Consumer Credit Risk

FM 9528 Banking Analytics

Miles Xi

Executive summary

In this project, we analyzed the Single-Family Loan-Level Dataset from the U.S. Freddie Mac. This extensive dataset consists of two data files, a mortgage application file (61,404 rows, 32 columns) and a repayment performance file (1,759,940 rows, 32 columns). Our first objective is to build a behavioral scorecard predicting one-year-ahead default probability for each observation.

We first gained a robust understanding of the data by reading the General User Guide [1]. The response variable is a binary indicator created based on the delinquency status in the next 12 months. If the observation window is shorter than 12 months, due to, for example, early repayment or reasons arising from default, the target value can still be determined. If there was neither enough observation nor relevant information, the target could not be determined and later the observation was removed. Useful features such as the rate of change in unpaid balance were created. Then two datasets were merged and the joined data set was split into a training set, a test set, and an OOT sample. Finally, missing values in each feature were handled through imputation; if there were a substantial portion of missing values due to unknown reasons, that feature was deleted. Outliers were handled by removing the case or were left as is. The data cleaning procedures were done separately and consistently on the training and test sets.

We binned the remaining features using Weight of Evidence binning, creating a more meaningful interpretation for each feature and introducing non-linearity into the linear model. We used logistic regression with the ElasticNet regularization and trained the model with the 3-fold cross-validation to tune the hyperparameters. The model achieved an ROCAUC score of 0.940 – 0.944 at a 95% confidence level. Given the assumption that each loan could yield a profit margin of 30% of the interest or suffer 40% of the house price, with the predicted probability, we computed the total expected profits of the loans to determine the best cutoff for the loan approval decision-making.

The second task is to develop a time series model of the PD for ratings of loans that have distinct risk profiles from low to high. We obtained a series of breakpoints (which are the False Positive Rate) by approximating the ROC curve of the logistic regression model using piecewise linear functions and mapped these breakpoints to the probability cutoffs, resulting in eight ratings. With these ratings, we put observations in the test set into their corresponding bins and time slots, getting eight time series of the default probability for each state, which was modeled using the SARMIAX model where exogenous variables are macroeconomic factors including the monthly unemployment rate and the House Price Index (HPI) obtained from St Louis Fed. Finally, we obtained the long-run unemployment rate and estimated the HPI to forecast the loan's performance in the OOT sample using the fitted time series model.

Under the Advanced IRB approach to credit risk, all three parameters are estimated internally. Thus, for defaulted loans, we derived their LGD using the features of loss calculation and unpaid balance in the dataset. We performed similar data cleaning procedure to the training, test, and OOT samples and then trained and fine-tuned an XGBoost model to predict LGDs. SHAP were employed to explain the feature importance. Finally, we used the model to predict the LGD on the whole OOT set and trimmed the predicted values so that they satisfied the Basel III floor of 5%.

We simulated recovery rates using a uniform distribution to derive EAD, and with PD and LGD obtained above, we calculated the provision for a portfolio of loans in the OOT sample. We found that the provision accounts for 0.1878% of the portfolio value and is not sensitive to the recovery rates. This is likely due to low long-run PDs and small and stable predicted LGDs in the sample.

Table of Contents

1. Data cleaning

The two datasets are a subset of the standard Single-Family Loan-Level Dataset provided by the U.S. Freddie Mac. The first is mortgage application data that has information about the property and the loan. It contains 61,404 rows and 32 variables, including *creditScore*, *firstPaymentDate*, and *firstTimeHomebuyerFlag*, etc. The second is a performance file, where borrowers' repayment behavior is recorded in a panel data format between 2021-01 and 2024-06. It includes 1,759,940 observations and 32 features, such as *monthlyReportingPeriod*, *currentActualUpb*, and *currentLoanDelinquencyStatus,* etc. The two datasets both have a *loanSequenceNumber* that identifies each loan. For the definition of these variables, please refer to the General User Guide [1].

Creating the target variable and behavioral variables

We created a target variable *target* based on the loan payment behavior *currentLoanDelinquencyStatus* and the condition that it was ultimately paid off. For example, suppose for a given loan there are 15 months of observations. It is easy to determine the *target* values for the first 3 months as there are a full 12 months of data to observe the behavior. In the fourth month, however, we only have 11 months to observe. The *target* was then determined by using the *zeroBalanceCode*. If it is 01 (meaning prepaid), the fourth month's *target* should be non-default. If it is anything other than 01, such as 02 (third party sale), the fourth month should be a default. If it is NA, then the *target* would not be populated. This leads to 511,540 (29%) NAs, 1,192,895 (67%) non-default cases, and 55,505 (3.1%) default cases.

We created a set of meaningful behavioral variables as follows:

- *upbPctChange* is the rate of change between the current month's UPB and the last month's. This is used to measure the most recent behavior and eliminate the effect of the scale of the loan amount.
- *nonPmts_3m* counts the number of times there has been no decrease in unpaid balance in the recent three months (including the current month). This measures a borrower's payment habit.
- *delinquencyDueToDisaster_hist* is 'Y' (otherwise 'N') if a person defaulted due to a disaster.
- *interestBearingUpb_ratio* is defined as the ratio of *interestBearingUpb* to *currentActualUpb.*

Train, test, and OOT samples

We then joined two datasets using *loanSequenceNumber*, removed cases where *the target* was null, and created an out-of-time (OOT) sample which includes all mortgages with an active record (still under repayment) in the last observed month, 2024-06, resulting in 257 observations/loans. Then we split the dataset into a train (70%, 873,700 rows) and a test set (30%, 374,443). The temporal order was not preserved. The target class distribution in both sets is maintained.

Data cleaning

We explored the training set and cleaned the variables originally in the application dataset.

- *creditScore* and *originalDebtToIncomeRatio* had missing values encoded as 9999 or 999. We replaced them with the median.
- *areaCode* and *postalCode* had a high standardized mutual information of 0.81. Therefore we dropped the *areaCode*.
- *originalCombinedLoanToValue* and *originalLoanToValue*. Both are a ratio of the loan amount to the property's appraised value. The former was dropped due to a high correlation (0.98) with the latter.
- *sellerName* and *servicerName* were considered useless, so both were dropped.
- *superConformingFlag*. The missing values meant "not super conforming", so they were filled with 'N' (another category is Y).
- *preReliefRefinanceLoanSeqNumber* and *reliefRefinanceIndicator*. The first variable was a sequence number with no predictive power, so we dropped it. NAs in the second variable meant loans were not part of Freddie Mac's Relief Refinance Program, so we filled them with 'N'.

We cleaned the variables originally in the performance dataset.

- *zeroBalanceCode* and *zeroBalanceEffectiveDate* indicated the way and the month a loan's balance was reduced to zero. They only take on values at the end of the performance period and have been used to create the target variable. At the prediction time they will not be available, so they were dropped to avoid leaking future data.
- *modificationFlag*, *stepModificationFlag*, *paymentDeferral*, and *borrowerAssistanceStatusCode* had a substantial portion of nulls. However they are a category by themselves, thus they were replaced with values.
- *delinquencyDueToDisaster*. Same as above
- *defectSettlementDate*, *miRecoveries*, *netSaleProceeds*, *nonMiRecoveries*, *zeroBalanceRemovalUpb*, *delinquentAccruedInterest*, and *actualLossCalculation*. The second to last features were only populated on the *defectSettlementDate*, which was a date when a service or underwriting defect was settled. At the prediction time they will not be available, so they were dropped.
- *totalExpenses*, *legalCosts*, *maintenanceAndPreservationCosts*, *taxesAndInsurance*, and *miscellaneousExpenses*, too, depended on *defectSettlementDate*. Thus they were dropped.
- *cumulativeModificationCost* and *currentMonthModificationCost*. For *cumulativeModificationCost*, only the last month's observation was populated for modified loans, so it was dropped. The NAs in *currentMonthModificationCost* were replaced with 0, meaning the cost was not applicable and thus none.
- *dueDateOfLastPaidInstallment* was only populated for the last observation. The variable's information is captured by *maturityDate*, so it was dropped.
- *estimatedLoanToValue*. Unknown cases (5.83%) were represented as 999. We imputed them using the median.

We finally cleaned the variables just created.

- *upbPctChange* was missing for every loan's first observation, as expected. We filled them with 0, assuming no change in unpaid balance.

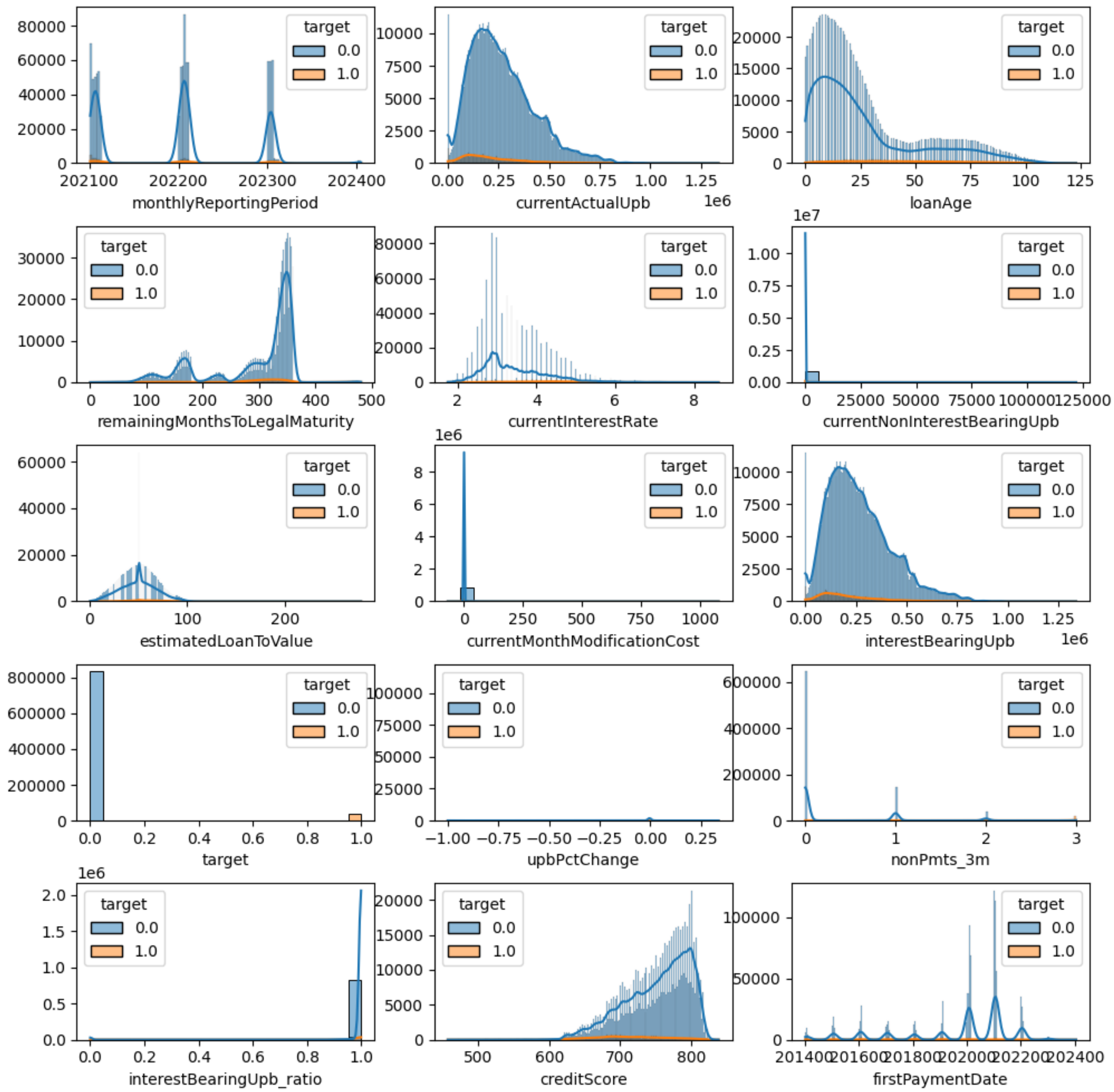

Fig 1. The Distribution of Numeric Features

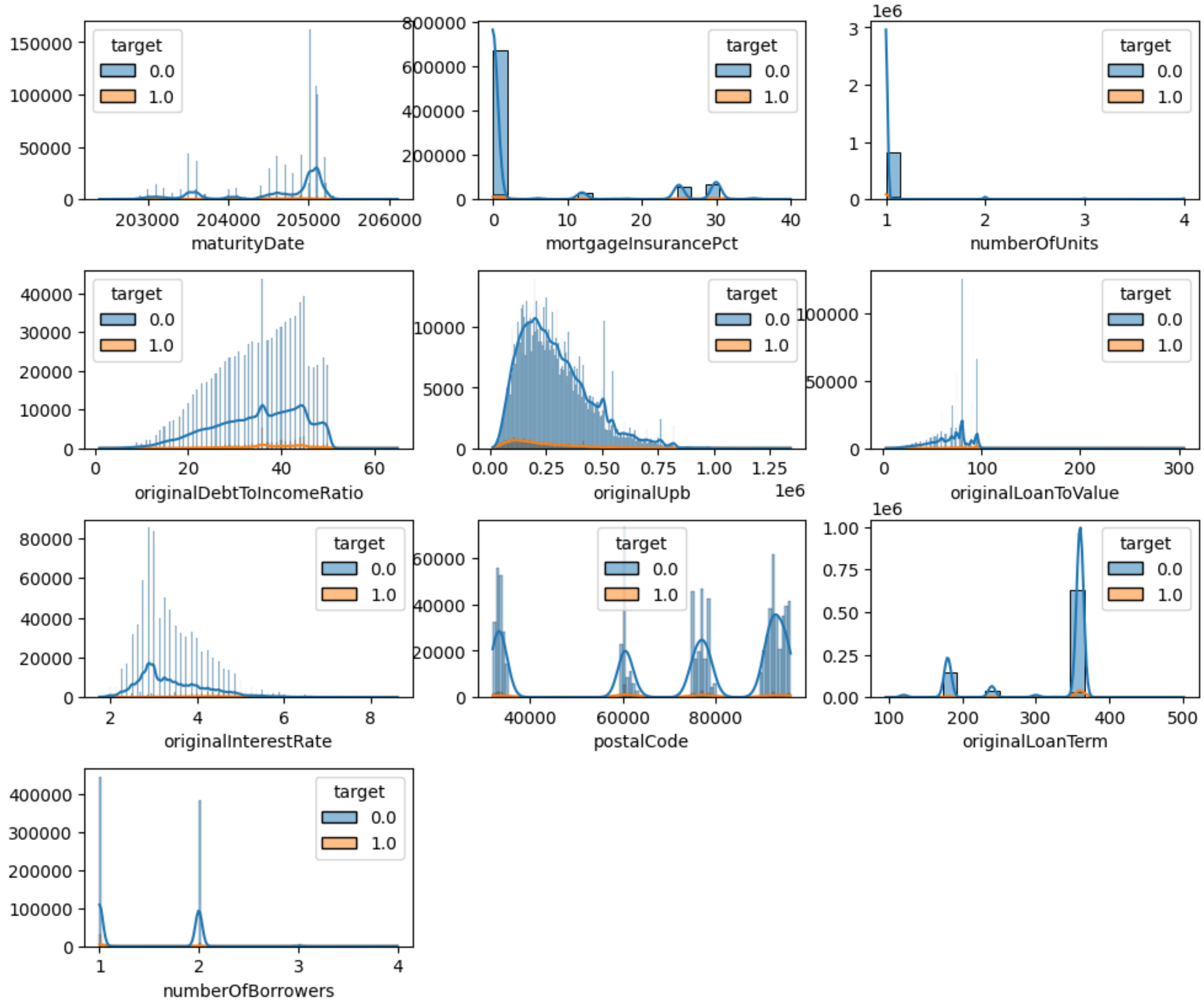

Fig 1. The Distribution of Numeric Features (continued)

After the steps above we had 47 variables, none of which had missing values.

We then handled outliers. From Fig 1, we can see the following:

- *interestBearingUpb_ratio.* A few outliers were below 0.4 and we only kept those above 0.4.
- *currentActualUpb*, *interestBearingUpb*, and *originalUpb* looked identical in distribution and thus all had the same large outliers. Several outliers were greater than 1,250,000 and we only kept the records below this threshold.
- *currentNonInterestBearingUpb* had large valid outliers. We kept it as is.
- *currentMonthModificationCost*. Outliers were those with a high modification cost. They became outliers because in the previous step missing values had been replaced by 0. We decided to keep as is these valid outliers.
- *creditScore*. We did nothing to them as low values might have been indicative of default. Removing them would have changed this signal.

The cleaned training dataset had 861,593 rows and 47 columns. We applied the same cleaning process to the test set.

2. A behavioral Scorecard

WoE coding

We first used Weight of Evidence (WoE) to bin and encode the features, managing non-linearity in a linear logistic regression that would be used. We removed features that had an information value (IV) less than 0.02. They were *superConformingFlag*, *programIndicator*, *channel*, *currentMonthModificationCost*, *interestBearingUpb_ratio*, *reliefRefinanceIndicator*, *paymentDeferral*, *stepModificationFlag*, *modificationFlag*, *numberOfUnits*, *borrowerAssistanceStatusCode, currentNonInterestBearingUpb, currentLoanDelinquencyStatus*, and *delinquencyDueToDisaster*.

Three features (*prepaymentPenaltyMortgageFlag*, *amortizationType*, and *interestOnlyIndicator*) only had one unique value, so they were dropped as well.

We then applied the WoE coding, which uses a one-level decision tree to create bins. Then we manually adjusted the bins for explainability; *OriginalDebtToIncomeRatio* was adjusted to have fewer bins and a more linear trend. The same bins were applied to the test set. The variables constructed and their bins are shown in Fig 2.

As a final correlation analysis, we found that the binned *interestBearingUpb*, *postalCode*, *and originalInterestRate* were highly correlated (> 0.8) with *currentActualUpb*, *propertyState*, and *currentInterestRate*, respectively. The former set of variables was dropped as a result.

The 23 (excluding the response) features would be included in the logistic regression model for (a) they met a minimum threshold (0.02) of IV, (b) they were not highly correlated with other predictors, and (c) the reasons explained in data cleaning.

The operational limits of the model included:

- using median for a feature's unknown value, for example, *creditScore*
- *currentActualUpb* and *originalUpb* needed to be less than 1,250,000 as observations were removed before WoE coding transformation.

Model fitting

Using sci-kit learn in Python, the final logistic regression model with an ElasticNet penalty was estimated and tuned using 3-fold cross-validation as follows:

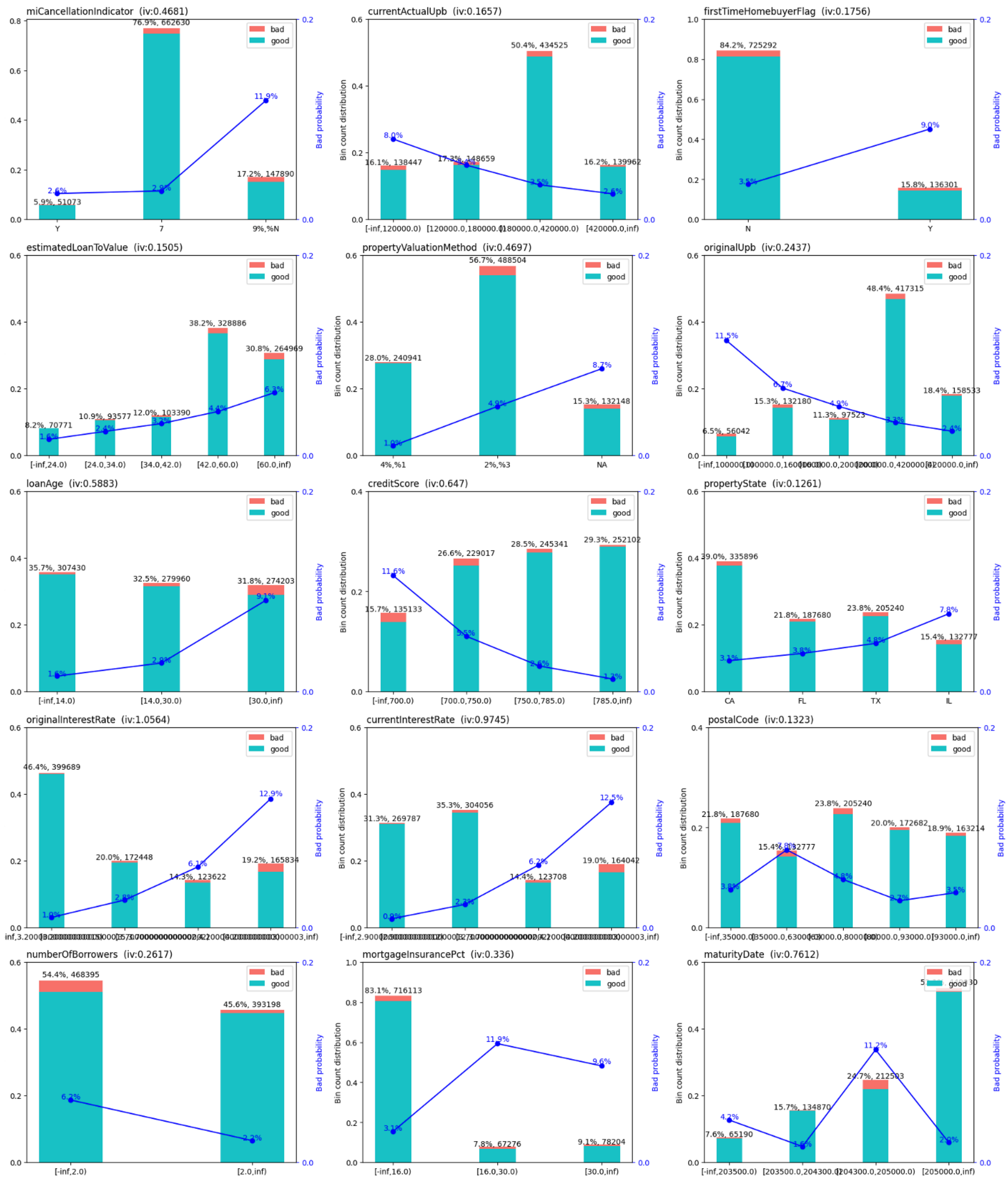

Fig 2. Weight of Evidence Coding

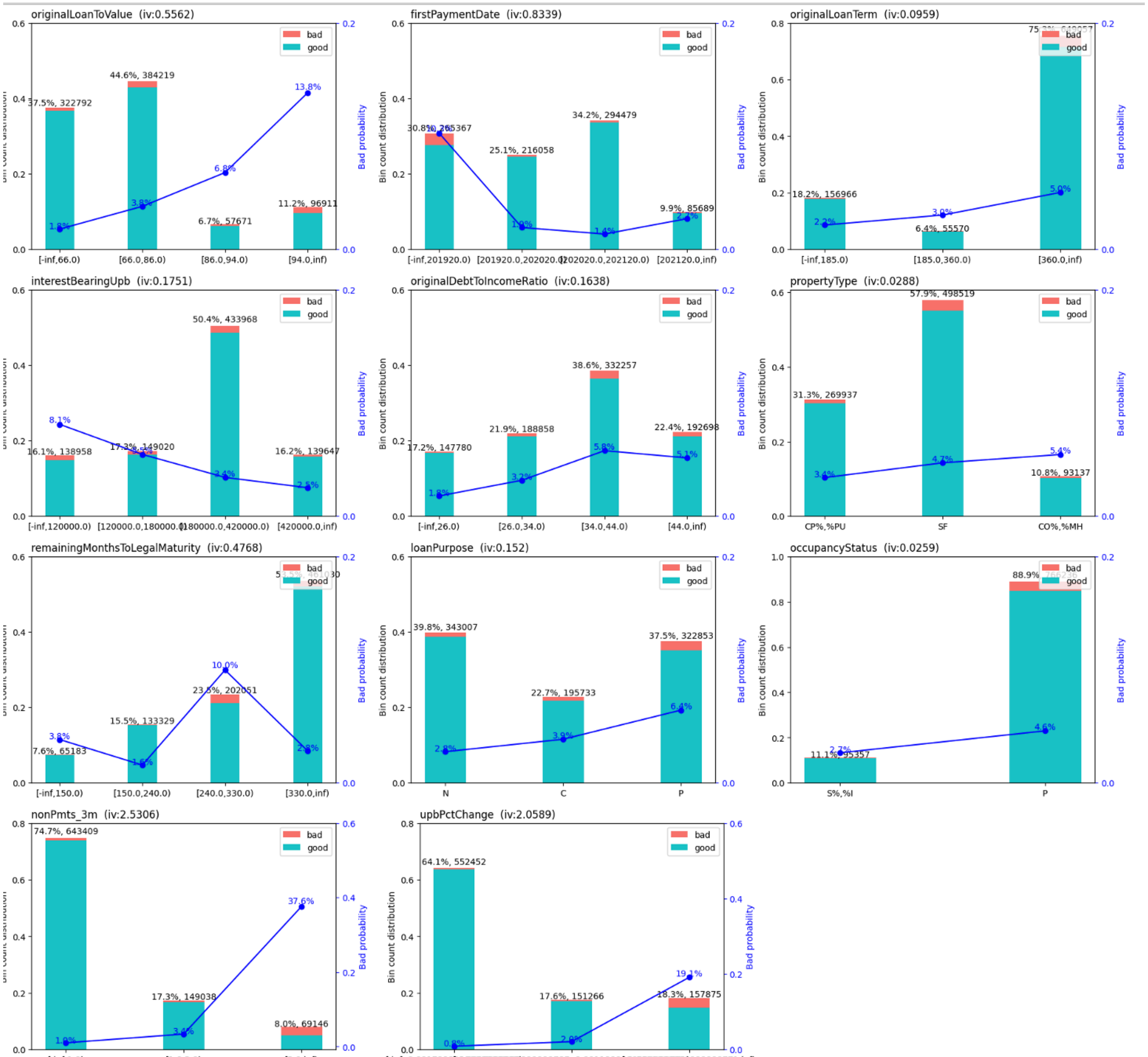

Fig 2. Weight of Evidence Coding (continued)

$$\begin{aligned} ln\left(\frac{p}{1-p}\right) = 0.0849 - 0.1943 \times mortgageInsurancePct \\ - 0.7825 \times remainingMonthsToLegalMaturity \\ - 0.1160 \times firstPaymentDate + 0.6686 \times nonPmts\_3m \\ + 1.1313 \times loanAge + 0.4530 \times upbPctChange \\ - 1.0531 \times currentActualUpb + 0.4864 \times maturityDate \\ + 0.7661 \times creditScore + 0.9502 \times originalUpb \\ + 0.5910 \times originalDebtToIncomeRatio + 0.8968 \times occupancyStatus \\ + 0.4837 \times miCancellationIndicator + 0.4401 \times propertyType \\ + 0.1490 \times currentInterestRate + 0.1848 \times propertyValuationMethod \\ - 0.0775 \times loanPurpose + 0.1604 \times originalLoanTerm \\ + 0.7784 \times numberOfBorrowers - 0.3234 \times propertyState \\ + 0.8163 \times estimatedLoanToValue - 0.1037 \times firstTimeHomebuyerFlag \\ + 0.2052 \times originalLoanToValue \end{aligned}$$

where p is the default probability. Elastic-Net mixing parameter *l1_ratio_* was 0.2 and regularization strength *C_* 21.5443. Note this method ignores the correlation within each subject (i.e., client) as is common with longitudinal datasets, but Biron and Bravo [2] showed that violating the independence assumption in logistic regression in this setting does not impact the model's discriminative power strongly.

Model evaluation

We applied the same data preprocessing procedure, i.e., handling missing values and outliers, to the test set, and then evaluated the model on it. Fig 3 displays the confusion matrix; the model achieved a ROCAUC score of 0.940 – 0.944 at a 95% confidence level. Note that the interpretation of the sign of the estimated coefficient depends on the relationship between the original variable and the WOE-transformed variable. For example, because the WOE-transformed *creditScore* decreases as the original *creditScore* increases, the coefficient estimate is positive.

We then estimated a cutoff point using an OOT sample. When processing the raw OOT data, we imputed the missing values the same way but did not do anything to the outliers, because (a) there is no observation with *currentActualUpb* and *originalUpb* greater than 1,250,000, and (b) while all observations had *interestBearingUpb_ratio* less than 0.4, this feature was not included in the final model. We then applied the model to the binned OOT data; it achieved a ROCAUC of 0.992.

Determining the best cutoff

Table 1. Cutoff and Total Expected Profit

Cutoff	Total Profit
0.1000	49,991,554.7
0.1474	55,381,122.9
0.1947	57,252,875.8

0.2421	58,719,025.8
0.2895	59,257,896.5
0.3368	60,320,875.6
0.3842	60,517,183.3
0.4316	60,517,183.3
0.4789	60,785,965.1
0.5263	**60,870,216.9**
0.5737	60,684,131.6
0.6211	60,676,274.6
0.6684	60,537,242.2
0.7158	60,426,245.3
0.7632	60,380,785.8
0.8105	60,041,635.1
0.8579	59,279,723.2
0.9053	58,875,637.6
0.9526	57,860,401.0
1.0000	57,385,489.3

To determine the best cutoff for classification and loan approval decision-making. We estimated the default probability of loans in the OOT sample and computed each loan's expected profit using the loan's amount (*originalUpb)*, interest rate (*currentInterestRate*), and the property's value (*estimatedLoanToValue*), assuming a profit margin of 30% of the interest rate and a haircut of 40% of the house price. Specifically the profit margin, the loss, and the expected profit for each loan were calculated as follows:

$$profit\ margin\ =\ 0.30 \times current\ interest\ rate \times original\ UPB$$

$$loss = 0.40 \times \frac{original\ UPB}{estimated\ loan\ to\ value}$$

$$expected\ profit = (1 - \hat{p}) \times profit\ margin + \hat{p} \times (-loss)$$

We selected the cutoff by maximizing the total expected profit of the portfolio. The best cutoff was 0.5263, as shown in Table 1. The distribution of the predicted default probabilities is shown in Fig 4.

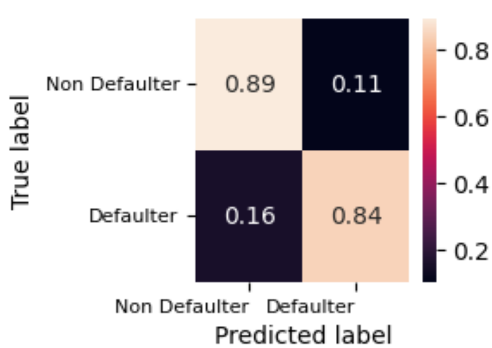

Fig 3. Confusion Matrix

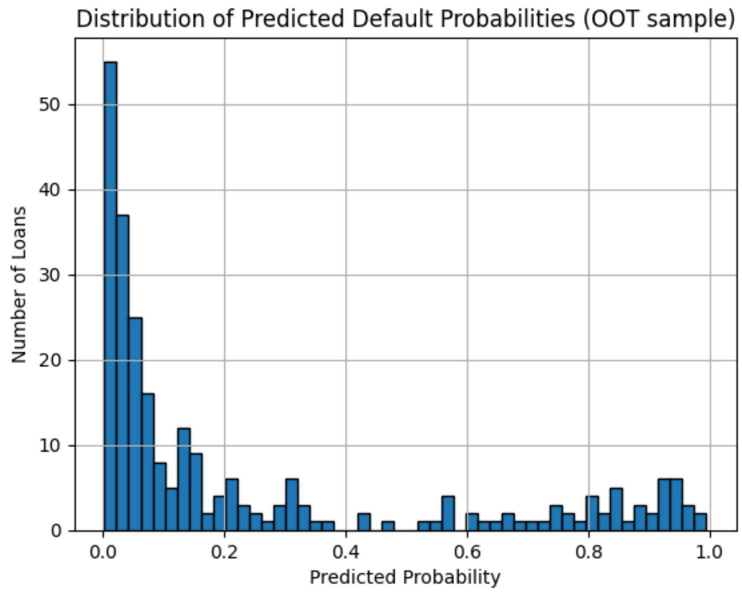

Fig 4. The Distribution of Predicted Default Probabilities

3. PD Ratings and Calibration

In this section, we performed PD calibration which is a process where the predicted default probabilities are aligned with the observed proportion of defaults in the data. We then modeled these default rates using ARIMAX with macroeconomic regressors and got long-term default rates for provision calculation.

To do so, we grouped loans with similar predicted default probabilities into ratings, creating a more stable risk measurement, by approximating the ROC curve piecewise linearly using seven line segments.

As shown in Fig 5, the approximated ROC followed the original ROC curve closely. The locations of the breakpoints are specific FPR values that correspond to probability cutoffs. Thus,

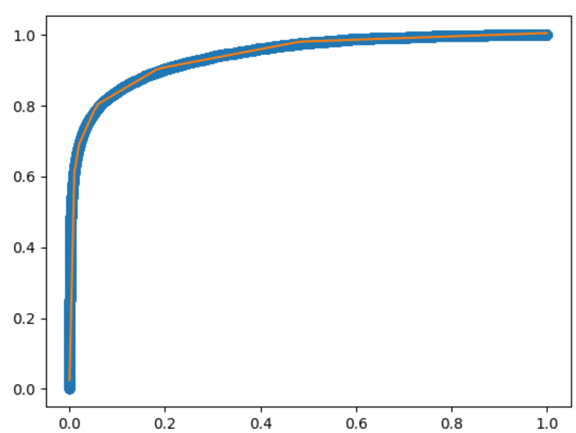

Fig 5. Piecewise Linear Approximation to the ROC Curve

we obtained 7 cutoffs or 8 bins into which we put a predicted probability of default, as shown in Table 2. The monotonicity of the default rate per rating is well preserved in that the default rate increases as the bin becomes worse. We consider these bins reasonable as they met the Basel requirement that there should be 7 to 15 ratings and each bin reflects a unique risk profile where risk increased with worse bins, although the predicted probabilities of default failed to reflect the actual default rates in the bins.

Table 2. Putting predicted PDs to ratings

Bin	# non-defaults	# defaults	Default rate
(0.0, 0.000616]	4	0	0.000000
(0.000616, 0.102]	182848	427	0.002330
(0.102, 0.335]	104583	1271	0.012007
(0.335, 0.666]	45636	1729	0.036504
(0.666, 0.885]	14932	2146	0.125659
(0.885, 0.96]	4230	2868	0.404057
(0.96, 0.985]	737	3721	0.834679
(0.985, 1.0]	177	4030	0.957927

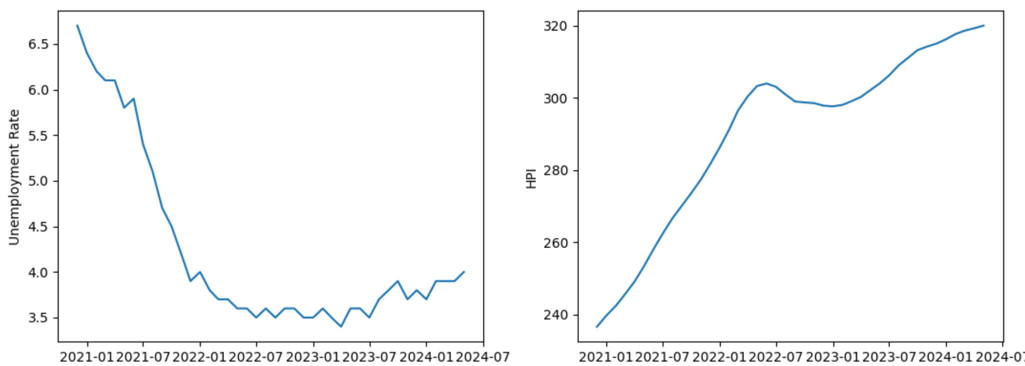

Fig 6. Macroeconomic Variables

Macroeconomic data

We collected macroeconomic factors for each state using fredapi to model PD. We chose two economic factors: the monthly seasonally adjusted unemployment rate and the monthly seasonally adjusted Case-Shiller U.S. National Home Price Index (HPI). The unemployment rate, associated with GDP, measures the economic activity directly without the need for price adjustment. Recession and depression would increase the risk of mortgage defaults. HPI reflects the property value. If it falls, borrowers are likely to walk away and default. The two variables are measured in the national level. Time series plots are shown in Fig 6.

Time series models for PD

Now the predicted PD of each observation in the test set was put into one of the eight PD bins and is associated with a time index (*monthlyReportingPeriod*) and a default-or-not indicator (*target*). For each rating, at a given month, we computed the proportion of defaults as the observed default rate, as shown in Fig 7. They were then modeled using the ARIMAX model, which included two exogenous variables mentioned above and their lags.

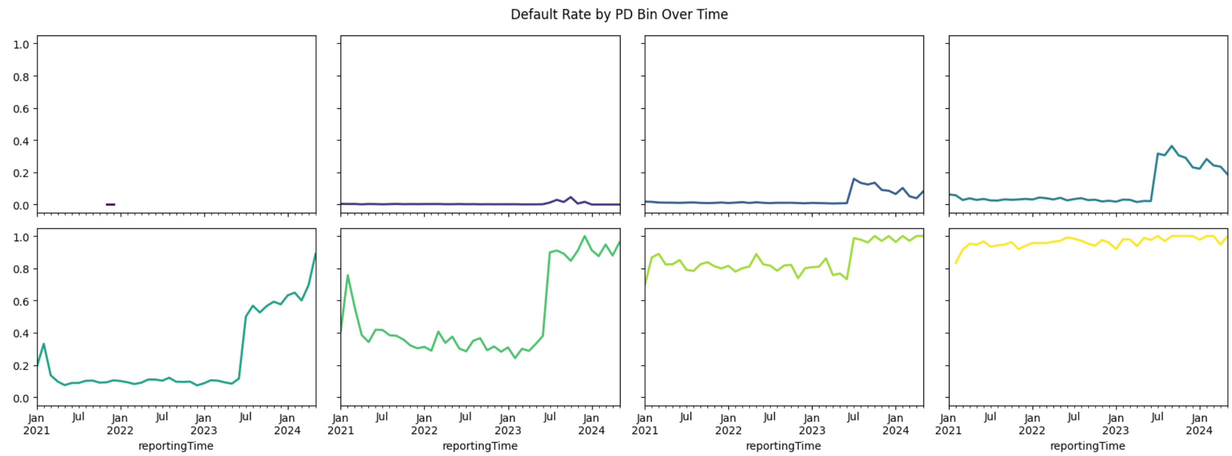

Fig 7. Default Rate by PD Bin over Time

There was only seven PD models due to a lack of sufficient observations in the first bin. Having set up a parameter grid that contains different orders of AR, MA, Integration, and the Seasonal components, we selected the best models based on AIC.

Long-term estimates of the macroeconomic variables

The long-run estimate of the unemployment rate should be some unemployment benchmarks or natural rates of employment that eliminate all the shocks that cause a current business cycle and that are determined by labor market dynamics that slowly change over time 32]. There are three approaches to estimating this value. The most commonly used estimate is the Congressional Budget Office's (CBO) noncyclical rate of unemployment which can be retrieved from St. Louis Fed [4].

According to the website, nationwide, in Q1 2025 the rate is 4.32% and is projected to decline to 4.11% in Q4 2035. Thus we consider an approximate value to be 4.20%.

For HPI, there is no official long-term forecast. Therefore, we fitted a simple SARIMA model, using the parameters that minimize AIC. We then generated forecasts for the next 12 months and take the average. The result was 286.55.

Forecasting the long-term PD for the OOT sample

We plugged in these long-term exogenous variables to the fitted model, which depended on the bin, to get long-term PDs for the observations in the OOT sample. The distribution is shown in Fig 8.

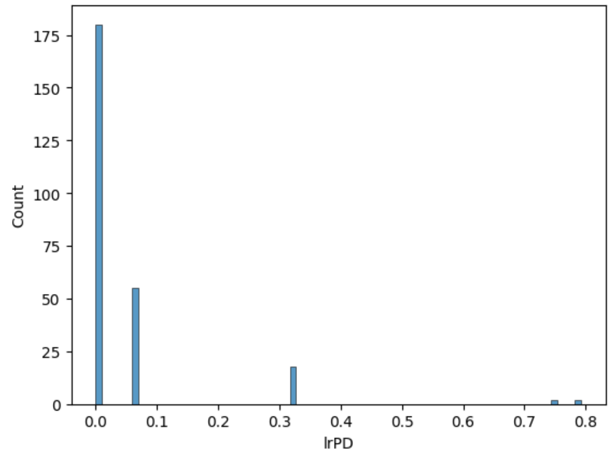

Fig 8. The Distribution of the Long-Run PD for the OOT Sample

4. LGD Modeling

Modeling LGD using XGBoost

The response variable, *LGD* (Loss Given Default), was constructed as follows

$$\mathrm{LGD} = \frac{\text{actual Loss Calculation}}{\text{zero Balance RemovalUpb}}$$

where *zeroBalanceRemovalUpb* is the amount of unpaid balance prior to the application of the zero balance code, and *actualLossCalculation* is this amount plus sale proceeds and minus delinquent accrued interests and other adjustments. We trimmed *LGD* less than 0 to 0 to indicate no loss from the default, and selected observations with non-NA *LGD* values in the training set, which has 964 rows corresponding to 964 defaulted loans.

We used XGBoost to model it because LGD is bounded between 0 and 1, is non-normally distributed, and has multiple modes. During data cleaning, features not relevant, such as *loanSequenceNumber*, were dropped, and so were the features not known at the prediction time such as *miRecoveries* to avoid leaking future data. Features that took on only one unique value were also dropped. Variables with many missing values (*cumulativeModificationCost*, *stepModificationFlag*, and *areaCode*) were dropped; otherwise missing values were kept. Potentially useful features were created. The steps above were applied to the test set and the OOT sample consistently. There were 35 columns in the training set.

We trained the model and tuned the hyper-parameters using 3-fold cross-validation. They are shown in Table 3. We experimented with 500 combinations of these hyperparameters using RandomizedSearchCV which yielded the set of hyperparameter values in Table 3. The final model had a RMSE of 0.0673 on the test set. This was not very useful as LGD usually varies between 0.0 to 0.20.

Table 3. Tuned parameters for XGBoost

Hyperparameter	Value
colsample_bytree	0.8479
gamma	0.0100

learning_rate	0.0356
max_depth	4
n_estimators	106
reg_alpha	0.6000
reg_lambda	0.2173
subsample	0.9970

We then computed the SHAP value contribution of the variables for the test set, as shown in Fig 9. *currentLoanDelinquencyStatus_RA*, which means whether the delinquency status is REO (Real Estate Owned) Acquisition, has the largest impact on the predicted LGD values. When a borrower defaulted and the property failed to sell at an auction, the property reverts to the lender and now the bank has to recognize the loss. Therefore, a loan with this characteristic would contribute to a much higher LGD than others. The next four important feature are interpreted as follows:

- *estimatedLoanToValue* is a ratio of the loan to estimated current value of the property. A large ratio indicates a high loan amount and a low value property, which should be associated with a high LGD for defaulted loans since other parties tend not to take over the property and the bank has to suffer the loss. However, the SHAP value suggests otherwise; the scatterplots in Fig 10 reflect this situation.
- *originalDebtToIncomeRatio*. The impact of *originalDebtToIncomeRatio* was unexpected, as high values were associated with a positive SHAP value, leading to more loss given default.
- *propertyType_SF* takes on 1 if the property type is Single Family (SF), otherwise 0. It was shown that single family properties are associated with low predicted LGD.
- *mortgageInsuarancePct*. Compared to the uncertain impact of high mortgage insurance percentages, a low insurance percentage is associated with a high LGD, which is expected.

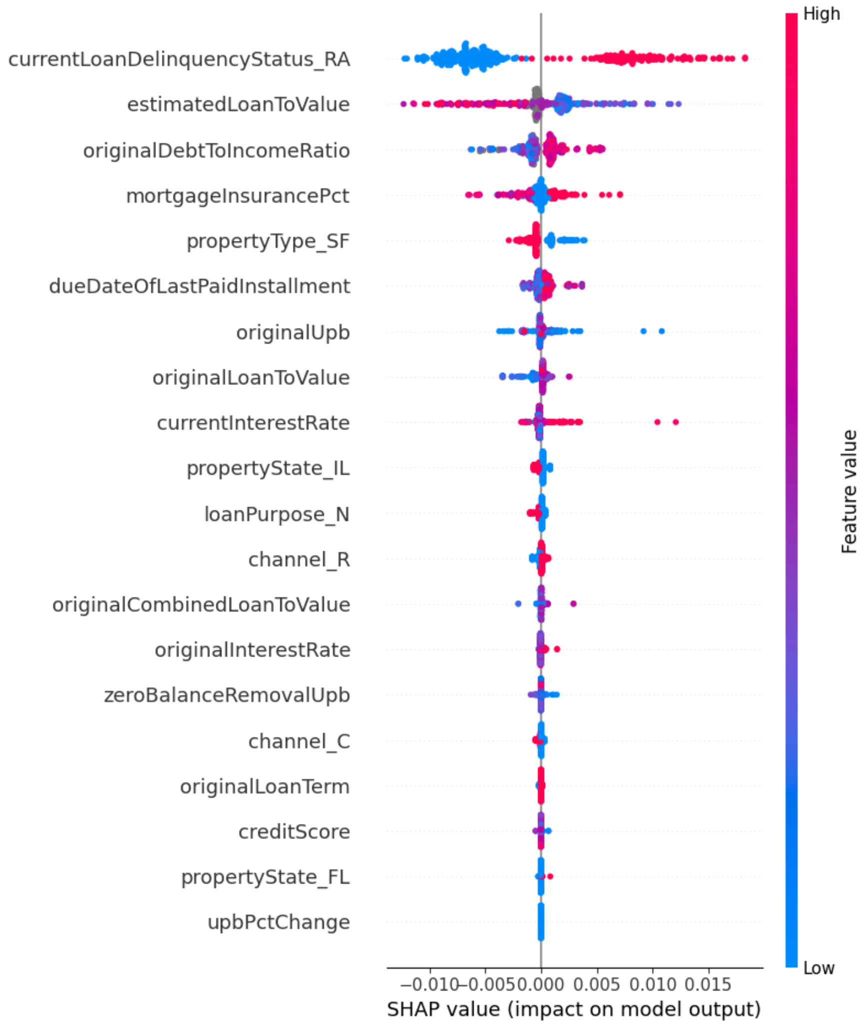

Fig 9. A Summary Plot for the Model

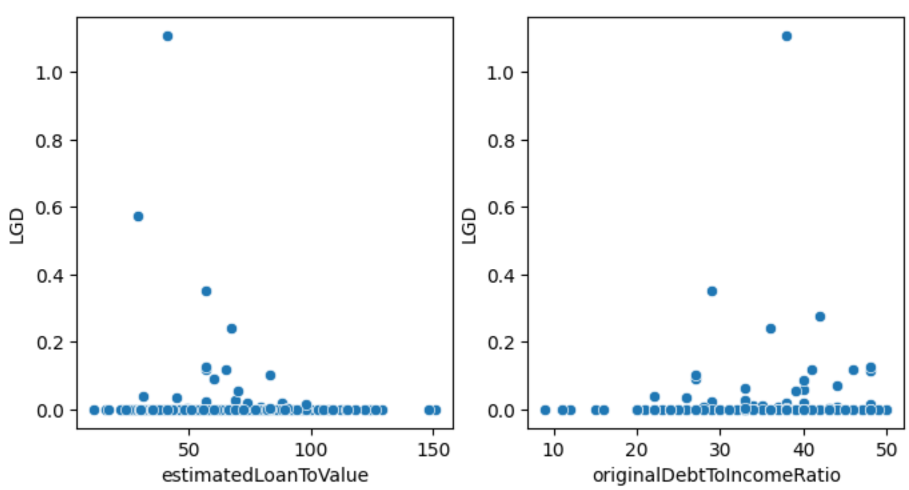

Fig 10. The Relationship between LGD and Two Predictors

Prediction on the OOT sample

Finally, we used the final model to predict the LGD in the whole OOT sample. For simplicity, we used the point estimates directly for the following provisions calculation, although LGD bins should have been created and time series modeling for downturn LGD done. We increased predicted values from less than 0.05 to 0.05 to satisfy the Basel III LGD floor of the retail mortgages [5].

5. EAD Simulation and Portfolio Provisions Calculation

With the long-run PDs and the LGDs for a portfolio of loans in the OOT sample, we derived a distribution of the provision for the portfolio by simulating a recovery rate with a uniform distribution between (0.40, 0.60). That is, EAD is computed as follows

$$\text{EAD} = \text{mortgage value} - \text{property value} \times \text{recovery rate}$$
$$\text{recovery rate} \sim \text{Uniform}(0.4, 0.6)$$
$$\text{provision} = \text{Long} - \text{run PD} \times \text{LGD} \times \text{EAD}$$

We used *zeroBalanceRemovalUpb* for mortgage value and the ratio of *zeroBalanceRemovalUpb* to *estimatedLoanToValue* for property value. The distribution of simulated provision is shown in Fig 10, and it is not sensitive to the varying recovery rate, likely due to small long-run PDs and small, stable predicted LGDs in the data. The provision has a mean of 115,833.40, with 95% confidence interval between 115,832.17 and 115,834.62. This corresponds to 0.1878% of the total amount of the loans.

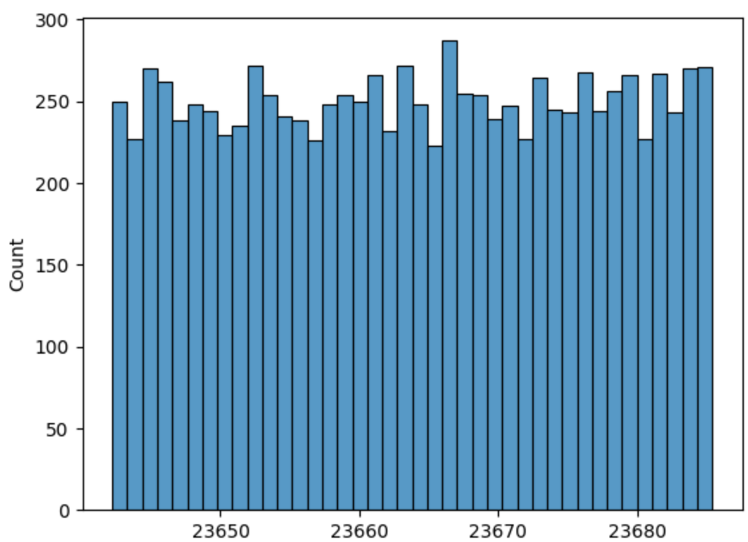

Fig 10. The Distribution of Provision

References

[1] Single-family loan-level dataset general user guide, https://www.freddiemac.com/fmac-resources/research/pdf/user_guide.pdf (accessed Apr. 14, 2025).
[2] Biron, M., & Bravo, C. (2013). On the discriminative power of credit scoring systems trained on independent samples. *Studies in Classification, Data Analysis, and Knowledge Organization*, 247–254. https://doi.org/10.1007/978-3-319-01595-8_27
[3] B. Bok, N. Petrosky-Nadeau, C. Nekerda, and R. Crump, Estimating Natural Rates of Unemployment: A Primer, https://www.frbsf.org/wp-content/uploads/wp2023-25.pdf (accessed Apr. 14, 2025).
[4] “Noncyclical rate of unemployment,” FRED, https://fred.stlouisfed.org/series/NROU (accessed Apr. 14, 2025).
[5] Bank for International Settlements, High-level summary of basel III reforms, https://www.bis.org/bcbs/publ/d424_hlsummary.pdf (accessed Apr. 15, 2025).