

Explainable Credit Risk and Default Prediction with Robustness, Fairness, and Governance Constraints

Project Proposal (Example)
TC2038 – Analysis and Design of Advanced Algorithms
January 2026

Santiago Arista Viramontes (A01028372)

Diego Vergara Hernández (A01425660)

José Leobardo Navarro Márquez (A91541324)

I. PROBLEM STATEMENT

Credit risk assessment is a central problem in financial services, where lenders must decide whether to approve loan applications while minimizing the risk of future defaults. Traditional machine learning models (such as gradient-boosted trees or deep learning) have strong predictive performance but tend to behave as black boxes, making it difficult for end users and regulators to understand the basis of a prediction. This lack of transparency undermines trust, hinders regulatory compliance, and limits the practical adoption of AI in high-stakes settings like credit underwriting [9].

This project addresses the problem of predicting the probability that a loan applicant will default within 12 months using supervised machine learning, while simultaneously producing interpretable explanations for each prediction, with explicit guarantees about when the explanations are faithful to the underlying model. Explainability is treated as a first-class algorithmic objective: for model classes where exact additive attributions are available (e.g., tree ensembles with TreeSHAP), explanations are computed to be consistent with the trained model's output decomposition; for other model classes, explanation limits and approximation error are stated and evaluated. This framing supports transparency, regulatory compliance, and informed decision-making. By integrating explainability into the predictive pipeline, the system aims to support transparency, regulatory compliance, and informed decision-making.

From a software engineering and algorithmic systems perspective, the problem involves:

- modeling credit decision-making as an optimization problem,
- designing reproducible and auditable machine learning pipelines,
- enforcing algorithmic constraints such as robustness, explainability, and fairness, and
- documenting governance, traceability, and system behavior for production-level readiness.

II. GENERAL OBJECTIVE

The general objective of this project is to design, implement, and evaluate an algorithmic credit risk decision pipeline that

integrates predictive modeling with explainability, robustness testing, fairness auditing/mitigation, and governance controls appropriate for regulated decision making.

Specifically, the system aims to:

- 1) estimate borrower repayment willingness and capacity through supervised learning models using structured financial and behavioral features;
- 2) generate faithful and interpretable explanations for individual credit decisions, suitable for both end users and regulatory review;
- 3) incorporate robustness mechanisms, including regularization techniques and anomaly detection for anti-money laundering (AML) signals;
- 4) evaluate fairness and disparate impact across sensitive attributes, and conclude with a comprehensive governance dossier supporting sandbox or pilot deployment.

III. MODULE SCOPE (M1–M4)

A. M1 — Explainable Credit Risk Scoring (baseline)

- **Inputs:** structured loan application features (financial + behavioral + demographic when available) and binary label (default within 12 months).
- **Core algorithm:** supervised classifier producing calibrated default probabilities (e.g., logistic regression and/or gradient-boosted trees).
- **Outputs:** predicted probability \hat{y} , risk band / decision threshold output, and per-instance explanation artifacts.
- **Evaluation:** AUC-ROC, PR-AUC (if imbalanced), Brier score, and calibration error (e.g., ECE).
- **XAI artifacts:** local explanations per applicant and global feature importance summaries.

B. M2 — Robustness + AML/Anomaly Layer (prototype)

- **Robustness:** add explicit regularization and stress tests (noise injection, feature dropout, and mild distribution shift simulation).
- **Explanation stability:** quantify attribution stability under perturbations (e.g., correlation/top- k agreement and attribution-distance metrics).

- **AML/anomaly:** unsupervised anomaly score from engineered “unusualness” signals; outputs include anomaly score, alert flag, and reasons/metadata.
- **Logging:** store model version, data snapshot hash, explanation artifacts, and anomaly outputs for traceability.

C. M3 — Fairness Audit + Mitigation

- **Decision policy:** define how \hat{y} becomes an approval/deny decision (threshold or risk bands).
- **Audit metrics:** disparate impact (selection rate ratio), TPR/FPR gaps, and calibration within groups (as feasible).
- **Mitigation:** implement one strategy (pre-, in-, or post-processing) and quantify trade-offs vs. baseline.

D. M4 — Governance + Deployment Readiness Dossier

- **Reproducibility:** fixed train/validation splits, experiment tracking, and model/data versioning.
- **Traceability:** per-decision record linking input features, model parameters/version, prediction, explanation, and anomaly outputs.
- **Documentation artifacts:** model card + data sheet + risk/compliance matrix + monitoring/limits for sandbox execution.
- **Closure:** cost–benefit matrix comparing algorithmic pipeline vs. traditional process.

IV. ALGORITHMIC METHODOLOGY (TC2038 FOCUS)

The project is formulated as a decision and optimization pipeline that integrates predictive modeling with explainability, robustness, fairness, and governance constraints. In contrast to traditional machine learning pipelines where explainability is treated as a post-hoc visualization step, this work frames Explainable Artificial Intelligence (XAI) as a core algorithmic problem whose correctness, complexity, and limitations must be analyzed alongside the predictive model itself [5].

The overall system is decomposed into four modules (M1–M4), corresponding to incremental algorithmic capabilities and aligned with the midterm and final project deliverables.

TABLE I
WORK PLAN BY MODULES AND DELIVERABLES

| Module | Topic | Deliverable |
|--------|---|-------------|
| M1 | AI + XAI (Shapley-style explanations) | Midterm 1 |
| M2 | Risk + AML (regularization, anomaly detection) | Midterm 1 |
| M3 | Strategy + Fairness (disparate impact audit) | Midterm 2 |
| M4 | Governance + Final Project Report (master file) | Midterm 2 |

A. Modeling

Let $X = \{x_i\}_{i=1}^n$ denote a dataset of historical loan applications, where each instance $x_i \in \mathbb{R}^d$ contains structured financial, behavioral, and demographic features. Let $y_i \in \{0, 1\}$ indicate whether applicant i defaults within a 12-month horizon. The objective is to learn a scoring function:

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

that outputs a calibrated default probability:

$$\hat{y}_i = f(x_i) = \Pr(y_i = 1 | x_i).$$

Candidate models include logistic regression and tree-based ensembles, which are widely used in credit scoring due to their favorable trade-off between predictive performance and interpretability [6]. These models also support efficient post-hoc explanation methods required for regulatory transparency.

B. Regularized learning

Model training is formulated as a regularized empirical risk minimization problem:

$$\min_f \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \lambda \Omega(f), \quad (1)$$

where $\ell(\cdot)$ is a loss function (e.g., log-loss or squared error), $\Omega(f)$ is a complexity penalty (e.g., L2 for linear models), and $\lambda > 0$ controls the bias–variance trade-off. Regularization improves generalization performance and mitigates overfitting, particularly under limited or noisy data regimes [3]. Additionally, prior work has shown that smoother decision boundaries induced by regularization lead to more stable and consistent explanations, improving robustness of XAI outputs under small perturbations [1].

C. Explainability

Each prediction is decomposed into per-feature additive contributions using Shapley-style attributions:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^d \phi_j(x_i), \quad (2)$$

where ϕ_j denotes the marginal contribution of feature j . Shapley-based methods satisfy desirable axioms such as efficiency, symmetry, and consistency, making them suitable for high-stakes decision settings [7]. Both local explanations (per applicant) and global explanations (aggregated feature importance) are generated to support regulatory and business interpretation.

D. AML anomaly detection

In parallel with default risk estimation, an anomaly detection module is incorporated to identify unusual behavioral or transactional patterns associated with potential money laundering risks. This component operates independently of the supervised credit scoring model and relies on unsupervised anomaly detection techniques.

Formally, an anomaly score is computed as:

$$a_i = g(x_i), \quad (3)$$

where higher values of a_i indicate deviations from typical behavior. Such approaches are commonly used in financial surveillance systems where labeled anomalies are scarce [4]. By decoupling AML detection from credit scoring, the system avoids contaminating default predictions while still supporting compliance objectives.

E. Fairness auditing

Fairness is evaluated on the model's operational decision \tilde{Y} (e.g., approve/deny derived from \hat{y} via a fixed threshold or risk bands). Let G denote a sensitive attribute and \tilde{Y} the decision. Disparate impact is assessed via selection rates across groups:

$$\Pr(\tilde{Y} = 1 | G = g_1) \text{ vs. } \Pr(\tilde{Y} = 1 | G = g_2)$$

and summarized with a selection-rate ratio. Where feasible, additional diagnostics include TPR/FPR gaps (error-rate balance) and calibration within groups to understand trade-offs between accuracy and parity. Such metrics are widely used to diagnose algorithmic bias and unequal treatment in automated decision systems [2]. This auditing framework establishes a quantitative baseline for fairness mitigation strategies implemented in M3.

F. Governance

The final module addresses governance and operational constraints required for safe deployment. Governance mechanisms include dataset and model versioning, experiment logging, explanation traceability, and reproducibility guarantees.

Each prediction is associated with an auditable record linking input data, model parameters, explanation artifacts, and decision outputs. Treating governance as an algorithmic systems concern aligns with emerging best practices for accountable AI in regulated domains [8].

V. DELIVERABLE: MIDTERM 1 (M1–M2)

A. Objective

The objective of Midterm 1 is to develop a first operational prototype of an explainable and robust credit risk scoring system, integrating explainable artificial intelligence techniques with regularization and an initial anti-money laundering (AML) anomaly detection layer. This deliverable focuses on establishing the algorithmic foundations and system structure required for a compliant, auditable decision pipeline, while laying the groundwork for fairness and governance extensions in Midterm 2.

B. Required Content

1) *Explainable scoring model (M1):* The first component consists of an explainable credit risk scoring model that predicts the probability of default within a 12-month horizon. Features are defined and justified from a systems and algorithmic perspective, incorporating traditional financial indicators (e.g., income, credit utilization, debt-to-income ratio) alongside alternative behavioral attributes when available. A baseline supervised learning model is trained and evaluated using AUC-ROC and calibration quality (e.g., ECE), and additionally reported with Brier score (probability accuracy). If the dataset is imbalanced, PR-AUC is also reported. Explainability is achieved through SHAP-style feature attribution methods, producing both local explanations (per applicant) and global summaries. Summary plots are analyzed to identify the most influential drivers of default risk and to assess whether model behavior aligns with domain expectations and regulatory constraints.

2) *Regularization and stress testing (M2):* To improve robustness and generalization, the scoring model is extended with explicit regularization. A regularization parameter λ is selected through controlled experimentation under predefined stress scenarios, such as feature noise injection, partial feature removal, or simulated distribution shifts. Performance and explanation stability are compared between unregularized and regularized models, highlighting trade-offs between accuracy, robustness, and interpretability. This analysis demonstrates how structural penalties contribute not only to improved generalization but also to more stable and reliable explanations under perturbations.

3) *AML detection prototype (M2):* An initial AML anomaly detection layer is implemented alongside the credit risk model. This module defines a set of anomaly signals derived from structured inputs, such as unusual feature combinations, extreme values, or atypical behavioral patterns. Thresholds are selected based on statistical deviation criteria rather than supervised labels.

The AML detector produces a structured output consisting of an anomaly score, alert flag, and associated metadata, and is integrated with logging mechanisms to ensure traceability. Importantly, this component operates independently of the default prediction model, ensuring that compliance signals do not distort credit risk estimation while still supporting regulatory monitoring requirements.

C. Evidence to Submit (Midterm 1)

- A technical report (5–7 pages) describing the scoring model design, explainability methodology, regularization strategy, stress testing results, and AML detection logic.
- A Git repository containing the end-to-end scoring pipeline, explainability notebooks, and AML detection scripts.
- A reproducible README detailing environment setup, experiment execution, and instructions to reproduce results from raw data to final explanations.

Together, these artifacts demonstrate a functional, explainable, and robust prototype that satisfies the objectives of modules M1 and M2 and prepares the system for fairness and governance extensions in Midterm 2.

VI. DELIVERABLE: MIDTERM 2 (M3–M4)

A. Objective

The objective of Midterm 2 is to complete the algorithmic lifecycle of the system by incorporating fairness analysis, bias mitigation strategies, and governance mechanisms required for sandbox-ready deployment. This stage extends the explainable and robust prototype developed in Midterm 1 toward a compliant, auditable, and operational decision system.

B. Required Content

1) *Disparate impact audit (M3):* A fairness audit is conducted to assess whether model decisions induce disparate impact across predefined sensitive attributes. Relevant attributes and comparison groups are explicitly defined based on

data availability and regulatory relevance. Approval rates and selected fairness metrics are computed and compared across groups, enabling a quantitative assessment of algorithmic bias. This audit establishes a baseline against which mitigation strategies are evaluated.

2) *Mitigation strategy (M3)*: Based on the audit results, a bias mitigation strategy is proposed and justified at the algorithmic level. Possible approaches include pre-processing adjustments, in-processing constraints, or post-processing calibration. Model performance and fairness metrics are re-evaluated after mitigation to analyze trade-offs between predictive accuracy, explainability, and fairness. The objective is not to eliminate all disparities, but to demonstrate controlled, measurable improvements under explicit constraints.

3) *Governance and master dossier (M4)*: Governance mechanisms are formalized to support traceability and controlled execution in a sandbox environment. These include model and data versioning, decision logging, and structured documentation describing model behavior, limitations, and intended use. Operational limits and monitoring rules are defined to constrain system behavior during sandbox execution. A transition plan outlines the requirements and risks associated with scaling the system toward full deployment.

4) *Cost–benefit decision matrix (closure)*: A cost–benefit matrix is constructed to compare the algorithmic approach with traditional credit decision processes. The analysis considers operational efficiency, explainability, compliance overhead, and regulatory risk, supporting an informed decision on system adoption.

C. Evidence to Submit (Midterm 2)

- An executive whitepaper (maximum 10 pages) summarizing technical results, fairness analysis, mitigation outcomes, and governance considerations.
- A Git repository containing fairness metrics, mitigation experiments, and the complete, runnable decision pipeline.
- A risk and compliance matrix identifying key risks and corresponding controls.
- Short video pitch (3 minutes) explaining the system architecture and value proposition.

VII. EVALUATION CRITERIA

The project is evaluated based on both algorithmic rigor and system-level completeness, reflecting the requirements of explainable artificial intelligence in regulated decision-making contexts. Assessment emphasizes correctness, transparency, robustness, and operational readiness.

The evaluation criteria are weighted as follows:

- Algorithmic correctness and system design quality: 35%.
- Explainability and interpretability clarity: 20%.
- Robustness and AML integration: 20%.
- Fairness analysis and mitigation effectiveness: 15%.
- Governance, documentation, and reproducibility: 10%.

VIII. ASSUMPTIONS AND RISKS

This project is developed under a set of explicit assumptions regarding data availability, system scope, and operational constraints. These assumptions, associated risks, and mitigation strategies are summarized below to ensure transparency and to contextualize the validity of the results.

A. Assumptions

The system assumes access to synthetic or properly anonymized loan application data that preserves the statistical properties of real-world credit portfolios while complying with privacy requirements. It further assumes a clear and stable definition of applicant profiles, feature semantics, and outcome labels (default within 12 months). Feature distributions are assumed to be sufficiently representative to support supervised learning and fairness evaluation.

B. Risks

Key risks include sampling bias due to non-representative or imbalanced datasets, which may distort both predictive performance and fairness assessments. The AML anomaly detection component may suffer from weak or ambiguous anomaly signals, particularly in the absence of labeled illicit behavior. Additionally, overfitting on small or rare subpopulations poses a risk to both robustness and explanation stability.

C. Mitigation Strategies

These risks are mitigated through the use of regularization, cross-validation, and controlled stress testing to improve generalization. Sensitivity analyses are applied to assess stability under perturbations, while periodic audits of model performance, explanations, and fairness metrics are planned to detect drift and unintended behavior over time.

REFERENCES

- [1] D. Alvarez-Melis and T. Jaakkola, “On the robustness of interpretability methods,” 2018.
- [2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*, 2019. Retrieved from <https://fairmlbook.org/pdf/fairmlbook.pdf>.
- [3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. Retrieved from <https://github.com/Benlau93/Data-Science-Curriculum/blob/master/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>.
- [4] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, 2009.
- [5] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017. Retrieved from <https://arxiv.org/pdf/1702.08608.pdf>.
- [6] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, 2001. Retrieved from https://www.researchgate.net/publication/2424824_Greedy_Function_Approximation_A_Gradient_Boosting_Machine.
- [7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *NeurIPS*, 2017. Retrieved from https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions.
- [8] I. D. Raji et al., “Closing the AI accountability gap,” in *FAT*, 2020. Retrieved from <https://dl.acm.org/doi/pdf/10.1145/3351095.3372873>.
- [9] A. Taneja, “Explainable Machine Learning for Loan Default Prediction: Enhancing Transparency in Banking,” 2021. Retrieved from <https://ijaibdcms.org/index.php/ijaibdcms/article/view/200>.