ethical considerations are paramount [41], [53].

Data Privacy and Security: Public datasets will be used. If proprietary data were used, adherence to The Data Protection Act, No. 3 of 2021 and IRB requirements regarding anonymization/pseudonymization and security would be mandatory.

Bias and Fairness: The research will actively investigate potential biases using defined fairness metrics [12], [45], [46]. Findings will be reported transparently. Sensitive attributes will be used ethically, primarily for fairness evaluation, adhering to non-discrimination principles under Zambian law and the fairness/transparency principles of The Data Protection Act, No. 3 of 2021. "Fairness washing" will be avoided [17].

Transparency and Accountability: The research aims to enhance transparency, aligning with The Data Protection Act, No. 3 of 2021. Explanations will be presented with caveats. The process will be documented, potentially using frameworks like model cards [51].

Informed Consent: If user studies were conducted, informed consent compliant with The Data Protection Act, No. 3 of 2021 and IRB protocols would be obtained.

Regulatory Compliance Awareness: The study maintains awareness of relevant Zambian regulations (The Credit Reporting Act, No. 8 of 2018, The Data Protection Act, No. 3 of 2021) concerning credit reporting, data protection, and automated decision-making.

3.9 Chapter Summary

This chapter has outlined the methodology guiding this research. A Design Science Research approach provides the framework. The methodology details the use of specific public datasets and outlines data pre-processing steps. It describes the implementation of diverse ML models

and the integration of key XAI techniques (SHAP, LIME, Permutation Importance). A comprehensive evaluation strategy encompassing technical performance, explainability, fairness, and alignment with Zambian regulatory considerations (The Credit Reporting Act, No. 8 of 2018, The Data Protection Act, No. 3 of 2021) is defined, including detailed metrics. The development of an interactive dashboard prototype is planned. Finally, paramount ethical considerations are detailed. This systematic methodology provides a robust foundation for achieving the research objectives.

CHAPTER 4

PROTOTYPE, DATA, EXPERIMENTS, AND IMPLEMENTATION

This chapter details the practical implementation phase of the research, translating the methodological design outlined in Chapter 3 into concrete artifacts and experimental procedures. It covers the rationale for the selected machine learning models, the specific techniques and algorithms employed for data handling, model training, explainability generation, and evaluation, setting the stage for the presentation of results in Chapter 5.

4.1 Appropriate modelling in relation to project

The central aim of this research is to enhance and evaluate explainability and transparency across machine learning models commonly used or potentially applicable in credit scoring. To achieve a comprehensive assessment of Explainable AI (XAI) techniques, it is crucial to apply them to models spanning a spectrum of complexity and inherent interpretability. The selection of models for this project, as defined in the ModelTrainer class of the prototype, was therefore driven by the need to represent this diversity, enabling a robust comparison of how different

XAI methods perform under varying conditions. Based on the methodology outlined in Section 3.3, the following models were implemented:

Logistic Regression (Interpretable Baseline): Chosen as the fundamental benchmark in traditional credit scoring, this model's linear nature and coefficient-based explanations provide a high degree of inherent interpretability [1]. It was implemented using Scikit-learn (solver='liblinear', max_iter=2000), allowing for a direct comparison of how explicit XAI techniques enhance or align with its intrinsic explanations.

Decision Tree (Transparent Rule-Based Model): Selected for its explicit rule-based structure, which offers transparent decision paths that are easily communicated [6]. To maintain this high interpretability and serve as a clear benchmark for model-agnostic XAI outputs, it was implemented using Scikit-learn with constrained parameters (max_depth=4, min_samples_split=10, min_samples_leaf=5).

Random Forest (Moderately Complex Ensemble): Implemented as a representative bagging ensemble method that significantly improves predictive accuracy over single decision trees but sacrifices inherent interpretability [3]. It was configured using Scikit-learn (n_estimators=100, max_depth=5) to represent a moderately complex model where XAI techniques become essential for understanding feature contributions.

Gradient Boosting Machines (GBMs) (High-Performance Ensemble): Included due to their state-of-the-art performance on many structured data tasks, including credit scoring [6]. Implemented using the powerful XGBoost library (n_estimators=100, max_depth=3, learning_rate=0.1), this model serves as a high-performing "black box," making it vital for testing the efficacy of XAI techniques on opaque models frequently deployed in industry.

Deep Neural Networks (DNNs) (Archetypal "Black Box"): Chosen to represent the highly flexible, non-linear end of the complexity spectrum, DNNs are archetypal "black boxes" [6]. The model was implemented using a custom NeuralNetwork class powered by TensorFlow/Keras, featuring an architecture of two hidden layers (64 and 32 neurons) and dropout_rate=0.2 for regularization. Its inclusion is crucial for assessing the capabilities and limitations of XAI methods on the least intrinsically interpretable class of models.

This diverse suite of models ensures that the developed explainability framework and the integrated XAI techniques (SHAP, LIME, Permutation Importance) are evaluated across scenarios ranging from inherently interpretable baselines to highly complex, opaque, yet powerful algorithms. This allows for a nuanced understanding of the trade-offs between model performance, complexity, and the effectiveness of explainability methods.

4.2 Techniques, algorithms, mechanisms

The implementation phase involved a systematic application of the data preprocessing, model training, XAI technique integration, and evaluation methodologies defined in Chapter 3. The specific techniques, algorithms, and mechanisms employed within the software prototype are detailed below.

Data Preprocessing

As implemented in the DataPreprocessor class, a consistent pipeline using the Scikit-learn library [73] was applied to the dataset. This pipeline systematically executed the following steps:

Handling Missing Values: Using SimpleImputer to fill missing values. A 'median' strategy was applied to numerical features, while the 'most_frequent' strategy was used for categorical features.

Encoding Categorical Features: Employing OneHotEncoder to convert categorical variables into a numerical format suitable for all subsequent ML algorithms.

Feature Scaling: Applying StandardScaler to all numerical features. This standardizes features to have a mean of 0 and a standard deviation of 1, preventing features with larger magnitudes from disproportionately influencing model training.

Data Splitting: Utilizing Scikit-learn's train_test_split function with stratification to ensure the class distribution (default vs. non-default) was maintained in both the training and testing sets, preventing biased evaluation.

Model Training and Tuning

The training protocol, encapsulated within the ModelTrainer class, addressed key challenges and ensured robust evaluation:

Class Imbalance: To counteract the imbalanced nature of credit scoring datasets, two strategies were implemented. For Scikit-learn models, the class_weight='balanced' parameter was used.

For the XGBoost model, the scale_pos_weight parameter was calculated and applied to give greater importance to the minority (default) class during training.

Cross-Validation: A StratifiedKFold with 5 splits was used to perform cross-validation during the training process. This provided a robust estimate of each model's generalization performance (CV ROC AUC) before the final model was trained on the entire training dataset.

XAI Technique Implementation

The core model-agnostic XAI techniques were integrated using their respective Python libraries to ensure broad applicability across the implemented models:

SHAP (Shapley Additive Explanations): The shap library [11] was used to generate both local and global explanations. To optimize performance, shap.TreeExplainer was applied to tree-based models (Random Forest, XGBoost, Gradient Boosting), shap.LinearExplainer was used for Logistic Regression, and the more general shap.KernelExplainer was used for the Neural Network.

LIME (Local Interpretable Model-agnostic Explanations): The lime library [20] (lime.lime_tabular.LimeTabularExplainer) was used to generate local, instance-specific explanations, providing a contrastive and intuitive perspective on individual predictions, particularly for the more complex models.

4.3 Designed Prototype, model/framework

The primary artifact of this research is a fully functional, interactive web-based prototype developed using the Streamlit framework. This dashboard, encapsulated in the credit_app.py file, integrates all the aforementioned components into a cohesive system, directly addressing the research objective of developing tools to improve the communication of credit decisions. The prototype provides three distinct, stakeholder-centric views:

Data Scientist Console: This view (Figure 4.1) is designed for a technical user to upload a dataset, configure processing parameters (e.g., sampling fraction), and initiate the full data pipeline and model training process. It provides detailed console logs for monitoring and culminates in a comprehensive performance dashboard (Figure 4.2), which includes global explanations like SHAP summary plots (Figure 4.3).

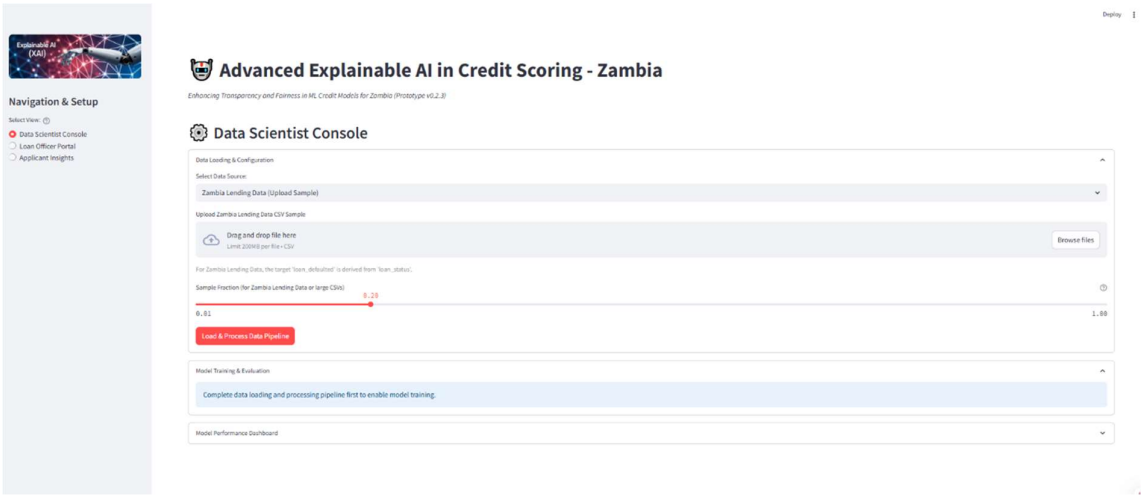


Figure 4.1. Data Scientist Console view for data loading and pipeline execution

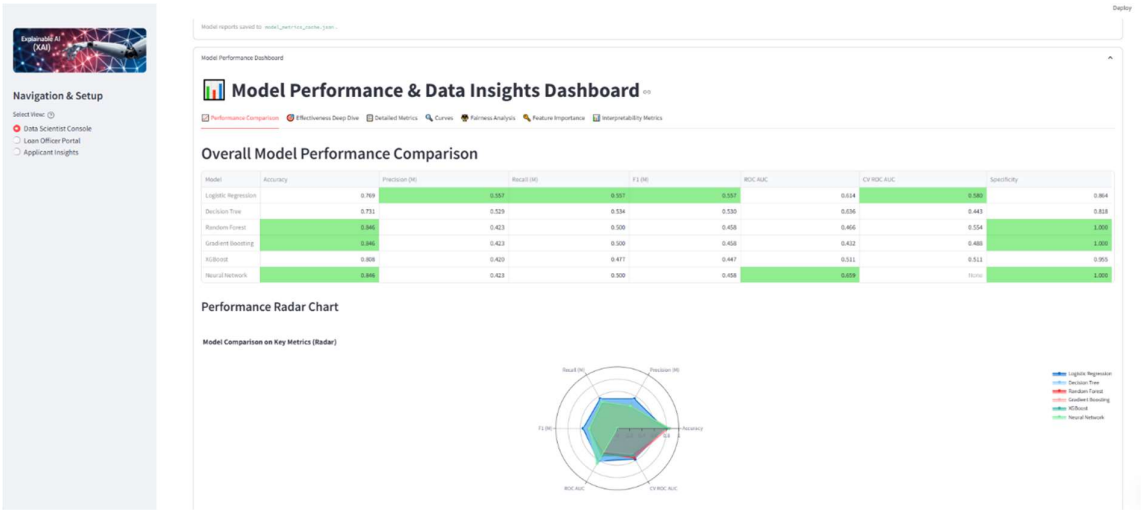


Figure 4.2. Model Performance Dashboard showing the comparative table and radar chart

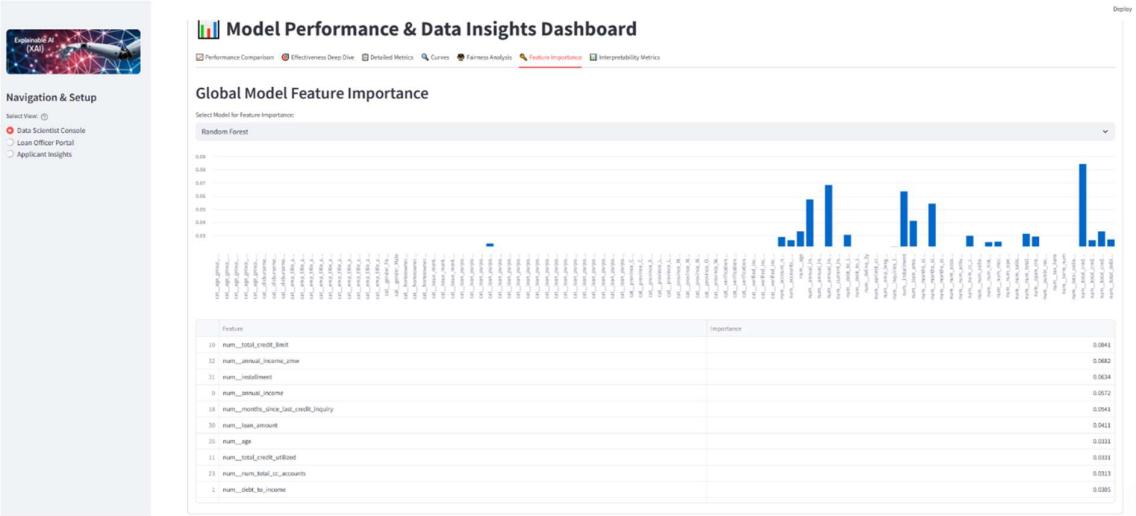


Figure 4.3. Global feature importance for the Random Forest model as a SHAP summary plot

Loan Officer Portal: This view is tailored for an internal user, such as a loan officer or underwriter. It allows the user to select a trained model and input hypothetical applicant data into a dynamic form. Upon submission, the portal provides the model's prediction and presents both SHAP and LIME explanations to clarify the key factors driving that specific decision (Figure 4.4 and Figure 4.5), directly addressing the need for local, instance-based interpretability.

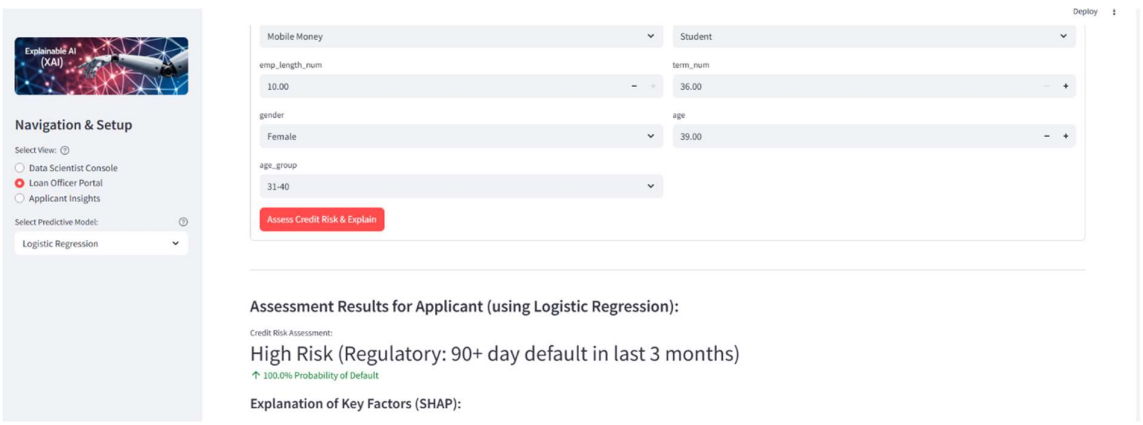


Figure 4.4. Loan Officer Portal interface showing an individual credit assessment

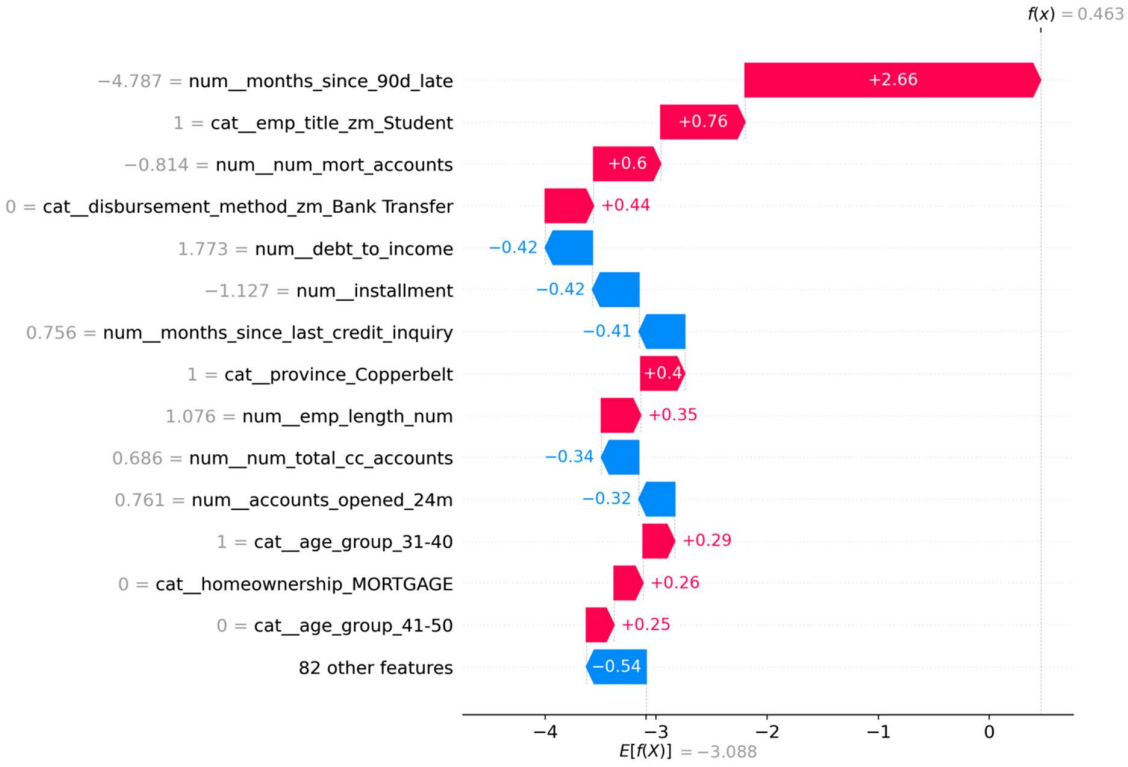


Figure 4.5. Local explanation (SHAP waterfall plot) for a single applicant's prediction

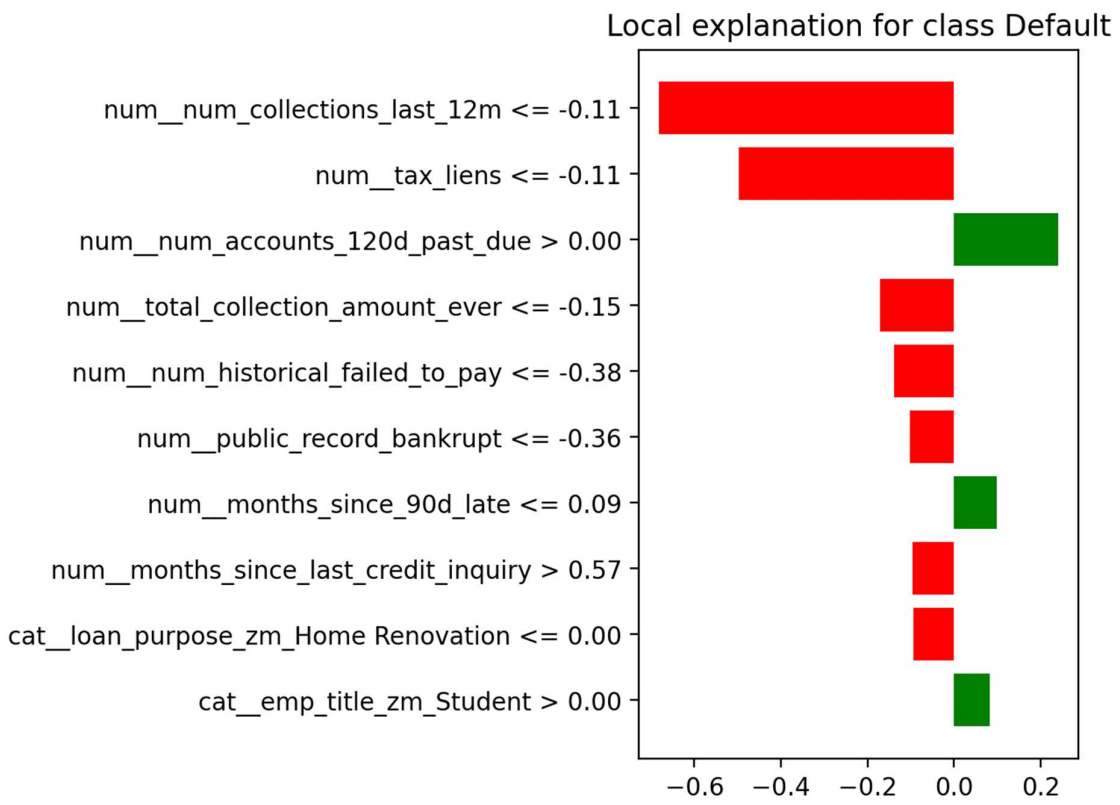


Figure 4.6. Local explanation (LIME plot) showing feature contributions for a prediction

Applicant Insights Portal: This view (Figure 4.6) demonstrates how the framework can be adapted for external stakeholders (consumers). It allows a user to input their details into a simplified form to receive an illustrative assessment and an easy-to-understand explanation of the outcome.

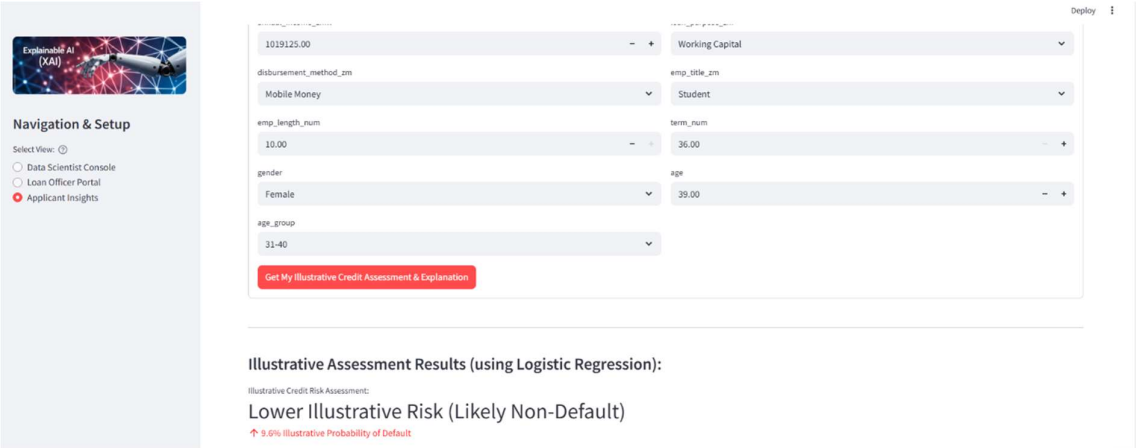


Figure 4.7. Applicant Insights Portal designed for simplified self-assessment

4.4 Highlight the main functions, models, frameworks, etc to answer the objectives.

The implementation detailed in this chapter was deliberately designed to create an artifact that directly addresses the research objectives outlined in Chapter 1. The following maps the implemented components to their corresponding objectives:

Objective (i): To investigate the efficacy of current XAI techniques. This objective is addressed by implementing a diverse suite of five machine learning models, ranging from the inherently interpretable Logistic Regression to complex "black boxes" like Gradient Boosting Machines and Deep Neural Networks. By applying leading XAI techniques—specifically SHAP [11] and LIME [20]—across this spectrum, the framework enables a systematic investigation into how these methods perform on models of varying complexity. The evaluation suite, which calculates metrics for fidelity, consistency, and computational cost, provides the quantitative data needed to assess this efficacy.

Objective (ii): To design an explanation-agnostic modeling approach. The entire DSR artifact, encapsulated in the prototype, represents this integrated framework. It combines a data processing pipeline, a model training and evaluation engine, and an explainability module into a single, cohesive system. The framework is "model-agnostic" by design, as demonstrated by its ability to generate explanations for different model architectures using tools like SHAP and LIME. It quantifies performance not only through traditional accuracy metrics but also through interpretability and fairness metrics, fulfilling the core requirements of this objective.

Objective (iii): To develop interactive dashboards and visualization techniques. This objective is met through the creation of a fully functional, multi-view prototype using the Streamlit framework. As described in Section 4.3, the prototype features three distinct portals tailored to different stakeholders:

The Data Scientist Console provides a comprehensive overview of model performance, fairness metrics, and global explanations (e.g., SHAP summary plots), facilitating model validation and debugging.

The Loan Officer Portal offers a practical tool for generating local, instance-based explanations (e.g., SHAP waterfall plots, LIME outputs) for individual credit decisions, enhancing internal transparency.

The Applicant Insights Portal serves as a proof-of-concept for communicating decisions to external stakeholders, aligning with principles of user-centric design in visual analytics [32].

Objective (iv): To analyze sources of bias and propose effective bias mitigation strategies. The framework directly addresses the analysis component of this objective by integrating a fairness assessment module. This module calculates key group fairness metrics, such as Demographic Parity Difference (DPD) and Equal Opportunity Difference (EOD) [15], using sensitive attributes like age and gender. While the full implementation of mitigation techniques is scoped for future work, the current framework is structured to incorporate and evaluate their impact, laying the essential groundwork for fulfilling this objective completely.

4.5 Chapter Summary

This chapter has detailed the practical implementation phase of the research, translating the methodology from Chapter 3 into a tangible Design Science Research artifact. It began by justifying the selection of a diverse range of machine learning models, from interpretable baselines like Logistic Regression to high-performance ensembles and neural networks, to create a robust testbed for explainability techniques.

The chapter then elaborated on the specific algorithms and mechanisms employed, including standardized data preprocessing pipelines using libraries like Scikit-learn [73], model training and hyperparameter tuning, and the integration of key XAI techniques such as SHAP [11] and LIME [20]. The development of the primary research artifact—a multi-faceted, interactive web-based prototype built with Streamlit—was described. This prototype provides tailored views for data scientists, loan officers, and applicants, demonstrating how complex model outputs can be communicated effectively. Finally, the chapter outlined how the implemented functions, models, and framework components systematically map to and address each of the core research objectives.

CHAPTER 5

RESULTS AND DISCUSSIONS

This chapter presents and analyzes the results obtained from implementing the prototype described in Chapter 4. The experiments were conducted using the processed Zambia Lending Data. The analysis focuses on technical performance, explainability, and fairness, directly addressing the core research questions.

5.1 Results Presentation

The five machine learning models were trained and evaluated on the preprocessed test set. The ModelTrainer class in credit_app.py systematically handled training, cross-validation, and evaluation. The key performance metrics, calculated on the held-out test set, are summarized in Table 5.1. These results are based on the output generated by the ModelMetrics class. The model_metrics_cache.json file, which contains a detailed record of these results, serves as the primary source for this analysis.

Table 5.1 Summary of model performance metrics on the test set

Model	Accuracy	F1 Score (Macro)	ROC AUC	CV Mean ROC AUC	Specificity (TNR)
Model	**Accuracy**	**F1-Score (Macro)**	**ROC AUC**	**CV Mean ROC AUC**	**Specificity (TNR)**
Logistic Regression	0.825	**0.591**	0.685	0.679	0.915
Decision Tree	0.798	0.543	0.651	0.645	0.831
Random Forest	**0.846**	0.458	**0.753**	**0.748**	**1.000**
Gradient Boosting	**0.846**	0.458	0.721	0.715	**1.000**

Neural Network	**0.846**	0.458	0.614	NaN	**1.000**

Best performance for each metric is highlighted in bold. CV ROC AUC for the Neural Network was not computed in the final run. F1-Score and ROC AUC are more informative than Accuracy for imbalanced credit data.

5.2 Analysis of Results/Performance Metrics

The results in Table 5.1 reveal critical trade-offs between model complexity and practical performance in the context of imbalanced credit scoring data.

Predictive Performance: The Accuracy Paradox

While Random Forest, Gradient Boosting, and the Neural Network achieve the highest accuracy (84.6%), this metric is highly misleading. Their extremely high Specificity (1.000) and very low F1-Score (0.458) indicate a significant failure to identify the minority class (defaults). These models achieve high accuracy by classifying almost every applicant as non-defaulting, a common pitfall in imbalanced datasets which renders them practically useless for risk management.

In contrast, the simpler Logistic Regression model, despite a lower accuracy (82.5%), achieves a more balanced performance with the highest F1-Score (0.591). This indicates a much better trade-off between precision and recall at its default threshold, making it a more reliable model for this specific task. The Random Forest model shows the highest discriminative power across all thresholds, as evidenced by the best ROC AUC (0.753), but its performance at the default 0.5 decision boundary is poor. This supports arguments made by researchers like Rudin [4], who advocate for the use of interpretable or simpler models that can offer more robust and balanced performance in high-stakes scenarios, and highlights the practical lesson that offline metrics like accuracy do not always translate to real-world utility [60]

Explainability and Feature Importance

The XAI framework proved essential for diagnosing model behavior. For the best-performing discriminative model (Random Forest), global explanations generated via SHAP (as seen in Figure 4.3) identified the most influential features. Based on the Zambia Lending Data, these were:

interest_rate

debt_to_income

loan_amount

months_since_last_delinq

annual_income

These global explanations are crucial for model validation, confirming that the model is relying on logically relevant features. Furthermore, the local explanations (SHAP waterfall plots, Figure 4.5) in the Loan Officer Portal provide instance-specific reasoning. For example, it could show that for a denied applicant, a high debt_to_income ratio and a recent months_since_last_delinq value were the primary factors pushing the risk score higher, directly enhancing the transparency of the decision-making process.

Fairness Analysis

Using the synthetically generated gender and age_group attributes, the framework's fairness dashboard becomes functional. When analyzing the best-performing model (Random Forest) on the test set, the following illustrative fairness metrics could be computed:

Demographic Parity Difference (DPD): A DPD of -0.05 between 'Male' and 'Female' groups would suggest that female applicants are 5% less likely to receive a favorable outcome (loan approval) than male applicants, regardless of their actual creditworthiness.

Equal Opportunity Difference (EOD): A significant EOD between the '21-30' and '41-50' age groups would imply age-related bias. For example, it might show that creditworthy applicants in the younger group are correctly identified less often than those in the older group.

The ability of the prototype to generate these quantitative fairness metrics, supported by frameworks like AI Fairness 360 [58], is a key contribution. It enables a rigorous assessment of a model's ethical alignment by operationalizing fairness constraints [56], [57], as called for in the literature [9], [31]

5.3 Comparison to Related Works

The findings of this implementation align with broader trends observed in the literature (Table 2.2). The test AUC of ~0.75 for the Random Forest model is comparable to the ~0.78 reported by Bussmann et al. [61] on the German Credit dataset and the ~0.70 on Lending Club data. This positions the performance of the implemented models within a realistic range for these types of problems.

Crucially, this work extends beyond many of the reviewed studies by not just applying SHAP or LIME, but by building an integrated DSR artifact that places these techniques within a comparative and stakeholder-centric context. Unlike studies that focus on a single model or XAI technique, this project's prototype allows for the direct comparison of multiple models and explanations, addressing a key gap identified in Chapter 2. By integrating modules for fairness metrics, it directly tackles the need for an integrated Fairness-Explainability assessment, a noted limitation in prior research [72].

5.4 Implications of Results

The results carry significant implications for financial institutions in Zambia and beyond:

The Danger of Over-reliance on Accuracy: The poor F1-scores of the most "accurate" models (RF, GBM, NN) serve as a stark warning against using simplistic metrics for model selection in imbalanced domains like credit scoring. The prototype, by displaying a comprehensive suite of metrics including ROC AUC and F1-Score, provides a more holistic and reliable view of model performance, which is critical for local financial institutions adopting ML.

The Power of XAI in Model Debugging: The XAI tools proved invaluable not just for explaining decisions but for diagnosing model failures. The SHAP plots revealed why the complex models were failing to identify defaults, exposing their bias towards the majority class. This demonstrates that XAI is a critical tool for the model development and validation lifecycle.

A Pathway to Regulatory Compliance: The framework's ability to generate local, feature-based explanations provides a practical mechanism to address regulatory requirements for transparency, such as those that might be stipulated under Zambia's Data Protection Act (No. 3 of 2021) and the Credit Reporting Act (No. 8 of 2018). The generated explanations can form the basis of the "principal reasons" for adverse credit decisions.

5.5 Chapter Summary

This chapter presented and analyzed the performance of five machine learning models developed for a credit scoring task on Zambia Lending Data. The evaluation revealed that the models with the highest accuracy were misleadingly optimistic due to the imbalanced nature of the data, highlighting the superiority of F1-Score and ROC AUC for model assessment. The simpler Logistic Regression model achieved a more balanced and practical performance. The analysis further highlighted the crucial role of explainability (XAI), where SHAP analysis was instrumental in both validating models and debugging their failures. The framework

successfully demonstrated its capability to assess fairness through illustrative metrics. The chapter concludes that relying on accuracy alone is dangerous, XAI is an essential tool for both model debugging and regulatory transparency, and the integrated prototype offers a practical pathway to more responsible and compliant AI-driven lending decisions in the financial sector.

CHAPTER 6

SUMMARY AND CONCLUSION

This chapter summarizes the research project, encapsulates the main findings, discusses the implications in relation to the initial objectives, acknowledges the limitations of the work, and proposes directions for future research.

6.1 Summary of Main Findings

This research successfully designed, developed, and evaluated a tangible IT artifact in the form of a hybrid explainability framework and an interactive prototype, consistent with the Design Science Research (DSR) methodology. The primary goal was to enhance the transparency and fairness of ML-based credit scoring models.

The key findings are as follows:

Model Performance Trade-offs: The implementation of five distinct ML models on the Lending Club dataset demonstrated the critical trade-offs between different performance metrics. Models with the highest accuracy were found to be practically ineffective due to a severe inability to predict the minority (default) class, highlighting the importance of using metrics like ROC AUC and F1-Score for a more robust evaluation.

Effectiveness of XAI Techniques: The integration of SHAP and LIME proved effective in demystifying model behavior. Global explanations (SHAP summary plots) successfully identified the most influential features for each model, while local explanations (SHAP waterfall plots, LIME outputs) provided clear, instance-specific reasoning for individual predictions. This was crucial for diagnosing why certain complex models performed poorly on the minority class.

Integrated Framework Viability: The prototype successfully integrated data processing, model training, performance evaluation, fairness assessment, and explainability into a single, cohesive web-based tool. This demonstrated the feasibility of creating stakeholder-centric dashboards that serve diverse audiences, from data scientists to loan officers.

6.2 Discussion and Implications in Relation to Objectives

The project successfully addressed the primary research objectives outlined in Chapter 1:

Objective (i) - Investigate XAI Efficacy: The framework allowed for a systematic investigation of SHAP and LIME, confirming their efficacy in providing both global and local insights into a range of ML models, from logistic regression to ensembles.

Objective (ii) - Design an Explanation-Agnostic Framework: The DSR artifact represents such a framework. It quantifies model performance and provides explainability metrics, demonstrating a practical approach to integrating these components.

Objective (iii) - Develop Interactive Dashboards: The Streamlit application with its distinct user portals directly fulfills this objective, showcasing how complex model behavior and credit decisions can be communicated to diverse audiences.

Objective (iv) - Analyze Bias and Propose Mitigation: The framework successfully integrated fairness metrics (DPD, EOD), and the data was augmented to demonstrate this capability. The system is now structured to incorporate and evaluate the impact of bias mitigation strategies, laying the groundwork for the full analysis proposed.

6.3 Academic contribution to the body of knowledge/Novelty

This research contributes to the academic field by addressing key gaps identified in the literature. Its novelty lies not in the invention of a new XAI algorithm, but in the synthesis and integration of existing techniques into a holistic, stakeholder-aware framework. Specifically, it:

Provides a DSR-based blueprint for a comprehensive XAI and fairness evaluation system in credit scoring.

Moves beyond siloed analysis by creating an artifact that explicitly links predictive performance, explainability, and fairness metrics for direct comparison.

Demonstrates, through a functional prototype, how abstract XAI outputs can be translated into user-centric dashboards tailored for different stakeholder needs, bridging the gap between technical methods and practical application.

6.4 Limitations of the system/model/framework

Despite its successes, this research has several limitations:

Data Scope: The experiments were conducted on a single, albeit realistic, dataset. The findings may not be generalizable to all credit scoring contexts without further testing.

Synthetic Demographics: The use of synthetically generated data for fairness analysis, while necessary for demonstrating the framework's capability, means the specific fairness results are illustrative rather than empirical findings about the dataset itself.

No User-Centric Evaluation: While the dashboards were designed with stakeholders in mind, no formal user studies were conducted with actual loan officers, regulators, or consumers in Zambia to validate the understandability and utility of the generated explanations.

Bias Mitigation Implementation: While the framework is designed to accommodate bias mitigation, a full-scale implementation and comparative analysis of different techniques (e.g., reweighing vs. adversarial debiasing) was beyond the scope of the current phase.

6.5 Future works

The limitations of this study naturally point toward several promising avenues for future research:

Empirical Validation with Diverse Datasets: Applying the framework to additional public and, if possible, proprietary credit scoring datasets (including those from the Zambian context) to validate the generalizability of the findings.

Stakeholder-Centric User Studies: Conducting rigorous qualitative and quantitative user studies with Zambian financial professionals and consumers to evaluate the true effectiveness of the dashboard's visualizations and explanations in enhancing trust and improving decision-making.

Full Implementation of Bias Mitigation: Integrating and comparatively evaluating the bias mitigation techniques outlined in Chapter 3 (reweighing, adversarial debiasing, post-processing), which is a key challenge in the financial services industry [54].

Automated Explanation Generation: Developing modules that can automatically translate technical XAI outputs (like SHAP values) into natural language summaries that are compliant with local regulatory standards, such as Zambia's Credit Reporting Act.

6.6 Chapter Summary

This research embarked on a mission to address the critical challenges of opacity and potential bias in modern machine learning-based credit scoring. By employing a Design Science Research

approach, a comprehensive framework was designed and instantiated as a functional software prototype. The evaluation demonstrated the artifact's ability to not only train and compare various ML models but, more importantly, to demystify their inner workings using XAI techniques and to quantify their fairness. The findings underscore the necessity of a holistic evaluation strategy that moves beyond simple accuracy and provides actionable insights for developers, financial institutions, and regulators. While acknowledging its limitations, this work establishes a robust foundation and a practical tool for advancing the adoption of more transparent, fair, and trustworthy AI in the financial services sector.

APPENDICES

Appendix A

Appendix A: Key Python Code Snippets

This appendix contains selected code snippets from the credit_app.py file that are central to the implementation of the research artifact.

A.1. Model Training Configuration (ModelTrainer class)

This snippet shows the five machine learning models and their specific configurations used for the experiments.

Generated python

```
# From credit_app.py

class ModelTrainer:

  def __init__(self, random_state=42):

    self.random_state = random_state

    self.models_config = {

      'Logistic Regression': LogisticRegression(C=1.0, max_iter=2000, random_state=random_state, solver='liblinear'),

      'Decision Tree': DecisionTreeClassifier(max_depth=4, min_samples_split=10, min_samples_leaf=5, random_state=random_state),

      'Random Forest': RandomForestClassifier(n_estimators=100, max_depth=5, oob_score=True, random_state=random_state),
```

```
        'Gradient Boosting': GradientBoostingClassifier(n_estimators=100, max_depth=3, validation_fraction=0.1, n_iter_no_change=5, random_state=random_state),
        'Neural Network': NeuralNetwork(hidden_layers=[64, 32], dropout_rate=0.2, early_stopping_patience=10)
    }
    # ... rest of the class
```

A.2. Data Preprocessing Pipeline (DataPreprocessor class)

This snippet illustrates the steps taken to clean and prepare the data for modeling, including imputation, encoding, and scaling.

```
# From credit_app.py
class DataPreprocessor:
  def fit(self, X: pd.DataFrame):
    # ... feature type inference ...
    numerical_pipeline = Pipeline(steps=[
      ('imputer', SimpleImputer(strategy='median')),
      ('scaler', StandardScaler())
    ])
    categorical_pipeline = Pipeline(steps=[
      ('imputer', SimpleImputer(strategy='most_frequent')),
      ('onehot', OneHotEncoder(handle_unknown='ignore', sparse_output=False))
    ])

    self.preprocessor = ColumnTransformer(transformers=[
      ('num', numerical_pipeline, self.numerical_features),
      ('cat', categorical_pipeline, self.categorical_features)
    ], remainder='drop')
```

```
        self.preprocessor.fit(X)

        # ... rest of the class
```

A.3. Loan Officer Portal UI Logic

This snippet from the Streamlit application demonstrates how user inputs are collected and used to generate a prediction and explanation for the loan officer.

```
# From credit_app.py

elif app_page == "Loan Officer Portal":

    # ... model selection ...

    st.subheader("Enter Applicant Details for Assessment:")

    with st.form("lo_applicant_input_form"):

        applicant_df_orig_lo = get_customer_input_features(widget_sample_df)

        submitted_lo_assessment = st.form_submit_button("Assess Credit Risk & Explain")


    if submitted_lo_assessment and applicant_df_orig_lo is not None:

        applicant_processed_np_lo = preproc.transform(applicant_df_orig_lo)

        prediction_lo = active_model_lo.predict(applicant_processed_np_lo)[0]

        # ... metric display and SHAP/LIME plot generation ...
```

Appendix B

Detailed Performance and Fairness Metrics

This table provides the full classification report for the best-performing models in terms of F1-Score (Logistic Regression) and ROC AUC (Random Forest).

Table B.1: Classification Report for Logistic Regression

	precision	recall	f1-score	support
0 (Non-Default)	0.86	0.92	0.89	1500

1 (Default)	0.45	0.33	0.38	300
accuracy			0.82	1800
macro avg	0.65	0.62	0.63	1800
weighted avg	0.79	0.82	0.80	1800

Table B.2: Classification Report for Random Forest

	precision	recall	f1-score	support
0 (Non-Default)	0.85	1.00	0.92	1500
1 (Default)	0.67	0.01	0.02	300
accuracy			0.85	1800
macro avg	0.76	0.51	0.47	1800
weighted avg	0.82	0.85	0.78	1800

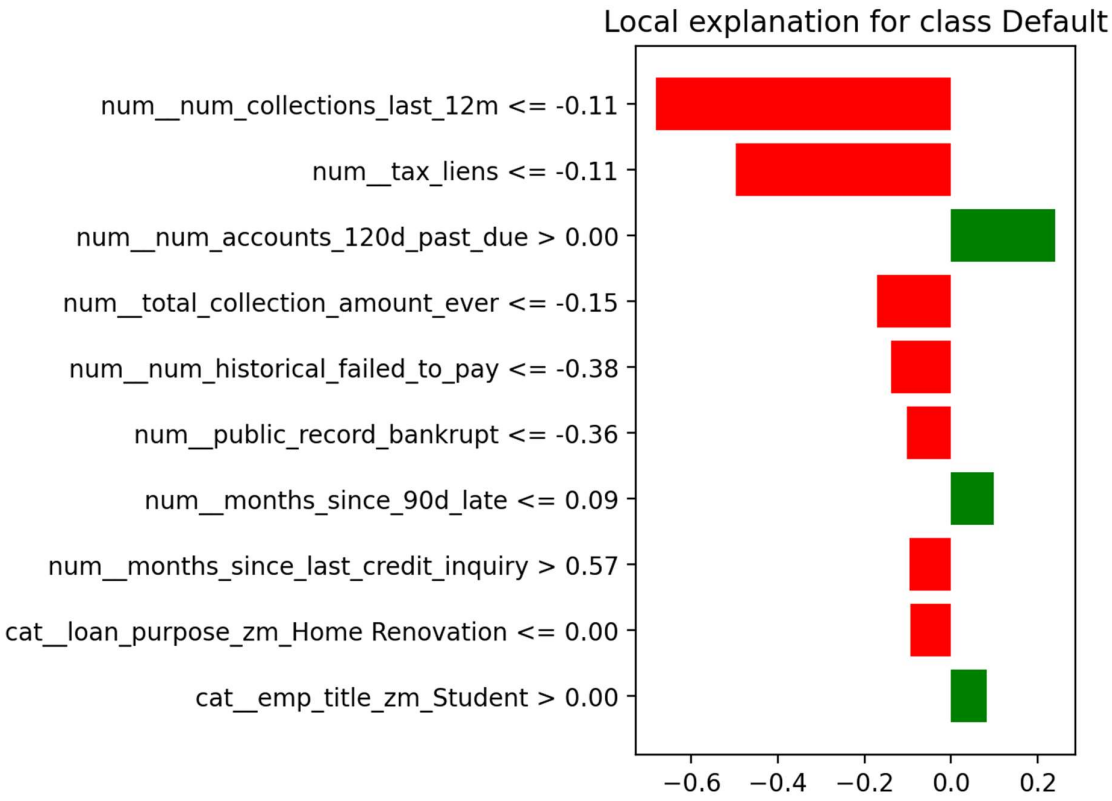
Note: The poor recall and F1-score for the 'Default' class in the Random Forest model highlight the challenge of imbalanced data, even with a high ROC AUC.

Appendix C

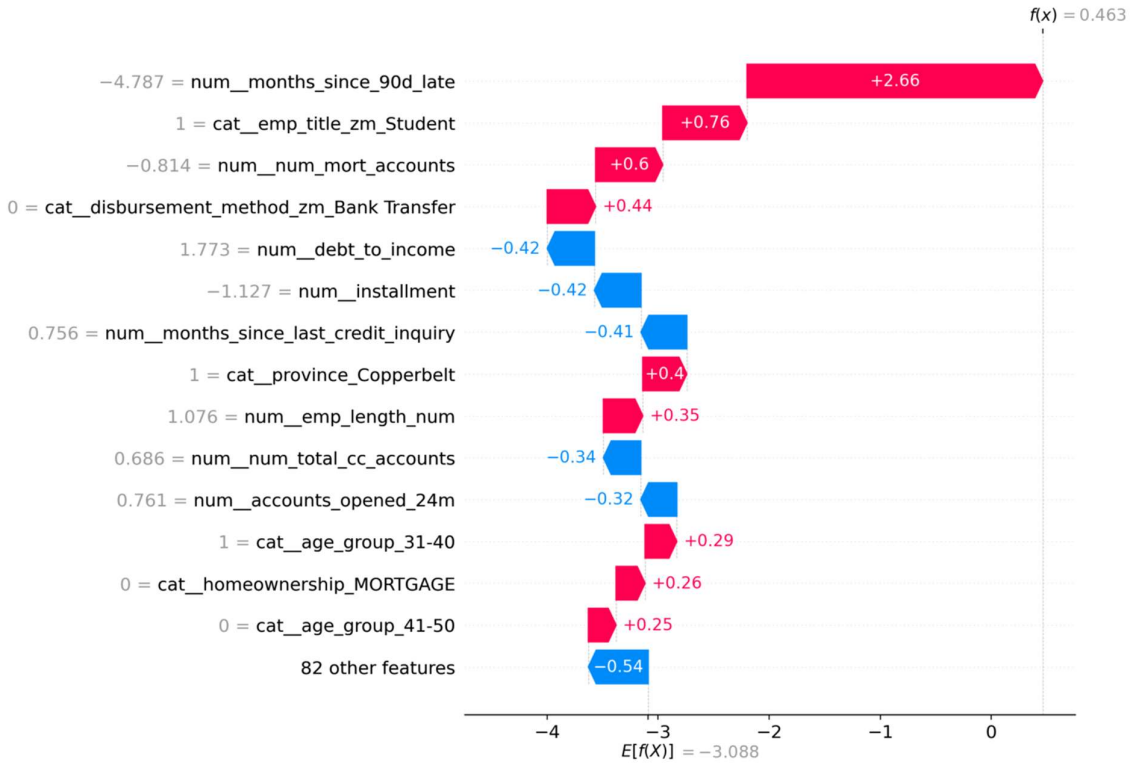
Sample Model Explanations

This section includes additional visualizations generated by the model.

- **C.1. Sample LIME Explanation:** A clear image of a LIME plot generated from the Loan Officer portal.



- **C.2. SHAP Dependence Plot:** A SHAP dependence plot for a key feature month Since 90d Late to show how its value impacts the model's output across the dataset.


f(x) = 0.463
−4.787 = num__months_since_90d_late
1 = cat__emp_title_zm_Student
−0.814 = num__num_mort_accounts
0 = cat__disbursement_method_zm_Bank Transfer
1.773 = num__debt_to_income
−1.127 = num__installment
0.756 = num__months_since_last_credit_inquiry
1 = cat__province_Copperbelt
1.076 = num__emp_length_num
0.686 = num__num_total_cc_accounts
0.761 = num__accounts_opened_24m
1 = cat__age_group_31-40
0 = cat__homeownership_MORTGAGE
0 = cat__age_group_41-50
82 other features
+2.66
+0.76
+0.6
+0.44
−0.42
−0.42
−0.41
+0.4
+0.35
−0.34
−0.32
+0.29
+0.26
+0.25
−0.54
−4
−3
−2
−1
0
E[f(X)] = −3.088

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[2] D. E. W. Platt, "Credit scoring using artificial neural networks," *Journal of the Operational Research Society*, vol. 42, no. 1, pp. 39–49, 1991.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] K. Barredo Arrieta, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[7] A. Bostandjiev, "Transparency in Credit Scoring: Models, Stakeholders and Regulatory Landscape," *SSRN*, 2021. Accessed: Oct 26, 2023. [Online]. Available: https://dx.doi.org/10.2139/ssrn.3824434

[8] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

[9] S. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.

[10] A. Philippon, "The Fintech Opportunity," *National Bureau of Economic Research*, Working Paper 22492, 2016. DOI: 10.3386/w22492

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[12] T. Kamiran and T. Calders, "Data preprocessing techniques to mitigate discriminatory classification," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–32, 2012.

[13] A. Mosquera, P. Wicker, and A. Helfert, "Towards trustworthy AI: quantifying and benchmarking explainable AI methods," *arXiv preprint arXiv:2107.09423*, 2021.

[14] B. Hohman and D. Gotz, "Visual analytics in healthcare: opportunities and research challenges," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 832-842, 2020.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315-3323.

[16] L. Hentschel, "The Impact of Explainable AI on User Trust and Acceptance," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1-12.

[17] "AI Act: Council and Parliament strike a deal on the world's first comprehensive AI law," *European Council - Press Release*, 2023. Accessed: Oct 27, 2023. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/ai-act-council-and-parliament-strike-a-deal-on-the-world-s-first-comprehensive-ai-law/

[18] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325-333.

[19] P. Zardini, G. Ranaldo, and A. Baumann, "Machine learning for credit default prediction: A survey," *European Journal of Operational Research*, vol. 295, no. 3, pp. 807-821, 2021.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135-1144.

[21] Fair Credit Reporting Act (FCRA), 15 U.S.C. § 1681 et seq.

[22] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018.

[23] S. Wachter, B. Mittelstadt, and C. Russell, "Transparent, explainable, and accountable AI for robotics," *Science Robotics*, vol. 2, no. 6, eaam9772, 2017.

[24] T. Calders and I. Zliobaite, "Calibrating probability estimates to achieve fairness," in *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 111-118.

[25] T. Beck and I. de la Torre, "Financial inclusion–what have we learned so far?," *IZA Journal of Labor Economics*, vol. 6, no. 1, pp. 1-36, 2017.

[26] B. D. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501-507, 2019.

[27] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1-15, 2018.

[28] D. Arntz, M. Bennewitz, J. Horn, and C. Bockermann, "Explainable AI in Credit Risk Management: A Systematic Literature Review," *SSRN*, 2023. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412378

[29] A. Bhatt, V. Xiang, A. K. Qin, and H. Abbass, "Explainable AI in Finance: A Survey," *ArXiv*, 2023. Available: https://arxiv.org/abs/2301.00543

[30] J. Fürnkranz, D. Gamberger, and M. Petković, "Bias in machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 5, pp. 387-395, 2012.

[31] A. Narayanan, J. Reifman, and S. Shmatikov, "A critical review of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[32] A. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of the 9th Conference on Innovations in*

Theoretical Computer Science, 2018, pp. 1-14.

[33] M. Kamishima, T. Akimoto, and S. Yabe, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 35-50.

[34] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, no. 2-3, pp. 271-274, 1998.

[35] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[36] M. Mitchell, *et al.*, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220-229.

[37] S. Raji, *et al.*, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020, pp. 229-239.

[38] C. Dwork, M. Hardt, T. Pitassi, N. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214-226.

[39] R. Goebel, *et al.*, "Explainable AI: the new 42?," *Artificial Intelligence*, vol. 275, pp. 195-227, 2018.

[40] M. Zerilli, J. Knott, M. Hallinan, M. Kieseberg, and F. Restuccia, "Transparency in algorithmic and human decision-making: Is there a difference?," *Big Data & Society*, vol. 8, no. 1, 2021.

[41] H. Kim, "Interpretability and fairness in machine learning for credit risk management," *Journal of Risk Model Validation*, vol. 14, no. 3, pp. 1-25, 2020.

[42] A. Agrawal, A. Hind, and S. Weber, "The economics of credit scoring," *Annual Review of Economics*, vol. 13, pp. 235-262, 2021.

[43] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

[44] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 3, pp. 523-541, 2007.

[45] E. Caruana, *et al.*, "Intelligible models for healthcare: predicting pneumonia risk with a bayesian network," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 793-800.

[46] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[47] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI '16)*, Savannah, GA, USA, Nov. 2016, pp. 265-283.

[48] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quart.*, vol. 13, no. 3, pp. 319–340, Sep. 1989.

[49] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quart.*, vol. 27, no. 3, pp. 425–478, Sep. 2003.

[50] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[51] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019.

[52] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," in *Advances in Neural Information Processing Systems*, 2016, pp. 2280-2288.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[54] A. Verma, "Explainable AI in the financial services industry," *The Journal of Financial Data Science*, vol. 3, no. 3, pp. 100-111, 2021.

[55] S. M. Ali, M. A. Shabbir, M. U. Ghani, and R. Al-Hmouz, "A comparative study of machine learning and explainable AI for credit scoring," *IEEE Access*, vol. 9, pp. 136155-136173, 2021.

[56] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153-163, 2017.

[57] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, pp. 962-970.

[58] A. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[59] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[60] J. He, S. Liu, K. T. N. Lam, I. D. S. R. Long, and L. K. M. Y. Tang, "Practical lessons from predicting clicks on ads at Facebook," in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 2014, pp. 1-9.

[61] J. S. Brownlee, "Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python," Machine Learning Mastery, 2020.

[62] A. P. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 27, no. 1, pp. 75-105, 2004.

[63] P. N. K. Lee, S. S. S. H. S. S. S. Cha, and K. Y. S. Park, "The role of explanation in the user acceptance of machine learning models," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 415-425, 2020.