Research Candidate Name: Ravichandran Balashanmugam

Title

“Explainable Machine Learning for Credit Risk Assessment and Fraud Detection in Banking”

Abstract

The rapid adoption of Artificial Intelligence (AI) and Machine Learning (ML) in banking has transformed credit underwriting, fraud prevention, and customer risk profiling. However, most high-performing models (e.g., gradient boosting, deep learning) operate as “black boxes,” limiting transparency for regulators and customers. This research proposes an explainable ML framework that balances predictive accuracy, interpretability, and fairness for credit risk and fraud detection. The project will investigate hybrid architectures combining tree-based ensembles, SHAP/LIME explanations, and counterfactual reasoning to generate human-readable insights while maintaining strict regulatory compliance.

Background & Problem Statement

Financial institutions must make thousands of risk-related decisions daily: loan approvals, credit line adjustments, fraud alerts, and anti-money-laundering checks. Traditional statistical scoring models (e.g., logistic regression) provide clarity but underperform in complex data settings, whereas advanced ML achieves superior accuracy but sacrifices explainability. This trade-off challenges compliance with Basel III, GDPR “right to explanation,” and the US Equal Credit Opportunity Act, which require transparency and bias mitigation in automated decisions.

Current research focuses on post-hoc explanation techniques or simplified surrogate models, yet these approaches often fail to preserve fidelity under production conditions. There is a clear need for an end-to-end pipeline that integrates interpretability into the model design, enabling practitioners to deploy high-performance systems that remain auditable, fair, and trusted.

Research Objectives

1. Design an explainable ML architecture for credit risk scoring and fraud detection that achieves >95% of the accuracy of state-of-the-art black-box models.
2. Evaluate interpretability using quantitative metrics (e.g., fidelity, stability) and qualitative assessments by risk officers.

3. Develop bias-detection modules to ensure fairness across demographic groups while preserving predictive strength.
4. Prototype a visualization dashboard that enables analysts and regulators to explore model reasoning in real time.

Proposed Methodology

- **Data Acquisition & Preprocessing:** Use anonymized banking datasets (loan applications, transaction histories, fraud logs) with proper consent and security controls. Handle imbalance via SMOTE/undersampling.
- **Model Development:** Compare tree-based ensembles (XGBoost, LightGBM), explainable neural networks (tabNet, self-interpretable attention layers), and hybrid scorecards.
- **Explainability Layer:** Apply SHAP values, Integrated Gradients, and counterfactual analysis to provide local & global explanations.
- **Fairness Auditing:** Implement fairness metrics (equal opportunity, disparate impact) and post-processing algorithms (reject-option classification).
- **Evaluation:** Measure AUC, precision-recall, explanation fidelity, and user interpretability through surveys of credit analysts.
- **Prototype Interface:** Build a lightweight dashboard (e.g., Streamlit) to visualize predictions, feature attributions, and fairness indicators.

Expected Outcomes & Contributions

- A validated framework for explainable credit scoring and fraud detection that meets regulatory standards.
- Practical guidance for banks on balancing accuracy, fairness, and interpretability.
- Open-source reference implementation for research and educational use.
- Enhanced customer trust and regulator confidence in AI-driven financial services.

References (Selected)

- Lundberg, S. & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions (SHAP).*
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?" Explaining Classifiers.*
- European Banking Authority (2023). *Guidelines on the Use of Machine Learning in Credit Risk.*