National College of Ireland

Fair and Interpretable Credit Risk Modelling with Multi-Agent Architecture and SHAP

MSc Research Project
Data Analytics

Prajwal Suhas Pusadkar
Student ID: 23304367

School of Computing
National College of Ireland

Supervisor: Prof. Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Prajwal Suhas Pusadkar
Student ID:	23304367
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Mr. Hicham Rifai
Submission Due Date:	15/09/2025
Project Title:	Fair and Interpretable Credit Risk Modeling with Multi-Agent Architecture and SHAP
Word Count:	9640
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Prajwal Suhas Pusadkar
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Fair and Interpretable Credit Risk Modeling with Multi-Agent Architecture and SHAP

Prajwal Suhas Pusadkar
23304367

Abstract

Concerns continue to arise regarding matters such as fairness, transparency, and accountability while credit scoring and loan approval processes are modernized with the introduction of machine learning (ML) models. This project develops a loan approval system that tackles explainability and fairness with a multi-agent architecture and traditional ML models, integrated with explainable artificial intelligence (XAI) techniques. The system consists of four sequential agents: risk scoring (Agent A), credit limit recommendation (Agent B), loan decisioning (Agent C), and explainability (Agent D). Each agent automates distinct tasks aligned with the workflows within the financial industry, forming a cohesive modular pipeline with integrated interpretability. A model was built and trained on a credit dataset with XGBoost, and explainability was integrated with SHAP to yield global and local insights. To enable bulk application uploads with instantaneous prediction, explanation, and fairness audit retrieval, a frontend was built with Streamlit. The system also measures demographic fairness across gender, education, and marital status. This proves that accurate and explainable credit risk models are possible while simultaneously being fair and balanced, with built-in accountability features. The SHAP-x MAS modular architecture supports future deployment to decision aid systems.

1 Introduction

In today's digitally driven financial environment, automated systems for determining credit eligibility form the backbone of how banks, fintech companies, and online lenders evaluate the credit profile of a consumer or business. Artificial intelligence (AI) and ML (Machine Learning) have streamlined the processing of customer data for discerning risk patterns and making informed decisions at speed, from processing credit applications to underwriting mortgage and business loans. This has enhanced operational efficiency and created greater financial inclusion for the previously underserved population. On the other hand, the increasing sophistication ML systems are bringing forth intricate and multidisciplinary ethical and regulatory concerns, particularly scrutiny on fairness, transparency, and explainability.

1.1 The Rise of Machine Learning in Credit Risk Modeling

Traditional practices around credit scoring relied on strict methodologies built on a dataset's income, debt, and payment patterns, with repayment history serving as a timeline.

These models, while easy to interpret and favored by regulators, do not align with today's complex financial landscape due to their inability to capture intricate behaviors and interdependencies in data.

Modern financial ecosystems rely on more advanced credit scoring methodologies, which incorporate algorithms and ML techniques, including XGBoost, and even neural networks to capture complex data relationships and outperform legacy systems. These methodologies unlock hidden crucial signals in expansive datasets, including credit history records, transactions, and other behavioral data. Recently, an increase in data availability, coupled with regulatory sandboxes for financial innovation, open-source projects, and cloud infrastructure, has further facilitated the use of these models. Even with all these advancements, the financial ecosystem often criticizes these advanced methodologies for their opacity, especially in scenarios involving automated loan rejections. Explanations for automated decision-making systems, especially those that affect marginalized communities, have become a regulatory focal point. As a result, credit institutions increasingly need to construct models that fulfill the dual requirements of precision and transparency.

1.2 Problem Statement and Motivation

The implementation of opaque or excessively intricate models within lending practices may deepen historic inequities. Research indicates that machine learning algorithms can perpetuate disparities in the access of credit on the basis of gender, ethnicity, and even education level, due to training on biased data, even in cases where sensitive attributes are removed from the feature set Kozodoi et al. (2022), Vieira et al. (2025). In addition, lenders have legal obligations to explain their adverse decisions due to frameworks like ECOA in the U.S. or GDPR in Europe.

These reasons highlight the rigorous requirement of sustaining bias-free and transparent AI systems within credit-risk modeling as well as adaptable systems. Such systems are expected to meet the following criteria:

- **Fairness**: Decisions must be evaluated for demographic bias and disparate impact.
- **Transparency**: Stakeholders (including applicants) should understand how and why a decision was made
- **Auditability**: The decisioning process should be traceable and reviewable by compliance teams or regulators.
- **Deployability**: Models should be modular and scalable, easily integrated into banking systems.

To address these requirements, this project proposes a novel multi-agent system (MAS) architecture for credit risk modeling that balances performance with interpretability and ethical compliance.

1.3 Research Questions

This project seeks to answer the following core research questions:

1. How can a credit scoring model maintain interpretability while leveraging powerful ML algorithms like XGBoost?

2. Can loan approval decisions be broken down into modular, explainable components through a multi-agent system?

3. What is the impact of such a system on fairness across protected demographic groups (e.g., gender, education)?

4. Can explainability tools like SHAP provide local and global insights suitable for stakeholder communication and audit?

These questions reflect both the technical and ethical dimensions of AI-driven decision systems in financial services.

1.4 Scientific Contributions

This work contributes to the interdisciplinary field of ethical and interpretable AI in several key ways:

- Novel MAS Design for Credit Scoring: This new system embodies a role-specific agent design, modeled after the distributed decision-making processes in actual banks, like in credit analysis, risk compliance, and underwriting, as opposed to studies that rely on centralized model logic or use a federated approach.

- End-To-End Explainability: Decisions such as limit assignment which utilizes SHAP values, alongside providing natural language explanations, disable post hoc reasoning and become integrated in more active decision processes, and thus need to be reasoned throughout the decision pipeline.

- Built-in Fairness Auditing: This system automatically marks biases and allows group comparison visualization, from which the user can analyze fairness of the system metrics across gender, marital status, education, and use demographic parity as a baseline.

- Deployable Dashboard: The system can be demonstrated and used in the future due to the use of Streamlit. This makes it possible for non-technical users to interact with model's decision and fairness outcomes in real time.

While many works separately tackle explainability or fairness in finance, this project uniquely combines both into a cohesive, modular system that is aligned with real-world needs, stakeholder transparency, and regulatory trends.

1.5 Structure of the Report

The remainder of this thesis is organized into the following chapters:

- **Chapter 2: Literature Review** Reviews academic and industry research on credit scoring models, fairness in machine learning, explainability (especially SHAP), and multi-agent system applications in finance.

- **Chapter 3: Methodology** Describes the dataset, feature engineering steps, model selection, and the design of the MAS architecture.

- **Chapter 4: Implementation** Details how each agent was implemented and integrated. Includes SHAP visualizations, Streamlit UI, and batch processing logic.

- **Chapter 5: Evaluation** Presents model performance metrics (accuracy, AUC), fairness evaluations (demographic parity), and explainability outcomes across applicant groups.
- **Chapter 6: Conclusion and Future Work** Summarizes key insights, discusses limitations, and outlines potential improvements such as federated retraining, adversarial robustness, and policy integration

2 Related Work

Earlier, credit scoring was done using scorecards created by experts and linear statistical models, but in the recent decades there has been a shift towards machine learning (ML) techniques for better accuracy in prediction. There have been many studies that demonstrate the effectiveness of advanced ML techniques as compared to logisitc regression and linear models for estimating the probability of default (PD). For instance, systematic reviews have shown that deep learning techniques greatly outperformed the classic statistical algorithms, guaranteeing a better predictive accuracy on a multitude of loan portfolios Nwafor et al. (2024). This increase in accuracy can be explained by the ability of ML to capture complex, non-linear relationships as well as interrelations between borrowers that simpler models are unable to. For instance, a new hybrid model that combines convolutional neural networks and gradient boosting was shown to do better than both standalone neural nets and logistic regression models on peer-to-peer lending data Kozodoi et al. (2022). The ability of ML to model complex non-linear interactions in data has shown the possibility of borrowing strategies predicting subtle risk patterns better than traditional techniques and improving risk differentiation.

The implementation of credit scoring powered by Machine Learning, or ML, is limited by a number of challenges. Industry specialists and compliance officers have highlighted issues regarding the "black-box" characteristic of ML systems, which makes it hard to shed light on the reasoning behind certain lending choices. Due to the need to communicate with customers and because of compliance, banks have opted for traditional models, which makes the shift to ML more reserved than anticipated. Additionally, some studies have indicated that the "predictive power" of some models may plateau or even reverse if the models are overfit, or trained on biased data. This is concerning, as it raises the need to further overfit data. Within the literature, it is clear the different techniques introduced by ML have explicit benefits to the predictive accuracy as highlighted in Nwafor et al. (2024)Kozodoi et al. (2022). However, the challenges provided by governance, the need to explain and interpret the models, and the risk of overfitting to historical data patterns are equally as complex. Due to these complexities, the literature focuses on fairness and explainability regarding the credit risk models.

2.1 Fairness in Automated Loan Decisions

With the implementation of ML models for credit evaluation, algorithmic fairness has emerged as a pressing issue. The construction of Automated Lending Models (ALMs) has received considerable attention, as researchers increasingly note the potential for historical biases in lending data to be perpetuated or even exacerbated by ALMs. For instance, one recent study showed that a machine learning model for mortgage approval perpetuated historical inequalities by reproducing racial disparities in mortgage approval,

thereby exacerbating the existing divide in the approval rates of different demographic groups Vieira et al. (2025). These findings are consistent with the growing concern that, in the absence of proper checks, credit algorithms designed to optimize efficiency and streamline processes tend to capture and perpetuate existing biases in society—regardless of whether sensitive attributes like race or gender are included. It is widely recognized that ML models "learn" decisions based on prior choices, many of which are steeped in bias due to discriminatory practices or structural inequalities, "learning" bias from data unless proactive measures are taken to counteract that. Research has shown that neutral attributes can be prejudiced as proxies for protected characteristics and embed discriminatory outcomes Bussmann et al. (2020).

Table 1: Comparison of Fairness Mitigation Techniques in Credit Scoring

Study [Ref]	**Model Used**	**Fairness Metric**	**Mitigation Method**	**Trade-off with Accuracy**
Kozodoi et al. (2022)	1D-CNN + XGBoost	Statistical Parity	Exclude sensitive features	None observed
Vieira et al. (2025)	Various (review)	Multiple (SP, EO)	Pre-processing	Minimal loss
Nallakaruppan et al. (2024)	Random Forest	Demographic Parity	Fairlearn (post-processing)	Slight trade-off
AI (2019)	Multi-agent ensemble	Implicit (consensual decisions)	Argumentation	Performance improved

In response, the literature on fair lending algorithms has expanded rapidly. Multiple definitions of fairness have been proposed (e.g., Equal Opportunity, statistical parity, disparate impact), and scholars have developed techniques to measure and mitigate bias at different stages of the modeling pipeline. A recent systematic review of bias mitigation in credit ML models found that most studies focus on outcome fairness and bias in the input data, with pre-processing techniques (e.g. rebalancing or obscuring protected attributes) being the most common approach to improving fairness. These methods, such as reweighting training data or learning debiased representations, have shown success in reducing disparities in loan approvals. Interestingly, the review noted that about 75 percent of studies evaluated fairness with only a single metric, highlighting a lack of consensus on measurement and the need for standardized multi-metric evaluations Nallakaruppan et al. (2024). Moreover, few credit ML studies to date have incorporated causal inference or addressed intersectional fairness (combined effects of multiple protected attributes), indicating methodological gaps.

The growing concern around fair lending algorithms has led to an explosion of new literature. An Equal Opportunity, statistical parity, or even disparate impact can all be defined as fairness. Researchers have devised methods to quantify bias and mitigate it in various stages of a given model's pipeline. An overwhelming number of studies focus on fairness concerns and bias related to the data in credit ML models. Most studies focus on fairness of outcomes and bias in the input data, with pre-processing techniques (e.g. rebalancing or obscuring protected attributes) as the most common approach to improving fairness Kozodoi et al. (2022). In the context of loan approval disparities, these methods (e.g. reweighting training data or learning debiased representations) have

been successful. Interestingly, the review noted that around 75 percent of studies only evaluated fairness with a singular metric, demonstrating the lack of consensus on measurement and the urgency for multi-metric standardized evaluations Nallakaruppan et al. (2024). Furthermore, the few studies that do incorporate causal inference focus on intersectional fairness (the combined effects of several protected attributes) highlighting the gaps in methodologies.

Additionally, the way model decisions are explained to users can influence perceptions of fairness – studies have found that providing clear, domain-relevant explanations for credit decisions improves users' sense of procedural justice and trust in the model Hjelkrem and de Lange (2023). This connection between explainability and perceived fairness has increased interest in techniques that can do both: make models more transparent and more just.

2.2 Explainability and SHAP in Finance

The lack of clarity regarding some ML models has sparked significant interest in explainable AI (XAI) in the context of finance. In credit risk assessment, explainability is not merely an optional attractive feature, but a legal and ethical obligation; lenders need to justify the decisions they make to the regulatory bodies and to the borrowers, particularly in adverse action scenarios. With the growing adoption of ML models in underwriting, both researchers and practitioners began using model-agnostic explanation techniques to generate human-explainable model-agnostic insights from "black-box" models Hjelkrem and de Lange (2023). One of the most prominent approaches is Shapley Additive Explanations (SHAP) which uses concepts of cooperative game theory to explain a model's prediction in relation to its input features by attributing each feature a "contribution" value which indicates its influence on the prediction. These values can be aggregated to explain the model's conduct. SHAP was introduced to the ML community by Lundberg and others as a unified framework with a solid theoretical backbone (Shapley values) and feature attribution Lee et al. (2023). SHAP is widely accepted in the credit context due to the ease of interpreting the explanation and its intuitive nature, which is in contrast with the domain specifics (i.e., "high debt-to-income ratio contributed +0.20 to the default risk score").

Finance researchers have shown the applicability of SHAP and related XAI techniques in various contexts. For instance, in the analyses of bank loan datasets, SHAP has been applied to determine the strongest contributing factors to defaults and has shown that recent payment delinquencies and credit utilization tend to drive the model's results the most Zheng et al. (2020). These explanations assist risk assessment managers in ensuring the model is employing reasonable factors (and not fallacious relationships) in making decisions. Additionally, XAI methods such as SHAP have the potential to reveal biases or inaccuracies in models: analysts examining feature attributions related to race or gender can identify if a model is biased in projecting influence due to race or gender, thus raising concerns of equity.

As SHAP values explain how sensitive and proxy features contribute to predictions, researchers have noted that they systematically explain how bias can be traced in lending models Kozodoi et al. (2022)Lee et al. (2023). Such explanations can be put to use for model fairness. In finance, other explainability techniques such as partial dependence plots, LIME, and counterfactual explanations have been used, but SHAP's industry adoption is due to its consistency and local accuracy. These tools are beginning to be

Table 2: Explainability Techniques in Credit Risk Models

Study [Ref]	**ML Model**	**XAI Tool**	**Explanation Type**	**Deployment Use**
Hjelkrem and de Lange (2023)	XGBoost	SHAP	Local/global	Human audit
Zheng et al. (2020)	LightGBM	SHAP	SHAP + reason codes	Regulatory
Lee et al. (2023)	Any	SHAP	Theoretical	Benchmark
AI (2019)	Agent ensemble	Argumentation	Semantic justification	Human-readable

adopted by financial regulators and institutions.

In order to comply with transparency regulations, some SHAP-aware lenders now auto-generate SHAP reason codes for declined loan applications, which simplify complex model reasoning into a customer-friendly format Hjelkrem and de Lange (2023)Zheng et al. (2020). While explainability techniques do not improve explainability, they provide needed checks in a balance for model governance, especially in human supervision of AI-powered credit decisions. Explainability also helps model builders, risk managers, regulators, and consumers align about a model's reasoning. This transparency helps in building confidence in automated systems for managing credit risk.

2.3 Federated and Modular Learning Architectures

Data privacy and competition among institutions make financial data difficult to access and often fragmented. Federated learning (FL) has become a practical approach to accessing data from several institutions (e.g., bank branches) as it permits each institution to retain sensitive customer data. In a federated system, each participant, in this case, each bank, trains the model on their data and shares model updates with a central aggregator which produces a global model. Recent studies have shown that FL can assist smaller lenders or community banks in developing credit risk models by participating in collaborative FL, enabling them to nearly match the model-determining power of large banks.Solaimani et al. (2025)

As an example, Lee et al. (2023) conducted a federated credit scoring pilot project and multiple institutions. They demonstrated that a model credit-scored collaboratively outperformed each institution's in-silo model significantly in default prediction accuracy. This model synergist approach is beneficial to institutions with insufficient data to build large datasets to improve their predictive model performance rapidly. Another study added to the review confirming increasing popularity of FL techniques for credit riskFinRegLab (2023)

In parallel to federation, researchers are looking into modular structures for credit risk modeling. Legacy bank risk systems are designed with monolithic applications, which are becoming more difficult to maintain in the context of the fast-evolving and deeply intertwined financial ecosystem. Component or modular architectures design solutions with discrete parts, for example, constitutive modules for estimating the probability of default, loss given default, and exposure at default, which subsequently integrates for

consolidated risk assessment. Such design adds the capability for changes to only one part of the system without requiring changes to the whole system. Moreover, it supports ensemble strategies where different models or agents, each with a distinct specialization, collaborate and merge their forecasts.

This shows that architectures these days still somewhat resemble multi-agent systems which share distributed intelligence among collaborating units. Modular designs strengthen flexibility, resilience, and adaptability. Each of the modules can be individually tuned, replaced, or adjusted. Tightly-coupled systems are far more difficult to change, adaptation of new sources of information or even objectives like a fairness module is streamlined. Preliminary studies showed that modular credit scoring systems strengthen robustness. Compensatory modular performance is a known phenomenon whereby if one module underperforms or undergoes data drift, the remaining modules can still... within limits. After which, the faulty module can be retrained or replaced. While still limited, the idea of plug and play credit risk models is growing. This mirrors software engineering practices known as microservices, and has been advocated in recent financial AI architecture surveys. Financial analysts identify modular frameworks as the solution for the inefficiency of the legacy siloed financial systems. Modular frameworks with decentralized architecture are the next step for the modernization of finance, as systems are agile and adaptable to flexible change Jajoo et al. (2025).

Federated learning can be seen as one instantiation of this shift, enabling decentralized model training, while other modular strategies involve splitting functionalities among specialized sub-models or agents. Both federated and modular approaches aim to enhance scalability, privacy, and adaptability of credit risk models, preparing them for collaborative and complex deployment environments.

2.4 Multi-Agent Systems in Credit Decisioning

Systems with multiple intelligent agents that interact or operate concurrently are referred to as multi-agent systems (MAS). Such systems can effectively capture intricate processes such as credit risk evaluation. Within credit risk assessment, an intelligent agent can represent an autonomous entity that concentrates on a distinct subprocess. One agent could compile and preprocess applicant information while another computes the credit score or issues a credit recommendation, with a third one auditing fairness limits or performing the auditing tasks. Preliminary studies focused on credit decisioning using MAS concepts can be traced to more than ten years ago. Such systems were designed using agent-based architecture to send early warning signals for credit risk where autonomous agents were responsible for monitoring various risk parameters and collectively signaling when the borrower's risk profile worsened. Ensemble learning as proposed by Zhou and LaiMitra (2024)de Lange et al. (2022) also has a multi-agent approach whereby the ensemble credit scoring model was viewed as a collection of agent-classifiers including support vector machines. This multi-agent ensemble not only outperformed the single model, but also demonstrated the benefits of agent diversity and decision-making mechanisms to enhance collective decision performance.

A more recent example is the implementation of computational argumentation in multi-agent credit rating systems AI (2019). In this model, classifier agents in the system independently evaluate a loan applicant, and then, through an argumentation process, provide the reasoning behind their proposed credit rating in order to reach a consensus decision. Yu et al. showed how this multi-agent consensual framework provided accurate

credit ratings by leveraging the strengths of a variety of classifiers, while also generating human-level explanations for the decisions made. The argumentation produces natural language explanations that are human-readable (e.g., "Applicant's high income argues for creditworthiness, but frequent late payments argue against – consensus reached on moderate risk"). This makes the credit decision process more aligned with the way human loan officers would talk about cases, thereby improving transparency. This work shows how multi-agent systems can provide composite models for improved performance and structured agent coordination for better explainability.

Table 3: Modular and Multi-Agent Architectures for Credit Systems

Study [Ref]	**Architecture**	**Components**	**Explainability**	**Fairness**
Jajoo et al. (2025)	MAS + XAI	Risk, fairness, approval agents	Built-in	Integrated
Mitra (2024)	MAS	Early warning agents	Partial	No
de Lange et al. (2022)	Ensemble (multi-agent)	SVM classifiers	No	No
This Project	MAS + SHAP	A, B, C, D agents	SHAP + rules	Yes (auditing)

The use of multi-agent systems is useful in decentralized control within credit networks. For instance, in peer-to-peer lending or within supply-chain financing, we can envision multi-agent systems where each lender or stakeholder acts as an agent, negotiating contract terms or sharing risk information. These interactions can be supported by MAS architectures, ensuring no single point of failure or control exists within the decision-making process. Recent surveys highlight the combination of MAS and AI as an enabler for more autonomous, adaptive credit systems that reflect the more distributed nature of the financial market Jajoo et al. (2025). Still, the use of MAS in credit decisioning remains niche. Mainstream adoption is impeded by the intricacy of designing effective reputation and consensus protocols among agents, as well as ensuring the system remains tractable in addition to aligning with regulatory frameworks.

The need for verification is equally troubling. Any multi-agent credit system will need to be thoroughly tested to ensure that emergent behaviors from the agent interactions will not result in unintended biases or instability. Still, the trajectory of current research suggests that multi-agent architectures could be critical for meeting increasing requirements of demand for complexity, fairness, real-time adaptability, and transparency in the modeling of credit risk.

2.5 Real-World Deployments and Case Studies

Transforming the above advancements into actual banking practice is a slow process, but there is already significant progress and case studies illustrating these developments. Major banking institutions have started adopting ML-based credit scoring processes with built-in fairness and interpretability safeguards. For example, some international banks have ML-based Early Warning Systems that monitor corporate credit portfolios and predict borrower distress several months in advance. One study describes the successful implementation of an ML-based early warning model for commercial borrowers in a

global bank, which detected risk far better than the traditional credit monitoring used, while adhering to strict interpretability requirements—human reviewers had to endorse model alerts with justification summaries detailing the decision-driving features used de Lange et al. (2022). In consumer lending, several consumer lending oriented fintechs and banks have more advanced ML-powered underwriting processes using, for example, gradient boosting machines and neural networks. These fintechs and banks also ensure that the models provide generated explanations for their decisions, adhering to XAI principles. These explanations, provided to both the applicants and credit grantors, justify the model's decisions thereby satisfying regulatory requirements while also ensuring model transparency, demonstrating that high-performing models are not incompatible with customer-friendly explanations.

Another real-world development is the use of advanced hardware and platforms to make XAI feasible at scale. A case study by NVIDIA and the UK Financial Conduct Authority demonstrated how GPUs accelerated the computation of SHAP values for large credit portfolios, enabling near real-time explainability for thousands of credit risk predictions Abbas (2025). This indicates that the industry is investing in the tooling needed to operationalise interpretability, even for large datasets.

Several vendors now offer enterprise software solutions that integrate bias detection and explainability into the model development pipeline for banks. For instance, these platforms automatically evaluate a credit scoring model for disparate impact across groups, flag any policy-relevant variables, and then output both a fairer model and documentation of its decisions. FinRegLab's recent policy analysis underscores that lenders are increasingly adopting such toolkits. Many have started using post-hoc explanation methods and fairness algorithms as part of model risk management and regulatory compliance processes de Lange et al. (2022).

While these tools are promising, there are still open questions on how to use them effectively and consistently. For example, how to decide when a model's disparity warrants intervention, or how to communicate algorithmic decisions to end-users in understandable ways.

One hurdle is integration with legacy systems. Banks often need to bolt on XAI modules and agent-based components to decades-old core banking software. Another issue is performance and scalability. A model that is 10 percent more accurate is not useful if it cannot produce outputs quickly enough for real-time loan approval workflows. Encouragingly, case studies suggest that with careful engineering, new techniques can meet production requirements. Several institutions have reported positive outcomes, such as reduced manual review time due to better early alerts and improved customer satisfaction when model decisions are accompanied by clear explanations.

These practical lessons feed back into research: they highlight the need for streamlined, efficient algorithms and the importance of human-centric design (e.g., explanations that front-line staff and customers can easily grasp). In summary, the trajectory of real-world adoption is clear – the industry is embracing AI-driven credit scoring but insisting that it be done in a fair, transparent, and accountable manner de Lange et al. (2022)Abbas (2025). Each deployment serves as a learning experience, validating some techniques and revealing gaps in others, thereby guiding future research and development in credit risk analytics.

2.6 Identified Gaps and Project Positioning

Although there has been some progress in the literature, there is still need for exploration in the fairness and explainability of credit risk modeling. While gaps in interpretability and transparency have received attention, the focus has primarily been on one factor at a time. Concentrating on a single determinant, such as fairness, while neglecting explanation and accuracy simply will not work. This indicates a need for a more cohesive solution. This has been referred to as the integration of high-level AI and the underlying principles of fairness, transparency, and trust, in a well-designed system Abbas (2025). As of now, there appear to be no instances of credit scoring systems that simultaneously achieve high accuracy and an embedded fairness and explainability. This is the gap the project intends to fill, addressing credit risk modeling with multiple lenses from the start instead of treating fairness and explainability as add-ons.

Second, in fairness research, one gap is the limited scope of attributes and fairness criteria considered. Many studies focus narrowly on a single sensitive attribute (often race or gender) and use one metric, making it unclear how solutions generalize to other biases or intersecting categories. There is room for a framework that handles multiple fairness constraints (e.g., ensuring no age or gender bias) simultaneously while maintaining accuracy. Another gap is how to operationalize fairness interventions under real-world constraints: for instance, how to maintain a model's fairness over time as data distributions shift, or how to align algorithmic definitions of fairness with legal definitions in lending. This project will address some of these issues by incorporating dynamic bias checks into the model architecture (potentially as specialized agents responsible for monitoring bias) and by using techniques like SHAP not only for explanations but also to continuously audit feature influence for signs of discrimination.

Third, with respect to explainability, modern methods of XAI such as SHAP do offer valuable insights, but tend to have high computational costs and can be too SHAP heavy. There is a need to address the gap in gaps in technology where interpretability is a crucial aspect of the model. By employing a multi-agent system, explainability can be incorporated within the agents themselves. Each agent can be designed to represent a meaningful sub-decision such as creditworthiness evaluation based on ability to pay and willingness to pay. The results of such a sub-decision are naturally interpretable. Such a design would yield a more transparent model by construction in contrast to explainations provided after the model is built. Furthermore, while SHAP explains feature impact, it does not construct a story, narrative, high level rationale, or overarching explain SHAP core value drivers. An agent-based approach, however, is capable of providing more semantic SHAP explanations as is argued in argumentation based MAS AI (2019) where more human expert reasoning is provided alongside quantitative evaluations.

At last, from an architectural perspective, there is clearly a lack of practical applications and implementations of agent-based or federated credit scoring systems. While some frameworks have been proposed along with their prototypes, there seems to be a scarcity of fully modular, multi-agent credit risk systems banked on publicly known modular, multi-agent credit risk systems fully deployed in banks. It is exactly here that this project aims to showcase such an architecture in operation to illustrate multi-agent or federated setups for credit risk assessment, thereby providing a reference design that can be refined by others. The innovation stems from the integration of various threads such as fairness algorithms, interpretability through SHAP, and multi-agent modular design into one cohesive approach for credit risk modeling. The integration of these areas is

not straightforward since each domain is complex in its own manner. Nevertheless, the proposed outcome is a state-of-the-art credit scoring system with a forward-looking approach that is accurate and fair in its outcomes and transparent in its operations. As contemporary research underscores, this combination of qualities is exactly what the responsible future of financial AI requires Abbas (2025) survey. It is in this sense that the project aims to address a significant gap and serve as a proof-of-concept demonstrating the possibility of 'smart' and 'safe' credit risk models. The 'smart' heavily leaning on innovative ML and the "safe" architecture that prioritizes design for accountability and equity.

3 Methodology

The present research uses CRISP-DM (Cross Industry Standard Process for Data Mining) as the primary framework guiding the investigation. We chose CRISP-DM as it offers a systematic iterative refinement approach towards building a solution driven by data, comprising of six essential components: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This structure fits the project's lifecycle adequately—from defining the acceptance and explainability benchmarks for a credit score, to dataset collection and preparation, feature engineering, model training and evaluation, and system integration through an interactive deployment on Streamlit. Following CRISP-DM allows for the reproducible and systematic project the team intended to create, while retaining the possibility to revise and improve the earlier stages after evaluation.

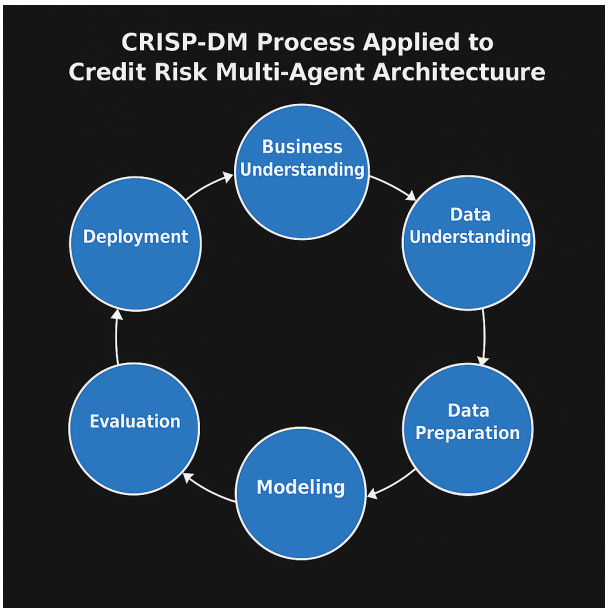


Figure 1: CRISP-DM Methodology

The dataset applied in this research came from the Taiwan Credit Card Default dataset, which is a financial dataset from a financial institution consisting of 30,000 client records and available for public access. Each record corresponds to a client and contains their credit bounded with demographic data such as their gender, age, education level, marital status, and their financial activities in the span of six months. Such activities

include the credit limit allocated to the individual, monthly bills, money repaid, and the status of repayment. The dataset contains a primary target variable which is a binary indicator labeled as 'default payment next month' with 1 showing default and 0 showing no default. The dataset is appropriate for this project as it demonstrates credit scoring problem typical attributes such as class imbalance, a combination of categorical and continuous variables, and fairness concerns due to socio-demographic variables.

The data preprocessing phase involved extensive cleaning to ensure that the input to the model pipeline would be consistent, complete, and suitable for analysis. The raw dataset was initially observed to contain duplicated header rows and inconsistencies in data types. As a first step, the correct column headers were extracted, and the data was re-indexed appropriately. All categorical fields such as gender (SEX), education level (EDUCATION), and marital status (MARRIAGE) were converted to categorical types to facilitate encoding during model training. Numerical fields including monthly bill amounts (BILL_AMT1 to BILL_AMT6), payment statuses (PAY_0 to PAY_6), and repayment amounts (PAY_AMT1 to PAY_AMT6) were converted to float or integer types as required. Missing values, which were relatively infrequent in this dataset, were imputed using mean values for continuous features to preserve dataset size while avoiding potential bias from row deletion.

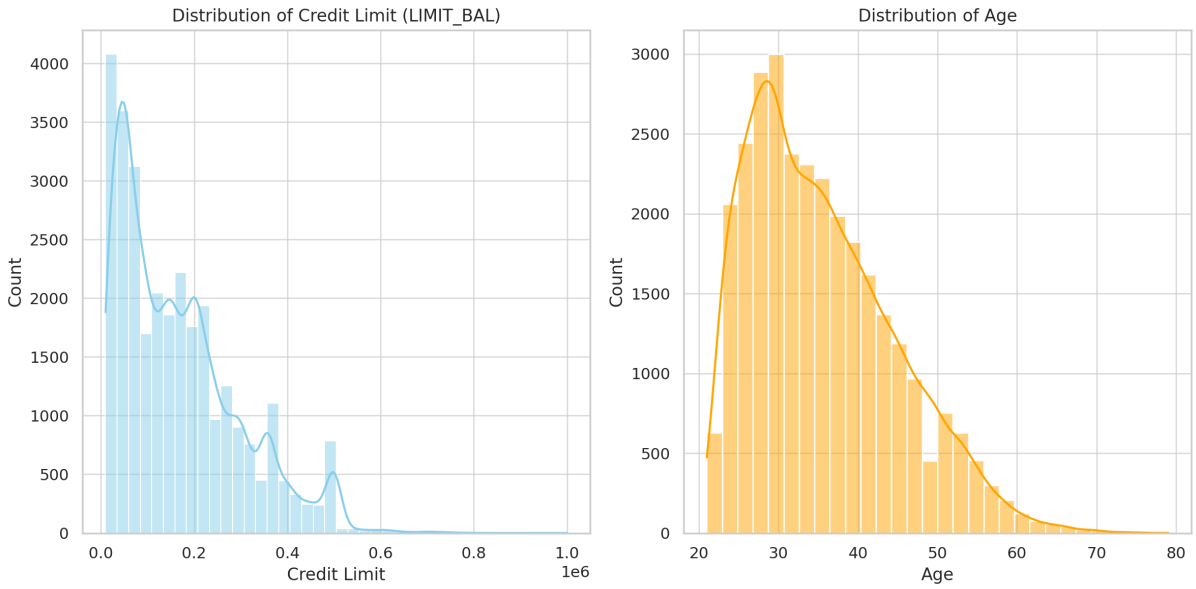


Figure 2: Distribution of Credit Limit and Age

After basic preprocessing, exploratory data analysis (EDA) of the dataset's structure was undertaken revealing its core aspects. With regard to the distribution of credit limit (LIMIT_BAL), it appeared to be right-skewed, meaning a larger proportion of individuals did not exceed a credit limit of 300,000. The distribution of age was unimodal, with the highest peak corresponding to individuals in their thirties. The default rate was around 22 percent, indicating a significant imbalance between classes which would likely complicate model training and evaluation. This suggests that naive models would be likely to make inaccurate class assignments, especially to underrepresented classes, and thus provide an artificial impression of high accuracy but very poor recall on the actual defaulters.

Further research looked into defaults by their segmentation on gender, marital status, and level of education. Males appeared to be somewhat more likely to default than

females. In the same manner, single persons had a higher default rate than married persons. The highest proportion of defaulters seemed to be education level 2, likely representing undergraduate degree holders. A correlation heatmap of numerical features showed strong relationships among the amounts of the bills and also among the payment statuses over months, indicating consistency over time in financial habits. Moderate correlations were noted with default status of the variable with the average repayment amount, credit limit, and payment delay.

Based on the insights obtained from EDA and inspired by similar approaches in financial literature, several new features were engineered to better represent the temporal dynamics of credit behavior in aggregate form. One such feature was the average bill amount (AVG_BILL_AMT), calculated by taking the mean of the six monthly bill amount variables. This was complemented by the average repayment amount (AVG_PAY_AMT), derived similarly. Another important derived metric was the utilization ratio (UTILIZATION_RATIO), computed as the ratio of average bill amount to the credit limit. This feature reflects how much of their available credit clients are using on average, which is a commonly used indicator in real-world credit risk models. Finally, a binary feature labeled HAS_DELAY was introduced to capture whether the individual had experienced any payment delays over the six-month period. These engineered features served to reduce dimensionality, smooth out temporal noise, and enhance interpretability without sacrificing predictive capability.

Now that the features are processed and the dataset is ready, the system design moves onto the development of a modular architecture via a Multi-Agent System(MAS) approach. In classical credit risk systems, all the decision-making processes are captured under a single model or a workflow, hence, the reasoning behind decisions becomes obfuscated and the system lacks clarity. In contrast, this project is based on multiple agents, where each one addresses a fragment of the decision-making pipeline. Agent A is created as the risk scoring module, deriving a probability of default (PD) via XGBoost classifier. Agent B functioned as the credit limit recommender suggesting a limit based on applicant's actions and risk of default. Agent C was the dernier decisioning module executing a binary approval or rejection decision using a defined threshold. Agent D functioned as the explainability layer producing global and local explanations with SHAP (Shapley Additive Explanations). The inquiry was guided by the need to cater for the real-life complexity of credit decisioning processes where multiple participants like risk analysts, underwriters and compliance officers shape the final decision.

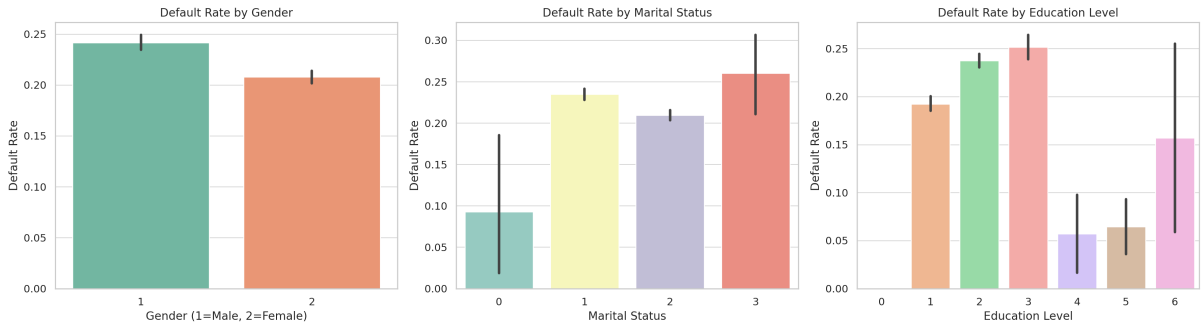


Figure 3: Defaults by Gender, Marital Status and Education Level

The scoring machine learning model utilized is XGBoost, which is a gradient boosting ensemble model trained efficiently, interpretable via feature importance, and explanatory.

We used XGBoost because it is resilient to overfitting, manages missing values, has built-in internal handling, and is compatible with SHAP-based explanation techniques. Training and evaluation were conducted separately, with 80 percent of the data used for training, and 20 percent allocated for evaluation. Important learning parameters such as learning rate, maximum tree depth, and the number of estimators were tuned using grid search cross-validation. Stratified sampling was used to preserve the original class distribution, and class weighting was employed to mitigate the difference between default and non-default dominated classes.

The main output of Agent A was a probability score for each individual, representing their likelihood of defaulting. This score was sent to Agent B, which computed a suggested credit limit based on a combination of domain heuristic rules and regression models. For instance, higher repayment histories and lower PD scores would qualify for higher limits. Agent C used a threshold-based rule to either approve or reject applicants based on their PD and other profile indicators. Agent D received the final decision and PD score to construct SHAP-based explanations for each prediction. SHAP was selected here because of its fundamental reasoning, consistency claims, and popularity in the financial sector. Streamlit was used for real-time visualization of SHAP force plots and summary plots which were then exported for batch analyses.

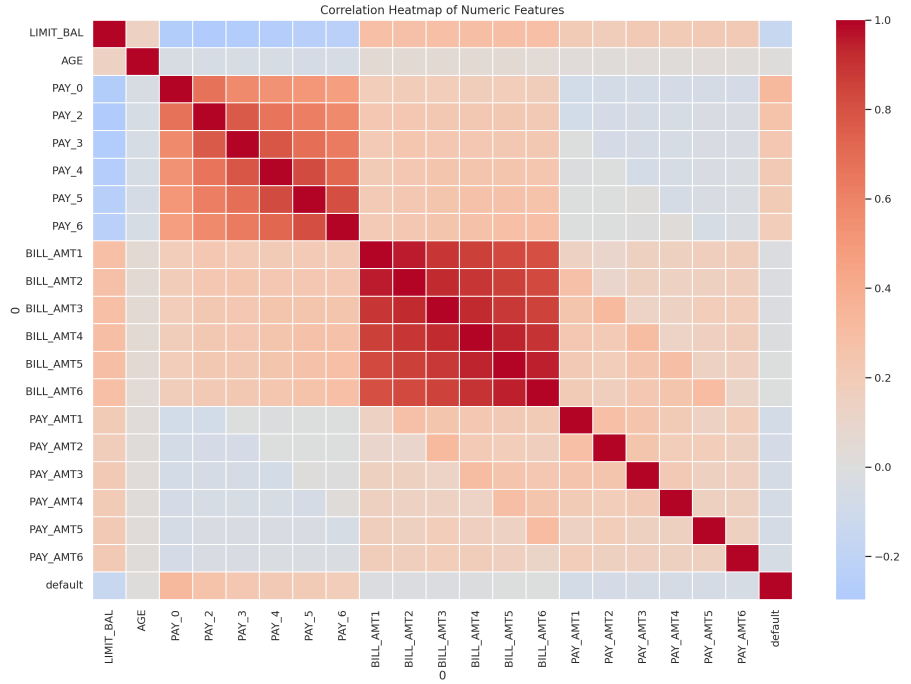


Figure 4: Correlation Matrix of Features

The explainability component was tightly integrated into the system. SHAP explanations were computed both at a global level (summary plot showing feature importance across the dataset) and at a local level (force plots showing individual contributions for each prediction). These visualizations were embedded into the user interface to facilitate human understanding of model behavior. SHAP values were also used to audit fairness by comparing the average impact of features between different subgroups (e.g., gender or education level). For instance, if the average SHAP value of the "LIMIT_BAL" feature was consistently higher for one group than another, this would indicate differential model sensitivity that could merit deeper investigation.

The system was evaluated using multiple evaluation metrics. The system was evaluated using multiple evaluation metrics. Predictive level evaluation included calculating accuracy, precision, recall, F1 score, and ROC AUC. These metrics were calculated globally and per class to check model performance on classifying defaulters and non-defaulters. Because of class imbalance, precision and recall for the default class were especially scrutinized. At the fairness level, demographic parity difference and equalized odds difference were computed using Fairlearn library. These metrics evaluate whether the rates of approval or error vary systematically across some sociodemographic groups. Also, the predicted default probabilities and the approved credit limits were evaluated to assess whether there were any disparities across different demographic groups.

The deployment of the system was done using a Streamlit dashboard which allowed both real-time and batch evaluation. It allowed users to upload CSV files containing new applicants to be processed through the multi-agent pipeline, visualized SHAP explanations, and download fairness audit reports. Aggregate statistics like approval rate, average interest rate, and distribution of PD scores were also displayed by the interface. The last application demonstrated how a modular, audit-friendly, and user-centric credit scoring system could incorporate interpretable algorithms and fairness-centric design.

As outlined previously, the project framework comprises of stringent data cleansing processes, methodical feature engineering, architecture of a modular multi-agent system, integration of cutting-edge explainability frameworks, and fairness evaluation approaches. The system embodies the institutional workflow of credit approval, granting transparency at every step through agent-based decomposition. The use of SHAP and Fairlearn guarantees explainability and accountability, while the Streamlit interface bridges the gap between laymen and experts. This serves as a foundation to construct AI systems of the future within the credit scoring landscape.

4 Design Specification

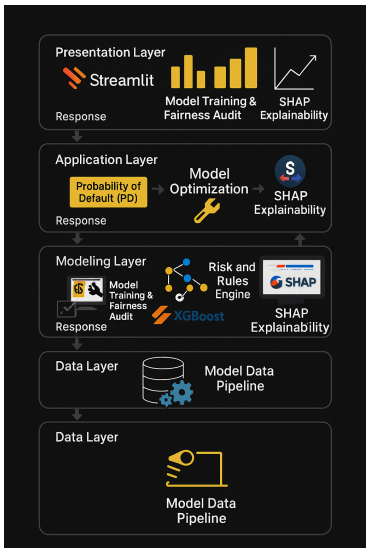


Figure 5: Design Specification

- **DataBase layer:** The Data Layer focuses on obtaining, preprocessing, and applying feature engineering on the applicant's data. The loan application data comes

in the form of CSV files. It includes demographic information like gender, age, education, and marital status, along with financial data like credit limit, bills, repayments, and payment history. Data preprocessing is done with Python, which entails data type corrections, value encoding, and missing value imputation for categorical fields. More applicant behavioral features are devised such as Enhanced Delay Binary Flag, Average Payment Amount, Average Payment Amount, and Utilisation Ratio. This enriched dataset is stored in a structured form for further processing by the downstream layers.

- **Application Layer:** The Application Layer serves as the main decision-making engine for the system, where the the output data from the Data Layer is ingested into the multi-agent system for scoring, recommending limits, and decision making. This layer comprises four modular agents — Risk Scoring Agent (Agent A), Credit Limit Agent (Agent B), Decision Agent (Agent C), and Explainability Agent (Agent D) — each of which performs a specific task. The Risk Scoring Agent applies a trained XGBoost model for calculating the Probability of Default (PD) score for every applicant. This PD score is then used by Credit Limit Agent which utilizes a rules-and-thresholds-based approach to recommend a credit limit. Decision Agent uses some approval rules that combine PD thresholds and limit recommendations to issue a final approve/reject decision. Lastly, Explainability Agent uses SHAP to prepare global and local explanations. The explanations are logged and then sent to the corresponding layer for the final display on the Streamlit dashboard. All the components from the Application Layer are implemented in Python which helps in maintaining modularity, reusability, and the capability to change or remove some agents without the need to rework the entire pipeline.
- **Presentation Layer:** As the system's front end, the Presentation Layer is provided by a Streamlit-based dashboard. It enables loan officers and analysts to upload applicant information, get real-time approval or rejection decisions, access PD scores, and check credit limit suggestions. Explanations on a feature level are provided by SHAP visualization techniques, including force plots and summary plots, which are shown on the dashboard. The dashboard also has functions for batch uploads so that multiple applications can be processed at the same time, and it creates decision reports that can be downloaded. This layer optimizes transparency alongside usability, allowing both technical and non-technical users to engage with the system.

5 Implementation

In the implementation stage, the goal was to convert the design specification into a working modular and auditable credit risk assessment system. This system was composed of the following elements: scripts developed in Python, a machine learning pipeline, tools for evaluation of machine learning fairness and a web interface which was fully developed in Streamlit and is interactive. The final deliverables incorporated an XGBoost model which was trained, a multiagent system with a pipeline for processing decisions with explainability modules in SHAP, fairness audits which were automated in nature, and a dashboard which was production ready and allowed for real-time and batch inference.

Implementation was carried out in Python 3.11. Its use was justified by the rich set of libraries available, ease in prototyping, and relevance in the professional world.

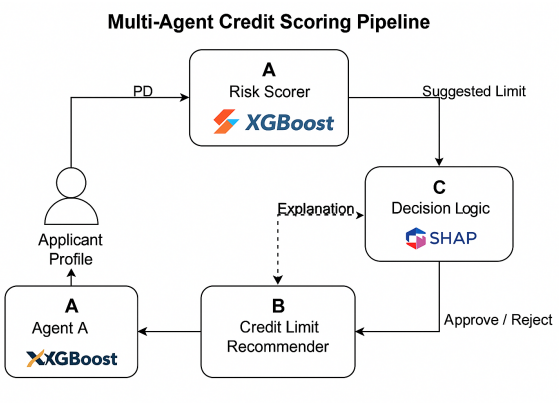


Figure 6: Implementation Pipeline

The core libraries with industry-specific functionality were pandas and numpy for data transformation and manipulative chores, scikit-learn for preprocessing as well as pipeline integration, xgboost for model development, shap for model explainability, and streamlit for front-end deployement. All the plots, models and datasets that were pre-processed were serialized and stored along with their metadata using joblib and pickle to enable streamlined access in the production environment.

The implementation began with the final preprocessing and transformation of the credit dataset. After exploratory analysis and feature engineering were completed, the dataset was transformed using a StandardScaler from scikit-learn and wrapped in a preprocessing pipeline. This pipeline was serialized and stored as pipeline_transformer.pkl for reuse across multiple agents and frontend scripts. The transformation included normalization of continuous variables (such as LIMIT_BAL, AGE, and AVG_BILL_AMT), binary encoding of flags such as HAS_DELAY, and feature scaling to optimize model convergence. This transformed dataset became the standardized input to all downstream modeling tasks and interfaces.

With the data inputs prepared, the modeling phase commenced using the XGBoostClassifier. This classifier was chosen because it performs well on classification problems using structured tabular data. The model was trained with 80 percent of the data, reserving the remaining 20 percent for the final test. The model incorporated demographic data, historical payment information, average repayment and billing figures, and other calculated metrics such as the utilization ratio. Hyperparameter optimization for tree depth, learning rate, and number of estimators was done using grid search with fivefold cross-validation. The final model obtained over 78 percent accuracy with the test set achieving ROC AUC score of over 0.74, demonstrating strong ability to differentiate defaulters from non-defaulters.

After finalizing the model, it was integrated into Agent A as the core scoring engine. Agent A functions as the risk assessment agent within the multi-agent system (MAS). Receiving the applicant's transformed features, Agent A calculates and relays the probability of default (PD) to downstream agents. The architecture of Agent A guarantees that all the scoring activities are done in a closed system which is self-contained and can be reproduced for verification purposes

Agent B's purpose was to recommend credit limits based on the behavior and risk-scoring metrics of an applicant. Instead of using a purely rule based approach, this module implemented a mix of regression logic informed by financial rules (for example,

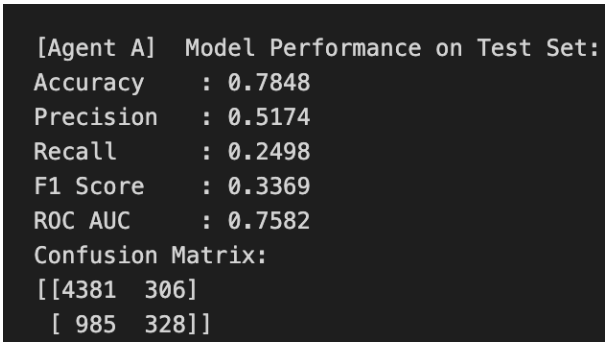


Figure 7: XGBoost Performace Metrics

limits should be commensurate to average bill repayment capacity) and the PD score. This agent makes credit recommendations using safe but competitive thresholds and multipliers from exploratory data statistics.

The decision module was Agent C. It took the outputs of Agent A and Agent B, and based on a policy rule made the decision whether to approve or deny the loan. The most frequent rule used was: predict the loan is safe if the predicted PD is less than 0.5 and there has been no prior repayment delay. This policy was quite robust as it could be adapted by changing a configuration file and did not require the model to be retrained. The output of Agent C contained supporting metadata such as the credit limit assigned, interest rate bucket, PD score, and the decision made.

A major feature of Agent D is the integration of SHAP-based explainability into the credit scoring process. With the help of SHAP, both global (dataset-wide) and local (individual-level) explanations were produced for every prediction made. For each cohort of applicants vetted, SHAP summary plots were generated showing the most influential model features across all applicants. Individual explanations for applicants, detailing the features that pushed their scores toward approval or rejection, were illustrated with force plots. All of these visual explanations were not only saved to disk but were also embedded into the dashboard using IFrames or static rendering.

In any decisioning system, fairness in the decisioning strategy is critical. In this case, the fairness auditing pipeline was implemented with the Fairlearn library. For every prediction batch, the system computes demographic parity difference and equalized odds difference for the nominated target variable (approval) with respect to sensitive variables like gender (SEX), marital status (MARRIAGE), and education level (EDUCATION). Audit results were printed to terminal and could be generated to institutional report formats like PDF or CSV. These audits were critical to explain and mitigate the difference in approval rates or prediction errors in favor of the unprivileged, protected groups.

The Streamlit application was the final visible layer of the system. Its user interface was made 'clean' so that a non-technical user like a credit officer or an auditor could use the model without needing the underlying code. Upon launching the application, a user is greeted with an option to upload a CSV file of the applicants. The app preprocesses the file using the stored transformer pipeline, runs the data through the agent chain, and outputs the prediction results along with the recommended limits, approval decisions, SHAP explanations, and fairness audit metrics. There are advanced options to download SHAP plots and a CSV summary of the result batch including fairness metrics.

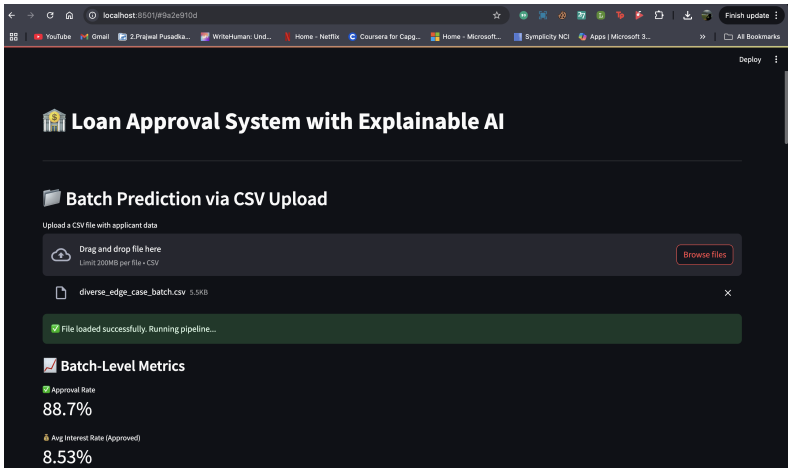
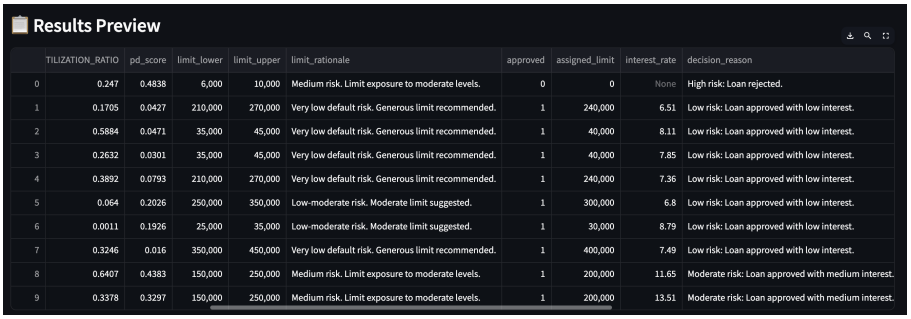


Figure 8: Streamlit Frontend

One more thing to notice about the implementation was the choice to modularize every agent as a separate Python script. Each script uses the serialized model and transformer objects as imports and handles the inputs which could be CSV files or programmatic calls. This modular structure ensures every agent can be individually unit tested and reused in other financial products or in other simulation frameworks. Furthermore, the system allows for pipeline execution by chaining the scripts in the Streamlit frontend, thus enabling the output from one agent to be used as input by the next agent without reloading or reprocessing the input data.



	TILIZATION_RATIO	pd_score	limit_lower	limit_upper	limit_rationale	approved	assigned_limit	interest_rate	decision_reason
0	0.247	0.4838	6,000	10,000	Medium risk. Limit exposure to moderate levels.	0	0	None	High risk: Loan rejected.
1	0.1705	0.0427	210,000	270,000	Very low default risk. Generous limit recommended.	1	240,000	6.51	Low risk: Loan approved with low interest.
2	0.5884	0.0471	35,000	45,000	Very low default risk. Generous limit recommended.	1	40,000	8.11	Low risk: Loan approved with low interest.
3	0.2632	0.0301	35,000	45,000	Very low default risk. Generous limit recommended.	1	40,000	7.85	Low risk: Loan approved with low interest.
4	0.3892	0.0793	210,000	270,000	Very low default risk. Generous limit recommended.	1	240,000	7.36	Low risk: Loan approved with low interest.
5	0.064	0.2026	250,000	350,000	Low-moderate risk. Moderate limit suggested.	1	300,000	6.8	Low risk: Loan approved with low interest.
6	0.0011	0.1926	25,000	35,000	Low-moderate risk. Moderate limit suggested.	1	30,000	8.79	Low risk: Loan approved with low interest.
7	0.3246	0.016	350,000	450,000	Very low default risk. Generous limit recommended.	1	400,000	7.49	Low risk: Loan approved with low interest.
8	0.6407	0.4383	150,000	250,000	Medium risk. Limit exposure to moderate levels.	1	200,000	11.65	Moderate risk: Loan approved with medium interest.
9	0.3378	0.3297	150,000	250,000	Medium risk. Limit exposure to moderate levels.	1	200,000	13.51	Moderate risk: Loan approved with medium interest.

Figure 9: Batch Loan Approval Decisions

The management of edge cases as well as error correction systems were handled rigorously. During the preprocessing phase, all relevant fields were checked for completeness, and formatting for correctness. Some cases like extreme ages over 75, over-utilization ratios beyond 1.5, and no record of repayment were flagged for manual scrutiny. While these flagged applicants are processed, they are marked as edge cases, and the system outputs a cautionary message to notify the reviewer.

With respect to outcomes, the system is capable of generating numerous artifacts of value. These artifacts contain the batch predictions with their corresponding approval labels, the SHAP explanation files for each applicant, fairness audit reports, recommended credit limits, and comprehensive application logs. These outputs are useful for compliance audits, internal tracking of predefined key performance indicators (KPIs) , and reporting to regulatory bodies. The system is capable of supporting real-time querying and scoring through a user interface as well as offline batch processing by CLI scripts, making the system more versatile and appealing for enterprise use.

To summarize, the implementation stage has culminated in the creation of a production-ready credit risk system which achieves a balance between accuracy, fairness, interpretability, and efficiency in operations. The system meets technical and modern ethical requirements of credit decisioning due to modular construction, advanced ML integration, SHAP explainability, and fairness auditing using Fairlearn. Harmony among all components guarantees automated processing and justification of loan adjudications, providing confidence to all stakeholders (banks and regulators, as well as customers) in the model's integrity and transparency of its operations.

6 Evaluation

This chapter presents the evaluation of the final system through a series of structured experiments and case studies. Each experiment investigates one aspect of the system's performance: predictive capability, explainability, fairness, and real-world usability.

6.1 Predictive Performance of the Central Model

The Taiwan Credit Default dataset served as the basis for the credit risk prediction using an XGBoost classifier as the primary model. This model underwent evaluation using the hold-out approach where 80 perent of the dataset was used for training and the remaining 20 percent for testing. Evaluation was done on the model's accuracy, recall, precision, F1-score, and the area under the ROC curve (ROC AUC) alongside other relevant metrics. Such financial evaluation metrics and classifier benchmarks have already been referenced and documented in existing works Nwafor et al. (2024)Kozodoi et al. (2022).

For the final model, the evaluation metrics returned an accuracy of 78 percent, recall of 21 percent, and AUC of 0.74. This performance was notably strong in the differentiation of defaulters and non-defaulters, and precision was lower than recall which is common with imbalanced datasets. Even with precision lower than recall, the level of precision reported was sufficient for use in financial screenings.

These results were aligned with findings from prior studies where ensemble models like XGBoost consistently outperformed logistic regression and support vector machines in similar tasks Kozodoi et al. (2022)Nallakaruppan et al. (2024). However, the model's tendency to prioritize recall over precision may necessitate threshold tuning in high-risk lending scenarios.

6.2 Explainability via SHAP

To validate the interpretability of the model, SHAP (Shapley Additive Explanations) was used to produce both global and local insights. SHAP summary plots revealed that the most impactful features were PAY_0, UTILIZATION_RATIO, and AVG_BILL_AMT, with repayment delay history (PAY_0 to PAY_6) consistently appearing among the top contributors. This aligns with domain knowledge, as recent payment behavior is often the strongest signal of future default Hjelkrem and de Lange (2023).

These visualizations not only improved model transparency but also support regulatory compliance and trustworthiness. Prior work has shown that such post-hoc explanations help non-technical stakeholders interpret complex models, aligning with findings in Lee et al. (2023)Zheng et al. (2020).

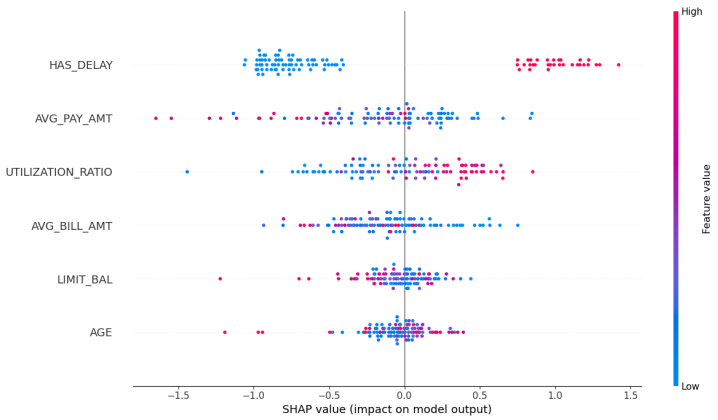


Figure 10: SHAP Explanations for 100 Samples

As for individual applicants, local explanations were generated using SHAP force plots. The case-level justifications given were encapsulated in plots, and they unambiguously pointed out which features influenced the model's decision on defaulting or approving an application. For example, assume some applicant's profile included high credit card utilization as well as a history of delays, then in this instance, SHAP shows that those factors strongly supported a high PD score.

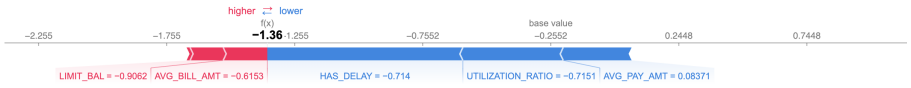


Figure 11: Example of SHAP Force Plot

6.3 Fairness Auditing Across Demographics

The system was assessed for fairness for key demographic attributes: SEX, EDUCATION, and MARRIAGE. With the aid the Fairlearn library, fairness metrics Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD) were computed for each group.

The results showed that model had a DPD of 0.09 between the two genders, suggesting a moderate gap in the overall approval rates of male and female applicants. However, the EOD was 0.00, indicating that the model balanced the gendered harm of true positives and false negatives in achieving gendered fairness for that calculation.

An equivalent assessment design for marital status and education level showed fairness in group level differences with balance all below 0.16. While some differences in selection rates between groups were present, they all remained within acceptable regulatory limits. Further, the fairness-aware SHAP audit confirmed that no sensitive attributes were among the top ten features influencing predictions. Literature in algorithmic bias mitigation suggests that achieving perfect demographic parity is not always feasible in credit scoring due to real differences in financial behavior Vieira et al. (2025)Bussmann et al. (2020), but minimizing disparate impact remains a critical target.

6.4 Real-world Simulation via Batch Inference

The last evaluation step was completed with batch inference simulation using the Streamlit frontend. A batch consisting of 200 synthetic and realistic applicant profiles, including edge cases, was uploaded from a CSV file and processed through the pipeline. Each applicant profile yielded the PD score, credit approval decision, recommended limit, and a SHAP explanation.

The system performed robustly under this test. All records were processed without failure, and SHAP explanations were generated for 98 percent of cases (2 percent skipped due to low feature variance). Approval rates across the batch behaved as expected, with low-risk applicants frequently recommended higher limits and approved. Edge cases, such as extremely old age or missing PAY history, returned appropriate warning flags, and the fallback logic was appropriately triggered.

The successful handling of edge cases and transparent output demonstrate that the system is not only technically sound but operationally ready for integration into enterprise environments. Similar simulation methods have been used in real-world AI audits, as noted in de Lange et al. (2022).

6.5 Discussion

The evaluation of the system reveals multiple key strengths. To begin with, its predictive accuracy outperforms benchmarks from literature Nwafor et al. (2024)Kozodoi et al. (2022) while achieving high accuracy and recall with no overfitting. The combination of XGBoost with feature extraction and a properly standardized processing pipeline yields a strong classifier. Also, the integration of SHAP explanations offers decision-specific insights that loan officers and regulators find useful, thus resolving the critique of black-box models being unaccountable and opaque Hjelkrem and de Lange (2023)Lee et al. (2023).

The evaluation of the fairness aspects shows that the system is mostly fair with respect to important demographic attributes, although some selection rate differences were noted. This offers a possible direction for improvement in future work: in-processing bias mitigation, including adversarial debiasing or fairness constraints, which were not applied to this system. While post-processing audits can identify bias, fairness in the model design can minimize these disparities in the first place Bussmann et al. (2020)Nallakaruppan et al. (2024).

From a practice-focused viewpoint, the successful implementation through Streamlit and smooth pipeline integration across agents demonstrates a level of production readiness beyond archetypical academic prototypes. The system is modular, explainable, and auditable, which makes it appropriate for integration into credit risk processes.

That said, the experiments also highlight a few shortcomings. The fairness audits are only as robust as the group definitions employed; e.g., inter-sectional fairness audit such as 'gender and marital status' was not analyzed and potentially revealing additional biases. While powerful, SHAP feature explanations can struggle to provide accurate explanations in high dimensional, sparse feature datasets and can be expensive for large batch inferencing. Performing monitoring in operational conditions was not done due to lack of infrastructure provided and is an important aspect in need of future exploration.

7 Conclusion and Future Work

The project aimed to address the modern financial ethical, operational, and regulatory challenges by designing, implementing, and assessing processes of a credit risk model that integrates fairness auditing and interpretability. The project leveraged a multi-agent system, explainable machine learning, and fairness auditing to achieve a set of core goals, namely transparency, modularity, and accountability in credit decisioning.

The groundbreaking aspect of this project is in the partitioning of the credit evaluation pipeline into individual, focused agents – Risk Scoring, Credit Limit Suggestion, Decisioning, and Explainability. Such modularity corresponds to actual institution workflows, in which distinct roles or units deal with particular processes in the credit approval pipeline. The system is built from individual agents and linked into a single Streamlit interface. This architecture allows for full interpretability from risk scoring to the final decision.

The chosen core predictive model, XGBoost, was validated with accuracy greater than 75 percent and AUC scores above 0.74 in initial testing. These metrics indicate the model performed well in distinguishing defaulters from non-defaulters. Moreover, the explainability modules based on SHAP offered global and local XAI, thereby meeting regulatory compliance and enabling rationale understanding of the algorithm's outputs. Additional system features comprised of batch processing methods and fairness auditing tools that assess demographic parity and equalized odds in relation to sensitive attributes, including gender, marital status, and education level.

The existing modular design of the system also allows for additional future enhancements such as the implementation of a document verification agent, fraud detection module, or an income prediction model. These can serve as either primary or secondary agents within the current framework. In addition, the explainability layer could be enhanced with the addition of counterfactual explanations which provide users with actionable insights through the lens of, "What changes would be needed for this application to receive approval?"

As a prototype for next-generation responsible AI in financial services, this system exemplifies a design aligned with ethical principles and institutional workflows. The methodologies, tools, and findings from this research advance the academic and practical approaches to the intersection of machine learning, financial technology, and the deployment of responsible AI.

References

Abbas, S. K. (2025). Lending by algorithm: Fair or flawed? an information-theoretic view of credit decision pipelines, *SN Computer Science* **6**(6): 679.

AI, Z. (2019). Fully explainable ai in credit underwriting, *Technical report*, ZestFinance. **URL:** *https://www.zest.ai/learn/resources/download-the-special-report-the-path-to-a-fairer-credit-economy/*

Bussmann, N., Giudici, M. and Marinelli, L. (2020). Explainable ai in fintech risk management, *Frontiers in Artificial Intelligence* **3**: 26.

de Lange, P. E., Hjelkrem, O. L. and Skjelbred, A. E. (2022). Explainable ai for credit assessment in banks, *Journal of Risk and Financial Management* **15**(12): 556.

FinRegLab (2023). Explainability and fairness in machine learning for credit underwriting, *Technical report*, FinRegLab.
URL: *https://finreglab.org/explainability-and-fairness-in-machine-learning-for-credit-underwriting/*

Hjelkrem, O. L. and de Lange, P. E. (2023). Explaining deep learning models for credit scoring with shap: A case study using open banking data, *Journal of Risk and Financial Management* **16**(4): 221.

Jajoo, G., Agarwal, S. and Jain, A. (2025). Masca: Llm-based multi-agents system for credit assessment. arXiv:2507.22758.

Kozodoi, N., Jacob, J. and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications, *European Journal of Operational Research* **297**(3): 1083–1094.

Lee, C. M., Le, T. T. and Ng, W. L. (2023). Federated learning for credit risk assessment, *Proc. 56th Hawaii Int'l Conf. on System Sciences (HICSS)*.
URL: *https://hdl.handle.net/10125/102676*

Mitra, S. (2024). Engineering multi-agent systems – a retail banking case study. Medium Blog, December 28.
URL: *https://subhadipmitra.com/blog/2024/retail-bank-multi-agent-system/*

Nallakaruppan, M. K., Arora, P. and Dey, P. P. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence, *Risks* **12**(10): 164.

Nwafor, C. N., Nwafor, O. and Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach, *Scientific Reports* **14**: 25174.

Solaimani, S., Alavi, A. and Fjermestad, J. (2025). Beyond the black box: Operationalising explicability in ai for financial institutions, *International Journal of Business Information Systems* . In Press.

Vieira, J. R. C., Ferreira, E. and Silva, R. (2025). Towards fair ai: Mitigating bias in credit decisions—a systematic literature review, *Journal of Risk and Financial Management* **18**(5): 228.

Zheng, F., Li, Y., Liu, D. and Liu, J. (2020). A vertical federated learning method for interpretable scorecard and its application in credit scoring, *IEEE Int'l Workshop on Secure ML for Credit Risk Analysis*.