CORAL-X: Contextual Risk Assessment for Loan default prediction using LLMs for Explainability

Authors: Kahan Sheth and Harsh Bhatt
Northeastern University
DS 5500 Capstone: Applications in Data Science
Spring 2025

March 10, 2025

Contents

1 Purpose of the Methodology

CORAL-X (Contextual Risk Assessment for Loan default prediction using LLMs for Explainability) unifies both structured financial data and unstructured economic/policy documents (e.g., federal lending regulations, regional statistics) to create a more comprehensive and **context-aware** loan default prediction system. By integrating traditional **machine learning** (ML) classifiers (such as XGBoost or LightGBM) with **LLM-driven retrieval** (e.g., Retrieval-Augmented Generation), the methodology effectively addresses the following:

1. **Enhanced Risk Assessment:**
   - Enriches borrower-centric features with broader macroeconomic indicators and regional policy factors, providing a deeper understanding of default risk drivers.
   - Reduces the chance of overlooking external variables (e.g., recession trends or local unemployment spikes) that can affect creditworthiness.
2. **Explainability and Transparency:**
   - Incorporates LLM-based reasoning to produce human-readable justifications for model decisions, ensuring **interpretability** for stakeholders such as financial analysts and regulators.
   - Supplements model-agnostic explanations (e.g., SHAP values) with contextual policy insights, thus bolstering trust and regulatory compliance in high-stakes finance.
3. **Mitigating Biases and Uncertainty:**
   - Merges structured and unstructured data sources to guard against data biases, where purely borrower-level indicators may not capture systemic or environmental effects.
   - Facilitates cross-verification of decisions using external knowledge bases (policy documents, economic reports), thus minimizing blind spots in credit risk modeling.

Why Compare Multiple Models?

Comparing distinct ML algorithms (e.g., **XGBoost**, **LightGBM**, and baseline approaches) is critical to:

- **Evaluate Performance Trade-offs:** Each model behaves differently with respect to accuracy, F1-score, and AUC-ROC, revealing which architectures most reliably detect high-risk borrowers under varying data conditions.
- **Assess Interpretability Gains:** Some models may excel at raw performance, whereas LLM-driven pipelines may produce superior explanation clarity. Balancing predictive strength with transparency is key.
- **Derive Actionable Insights:** Differences in feature importance highlight consistent default risk factors (e.g., policy- or region-specific), guiding further refinements to the final pipeline.

By benchmarking multiple approaches using both quantitative (accuracy, AUC-ROC) and qualitative (explanation clarity, domain relevance) metrics, CORAL-X ensures a methodology that is technically robust, transparent, and fair.

2 Problem Statement

2.1 Defining the Problem

Traditional loan default prediction systems often focus exclusively on individual borrower data (e.g., credit scores, employment history), which can lead to suboptimal decisions—especially among underserved populations lacking robust financial histories. These methods fail to account for contextual factors such as regional economic conditions, policy impacts, or income trends without extensive manual intervention by credit analysts.

CORAL-X bridges this gap by **integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG)** to dynamically incorporate insights from publicly available data, including regional macroeconomic metrics, regulatory documents, and policy guidelines. This combined approach:

- **Automates Contextual Extraction:** Pulls in relevant external information (e.g., unemployment rates, current local economic initiatives) without requiring human curation at every step.
- **Maintains Transparency:** Generates explainable credit decisions with natural-language justifications that policy makers and financial analysts can interpret.
- **Supports Fairness and Inclusivity:** Improves decisions for borrowers in regions with historically limited credit access by recognizing contextual nuances rather than relying solely on borrower-level attributes.

In doing so, **CORAL-X** transforms the *risk assessment workflow* from a narrow, borrower-centric paradigm into a **context-rich**, **automated**, and **transparent** framework, enabling more equitable lending outcomes across diverse socio-economic landscapes.

2.2 Significance

In the modern financial landscape, credit decisions carry significant socio-economic consequences. While advanced machine learning models have made strides in predictive accuracy, the lack of broader contextual knowledge can perpetuate biases and erode trust among borrowers. By integrating LLMs and RAG to harness external data sources (for example, macroeconomic indicators and regulatory guidelines), CORAL-X contributes to loan default prediction in ways that are:

- **Fair and Inclusive:** Recognizes critical external factors affecting underserved communities, ensuring that credit evaluations account for socio-economic disparities rather than only individual borrower history.
- **Transparent and Explainable:** Combines model-agnostic techniques (such as SHAP values) with natural-language justifications, clarifying how contextual elements including regional economic policies or income trends influence credit risk assessments.
- **Scalable and Automated:** Reduces reliance on human intervention by systematically retrieving and embedding relevant policy and economic data, accelerating risk assessment workflows while maintaining rigor.
- **Regulatory-Compliant:** Helps financial institutions align with fair lending requirements by clearly documenting how external policy information and economic signals inform lending decisions, thereby promoting adherence to evolving legal standards.

Taken together, CORAL-X not only strengthens predictive performance, but also addresses the practical need for equitable and transparent credit models in high-stakes lending environments.

3 Data Collection and Preparation

3.1 Data Sources

1. **Structured Data:**
   - **Lending Club Loan Dataset:** A large, public dataset from Kaggle containing over 2 million P2P loan records (loan amount, interest rate, borrower financial profiles, etc.).
2. **Unstructured Data:**
   - **Policy Documents / Regulatory Texts:** PDFs describing fair lending regulations, underwriting guidelines, or official economic reports (e.g., FDIC, CFPB).
   - **Economic Indicators:** Publicly available sources (U.S. Census Bureau, OECD) for regional unemployment, poverty rates, or income distributions.

3.2 Data Description

- **Volume:** Over 2 million structured loan records (≈500 MB) and multiple unstructured PDFs.
- **Features:** Include borrower demographics, loan attributes, and newly added macro-level indicators (e.g., regional data).
- **Target Variable:** `loan_status`, labeled as *Fully Paid* or *Charged Off*.

3.3 Data Preparation Steps

- **Data Cleaning:**
  - Missing values: Impute numerical columns with median or use tree-based models (XGBoost) that natively handle `NaN`s.
  - Outlier checks: IQR-based detection, generally leaving outliers in place if the model is robust to them.
  - Handling the data leakage issue by removing certain features.
- **Class Imbalance:**
  - Significant majority of *Fully Paid* loans.
  - Apply random oversampling or SMOTE to address minority (*Charged Off*) class imbalance.
- **Feature Engineering:**
  - One-hot encoding or ordinal encoding for categorical features.
  - Scaling of numerical features (`StandardScaler`).
  - Selection or removal of highly correlated variables (e.g., `loan_amnt` vs. `installment`).
- **Unstructured Data Processing:**
  - *PDF Parsing / Chunking:* Split large documents into 500–1,000 token chunks.
  - *Embedding:* Use `multilingual-e5-large` or similar model.
  - *Indexing:* Store embeddings in Pinecone for rapid retrieval.

4 Selection of Machine Learning and LLM Models

4.1 Model Consideration

1. **Traditional ML Models**
   - **XGBoost, LightGBM:** Known for strong performance on tabular data, can handle missing values, outliers, and large feature spaces efficiently.
   - **Random Forest or Logistic Regression:** Serve as baseline comparisons.
2. **LLM-Based Models**
   - **Mistral:** A streamlined large language model designed for efficient inference and balanced performance. Its compact size makes it a practical choice for generating concise explanations and handling moderate-scale policy documents.
   - **LLaMA 2:** Builds on the original LLaMA framework, offering robust language comprehension and improved parameter scaling. LLaMA 2 excels in generating coherent text over diverse contexts, making it highly adaptable to the variety of policy guidelines relevant to credit scoring.
   - **LLaMA 3:** A next-generation large language model that builds on Meta's open-source LLaMA architecture. It is designed to handle more extensive context windows and improve inference for domain-specific prompts, making it well-suited for tasks such as retrieving policy insights and generating detailed, domain-relevant responses in a credit-risk setting.
   - **RAG Pipeline:** Retrieval-Augmented Generation leverages vector databases to insert dynamically retrieved context into LLM prompts, ensuring that model outputs remain up-to-date and grounded in the latest policy or economic documentation.

4.2 Justification

- **XGBoost / LightGBM** excel with large structured datasets and demonstrate high AUC-ROC in initial experiments.
- **LLMs** add contextual reasoning by retrieving macroeconomic data or policy constraints, enabling more transparent, domain-specific explanations.

4.3 Final Selection

- Hybrid pipeline: **XGBoost** for numeric classification + **RAG** for contextual retrieval and textual justification.
- Evaluate on **accuracy, F1, AUC-ROC**, plus **explanation clarity** (SHAP attributions + LLM-based narratives).

5 Model Development and Training

5.1 Architecture

This section outlines the end-to-end workflow of CORAL-X, from data ingestion to model inference. The architecture components, including how structured data moves through a machine learning pipeline and how unstructured documents are embedded and retrieved via RAG-based LLMs, are depicted in Figure 1.

1. **Data Ingestion**
   - **Structured Data:** Readily available tabular datasets (e.g., Lending Club), which undergo cleaning, transformation, and feature engineering. These processed features then form the training input for XGBoost.
   - **Unstructured Data:** Policy or economic documents are chunked and embedded using a language model (for example, multilingual-e5-large). The resulting embeddings are indexed in Pinecone for later retrieval.
2. **XGBoost Training**
   - *Hyperparameter Tuning:*
     - **GridSearchCV**: Exhaustively searches specified parameter values (e.g., max_depth, learning_rate).
     - **RandomizedSearchCV**: Samples from specified distributions of hyperparameters, typically more efficient for larger search spaces.
     - **Bayesian Optimization with Optuna**: Iteratively refines hyperparameter selection by modeling the objective function and intelligently sampling parameters likely to yield better performance.
   - **Data Splits for Validation:** We split the dataset into a training and hold-out test set to assess out-of-sample performance. Each hyperparameter search process (Grid, Random, Bayesian) trains multiple candidate models and selects the best based on validation metrics (e.g., AUC-ROC, F1-score).
3. **LLM + RAG**
   - **Document Embedding and Indexing:** Unstructured documents are embedded and stored in Pinecone. During inference, relevant text chunks are retrieved based on semantic similarity to the query.
   - **Prompt Engineering:** The system combines partial model outputs (e.g., default risk probabilities or SHAP explanations) with retrieved text to form context-enriched prompts for the language model (e.g., GPT, LLaMA 3).
   - **Explanation and Reporting:** The LLM generates a user-friendly justification for the loan default risk decision, referencing both borrower-centric features (from the structured data) and contextual economic/policy insights.

5.2 Prompt Engineering for Explanation Generation

Rather than fine-tuning a large language model, CORAL-X employs carefully designed prompts to guide the model toward clear, policy-aware credit risk justifications. Prompt engineering involves:

- **Context Incorporation:** Relevant text chunks from the Pinecone vector database (e.g., economic indicators, policy statements) are inserted into the prompt to ensure the model references up-to-date domain information.
- **Structured Output Guidance:** The prompt may include instructions or templates (for instance, "Explain the default risk in terms of loan amount, DTI ratio, and regional unemployment data..."), which help the model produce concise, coherent responses tailored to credit risk analysis.
- **Iterative Refinement:** Prompt design is refined by testing outputs on sample loan applications and evaluating them for accuracy, clarity, and compliance. Adjustments are made to further emphasize critical features or highlight specific policy constraints.

By focusing on prompt engineering—rather than a full model fine-tuning—CORAL-X achieves robust, context-aware explanations that are easier to maintain, faster to update, and cost-effective to deploy.

6 Evaluation and Comparison

6.1 Quantitative Metrics

- **Accuracy, F1-Score:** General classification performance.
- **AUC-ROC:** Separability between the default vs. non-default classes.
- **Precision/Recall:** Assess model performance in detecting true defaulters.

6.2 Qualitative Metrics

- **Explainability (SHAP Values):** Identifies the contribution of each feature in a prediction.
- **LLM Explanation Relevance:** Human evaluation to gauge correctness of contextual references (policy docs, macro data).

6.3 Model Comparison

- **XGBoost vs. LightGBM:** Typically, XGBoost has shown slightly higher AUC–ROC in pilot tests, while LightGBM may train faster on certain hardware configurations.
- **LLM Variants:** In our evaluations, `Llama3` outperformed both `Llama2` and `Mistral`. Its responses demonstrated stronger domain-specific knowledge, yielding more coherent and contextually relevant outputs than the other models.

6.4 Conclusion of Comparisons

- **Performance**: Hybrid XGBoost + RAG outperforms baseline logistic regression or random forest approaches.
- **Interpretability**: Combined *SHAP* + LLM narrative provides a thorough, context-driven rationale for each prediction.

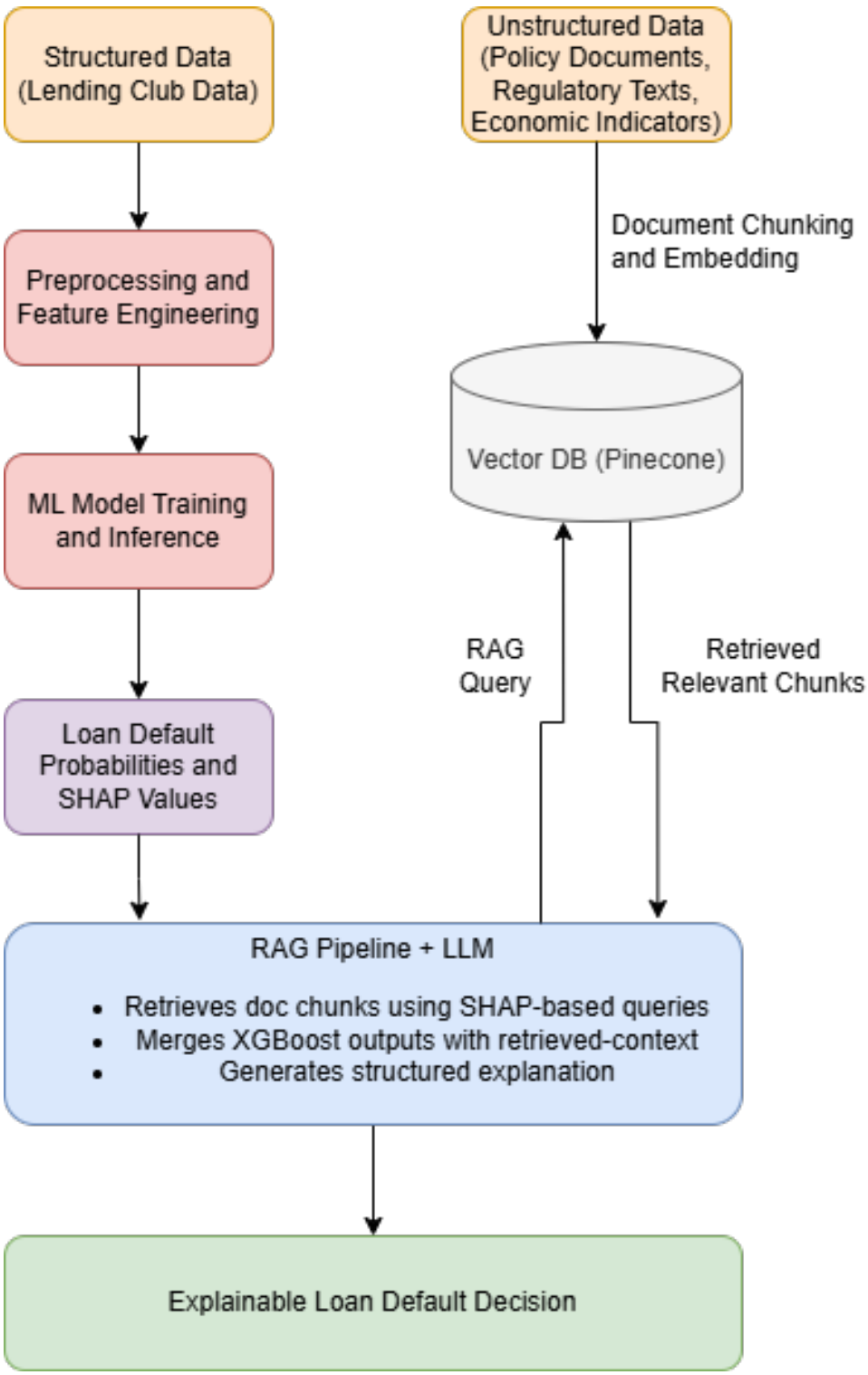

Figure 1: High-level architecture of CORAL-X demonstrating the full ML pipeline and RAG-based retrieval mechanism for LLMs.