Credit Risk Assessment with Uncertainty-Aware Decision Making and Human Escalation

Project Proposal

Group Name: Lazy Loaders

---

1. Background and Introduction

Credit risk assessment is a fundamental process in the financial industry where lending institutions evaluate the likelihood of a borrower defaulting on loan obligations. Traditional credit scoring systems, such as FICO scores, rely on rule-based approaches and statistical models. However, with the advancement of machine learning, financial institutions increasingly deploy sophisticated predictive models including neural networks, gradient boosting machines, and ensemble methods to make lending decisions.

Despite achieving high accuracy rates, these models often operate as "black boxes" that output binary approval/rejection decisions or probability scores without quantifying their confidence in these predictions. This lack of uncertainty awareness poses significant risks in several scenarios:

Edge Cases and Distribution Shift: Applicants with unusual financial profiles, recent immigrants with limited credit history, or individuals affected by economic disruptions (e.g., pandemic-related income changes) may fall outside the model's training distribution, leading to unreliable predictions.

High-Stakes Decisions: Loan decisions carry substantial financial implications for both lenders (risk of default) and borrowers (access to credit, interest rates). An incorrect decision can result in significant monetary losses or unfair denial of credit.

Regulatory Requirements: Financial regulations such as the Equal Credit Opportunity Act (ECOA) and Fair Lending laws require explainable and fair lending practices. Models must not only be accurate but also demonstrate appropriate confidence levels to justify their decisions.

Credit risk assessment qualifies as a safety-critical domain because:

- Financial safety: Poor credit decisions can lead to institutional insolvency and systemic financial risks
- Individual welfare: Erroneous rejections can deny individuals access to essential financial resources

- Fairness concerns: Overconfident models may perpetuate biases against underrepresented groups

Current industry practice typically involves fixed decision thresholds without consideration of prediction uncertainty, leading to two critical problems:

1. high-confidence incorrect predictions that should have been escalated.

2. uncertain predictions that are made automatically when human expertise would be more appropriate.

2. Literature Review

2.1 Traditional Credit Scoring and ML Approaches

Razaque et al. (2025) – "A Reinforcement Learning and Predictive Analytics Approach for Enhancing Credit Assessment in Manufacturing"

- Proposes a predictive-based reinforcement learning (PRL) model to enhance credit risk assessment for manufacturers and importers.
- Integrates predictive analytics with reinforcement learning to acquire more accurate and dependable credit risk evaluations.
- **Gap:** Does not provide uncertainty estimates or confidence intervals for predictions; treats all predictions as equally reliable.
- **Limitation:** No mechanism for handling out-of-distribution samples or identifying uncertain cases.

Link - https://www.sciencedirect.com/science/article/pii/S2772662225000165?ssrnid=4960837&dgcid=SSRN_redirect_SD

2.2 Deep Learning in Credit Risk

Sirignano et al. (2016) - "Deep Learning for Mortgage Risk"

- Applies deep neural networks to mortgage default prediction using Fannie Mae loan performance data
- Shows that deep learning models can capture complex non-linear relationships in credit data
- Gap: Does not provide uncertainty estimates or confidence intervals for predictions; treats all predictions as equally reliable
- Limitation: No mechanism for handling out-of-distribution samples or identifying uncertain cases

Link - https://arxiv.org/pdf/1607.02470

2.3 Uncertainty Quantification Attempts

Bussmann et al. (2021) - "Explainable Machine Learning in Credit Risk Management"

- Explores SHAP values and LIME for model interpretability in credit decisions
- Discusses the importance of transparency in lending decisions
- Gap: While interpretability is addressed, epistemic and aleatoric uncertainty are not quantified; no formal uncertainty estimation framework
- Does not propose rejection mechanisms for high-uncertainty cases

Link - https://link.springer.com/article/10.1007/s10614-020-10042-0

2.4 Machine Learning Applications in Credit Card Approval Prediction

Peiris (2019) - "Credit Card Approval Prediction by Using Machine Learning Techniques"

- Applies Artificial Neural Network (ANN) and Support Vector Machine (SVM) to predict customer eligibility for credit cards using demographic and transactional data
- Tests models using different batch sizes and learning rates under ANN, and evaluates both linear and nonlinear SVM
- Applies filter-based feature selection methods and uses SMOTE to handle class imbalance in the dataset
- Gap: While the study achieves high accuracy (0.88 with nonlinear SVM), it does not quantify prediction uncertainty or provide mechanisms for identifying when the model is uncertain about its predictions
- Limitation: The author notes that customer behavior may differ by country and suggests applying real banking datasets including COVID-19 impact as an area for future research
- Key finding: Nonlinear SVM outperformed ANN and linear SVM with accuracy of 0.88, precision of 0.88, and recall of 0.90

Link - https://dl.ucsc.cmb.ac.lk/jspui/bitstream/123456789/4593/1/2018%20BA%20026.pdf

2.5 Analytical Approaches to Credit Risk Using Machine Learning

Van der Plas et al. (2025) - "An analytical approach to credit risk assessment using machine learning models"

- Presents a novel Early Warning System for monitoring credit risk of commercial customers at a large international bank using machine learning algorithms including Random Forest, Gradient Boosting, and Neural Networks
- Employs SHAP values to enhance model explainability and support adoption in credit risk analysis
- Random Forest achieves the highest performance with strong F1 scores and successfully anticipates negative client transitions, helping prevent cases that would result in financial losses
- **Gap**: While the study emphasizes explainability through SHAP values, it does not address uncertainty quantification or provide confidence intervals for predictions

- **Limitation**: The study focuses on commercial banking customers rather than individual credit card applicants, and does not implement rejection mechanisms for high-uncertainty cases
- Key contribution: Demonstrates that explainable ML models can support data-driven decision-making in credit risk management

Link - https://www.sciencedirect.com/science/article/pii/S277266222500061X

3. Problem Statement

Core Problem: Existing credit risk assessment models provide point predictions (approve/reject or probability scores) without reliable uncertainty estimates, leading to overconfident decisions on edge cases where human judgment would be more appropriate and cost-effective.

Specific Deficiencies

3.1 Inability to Detect Out-of-Distribution Cases

- Models trained on typical applicants fail to recognize when they encounter unusual profiles (recent immigrants, gig economy workers, COVID-affected applicants)
- Confident predictions are made even when the applicant is unlike anyone in the training data

3.2 Lack of Uncertainty Quantification

- Current models cannot distinguish between:
  - **Aleatoric uncertainty:** Inherent unpredictability (e.g., unexpected life events causing default)
  - **Epistemic uncertainty:** Model's lack of knowledge due to insufficient similar training examples
- High epistemic uncertainty indicates cases requiring human expertise

3.3 No Rejection/Escalation Option

- Standard deployment makes predictions for all applicants regardless of confidence
- Results in errors on borderline cases that human loan officers could correctly assess
- Wastes human resources by routing easy cases (high confidence) to manual review while automating uncertain ones

3.4 Miscalibration

- Models often output poorly calibrated probabilities (e.g., predicting 80% approval probability but actual approval rate for such predictions is 60%)
- Miscalibration is severe for minority classes, underrepresented groups, and recent time periods

Impact: These deficiencies lead to financial losses (incorrect high-confidence predictions), unfair denials (uncertain predictions on creditworthy applicants), regulatory violations, and suboptimal resource allocation between automated and manual review.

4. Proposed Method

4.1 Base Predictive Model

Model Selection: Random Forest or XGBoost

- **Rationale:** Strong performance on tabular financial data, well-documented, relatively simple to implement
- **Output:** Binary classification (approve/reject) with probability scores

4.2 Uncertainty Quantification

Technique 1: Monte Carlo Dropout

- Enable dropout layers during inference
- Run 20-50 forward passes on each test sample
- Epistemic Uncertainty = Variance across predictions
- Advantages: Single model, minimal code changes, fast to implement

Technique 2: Bootstrap Ensemble

- Train 5-10 models on bootstrap samples of training data
- Epistemic Uncertainty = Disagreement among ensemble members
- Advantages: Straightforward, interpretable, works with any base model

Technique 3: Calibration with Temperature Scaling

- Post-hoc calibration to ensure probabilities reflect true confidence
- Learn temperature parameter on validation set
- Advantages: Improves reliability of probability estimates, simple to implement
- Evaluation: Expected Calibration Error (ECE), reliability diagrams

4.3 Human Escalation Strategy

Rejection Criteria (Send to human review if ANY condition met):

1. **High Ensemble Disagreement:** Variance > threshold (e.g., 0.15)
   - Indicates models fundamentally disagree about the applicant

2. **Low Maximum Probability:** max(P(class)) < threshold (e.g., 0.65)
   - Model is uncertain between approve/reject

3. **Conflicting Indicators:** Applicant has both strong positive and strong negative features
   - Example: High income but poor credit history

Threshold Selection:

- Use validation set to find optimal thresholds balancing automation rate vs. error rate
- Consider cost of manual review vs. cost of errors

4.4 Implementation Pipeline

- **Data Preparation & Baseline**
  - Load and explore dataset
  - Feature engineering and preprocessing
  - Handle missing values and outliers
  - Train baseline Random Forest/XGBoost model
  - Evaluate accuracy, precision, recall, AUC
  - **Deliverable:** Baseline model with performance report

- **Uncertainty Quantification**
  - Implement MC Dropout OR Bootstrap Ensemble
  - Implement Temperature Scaling calibration
  - Generate uncertainty estimates on validation set
  - **Deliverable:** Uncertainty-aware model with calibration

- **Human Escalation & Evaluation**
  - Define rejection thresholds using validation data
  - Implement decision logic (automated vs. escalated)
  - Comprehensive evaluation:
  - Accuracy on retained (confident) predictions
  - Automation rate (% cases handled automatically)
  - Error rate on escalated vs. non-escalated cases
  - Create visualizations (uncertainty distributions, calibration plots)
  - **Deliverable:** Full system with evaluation metrics

- **Analysis & Documentation**
  - Cost-benefit analysis (savings from avoiding errors vs. manual review costs)
  - Identify which types of applicants trigger escalation
  - Compare with baseline (no uncertainty) approach
  - Write final report and prepare presentation
  - **Deliverable:** Complete project report

5. Available Datasets

5.1 German Credit Data

- **Source:** UCI Machine Learning Repository
- **Size:** 1,000 applicants
- **Features:** 20 attributes (credit history, employment, demographics, loan purpose)
- **Target:** Binary (good/bad credit)
- **Advantages:** Small, well-studied, quick to train, good for prototyping
- **Limitations:** Small size limits generalization
- **Link -** https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

5.2 Lending Club Loan Data

- **Source:** Kaggle
- **Size:** ~400,000 loans (can use subset)
- **Features:** 100+ attributes (loan amount, interest rate, employment, income, credit history, loan status)
- **Target:** Loan status (Fully Paid, Charged Off, Default)
- **Advantages:** Large-scale real data, rich features, temporal aspect (can study distribution shift)

- **Limitations:** Requires more preprocessing

- **Link -** https://www.kaggle.com/datasets/urstrulyvikas/lending-club-loan-data-analysis

5.3 Kaggle Credit Card Default (Taiwan)

- **Source:** UCI/Kaggle
- **Size:** 30,000 customers
- **Features:** 23 attributes (payment history, bill amounts, demographics)
- **Target:** Default next month (binary)
- **Advantages:** Clean, moderate size, credit card specific
- **Limitations:** Single geography, specific to credit cards

- **Link -** https://www.kaggle.com/datasets/jishnukoliyadan/taiwan-default-credit-card-clients

6. Expected Contributions

1. **Empirical demonstration** that uncertainty-based escalation improves outcomes:
    - Higher accuracy on automated decisions
    - Fewer errors overall
    - Better resource allocation

2. **Cost-benefit analysis** showing when uncertainty-aware systems are worthwhile

3. **Implementation guide** and code repository for reproducibility

This research addresses a critical gap in deploying ML for financial decision-making by providing practitioners with tools to identify when models should defer to human expertise, ultimately enabling safer and more responsible AI adoption in credit risk assessment.