Credit Risk Modeling on Mortgage Loans Default Prediction

Joy Wu
11/25/2025

Agenda:

1. Motivation
2. Data Overview
3. Model Description
4. Results
5. Conclusion
6. Next Step

Motivation

This project is motivated by my interest in how commercial banks assess credit risk within their retained mortgage portfolios. After modeling prepayment risk in an MBS project this semester, I realized I needed a stronger understanding of the other key component of mortgage risk—default behavior. To build this complementary skill set, I use the Freddie Mac loan-level dataset (2020–2024) to develop a Logistic Regression default model. Its interpretability aligns with real PD modeling practices and allows for practical risk quantification for bank-held loans. This project deepens my mortgage risk expertise across both securitized and balance-sheet portfolios.

Data Overview

1. Data Source

The dataset used in this project is merge_data_2014_2024, a loan-level file originally prepared for an MBS risk-modeling project. It combines each loan's origination attributes with its latest available performance record. From this dataset, I extract loans observed between 2020 and 2024, forming a borrower-level snapshot suitable for default risk modeling.

2. Key Variable

- Target Variable
  i. Current Loan Delinquency Status (converts delinquency status to numeric; flags loans as Default = 1 if delinquency ≥ 3 or status equals "RA", otherwise assigns 0).
- Predictor Variable
  i. Origination Attributes
    - Credit Score
    - Estimated Loan-to-Value (ELTV)
    - Original UPB
    - Current Actual UPB
    - Loan Age
    - Number of Borrowers
    - Current Deferred UPB
    - Current Interest Rate

- Occupancy Status
- Property State

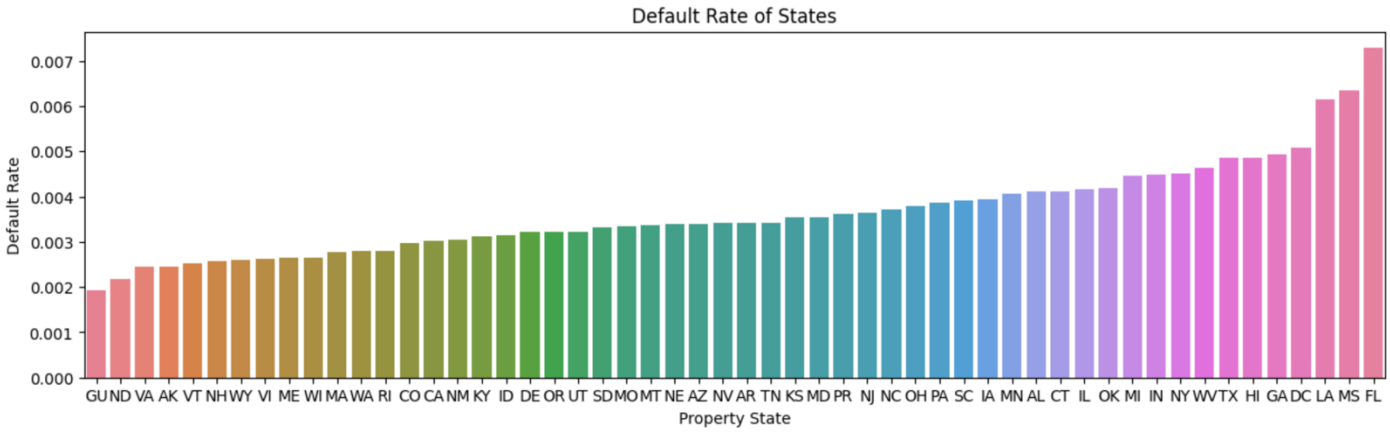
ii. Selection Rationale

Base on economic intuition and VIF

Features	VIF
Credit Score	61.583627
Original Loan-to-Value (LTV)	1177.395275
Original Combined Loan-to-Value (CLTV)	1179.414908
Original Debt-to-Income (DTI) Ratio	8.752397
Original Interest Rate	105312.881974
Original UPB	15.396219
Current Actual UPB	11.648417
Loan Age	44.262240
Remaining Months to Legal Maturity	1590.621502
Estimated Loan-to-Value (ELTV)	3.159415
Number of Borrowers	8.940008
Current Interest Rate	105336.238893
Original Loan Term	1992.128368
Current Deferred UPB	1.031555

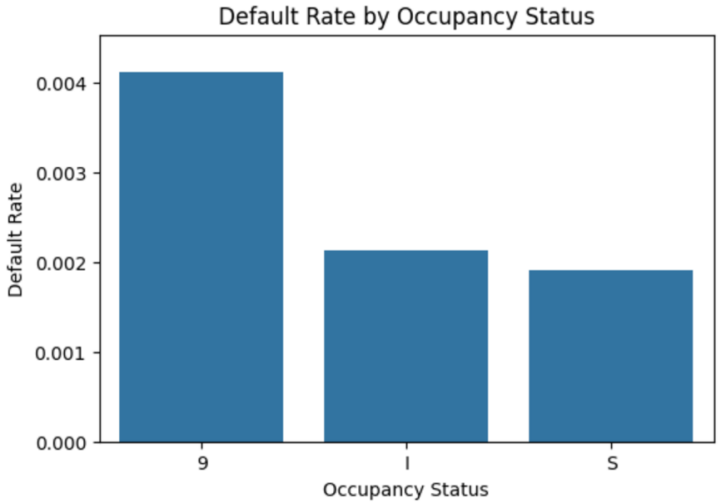
3. Single Variable Analysis (only variables used for model)

- **Property State vs. Default Rate**



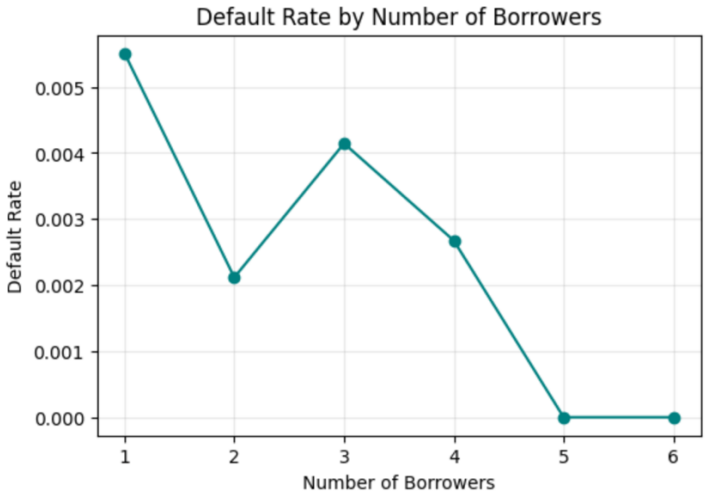
States show very low overall default rates, but variation exists, with FL, MS, and LA having the highest rates

- **Occupancy Status vs. Default Rate**



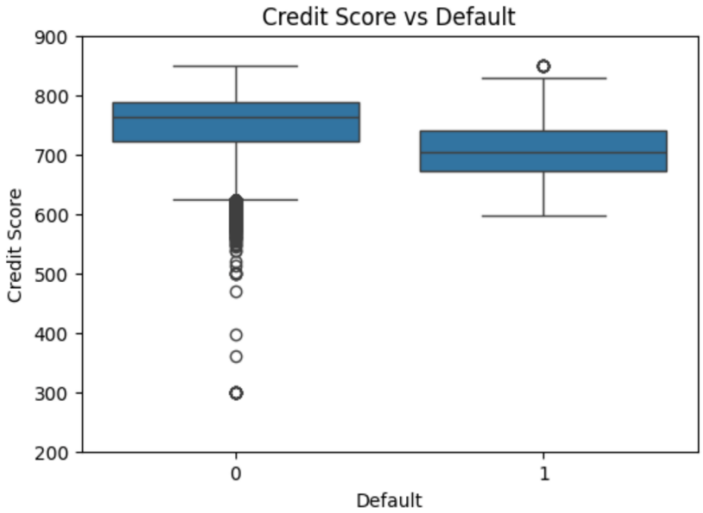
Loans with occupancy code “9”(Unknown) show the highest default rate, while investment (I) and second-home (S) properties have lower and similar default rates.

- **Number of Borrowers vs. Default Rate**



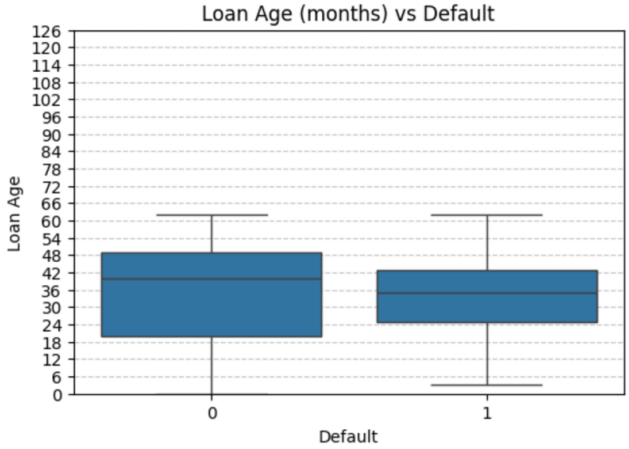
Default risk is highest for single-borrower loans and declines sharply as the number of borrowers increases.

- **Credit Score vs. Default**



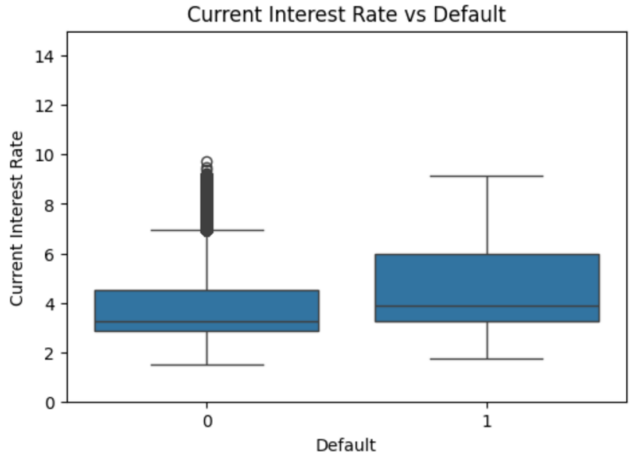
Defaulters generally have lower credit scores with a tighter range.

- **Loan Age vs. Default**



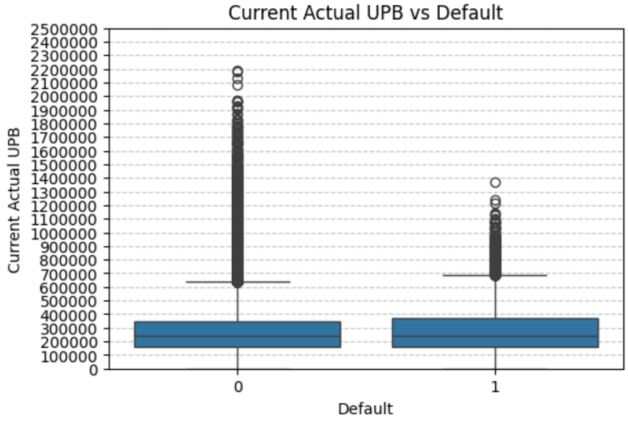
Defaulted loans tend to be slightly younger on average, but the overall loan-age for two groups remain broadly similar.

- **Current Interest Rate vs. Default**



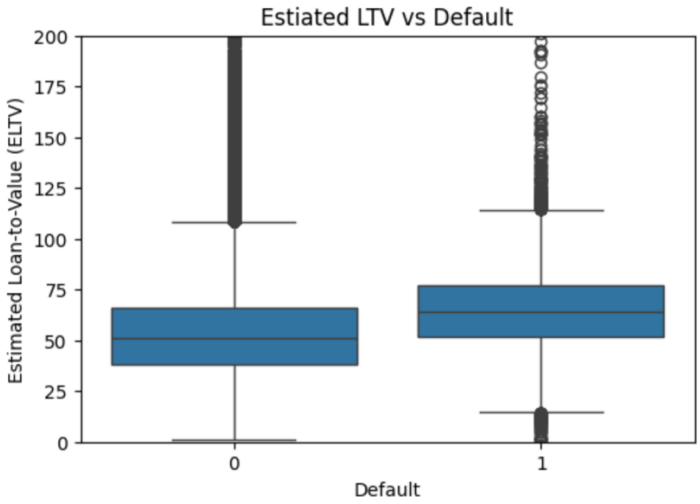
Defaulted loans generally carry higher current interest rates.

- **Current Actual UPB vs. Default**



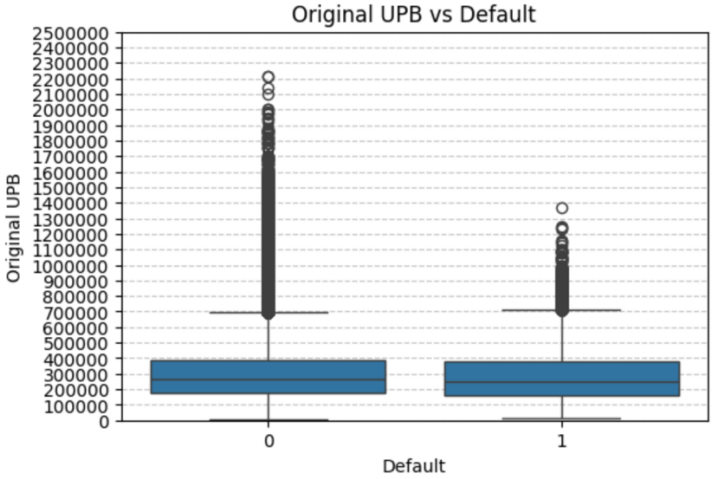
Defaulted loans appear across similar Current Actual UPB ranges as non-default loans.

- **Estimated Loan-to-Value (ELTV) vs. Default**



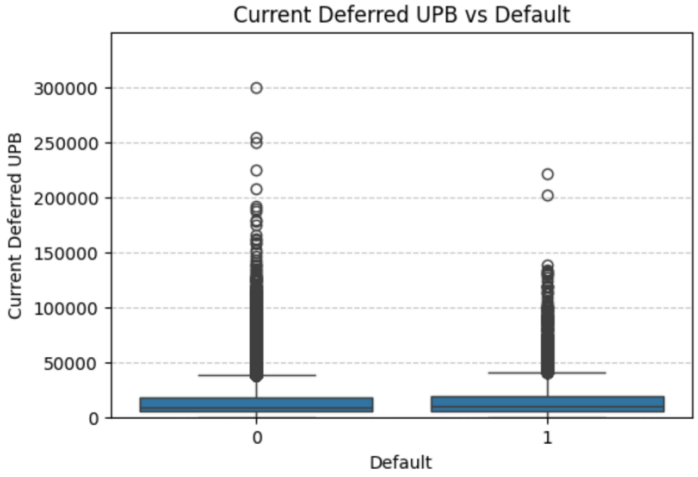
Defaulted loans show slightly higher typical UPB levels.

- **Original UPB vs. Default**



The default group shows a similar Original UPB distribution to non-default loans.

- **Current Deferred UPB vs. Default**



Current Deferred UPB shows very similar distributions for defaulted and non-defaulted loans.

Model Description

1. Objective

- To estimate the probability that a mortgage loan will default.

2. Model Type

- Logistic Regression

3. Model Specification

Target Variable

- Default (1 = 90+ days delinquent and 0 = non-default)

Features

- Consists of all cleaned and preprocessed borrower and loan attributes.

Train/Test Split

- A stratified sample of 100,000 observations is used for training.
- The remaining data is reserved for evaluating out-of-sample performance.

Model

- L2-regularized Logistic Regression
- Class weights are applied to address the imbalance data.

Hyperparameter Tuning

- Use grid search to optimize the strength of regularization.
- The primary tuning metric is **ROC-AUC**.

Threshold Optimization

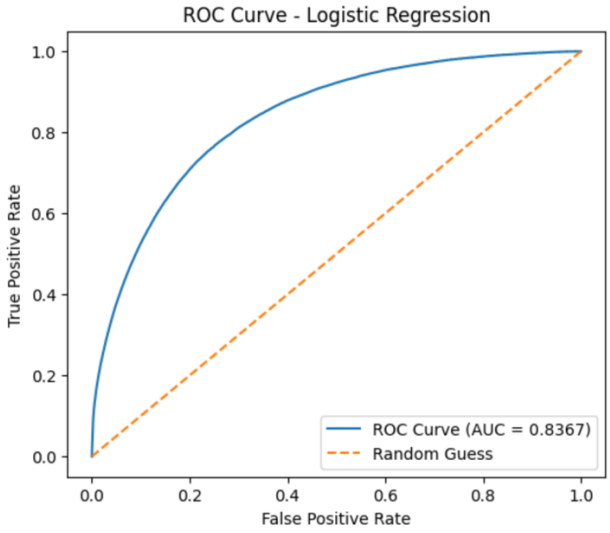
- Use Youden's J statistic to determine the optimal classification threshold.
- Best Threshold: 0.46
- ROC-AUC: 0.8367

Results

1. Performance Summary

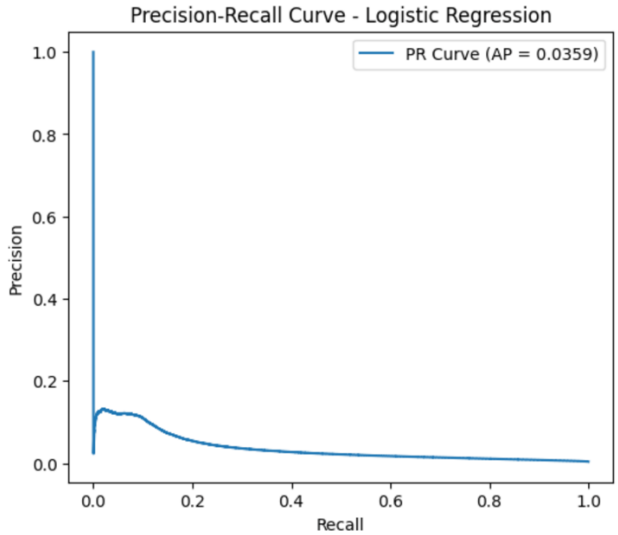
i. AUC Score:

- Cross-Validation: 0.8456
- Test: 0.8367



The ROC curve shows a clear separation between default and non-default classes, with an AUC of 0.8367, reflecting solid ability to rank borrowers by default risk.

ii. PR Curve:

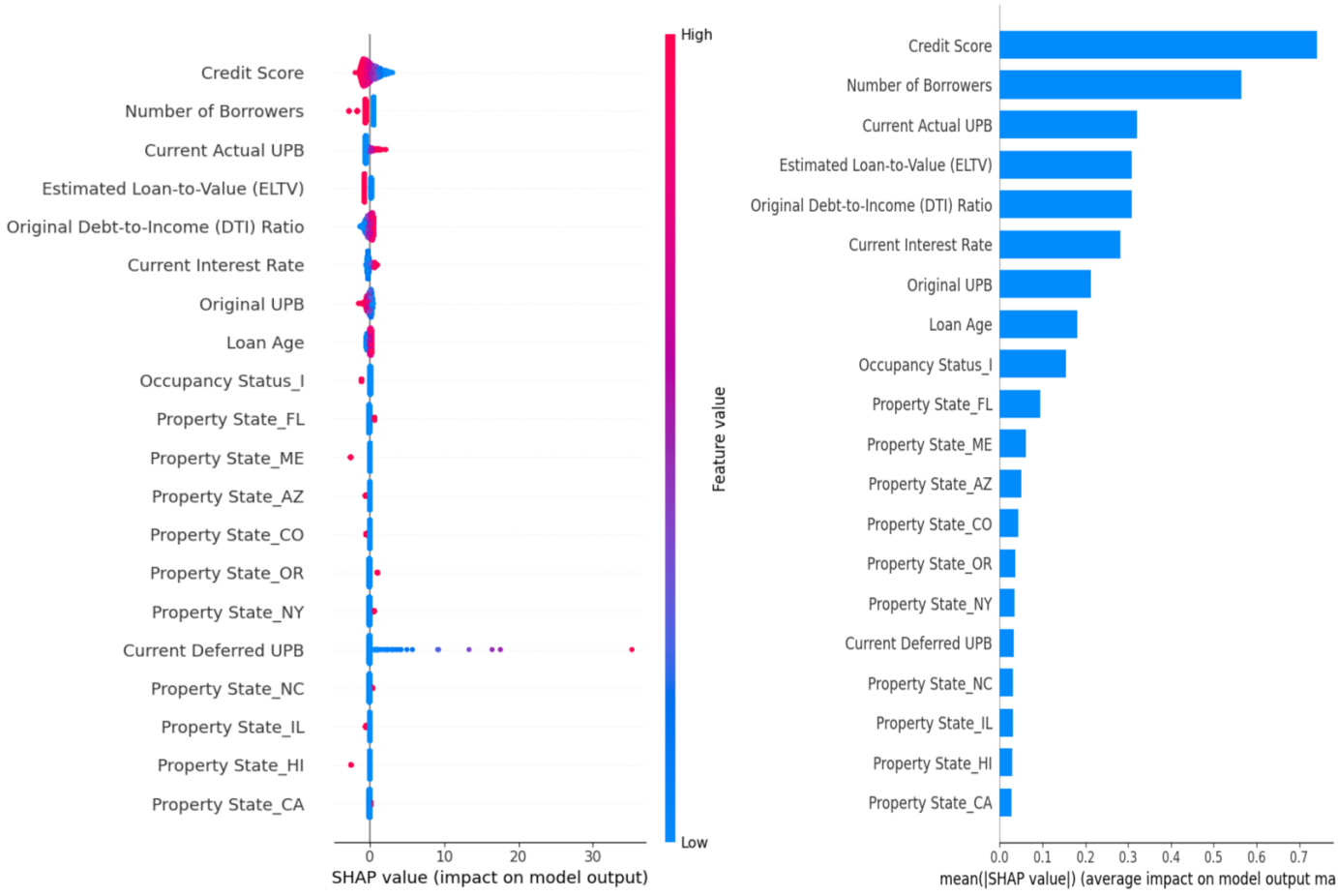


The Precision–Recall curve shows that, due to the highly imbalanced nature of mortgage defaults, the model achieves high recall but low precision overall, indicating that capturing more true defaults inevitably results in many false positives.

2. Feature Importance (SHAP Analysis)

i. Key Drivers of mortgage default risk:

- Credit Score: Most influential, lower credit score strongly increasing default risk.
- Number of Borrowers: Single-borrower loan are associated with higher risk.
- Current Actual UPB: Higher balance related to higher risk.
- Estimated LTV: Higher leverage increases default probability.
- Original DTI: Higher DTI ration leads to higher default probability.
- Current Interest Rate: Borrowers with higher interest rate are more likely to default.
- Original UPB: Larger loan has higher default probability.
- Loan Age: Older loans tend to have higher default rate.
- State-level Indicators: Geographic variation show smaller effects.



Conclusion

This project develops a borrower-level mortgage default model using Logistic Regression applied to Freddie Mac data from 2020–2024. The model demonstrates strong discriminatory power, with a test ROC-AUC of 0.8367 and stable cross-validation performance, indicating reliable out-of-sample generalization. Although precision remains low due to severe class imbalance, the model achieves high recall, making it effective for identifying borrowers at elevated risk of default. SHAP analysis further highlights economically intuitive drivers—such as credit score, leverage, repayment capacity, and loan size—strengthening the interpretability of results.

Next Step

1. Explore Machine Learning Model

Evaluate tree-based methods such as XGBoost or Random Forest to capture non-linear relationships and potentially improve predictive accuracy while comparing results with the interpretable logistic framework.

2. Incorporate Loan Performance History

Extend the model by adding multi-period delinquency, payment behavior, and balance dynamics, allowing the model to capture temporal patterns in borrower performance rather than relying solely on the most recent snapshot.

3. Develop an Integrated Credit Risk Framework

Extend the project by building a full credit risk measurement framework that incorporates Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD).