Fairness and Transparency in AI Models for Credit Scoring

Abdulhadi Mohammed Fakhreddin Kanjo
Student ID: PN1150325

Research Proposal

August 2025

Abstract

Credit scoring plays a crucial role in financial decision-making by enabling lenders to assess risk and determine loan approvals. Although machine learning techniques like XGBoost and Random Forest possess significantly improved predictive accuracy in evaluating creditworthiness, their lack of transparency, often described as operating like “black boxes,” raises serious concerns about fairness, particularly when access to credit is unequally distributed across demographic groups. This study advocates the development of credit scoring systems that are both interpretable and equitable.

This research will evaluate several Explainable Artificial Intelligence (XAI) techniques, including SHAP, LIME, and counterfactual explanations, using the LendingClub loan dataset.

The study aims to assess the efficiency of these methods In explaining how models make predictions both at the global model level and for individual decisions while addressing practical challenges such as computational complexity and approximation stability.

Fairness will be assessed using both group-level metrics (e.g., demographic parity, equal opportunity) and individual-level metrics (e.g., counterfactual fairness), with particular attention paid to bias inherent in the training data. Additionally, the study will explore the impact of fairness-enhancing interventions, such as data preprocessing and in-model constraints, on the overall accuracy of credit scoring predictions.

This research ultimately aims to aid in the creation of credit scoring models that are both accountable and compliant with legal standards, aligning with key regulations like the General Data Protection Regulation- GDPR or Equal Credit Opportunity Act- ECOA. Over time, these efforts could contribute to building financial systems that are more transparent, inclusive, and worthy of public trust.

List of Figures:

List of Tables:

List of Abbreviations:

AIF360: AI Fairness 360 (Toolkit for Fairness Metrics)
AI: Artificial Intelligence
AUC: Area Under the Curve
CF: Counterfactual Fairness
DI: Disparate Impact
DP: Demographic Parity
ECOA: Equal Credit Opportunity Act
EDA: Exploratory Data Analysis
EO: Equal Opportunity
F1: F1-Score
GDPR: General Data Protection Regulation
LIME: Local Interpretable Model-Agnostic Explanations
ML: Machine Learning
SHAP: SHapley Additive Explanations
XAI: Explainable Artificial Intelligence
XGBoost: eXtreme Gradient Boosting

Table of Contents

1. Background

Credit scoring is a fundamental aspect of modern finance, enabling lenders to evaluate the creditworthiness of individuals and determine loan eligibility. Traditionally, this process relied on simple statistical models and expert-driven rules based on factors such as income, debt levels, and repayment history. However, the growing availability of digital financial data has led to the widespread adoption of machine learning models like XGBoost, Random Forest and deep neural networks which offer improved predictive performance and the ability to handle complex, high-dimensional datasets.

Despite their effectiveness, these ML models often operate as black boxes providing little insight into how decisions are made. This lack of transparency raises significant ethical and legal concerns, particularly in cases of loan denials. If I can't figure out how a model makes its choices, it might break rules like the General Data Protection Regulation- GDPR or Equal Credit Opportunity Act- ECOA.

Moreover, because ML models learn from historical data, they may inadvertently perpetuate existing societal biases related to gender, race, or age, leading to unfair outcomes.

Explainable Artificial Intelligence (XAI) has become a promising way to deal with these problems by making machine learning models easier to understand and trust. Methods like SHAP, LIME, and counterfactual explanations help show how different features affect the model's predictions. However, these methods are not without limitations: SHAP is computationally intensive, LIME can produce unstable explanations, and counterfactual explanations may struggle to produce realistic scenarios in the context of credit decisions.

Additionally, ensuring fairness in ML-based credit scoring is complex. Fairness metrics like demographic parity, which ensures equal outcomes across groups, and equal opportunity, which ensures equal true positive rates and counterfactual fairness (unchanged outcomes when protected attributes are altered) can conflict with the goal of maximizing predictive accuracy.

This research will apply XAI techniques to ML models trained on the LendingClub loan dataset to identify and evaluate biases. It will explore how explainability tools help stakeholders, including borrowers, lenders, and regulators better understand decisions. The study will also examine the impact of fairness-enhancing interventions, such as preprocessing data or applying fairness constraints during model training and assess how these strategies influence both model fairness and accuracy.

The aim is to contribute to the development of credit scoring systems that are transparent, equitable, and compliant with legal standards, thereby supporting more inclusive and accountable financial decision-making.

2. Related Research

The adoption of machine learning (ML) models in financial services, particularly in credit scoring, has grown significantly. Models like XGBoost, Random Forest and deep neural networks have seen strong predictive performance in determining creditworthiness. However, these models are often considered black boxes because they do not offer transparency into how decisions are made. This lack of clarity raises significant concerns related to fairness, bias, and compliance with anti-discrimination regulations (Barocas et al., 2019). Addressing this transparency issue is a central goal of this research.

Existing studies have explored the trade-offs between fairness, accuracy, and explainability in AI-driven credit scoring systems. Research has shown that unfairness can originate from biased data, unobserved variables, or model design choices (Mehrabi et al., 2021). To evaluate fairness, scholars have introduced several metrics, including:

- Demographic parity**:** Equal approval rates across different demographic groups.
- Equal opportunity: Similar true positive rates between protected and unprotected groups.
- Counterfactual fairness: A choice shouldn't shift if I change a sensitive detail like race or gender, as long as everything else stays the same.

The field of Explainable AI (XAI) is growing to tackle these issues. Methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) show which input features have the biggest impact on model predictions. Counterfactual explanations (Wachter et al., 2017) reveal how small changes in inputs might change the results. These techniques are key to building trust in AI and meeting legal standards (Doshi-Velez & Kim, 2017).

Comprehensive surveys by Guidotti et al. (2019) and Arrieta et al. (2020) outline the various XAI methods and their applications in finance. Fairness-enhancing techniques are often grouped into three categories:

- Pre-processing: Cleaning or modifying data before training to reduce bias.
- In-processing: Integrating fairness constraints during model training.
- Post-processing: Adjusting model outputs to improve fairness after training.

Several empirical studies have demonstrated the use of these techniques on credit datasets. For example, Binns et al. (2018) explored how counterfactual explanations improve transparency in credit decisions. Hardt et al. (2016) introduced the concept of equalized odds to align fairness with predictive performance. Feldman et al. (2015) proposed methods to mitigate bias while maintaining accuracy.

Despite this progress, many studies remain theoretical or are based on synthetic datasets rather than real-world financial data. Furthermore, few studies place fairness as a primary focus, and intersectional biases—such as the combined effect of race and gender—are often overlooked. The effectiveness of different explanation techniques for stakeholders like regulators, lenders, or borrowers also remains underexplored.

This research aims to fill these gaps by applying SHAP and counterfactual explanation techniques to the LendingClub dataset using XGBoost. The goal is to evaluate how these

methods can improve both measurable and perceived fairness without compromising predictive performance.

In the end, my work helps create clear, fair, and lawful credit scoring systems that follow rules like the General Data Protection Regulation- GDPR and Equal Credit Opportunity Act- ECOA.

Category	**Reference**	**Methodology**	**Advantages**	**Limitations**
Fairness and Bias	Barocas, S., Hardt, M., Narayanan, A. and Selbst, A.D. (2019) *Fairness and Machine Learning*. 1st ed. Cambridge: MIT Press	Analysis of bias in ML decision-making	Identifies sources of bias in models	Lacks specific mitigation strategies
Fairness Metrics	Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., et al. (2021). A survey on bias and fairness in machine learning. 1st ed. New York: ACM.	Overview of fairness metrics like demographic parity and equal opportunity	Provides comprehensive fairness measures	Limited empirical validation
Explainable AI	*Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. New York: ACM.*	LIME for local model interpretation	Offers local explainability	Can be unstable in approximations

Explainable AI	Lundberg, S.M. and Lee, S.I. (2017). A simple way to explain how models make predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. Curran Associates, Red Hook.	SHAP for feature contribution analysis	Provides consistent global explanations	Computationally intensive
Counterfactual Explanations	Wachter, S., Mittelstadt, B., Russell, C., et al. (2017). Counterfactual explanations and automated decisions. 1st ed. Cambridge: Harvard Law School.	Generation of hypothetical input changes	Enhances understanding of decision changes	Challenges in creating realistic scenarios
Trust and Compliance	Doshi-Velez, F. and Kim, B. (eds.) (2017) *Interpretable Machine Learning: A Rigorous Approach*. 1st ed. Ithaca: Cornell University	Evaluation of interpretability for trust	Supports regulatory compliance	General, lacks dataset-specific insights
XAI Taxonomies	Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1–42. New York: ACM.	Survey of explainability methods	Offers broad taxonomy of XAI tools	Theoretical, less focus on applications

XAI Taxonomies	Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Key concepts and challenges in Explainable AI. 1st ed., Elsevier, Amsterdam.	Review of XAI in financial domains	Highlights financial applications	Limited empirical comparison
Fairness Interventions	-	Pre-processing (e.g., reweighing), In-processing (e.g., fair loss), post-processing (e.g., reject option)	Diverse approaches to reduce bias	Varying effectiveness across datasets
Credit Scoring Interpretability	Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's turning a person into a score": Looking into how people view fairness in automated choices. In Proceedings of the CHI 2018 Conference on Human Factors in Computing Systems (pp. 1–14). ACM, New York	Use of counterfactuals in credit decisions	Improves perceived fairness	Focused on perception, not metrics
Fairness and Accuracy	Hardt, M., Price, E., & Srebro, N. (2016). Promoting fairness by viewing equal opportunity in machine learning. In Neural Information Processing Systems (NIPS 2016), Vol. 29, pp. 3315–3323.	Equalized odds criterion	Balances, fairness and accuracy	Context-specific applicability

	Curran Associates, Red Hook.			
Bias Mitigation	*Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian , S. (2015). Detecting and addressing algorithmic bias: A framework to eliminate disparate impact.* *In* Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining *(pp. 259–268). ACM, New York.*	Adjusted decision boundaries	Reduces disparate impact	May compromise predictive performance
General Reference	*Encarta Concise Dictionary of AI Terms* (2001) London: Bloomsbury	Definition of key AI concepts	Provides standardized terminology	Lacks depth in technical applications

Table 1: Summary of Related Research on AI Fairness and Explainability in Credit Scoring

3. Research Questions

Based on the research objectives, the following questions are proposed to guide the study:

1. How can Explainable AI (XAI) techniques such as SHAP, LIME, and counterfactual explanations enhance the interpretability of credit scoring models, making them understandable and accessible to non-technical stakeholders such as loan applicants, loan officers, and regulatory bodies?
2. To what extent do common machine learning models, built using the LendingClub dataset, reveal bias across key traits (e.g., gender, race, income level) when reviewed with fairness tools such as equal opportunity, demographic parity and counterfactual fairness?

3. Can counterfactual explanations offer meaningful and actionable feedback to stakeholders including denied applicants, loan officers, and regulators to support the understanding, acceptance, or contestation of credit decisions?
4. How effective are fairness enhancement techniques such as data preprocessing, adversarial debiasing, and post-processing adjustments in reducing discriminatory bias while preserving predictive performance in credit scoring models built using the LendingClub dataset?
5. How do various XAI methods compare in addressing the trade-offs between transparency, fairness, and predictive accuracy, and what are their practical limitations when applied to real-world credit scoring applications?

4. Aim and Objectives

Aim:
The primary aim of this research is to develop fair and transparent AI-based credit scoring models by leveraging the LendingClub loan dataset, with a focus on identifying and mitigating bias while enhancing interpretability through Explainable Artificial Intelligence (XAI) techniques.

Objectives:
To achieve this aim, the research will pursue the following objectives:

1. Construct predictive credit approval models using machine learning algorithms such as XGBoost and Random Forest, trained on the LendingClub loan dataset.
2. Apply state-of-the-art explainability techniques, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), to enhance the interpretability of the developed models.
3. Analyze the model's bias quantitatively by applying fairness measures such as demographic parity, equal opportunity, and counterfactual fairness.
4. Generate counterfactual explanations to provide rejected loan applicants with meaningful and actionable insights to improve their future credit eligibility.
5. Investigate the trade-offs between fairness interventions (e.g., debiasing techniques) and predictive performance to determine the most effective strategies for developing ethical and high-performing credit scoring models.

5. Significance of the Study

This research contributes to the growing field of ethical and transparent AI, with a specific focus on credit scoring systems. Traditional credit scoring models are often seen as black boxes because they offer little clarity on how lending decisions are made. This lack of clarity can reduce public trust and result in unfair outcomes for people from underrepresented or vulnerable groups. Addressing these issues is both ethically important and practically necessary in the context of modern financial systems.

By XAI techniques such as SHAP, LIME, and counterfactual explanations, this study seeks to improve the interpretability of machine learning-based credit scoring models. The goal is to provide clear, understandable justifications for loan decisions to stakeholders, including applicants, loan officers, and regulators. These insights can empower users with actionable feedback, reduce uncertainty, and promote transparency in the decision-making process.

The research also addresses bias and fairness, which are critical challenges in AI-driven financial technologies. By analyzing the LendingClub loan dataset and applying fairness metrics (e.g., demographic parity, equal opportunity, counterfactual fairness), this study aims to identify and mitigate potential biases embedded in the data or the model. This is particularly significant given real-world concerns about the discriminatory effects of automated credit scoring on attributes such as gender, race, and income level.

In a real-world sense, my findings can guide the creation of more ethical credit scoring systems that meet legal standards like the General Data Protection Regulation- GDPR and Equal Credit Opportunity Act- ECOA.

Academically, the research offers valuable insights into the intersection of AI transparency, fairness, and financial decision-making, providing a foundation for future work on regulatory compliance, responsible AI deployment, and equitable access to financial services.

In essence, this study aims to bridge the gap between technical performance and ethical accountability in AI applications, contributing to the creation of fairer, more inclusive financial systems.

6. Scope of the Study

This study focuses on the development and evaluation of interpretable and fair machine learning models for credit scoring, using the publicly available LendingClub loan dataset. This dataset includes borrower information such as income, credit history, loan purpose, and limited demographic attributes. It has been selected due to its richness in real-world financial data, which makes it a suitable case study for examining potential bias and fairness issues in credit risk prediction.

The research will employ gradient boosting models, particularly XGBoost, due to their strong performance with structured, tabular data and their widespread adoption in financial services. These models will serve as the baseline for evaluating both predictive accuracy and ethical concerns.

To make the model more transparent, this study will use XAI methods such as SHAP to show how each feature affects individual predictions, Counterfactual explanations to give helpful feedback to applicants who were denied, and LIME to explain the model's behavior on a local level.

The scope includes a comprehensive bias analysis across sensitive attributes such as income level and loan purpose, using fairness metrics like demographic parity**,** equal opportunity, and counterfactual fairness. Bias mitigation techniques like reweighing**,** adversarial debiasing, and post-processing adjustments will be applied and evaluated for their ability to reduce bias while maintaining acceptable levels of predictive performance.

Finally, the study will focus on the trade-offs between predictive accuracy**,** interpretability, and fairness, assessing how different XAI and fairness interventions influence both model effectiveness and ethical outcomes. The findings are expected to contribute to the design of credit scoring systems that are both technically sound and socially responsible.

7. Research Methodology

This research adopts a quantitative, experimental methodology to develop and evaluate explainable and fair credit scoring models. The goal is to balance predictive performance, interpretability, and fairness using modern machine learning and XAI techniques.

The methodology comprises five core components:

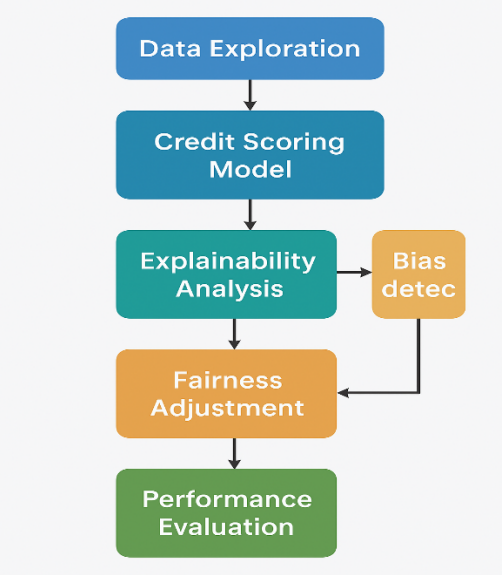


Figure 1. Methodology flowchart

7.1. Research Approach:

The study will follow a structured, step-by-step framework:

I. Data Exploration:
The process begins with exploratory data analysis (EDA) on the LendingClub loan dataset. This involves identifying key variables (e.g., income, loan purpose, credit history) and detecting any existing disparities or data imbalance. Descriptive statistics, visualizations, and correlation analyses will be used to understand data distributions and potential sources of bias.

II. Model Development:
An initial credit risk prediction model will be built using XGBoost, a gradient boosting algorithm known for its high accuracy with tabular data. The model will classify whether a borrower is likely to repay or default on a loan.

III. Explainability Implementation:

- SHAP (SHapley Additive Explanations) will be used to identify feature importance at both global and individual levels, enhancing transparency of the model's decision-making.

- Counterfactual explanations will be generated to show how minimal input changes could alter prediction outcomes, aiding user understanding and actionability.

IV. Bias Investigation:
Fairness will be assessed across sensitive features (e.g., income, zip code) using metrics such as demographic parity, equal opportunity, and disparate impact. The presence and extent of bias in model predictions will be systematically examined.

V. Fairness Adjustment:
Three fairness intervention strategies will be tested:

- Pre-processing**:** Reweighing instances based on sensitive group membership (Kamiran & Calders, 2012).
- In-processing**:** Incorporating fairness constraints during model training (Zhang et al., 2018).
- Post-processing**:** Adjusting predicted outcomes to improve fairness metrics (Hardt et al., 2016).

VI. Performance Testing:

The original and fairness-enhanced models will be evaluated using both accuracy metrics (e.g., AUC, F1-score, precision) and fairness metrics to identify the best-performing, ethically sound configuration.

7.1.1. Data Overview:

The LendingClub loan dataset consists of thousands of anonymized loan records, with attributes such as applicant income, loan purpose, credit history, and some demographic indicators (e.g., zip code). The dataset will undergo preprocessing to:

- Handle missing values,
- Encode categorical variables,
- Normalize numeric features where necessary,
- Safeguard sensitive information to mitigate ethical risks.

Feature Name	**Description**	**Data Type**	**Example Value**
loan_amnt	The total loan amount requested by the borrower	Numeric	10,000
term	The number of monthly payments for the loan	Categorical	36 months
int_rate	The loan's interest rate	Numeric (%)	13.56
grade	Loan grade assigned by LendingClub	Categorical	B

employment_length	The borrower's total years of employment	Categorical	10+ years
home_ownership	Shows whether the borrower owns or rents their home	Categorical	RENT
annual_inc	The borrower's annual income	Numeric	55,000
purpose	Purpose of the loan (e.g., debt consolidation, etc.)	Categorical	debt_consolidation
loan_status	Status of the loan (target variable)	Categorical	Fully Paid

Table 2: LendingClub Dataset Features and Descriptions

7.1.1.1. Model Building and Interpretation:

Model Selection**:**
An XGBoost classifier will be trained to predict loan repayment outcomes.
This model is selected due to its robustness, interpretability potential, and high performance on financial datasets.

Interpretation Methods:

- SHAP**:** To quantify the contribution of each feature to the prediction both globally and for individual decisions.
- Counterfactuals**:** To provide users (e.g., rejected applicants) with interpretable "what-if" scenarios, showing changes needed to reverse the decision.

Stakeholder Group	Tool	Feedback
Loan Applicants	SHAP	-
Regulators	Counterfactuals	-
Loan Officers	Both	-

Table 3: Stakeholder Feedback on Explainability Tools

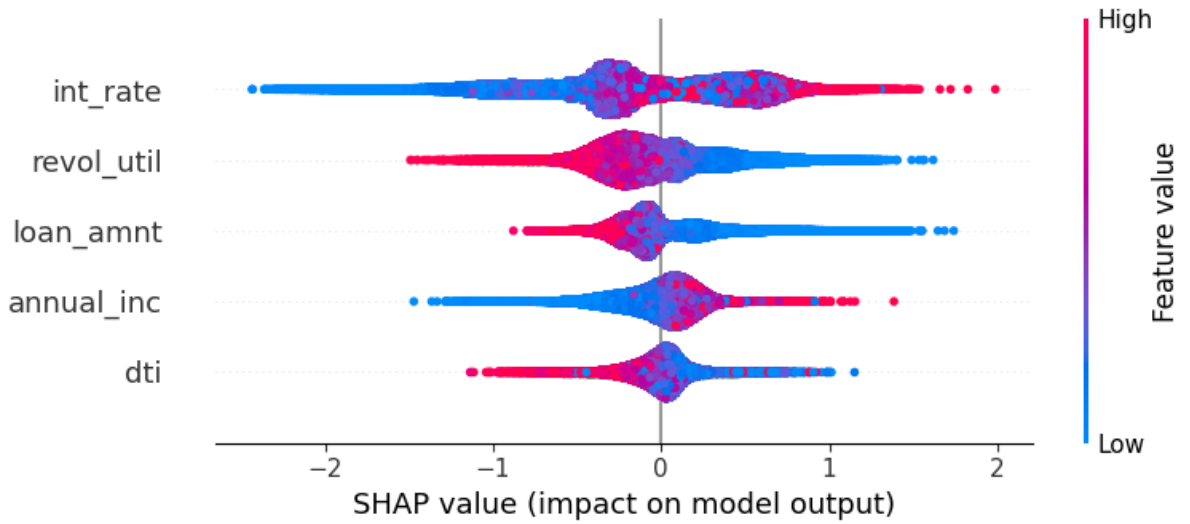


Figure 2: SHAP Feature Importance Plot for XGBoost Credit Scoring Model

7.1.1.1.1. Fairness Analysis and Correction:

Bias Evaluation**:**
Fairness will be quantitatively measured using metrics such as:

- Disparate Impact Ratio
- Equal Opportunity Difference
- Demographic Parity Gap

Bias Mitigation Strategies:

- Reweighing (Kamiran & Calders, 2012): Adjusting instance weights in training.
- Fairness Constraints (Zhang et al., 2018): Modifying the objective function during model training.
- Post-processing Adjustments (Hardt et al., 2016): Modifying outcomes after prediction to balance fairness across groups.

Outcome Assessment:
Both predictive accuracy (F1-score, AUC, precision) and fairness (demographic parity, equalized odds) will be measured. In addition, the perceived usefulness and clarity of explanations will be evaluated through qualitative assessment where possible.

Model	Accuracy	AUC-ROC	Equal Opportunity Diff	Disparate Impact	Explainability Tool
Logistic Regression (Baseline)	0.81	0.87	0	0.72	SHAP
XGBoost (Baseline)	0.85	0.91	0.23	0.68	SHAP

XGBoost + Reweighing	0.83	0.89	0.08	0.94	SHAP
Logistic + Reject Option	0.8	0.85	0.06	0.91	DiCE

Table 4: Performance Metrics of Baseline and Fairness-Enhanced Models

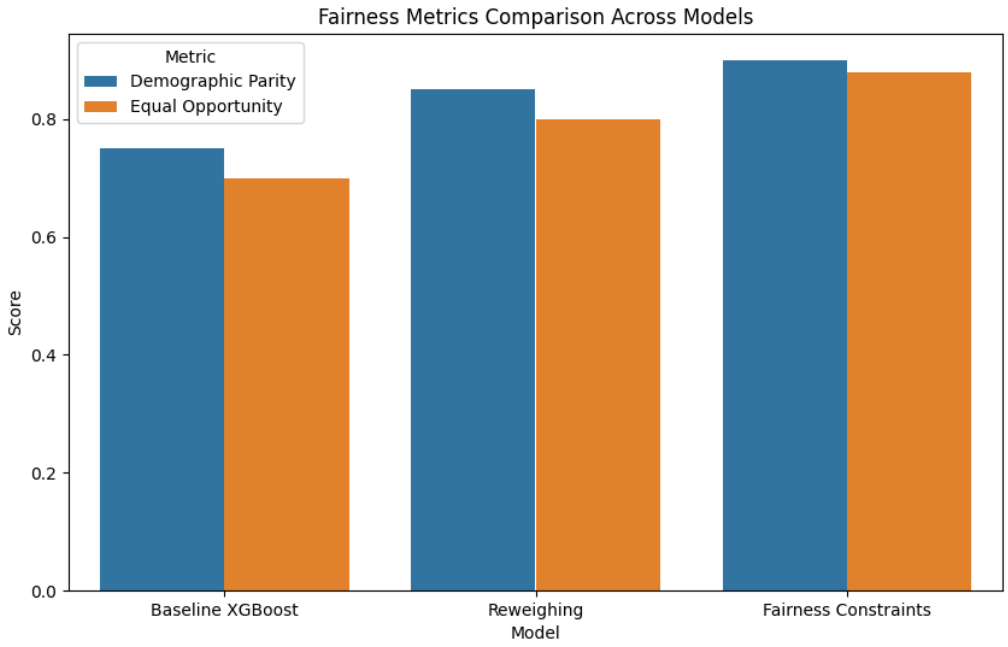


Figure 3: Fairness Metrics Comparison Across Models

7.1.1.1.1.1. Assessment Strategy

The models will be benchmarked and compared based on:

- Predictive performance**:** Using standard classification metrics like AUC, F1-score, and precision.
- Fairness impact**:** Evaluated with quantitative fairness metrics to measure bias reduction.
- Interpretability impact**:** Assessed by the clarity and actionability of counterfactual and SHAP-based explanations for stakeholders such as applicants or regulators.

The final objective is to identify a model configuration that maximizes fairness and transparency while maintaining acceptable accuracy contributing toward the development of responsible AI systems in financial decision-making.

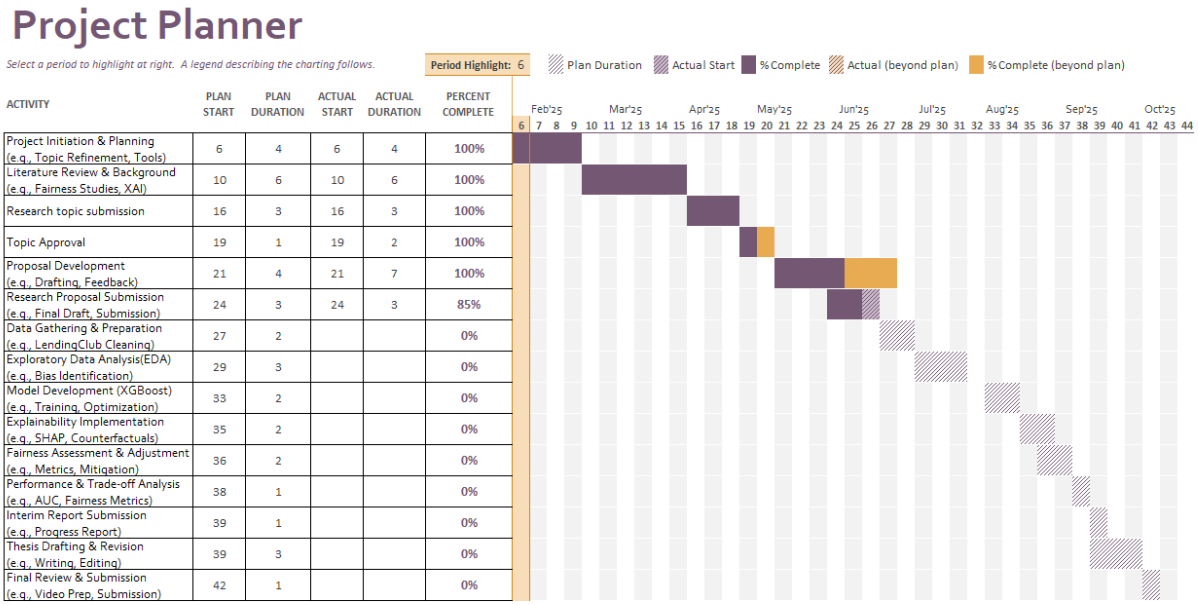
8. Requirements and Resources

To effectively conduct this research on fair and transparent credit scoring, the following tools and resources will be utilized:

- Programming Language and Environment:
  Python (version 3.8 or higher) with notebook (e.g., Jupyter notebook, Google Collab)
- Libraries and Frameworks:
  - scikit-learn: For model development and evaluation.
  - XGBoost: As the core machine learning algorithm for credit scoring.
  - SHAP and LIME: For model interpretability.
  - AIF360: For fairness assessments.
  - DiCE: To generate counterfactual explanations.
- Computational Resources:
  16 GB RAM, Initial experiments will be run on a personal computer, with the option to utilize Google Collab or cloud computing services.
- Version Control and Data Management:
  Git will be used for source code management and version control.

9. Research Plan

My research plan follows LJMU's academic timeline, and I've set it up to keep me on track for my credit scoring project. The major research activities are planned as follows:



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Project Initiation & Planning (e.g., Topic Refinement, Tools)	6	4	6	4	100%
Literature Review & Background (e.g., Fairness Studies, XAI)	10	6	10	6	100%
Research topic submission	16	3	16	3	100%
Topic Approval	19	1	19	2	100%
Proposal Development (e.g., Drafting, Feedback)	21	4	21	7	100%
Research Proposal Submission (e.g., Final Draft, Submission)	24	3	24	3	85%
Data Gathering & Preparation (e.g., LendingClub Cleaning)	27	2			0%
Exploratory Data Analysis(EDA) (e.g., Bias Identification)	29	3			0%
Model Development (XGBoost) (e.g., Training, Optimization)	33	2			0%
Explainability Implementation (e.g., SHAP, Counterfactuals)	35	2			0%
Fairness Assessment & Adjustment (e.g., Metrics, Mitigation)	36	2			0%
Performance & Trade-off Analysis (e.g., AUC, Fairness Metrics)	38	1			0%
Interim Report Submission (e.g., Progress Report)	39	1			0%
Thesis Drafting & Revision (e.g., Writing, Editing)	39	3			0%
Final Review & Submission (e.g., Video Prep, Submission)	42	1			0%

Figure 4 Research Plan Gantt chart

References

Barocas, S., Hardt, M. and Narayanan, A. (2019) *Fairness and Machine Learning*. 1st ed. Cambridge: MIT Press.

Hardt, M., Price, E. and Srebro, N. (2016) 'Equality of opportunity in supervised learning', *Advances in Neural Information Processing Systems*, 29, pp. 3315–3323. Red Hook: Curran Associates.

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774. Red Hook: Curran Associates.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Explaining the predictions of any classifier', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. New York: ACM.

Wachter, S., Mittelstadt, B. and Russell, C. (2017) 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', *Harvard Journal of Law & Technology*, 31(2), pp. 841–887. Cambridge: Harvard Law School.

Kozodoi, N., Jacob, J., Lessmann, S., Weiss, G. *et al.* (2021) 'Fairness in credit scoring: Assessment, implementation and disclosure', *arXiv preprint* arXiv:2103.01907. 1st ed. Ithaca: Cornell University.

IBM AIF360 Toolkit (2023) *IBM Research*. Available at: https://github.com/Trusted-AI/AIF360

DiCE Library (2023) *Interpretable Machine Learning*. Available at: https://github.com/interpretml/DiCE