Mortgage Credit Risk Prediction using Machine Learning Models

Github link : https://github.com/AlbertKong0827/DSGA1019Project.git

1. Introduction

This project aims to enhance mortgage credit risk prediction by integrating borrower, loan, macroeconomic, and geographic data through machine learning. Traditional ECL models, while regulation-compliant, often rely on limited variables and linear assumptions, making them less responsive to changing borrower behavior and external economic conditions. To address these limitations, we implement a unified ML pipeline that forecasts default probability, estimates ECL, models delinquency trajectories, and maps geographic risk. Leveraging Freddie Mac loan-level data and macroeconomic indicators from the FRED API, our approach combines supervised learning, time-series modeling, and spatial clustering to improve predictive accuracy and enable early detection of emerging credit risk patterns.

2. Methodologies

2.1 Expected Credit Loss Predictive Modeling

Developing a full-scale Expected Credit Loss (ECL) model that is production-ready, fully validated, and compliant with regulatory standards is a resource-intensive undertaking. It typically requires years of development, extensive historical data integration, expert judgement, robust infrastructure, and model governance frameworks. Given the time constraints of this project, our goal is to build a proof-of-concept (PoC) to evaluate the feasibility of using loan-level data and machine learning techniques to estimate 12-month ECLs. This PoC allows us to explore model architecture, data transformation pipelines, and preliminary model performance while identifying gaps and areas for refinement in a controlled, low-stakes setting.

Our approach breaks ECL into its core components: Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD). ECL is then computed using the standard formula: $ECL_{12M} = PD_{12M} \times LGD \times EAD$

For LGD, we adopted Tong et al.[1]'s zero-adjusted gamma mixture: a logistic regression for the chance of zero loss followed by a log-link Gamma GLM for positive losses . We used a reduced, domain-driven feature set to reduce the risk of overfitting, the Gamma GLM achieved an RMSE of $\$168,913$, an MAE of $\$129,825$ and a pseudo-R^2 of 0.09. We also tried XGBoost on the same task, its RMSE was slightly lower at $\$161,085$ but its R^2 was only 0.008, showing no gain in explained variance.

For PD, we labeled a loan "default" if it hit 60 days past due within the next 12 months, then fitted a logistic regression on the same six features. This model produced a PD_{12M} probability and achieved an ROC AUC of 0.62.

For EAD, we used two straightforward proxies: the prior-period unpaid balance and, for loans exiting the portfolio, the balance at removal. Multiplying PD × LGD by each EAD gave two ECL series; on average, our combined ECL prediction had an RMSE of $\$338,224$.

This PoC demonstrated that even straightforward, interpretable models and readily available loan-level features can capture meaningful credit-loss signals. The modest R^2 and AUC values are not surprising as dollar-loss outcomes are extremely heavy-tailed and defaults are rare. It also highlighted where further work is needed - richer feature sets, non-linear model forms and more sophisticated EAD projections - to move from feasibility testing toward a production-grade ECL engine.

2.2 Delinquency Risk Forecasting with Historical and Macroeconomic Data

To investigate temporal risk patterns in the mortgage market, we first conducted time series forecasting on the average delinquency period per month over the past 25 years. Features directly related to the loan—such as current UPB, interest rate, and credit score—were used as predictors. We implemented two forecasting models: LightGBM and LSTM. Both models performed well, confirming that the data exhibits strong sequential characteristics.

[1] Tong, E. N., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. International Journal of Forecasting, 29(4), 548-562.

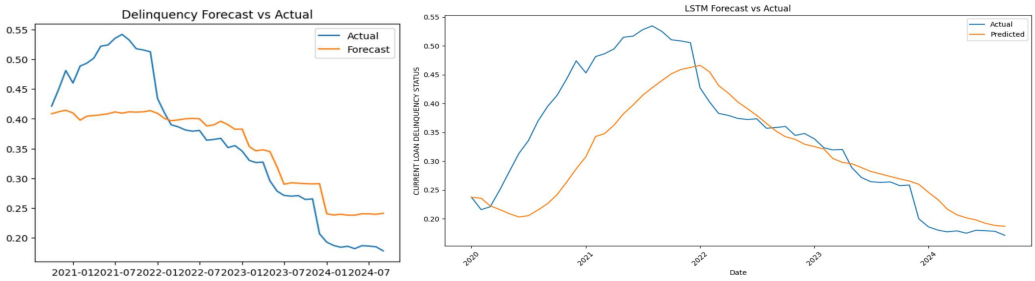


Figure 2.2.1. Model Prediction (LightGBM vs. LSTM)

2.2.1 Adjusted Delinquency Index Definition

However, we identified a limitation in interpreting the rising average delinquency time over the years. This increase does not necessarily reflect a growing number of delinquent loans but may instead indicate that older delinquent loans have accumulated longer unpaid periods. To address this, we proposed a more robust metric: the average number of days it takes for a loan to first become delinquent.Recognizing that this average metric alone lacks context, we developed a composite delinquency index (The Adjusted Delinquency Index) inspired by the Volume Price Trend (VPT)[2] indicator used in financial markets. By treating delinquency (i.e., days to 1st delinquency) as the "price" and loan count as the "volume," we constructed an adjusted index that reflects both the severity and scale of delinquency events. To ensure the index properly signals deteriorating market conditions, we inverted the average delinquency days (so lower quality results in higher index values) before combining it with the loan volume. As shown in Figure 3, the resulting indicator demonstrates improved seasonality and alignment with known historical trends, offering a more nuanced lens into evolving mortgage credit risk.

$$\text{Adjusted Delinquency Index} = \left(9999 - \frac{\text{Days to First Delinquency}}{\text{Number of Delinquent Loans}}\right)$$

$$\text{Adjusted Delinquency Index}_t = \text{Adjusted Delinquency Index}_{t-1} + \text{Number of Loans}_t \times \left(\frac{\text{Adjusted Delinquency Index}_t - \text{Adjusted Delinquency Index}_{t-1}}{\text{Average Delinquency Index}_{t-1}}\right)$$

Figure 2.2.2. Formula for Adjusted Delinquency Index

To further test its predictive utility, we trained an LSTM model using a range of macroeconomic indicators. The model achieved an RMSE of **0.0161** and MAE of **0.011**, indicating high predictive accuracy. These results underscore not only the sequential nature of delinquency risk but also the strong influence of macroeconomic conditions on mortgage market stability. This demonstrates the potential of our approach to serve as an early warning system for emerging credit risk, enabling data-driven intervention and proactive risk management.

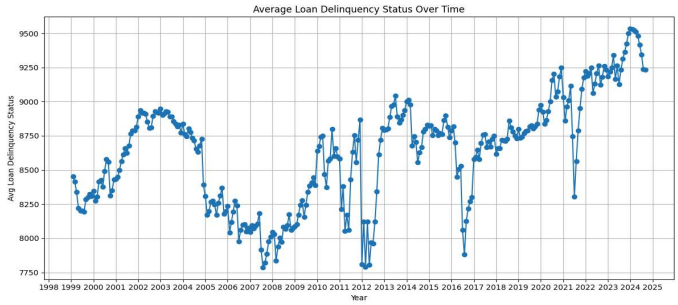


Figure 2.2.3. Adjusted Delinquency Index Over Time

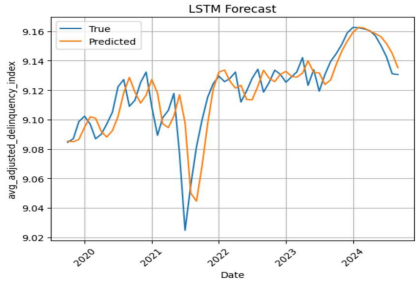


Figure 2.2.4. LSTM Model Prediction

2.2.2 Early Warning Macro Index Selection

Building on the above work, we then sought to identify macroeconomic indicators that can serve as early warning signals for elevated mortgage delinquency risk. While the LSTM model demonstrated strong predictive performance, it remains difficult to interpret which macro features offer early signs of stress in the loan market. To address this, we developed a dedicated pipeline that systematically screens over **4,000** macroeconomic time series from the Federal Reserve Economic Data (FRED) and evaluates their predictive power for our Adjusted Delinquency Index.

[2] Caginalp, G., & Desantis, M. (2009). Stock price dynamics: nonlinear trend, volume, volatility, resistance and money supply. Quantitative Finance, 11(6), 849–861. https://doi.org/10.1080/14697680903220356

We began by scraping and compiling monthly macro indicators from FRED, then aligned and merged them with the loan-level dataset by time. We cleaned the resulting panel data by dropping columns with excessive missing values, imputing missing values using the nearest available data, and standardizing each series to zero mean and unit variance to ensure comparability. To determine which indicators exhibit early-warning properties, we designed a set of interpretable metrics based on cross-correlation analysis. Unlike black-box feature attribution, these metrics offer intuitive insights into how each macro variable relates to future changes in loan delinquency.

We focused on the forward-only cross-correlation function (CCF) between each macro index **Xt** and the adjusted delinquency index **Y_t+h**, where **h ∈ [0,36]** months. The rationale is to simulate a realistic forecasting scenario in which we only use past macro information to predict future delinquency risk. From each correlogram, some metrics we extracted are:

- **Correlation at the forecast horizon**: **corr(Xt,Yt+h)**, indicating the signal strength at a policy-relevant horizon.
- **Maximum early correlation**: the strongest absolute correlation between **Xt** and any future **Yt+h.**
- **Lag of first threshold crossing**: the earliest lead time when the correlation exceeds a predefined threshold.
- **Area under the early window**: cumulative predictive strength across all lags.

To complement time-domain metrics, we applied a Fourier transform to each CCF to derive frequency-domain insights:

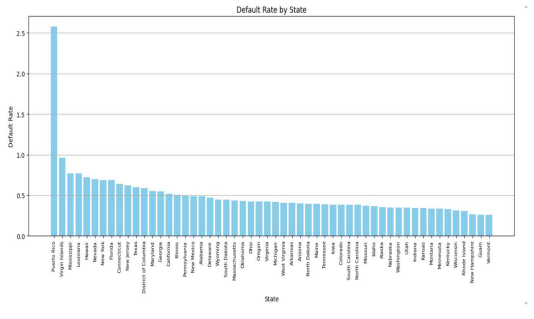
- **Spectral entropy:** a metric that quantifies the disorder in the correlation pattern. Low entropy implies signals.
- **Dominant frequency:** This reveals whether a macro variable's relationship with delinquency follows cycles.

Finally, we computed **Granger causality p-values** for each macro series to capture statistical causality. Although we treat it cautiously due to its assumptions, it serves as a complementary indicator of predictive relevance. To select the best indicators, we normalized each metric and computed a combined score based on the **geometric mean** of all of the above metrics. We ranked all macro series by this score and selected the top 15 as early-warning indicators. This approach balances statistical rigor with interpretability and offers a robust way to detect macroeconomic precursors to mortgage stress.

2.3 Geographic Risk Analysis: Default Rate and Average Loss by State

To uncover regional disparities in mortgage performance, we analyzed Freddie Mac loan-level data from 2014 to 2024 across two key dimensions: default frequency and loss severity. Specifically, we examined state-wise default rates and average realized loan losses, and visualized the results to support geographic segmentation.

In the first part of the analysis, we calculated the default rate for each state by identifying loans with delinquency status codes between 3 and 9 and dividing them by the total number of loans in that state.

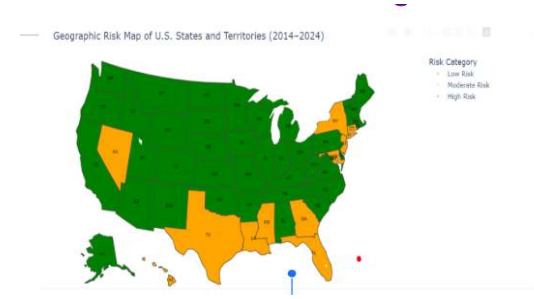


Equation 1. Default Rate per State

$$\text{Default Rate}_{\text{state}} = \frac{\text{Number of Seriously Delinquent Loans}}{\text{Total Number of Loans in the State}}$$

The bar chart in Figure 2.3.1 shows state-wise default rates in descending order, visually reinforcing PR's outlier status.Puerto Rico (PR) emerged as the only high-risk region, with a default rate of 2.57%—far exceeding all other states.

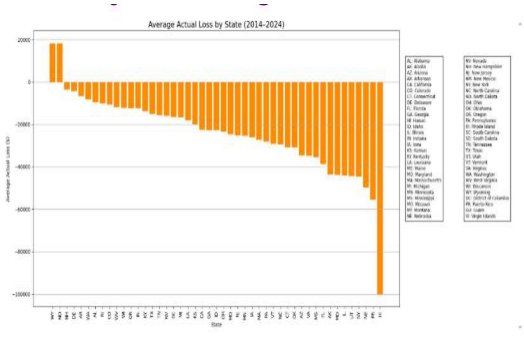
Figure 2.3.1 Default Rate by State



Next, we also applied KMeans clustering (k=3) to group states into Low, Moderate, and High risk categories based on their default rates. As shown in Figure 2.3.2, we assigned labels based on the default rate thresholds. Most mainland U.S. states clustered in the low-risk category, particularly across the Midwest and Mountain regions. The **moderate-risk group**, largely concentrated in the South and parts of the East Coast, reflects regions with elevated but not extreme default activity.

Figure 2.3.2 Geographic Risk Default Thresholds

Low Risk: Default rate < ~0.48 Moderate Risk: 0.48 ≤ Default rate ≤ 0.95 High Risk: Default rate > 0.95



In the second part, we evaluated average actual loss per state, using merged origination and servicing files. We retained loans with valid and non-zero ACTUAL_LOSS values and computed the mean by state. The results, shown in Figure 2.2.3, are displayed in descending order using a horizontal bar chart. Hawaii (HI) and Puerto Rico (PR) showed the highest average realized losses per loan. In contrast, Wyoming (WY) and North Dakota (ND) reported near-zero or slightly positive averages—possibly due to limited data points or unique servicing characteristics.

Figure 2.3.3 Average Actual Loss by State

The findings highlight how both default frequency and financial severity vary widely by geography. These insights emphasize the value of integrating location-based features into credit risk models and using unsupervised learning to support region-specific policy and lending decisions.

Together, these analyses provide a more complete view of geographic mortgage risk—capturing both the likelihood and financial severity of default. The clustering approach supports practical applications such as **risk-based pricing, underwriting adjustments, and targeted policy interventions**. Moreover, these results reinforce the importance of incorporating **location-aware features** into credit risk models and demonstrate how unsupervised learning methods can uncover actionable regional insights from historical loan data.

4. Python Optimization Methodologies

To efficiently process millions of loan-level records and support complex modeling tasks, our pipeline incorporates several advanced Python-based optimization strategies:

4.1 Profiling and Bottleneck Diagnosis

We began by identifying code inefficiencies using profiling tools such as cProfile, line_profiler, and memory_profiler. This allowed us to focus our optimization efforts on the most time- and memory-intensive functions in our pipeline, particularly during feature transformation and model inference.

4.2 Parallel Computing

Given the scale of computation (millions of rows and thousands of model evaluations), we implemented Multi-threading (ThreadPoolExecutor) for I/O-bound operations during scraping the macroeconomic series from FRED API.

4.3 Vectorization and Memory-Efficient Computation

Where possible, we replaced Python loops with vectorized operations using NumPy and Pandas.

5. Conclusion

In summary, this project progresses from a granular, loan-level analysis to broader regional and macroeconomic perspectives, offering a comprehensive view of mortgage credit risk. Beginning with a proof-of-concept ECL model using interpretable features, we demonstrated the feasibility of machine learning in capturing borrower-level risk. We then introduced an adjusted delinquency index to better quantify market stress over time, followed by a novel framework for selecting macroeconomic early-warning indicators through interpretable time- and frequency-domain metrics. Finally, our geographic clustering revealed spatial disparities in default severity. Together, these layered analyses form a unified approach to mortgage risk modeling—enhancing accuracy, interpretability, and early detection across scales.