Mortgage Credit Risk Prediction using Machine Learning Models

Team members: Yiming Jia (yj3229), Albert Kong (lk3189), Smruti Nalawade (ssn9062),
Peter Zhong (fz2411), Ziyue (Tom) Zhou (tz1307)

Introduction

The Expected Credit Loss (ECL) model is a cornerstone of modern credit risk assessment, mandated by regulatory frameworks such as IFRS 9 and CECL[1]. By estimating credit risk through a predefined formula—Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD)—ECL ensures consistency and compliance in financial reporting. However, despite its regulatory necessity, the model remains inherently limited in its ability to fully capture the complexity of credit risk. It relies on a fixed set of variables and linear assumptions, restricting its adaptability to the diverse and high-dimensional nature of real-world loan data.

With the rapid advancement of machine learning and AI-driven analytics, there is a growing opportunity to enhance credit risk assessment by leveraging richer data attributes and more sophisticated modeling techniques. In this project, we aim to explore innovative approaches that integrate machine learning into the credit risk evaluation process, complementing the ECL framework while maintaining regulatory compliance. Our approach seeks to cover a wider range of potential factors to provide a more comprehensive and adaptive understanding of credit risk. To be more specific, this project will focus on:

Objective

1. *Loan Default Prediction:* Utilizing historical loan performance and borrower attributes to improve default forecasting.
2. *Credit Loss Prediction*: Enhancing loss estimation by leveraging non-linear relationships and data-driven insights beyond traditional risk metrics.
3. *Forecasting Delinquency Trends and Macroeconomic Impacts*: Integrating macroeconomic indicators to assess how economic shifts influence credit risk.
4. *Identifying High-Risk Geographic Areas*: Leveraging spatial data analytics to enhance risk segmentation and proactive risk management strategies.

Proposed Prediction Methodologies

Our approach leverages machine learning techniques to enhance mortgage loan default and Expected Credit Loss (ECL) prediction while addressing broader credit risk factors such as delinquency trends, macroeconomic impacts, and geographic risk patterns. The methodology consists of the following key components:

1. Loan Default and ECL Prediction

[1] IFRS - IFRS 9 Financial Instruments

In our revised framework, Expected Credit Loss (ECL) is computed by multiplying the probability of default (PD) with the predicted actual loss. Since our dataset includes an actual loss column for each loan, the process is streamlined: we focus on developing a robust machine learning model to estimate PD based on borrower profiles and loan characteristics, and then derive ECL using the formula ECL = PD × Predicted Actual Loss.

2. Forecasting Delinquency Trends

To forecast potential risks and market shifts, we model delinquency trends as a high-dimensional time-series forecasting problem. Our approach aggregates historical records, macroeconomic indicators, and borrower-specific risk factors into a comprehensive multivariate time series. Although payment data is not available, we employ a combination of traditional machine learning models alongside deep learning techniques to capture sequential patterns and trend dependencies. This integrated methodology enables us to effectively forecast delinquency trends and identify early indicators of potential credit risks.

3. Assessing the Impact of Economic Conditions on Mortgage Default Rates

Since macroeconomic conditions significantly influence mortgage default rates, we incorporate external economic data such as interest rates, unemployment rates, and inflation into our models. Causal inference techniques, including SHAP (SHapley Additive exPlanations) and Granger Causality Tests, are used to quantify the relationship between economic changes and default risks. Counterfactual simulations are performed to explore how mortgage defaults would behave under various economic scenarios. This ensures that our approach not only captures historical trends but also provides forward-looking risk insights.

4. Identifying Geographic Risk Patterns

Geographic risk analysis is another critical component of our methodology. Using Geographic Information Systems (GIS) and spatial econometrics, we identify high-default regions and analyze state-wise credit risk variations. Clustering algorithms, such as K-Means and DBSCAN, are employed to detect geographic patterns in loan defaults, while interactive heatmaps and geospatial visualizations provide a clear representation of emerging risk areas. These insights help financial institutions refine regional credit policies and adjust lending strategies based on location-specific risks.

Dataset

1. Freddie Mac Dataset:

The Single-Family Loan-Level Dataset from Freddie Mac offers comprehensive mortgage data at the loan level, including origination details, monthly performance metrics, and actual loss information for fully amortizing fixed-rate single-family mortgages. Covering loans originated from 1999 to the present, the dataset encompasses critical variables such as borrower credit scores, loan-to-value and debt-to-income ratios, interest rates, delinquency status, and loss components, along with key events like modifications, prepayments, and foreclosures. In this study, a representative subset of the dataset is employed to develop machine learning-based predictive models for mortgage defaults and expected credit loss, while also examining delinquency trends, macroeconomic impacts, and geographic risk patterns.

2. St. Louis Federal Reserve Dataset:

Complementing the mortgage data, this research also leverages macroeconomic indicators from the Federal Reserve Bank of St. Louis via its FRED API. The FRED database provides timely and historical data on essential economic variables—including GDP, unemployment, inflation, and interest rates—offering a robust framework for understanding the broader economic environment. Integrating these high-frequency economic metrics enriches our analysis by contextualizing mortgage performance within prevailing economic conditions, thereby enhancing the overall evaluation of credit risk dynamics.

Python Optimization Methodologies

To ensure efficient computation while processing large-scale mortgage data, we employ various advanced Python optimization techniques. These methods enhance execution speed, improve memory efficiency, and allow seamless handling of high-dimensional data. The key methodologies implemented in this project include:

1. Python Performance Tuning and Profiling

Before implementing optimizations, we utilize profiling tools such as cProfile, line_profiler, and memory_profiler to identify bottlenecks in our code. By analyzing function execution times and memory usage, we pinpoint areas where performance improvements are needed.

2. Just-In-Time (JIT) Compilation

We leverage Numba, a JIT compiler that translates Python functions into optimized machine code at runtime, significantly accelerating numerical computations. This is particularly beneficial for iterative calculations involving large datasets, such as probability of default (PD) and expected credit loss (ECL) estimations.

3. Parallelization with Multi-Threading and Multi-Processing

Since mortgage data analysis involves complex computations across millions of records, we employ parallel processing to distribute tasks across multiple CPU cores. Depending on the workload, we utilize Multi-threading (ThreadPoolExecutor) for I/O-bound tasks and Multi-processing (ProcessPoolExecutor) for CPU-intensive computations like clustering and time-series forecasting.

4. Vectorization with NumPy and Pandas and Efficient Data Loading with Dask

Instead of using explicit loops, we harness vectorized operations via NumPy and Pandas, ensuring optimized performance when handling large datasets. Instead of loading large datasets into memory at once, we use Dask to handle out-of-core computations efficiently.

5. GPU Acceleration with CuPy

For extremely large-scale computations, we utilize CuPy, a GPU-accelerated library similar to NumPy. It enables offloading heavy matrix computations to the GPU, achieving substantial speedups.

Evaluation Methods

- For model evaluation, we employ standard metrics to assess predictive accuracy:
  - AUC-ROC
  - Precision-Recall
  - Mean Squared Error (MSE) and Mean Absolute Error (MAE)
- For visualization and interpretation:
  - Lime
  - SHAP
  - Heatmap

Outcomes and Benefits

This study aims to benefit researchers and financial institutions in these areas:

1. Improved Default Prediction

Machine learning–based modeling of default prediction enhances accuracy for both researchers exploring risk factors and financial institutions optimizing underwriting decisions.

2. Refined ECL Estimates

By leveraging predicted PD and actual loss information, this study produces more precise Expected Credit Loss (ECL) figures, aiding financial institutions in proactive provisioning and satisfying regulatory requirements.

3. Forward-Looking Risk Insights

Integrating macroeconomic indicators yields actionable forecasts of delinquency trends and geographic “hot spots,” assisting both researchers and financial institutions in identifying and mitigating emerging credit risks.

4. Model Interpretability and Efficiency

Techniques like SHAP and LIME enable transparent analyses crucial for regulatory compliance and stakeholder trust. Parallelization and GPU acceleration facilitate large-scale data handling for high-volume mortgage portfolios.

Proposed Project Timeline

- Stage 1: Data Acquisition & Preprocessing (Weeks 1–2)
- Stage 2: Exploratory Analysis & Feature Engineering (Weeks 2–3)
- Stage 3: Model Development – PD & ECL (Weeks 3–4)
- Stage 4: Delinquency Forecasting & Geographic Analysis (Weeks 4–5)
- Stage 5: Model Evaluation & Tuning (Weeks 5–7)
- Stage 6: Finalization & Deliverables (Weeks 7–8)