A purple and white logo

Description automatically generated

COURSE #: FM9528

BANKING ANALYTICS: CONSUMER CREDIT RISK

251467231

251420376

250739681

251135991

Word Count:

1. **Introduction**

Credit risk modeling is a cornerstone of modern banking supervision and strategic risk management, particularly under the regulatory frameworks established by Basel III. Financial institutions are required not only to assess the likelihood of borrower default but also to estimate potential losses in the event of default (Loss Given Default or LGD), the exposure at default (EAD), and to compute adequate provisions to cover expected losses. This report presents a comprehensive framework for consumer credit risk modeling, applied to US residential mortgage data sourced from Freddie Mac's Single-Family Loan Dataset (SFLD). The analysis follows industry-standard procedures and academic best practices for constructing and validating a Basel-III-compliant credit risk pipeline, incorporating data preprocessing, behavioral scorecard development, probability of default estimation, LGD modeling, and final portfolio-level provision analysis.

The primary objective of this project is to develop a behavioral scorecard capable of predicting one-year-ahead probabilities of mortgage default based on observed customer performance characteristics. This model is then used as a base to construct long-term point-in-time and through-the-cycle Probability of Default (PD) models, followed by LGD estimation using advanced machine learning techniques. Finally, we simulate portfolio-level expected losses using realistic assumptions about exposure recovery rates, providing a full-cycle credit risk provisioning framework. All data processing and modeling are performed in Python, ensuring replicability and computational robustness.

The dataset contains two key components: an application file with borrower and loan characteristics at origination, and a performance file documenting quarterly mortgage status over time. A subset of these datasets is used to construct training, test, and out-of-time samples. The out-of-time sample, consisting of loans active as of June 2024, serves as the basis for applying the final models and estimating expected portfolio-level provisions.

This report is structured in four main parts: the methodology details the theoretical and practical approach to model development and validation; the results section presents empirical findings from the models; the conclusion summarizes insights and policy implications.

1. **Methodology**

This section provides a detailed account of the data preprocessing steps, scorecard modeling techniques, segmentation strategies, macroeconomic modeling, LGD estimation, and provisioning simulation, aligned with the Basel III regulatory framework.

**2.1 Data Preprocessing**

The analysis begins by importing two datasets: the mortgage origination data and performance data. The two sources were merged on the ‘*LOAN SEQUENCE NUMBER’* variable to create a unified view of each loan's origination characteristics and monthly status history.

Time fields such as the ‘*MONTHLY REPORTING PERIOD*’ were converted to Python datetime format for temporal indexing. Records with reporting periods after June 2024 were isolated to define the out-of-time validation set.

The target variable for default classification was constructed using the ‘*CURRENT LOAN DELINQUENCY STATUS’*. Records coded with "*RA*" (repayment agreement) were converted to 0, representing non-default status. The remaining delinquency codes were converted to integers, and the default indicator was defined as:

This binary classification follows the Basel definition of default as an account being more than 90 days past due.

Subsequently, a behavioral dataset was constructed. For each loan, only the last available monthly record was retained to represent its most recent status prior to June 2024. This final snapshot included time-varying features such as: ‘*CURRENT ACTUAL UPB’*: outstanding unpaid balance; ‘*CURRENT INTEREST RATE’*: contractual interest rate; ‘*CURRENT LOAN DELINQUENCY STATUS*’: transformed delinquency status; ‘*MONTHLY REPORTING PERIOD*’: observation date. These records were merged with the corresponding origination attributes from the application data, creating a final modeling dataset with both behavioral and static borrower characteristics.

Data cleaning steps included removing duplicates, handling missing values (e.g., dropping rows with null balances or delinquency status), and ensuring the consistency of numerical data types across features. The cleaned dataset was split into training and testing partitions using a random 80-20 split. A fixed random seed was applied to ensure reproducibility.

**2.2 Behavioral Variable Construction**

Behavioral variables were engineered to capture temporal dynamics in borrower and loan performance over a rolling window of the four quarters preceding each prediction point. Features include:

**2.2.1 Delinquency**

A binary **delinquency flag** was first created to track whether a loan was in any state of delinquency:

To construct a robust behavioral feature, the number of months in delinquency over the past 12 months was computed using a rolling sum:

The last month’s delinquency status was also lagged by one period to prevent information leakage:

**2.2.2 Default history**

To monitor the cumulative default history, a cumulative sum of historical default flags was calculated for each loan:

This was used to remove all loan records following a default event, ensuring the predictive model only learns from periods prior to the first default. Specifically, only rows where were retained.

**2.2.3 Feature Engineering**

**Weight of Evidence**

These features were further binned and transformed using Weight of Evidence (WoE) to create monotonic relationships between independent variables and the log odds of default. WoE for a feature category k is defined as:

Information Value (IV) was computed to assess predictive strength:

Only variables with IV > 0.02 were retained for modeling.

**XGBoost**

In the modeling stage targeting XGBoost, a more refined pipeline is applied. First, columns with a high percentage of missing values, such as *'PRODUCT GROUP'*, are removed. Remaining missing values in numerical features are imputed using the median of each column to ensure completeness.

Categorical variables are transformed using one-hot encoding, producing binary indicator columns suitable for gradient boosting algorithms. The dataset is then split into training and testing sets using a group-aware strategy. Specifically, the train\_test\_split\_by\_group function ensures that all records associated with the same loan (*‘LOAN SEQUENCE NUMBER’*) remain within the same partition, preserving the temporal and relational structure of the data.

**2.3 Logistic Regression and Scorecard Scaling**

The behavioral scorecard was estimated using logistic regression:

Where is the probability that loan will default, given the set of features , is the intercept term of the logistic regression model, is the vector of coefficients associated with each feature in the vector

Model estimation was performed via maximum likelihood. The final model was transformed into a points-based scorecard using a linear scaling function:

Where the factor and offset are determined based on two reference points: score at odds = 1:50, and points-to-double-odds (PDO) = 20. The transformation constants are:

This allows the continuous logistic output to be converted into interpretable scores.

Cutoff optimization was based on profit-adjusted cost analysis, incorporating assumptions: Profit margin = 30% of interest rate; while Haircut = 40% of home value.

**2.4 PD Segmentation and Macroeconomic Modeling**

To forecast long-term PDs, we segmented the scorecard into quantiles of AUC (e.g., deciles) and mapped each segment to empirical default rates. We then incorporated macroeconomic variables at the state level, extracted from the FRED database, including:

* State unemployment rate
* House Price Index (HPI)
* GDP per capita

A time series model (e.g., ARIMA or linear regression with lags) was used to forecast these variables, and their impact on segment-level default rates was modeled via:

Long-run values were projected based on macroeconomic trend estimates and literature reviews, and these were used to compute the future PD distribution of the out-of-time sample.

**2.5 LGD Estimation Using XGBoost**

Loss Given Default was modeled using an XGBoost regressor trained only on defaulted loans. The target was:

Only defaulted loans with valid UPB and recovery values were included. Predictor variables included both origination and behavioral features such as interest rates, unpaid balances, credit scores, and debt ratios.

XGBoost minimizes an objective function composed of a loss term and a regularization term. The model’s final prediction is the sum of outputs from NNN regression trees:

where is the output of the tree and is its weight.

Each tree is fitted by computing gradients and Hessians from the loss function:

Optimal leaf weights are calculated as:

and predictions are updated using:

Hyperparameters such as *n\_estimators, max\_depth, learning\_rate*, and *subsample* were tuned via grid search. Model performance was evaluated using RMSE on a 20% test set.

SHAP values were used to interpret feature importance, identifying which variables contributed most to LGD. Finally, all predicted LGD values were floored at 10% to meet Basel regulatory requirements for secured exposures.

**2.6** **Expected Loss Simulation and Provisioning**

To estimate expected credit losses and calculate provisioning requirements, we applied the standard Basel formula:

Where PD is the predicted probability of default for each mortgage, LGD is the predicted loss given default from the XGBoost model, EAD is the current unpaid principal balance (*CURRENT ACTUAL UPB*).

The simulation incorporated uncertainty in the recovery rate. Specifically, we assumed the recovery rate follows a uniform distribution between 40% and 60%, consistent with industry haircuts on collateral values. For each loan, LGD was dynamically re-estimated in each iteration using:

where .

A **Monte Carlo simulation with 10,000 iterations** was conducted. In each iteration, the following steps were applied to the out-of-time sample: Sample a new recovery rate for each loan, recalculate LGD using the sampled recovery, compute expected loss using the PD and EAD from previous models, sum total portfolio-level expected loss across all loans.

The resulting distribution of simulated expected losses was used to evaluate central tendency (mean, median), dispersion (standard deviation), and sensitivity to varying recovery rates. This provided a robust estimate of the provisions required to cover potential credit losses under uncertainty.

1. **Results**

# 3.3 LGD

The XGBoost model for LGD was trained on defaulted mortgage observations using a regression objective. After grid search optimization, the best model was selected with a learning rate of 0.25 and a maximum depth of 2. This configuration achieved strong in-sample and out-of-sample performance, with a Mean Squared Error (MSE) of 0.0149 on the training set and 0.0399 on the test set. The relatively low test MSE indicates that the model generalizes well without significant overfitting.

Feature importance analysis revealed that the most influential predictors for loss severity were factors reflecting final settlement amounts, loan duration, borrower credit quality, accumulated unpaid charges, and original loan size (Figure 1). These drivers align with expectations in credit risk modeling, as older loans and those with higher original amounts tend to display more variability in recovery outcomes. Loans showing higher levels of accrued obligations before closure tended to be associated with more severe losses, while borrowers with stronger credit backgrounds were linked to lower estimated loss rates.

A graph with numbers and lines

Description automatically generated

Figure 1: Feature Importance

SHAP (SHapley Additive exPlanations) analysis further confirmed these results by quantifying each feature’s marginal impact on the LGD prediction. A closer look at the SHAP summary plot (Figure 2) revealed how these features influenced the output across different observations. For example, higher accumulated unpaid charges consistently pushed the predicted loss upward, as reflected by uniformly negative SHAP values for those with elevated balances. These effects were more pronounced when associated with specific loan purposes, suggesting that the context of the default (e.g., purchase vs refinance) may amplify the expected loss severity. Likewise, the magnitude of the final payout on a closed loan was shown to drive predicted losses upward, especially for large exposures. SHAP value dispersion in this feature highlighted a strong separation between loans that resulted in higher recoveries versus those with more substantial shortfalls.

Geographic effects were also evident. For instance, loans from certain regions had both high and low contributions depending on interactions with property type and collateral characteristics. The SHAP dependence plots confirmed that loans located in particular states contributed to higher loss expectations when combined with property valuation indicators or loan purpose flags. This suggests that geographic exposure, combined with underwriting practices and secondary market conditions, played a subtle but measurable role in determining loss outcomes.

A screenshot of a phone

Description automatically generated

Figure 2: SHAP Value Plot

The SHAP value analysis identified five key drivers of loss severity (in Appendix A): the balance removed at loan closure, accrued unpaid interest, borrower debt burden at origination, loan closure outcome category, and original loan amount. Loans with higher cleared balances and unpaid interest consistently showed greater predicted losses, reflecting distressed exits and prolonged delinquency. Borrowers with high initial financial strain were also associated with elevated loss severity, aligning with economic expectations of overextension risk. Loans closed through liquidation channels contributed negatively to recovery, particularly in certain regions, due to higher costs and lower collateral realization. Finally, larger original loans had greater SHAP impact, indicating increased exposure and recovery uncertainty, especially when tied to alternative valuation methods. Together, these features reflect the combined effects of borrower capacity, loan structure, and recovery environment on realized credit losses.

The trained loss severity model was applied to the out-of-time portfolio by predicting the expected loss given default for each active mortgage record. As shown in the results in Appendix B, the model output included a range of estimated loss values. To align with Basel III regulatory guidelines, all predicted values below the regulatory minimum of 10% were adjusted upward to meet the minimum floor. This step ensures that the model adheres to capital adequacy requirements for secured retail exposures, which stipulate a minimum loss estimate regardless of statistical prediction. After adjustment, the final set of loss estimates for the out-of-time sample ranged from 0.1 to approximately 0.75, reflecting a conservative view on recoveries in the event of default. This approach guarantees that all provision calculations are compliant with prudential standards and supports a risk-averse stance in capital allocation.

**3.4 Expected Loss Simulation and Provisioning**

To evaluate the capital required to cover future losses, a simulation-based framework was implemented. The simulation produced a distribution of portfolio-level provisions. As shown in Figure 3, the average provision required to cover expected losses was approximately 5.07, with a median of the same value. The results showed minimal skewness, with the 25th percentile at 5.00 and the 75th percentile at 5.15. The dispersion of the distribution was relatively narrow, as indicated by a standard deviation of only 0.11, suggesting stable estimates under moderate uncertainty.

A blue graph on a white background

Description automatically generated

Figure 3 Distribution of Portfolio Provisions

Further, a sensitivity analysis was conducted to examine how the total required provisions change under fixed assumptions about recovery rates. When a lower recovery assumption was used (equivalent to a 40% haircut), the total expected loss increased to approximately 6.09. In contrast, assuming higher recoveries (60% haircut) reduced the expected losses to 4.06. This difference of roughly 2.03 highlights the significant effect recovery assumptions have on provision calculations. Extending this analysis across a broader range of recovery rates in Figure 4, provisions ranged from just above 3 to slightly over 7, with a relative change of more than 130%. This finding underscores the importance of robust recovery modeling in the context of capital adequacy planning.

A graph with a line going up

Description automatically generated

Figure 4: Sensitivity Analysis

The simulation framework provided not only a central estimate of provisions but also captured the full distribution of potential losses. This enables more risk-sensitive planning and allows financial institutions to anticipate the capital required under both expected and adverse scenarios, supporting prudent and forward-looking risk management practices.

Appendix A Top Five Key Features in SHAP Plot

A screen shot of a graph

Description automatically generated

A screen shot of a white background with red dots

Description automatically generated

A white rectangular object with red dots

Description automatically generated

A graph with red and blue dots

Description automatically generated’ A screen shot of a graph

Description automatically generated

Appendix B: Estimated Loss Values

A number of numbers on a white background

Description automatically generated