Consumer Credit Risk

FM 9528 Banking Analytics

Students:
251417829

Word Count: 2719

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, and a repayment performance file, and 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 every variable in the datasets by reading the General User Guide. The response variable, a binary indicator, was created according to the delinquency status in the next 12 months. If a shorter observation window was due to early repayment or reasons arising from default, the corresponding 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 were created, such as the rate of change in unpaid balance and the number of times not paying on time but not default. 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 two 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.

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. The model achieved an RMSE of 0.050 on the training and 0.068 on the test set. SHAP were employed to explain the feature importance. Finally, we used to model to predict the LGD on the OOT set and trimmed the predicted values so that they satisfied the Basel III floor of 5%.

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. 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 null and then imputed using 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 observed 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. It was similar to *maturityDate*, so it was dropped.
- *estimatedLoanToValue*. Unknown cases (5.83%) were represented as 999. We replaced them with null and then imputed 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.

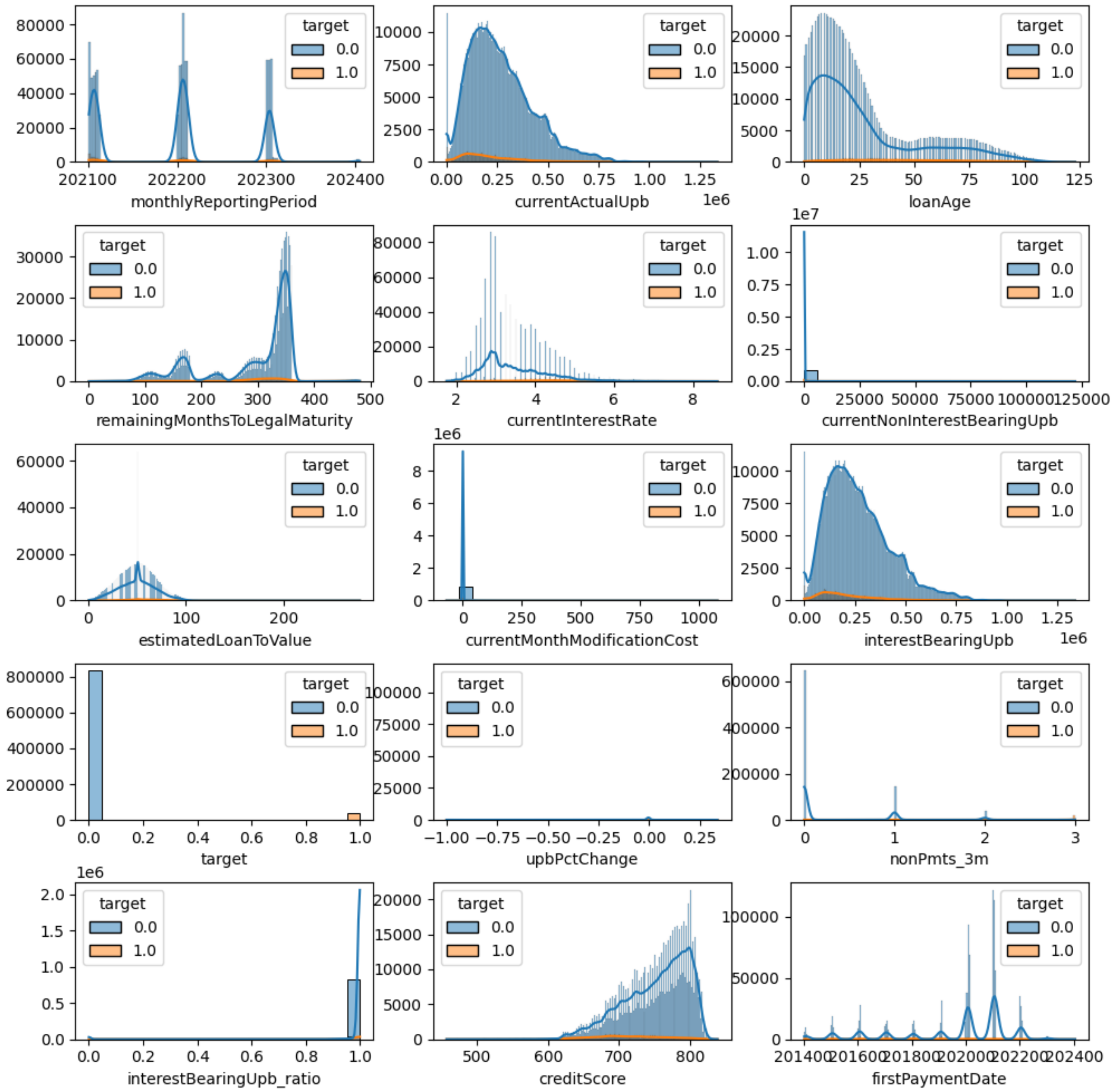

Fig 1. The Distribution of Numeric Features

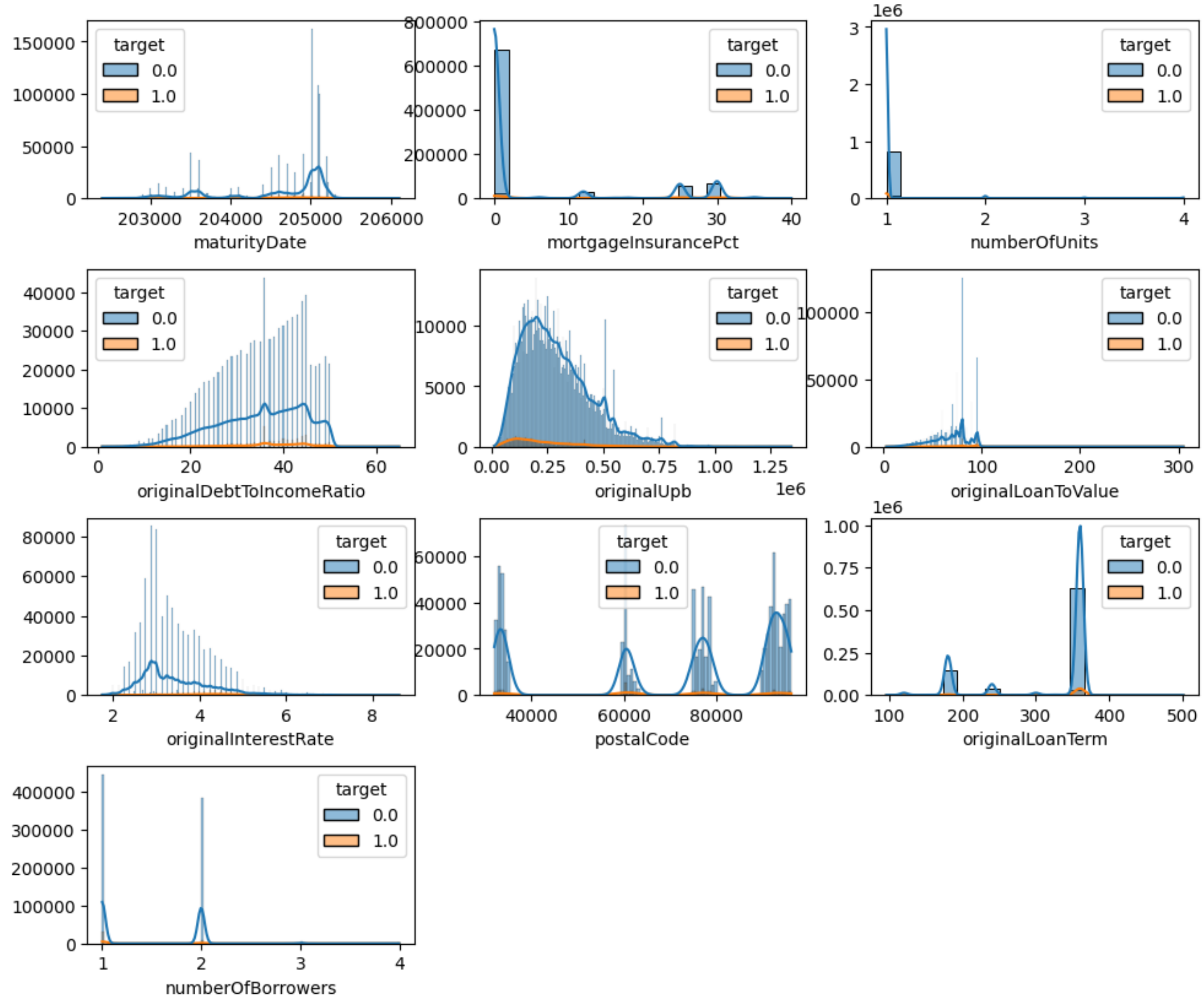

Fig 1. The Distribution of Numeric Features (continued)

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 and we only kept those 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 binning

We first used Weight of Evidence (WoE) to bin the features. 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 manually adjusted the bins. *OriginalDebtToIncomeRatio* was adjusted to have fewer bins and a more linear trend. Then the binning was applied to the training set and the test set separately. 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 as observations were removed before binning 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:

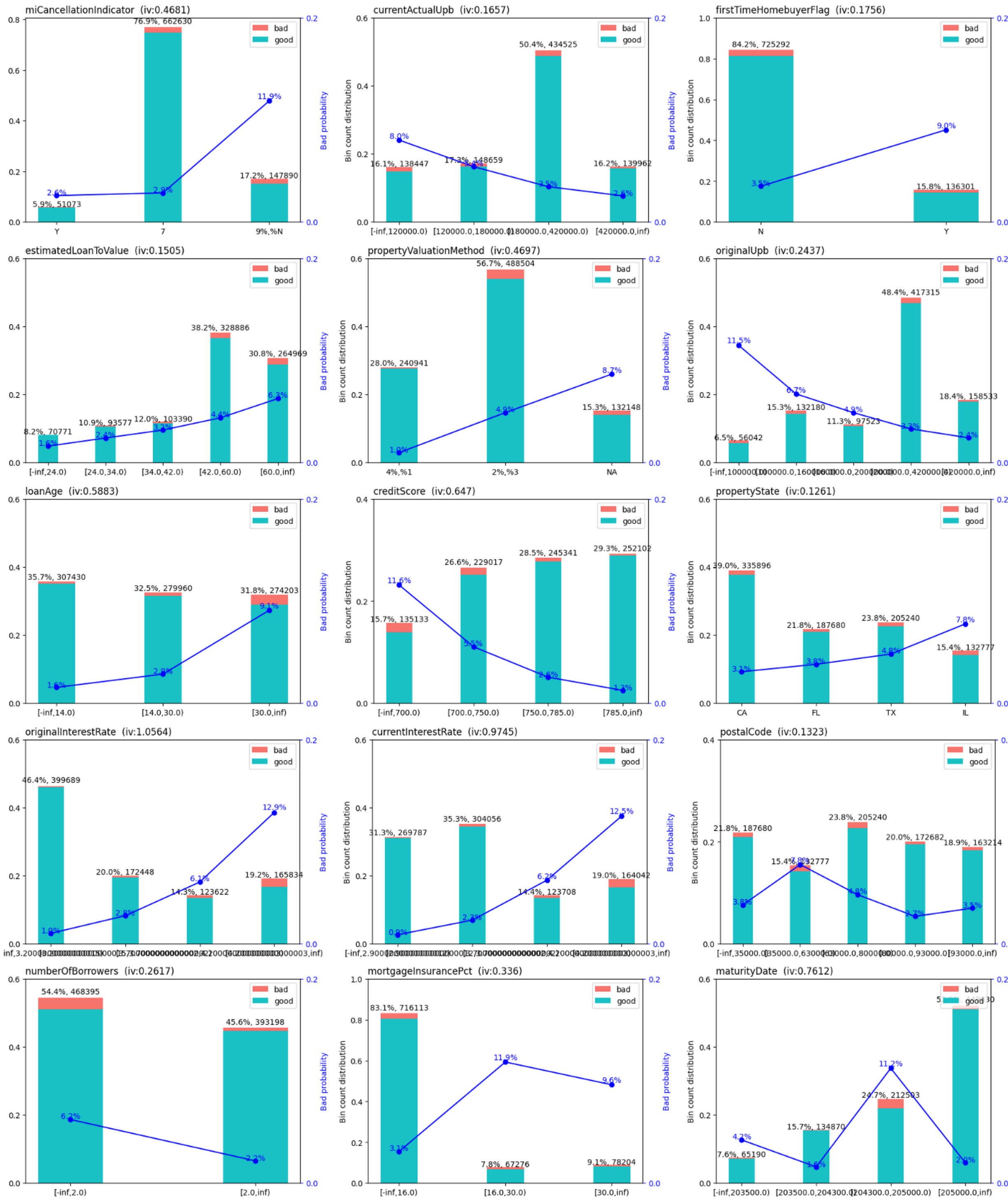

Fig 2. Weight of Evidence Binning

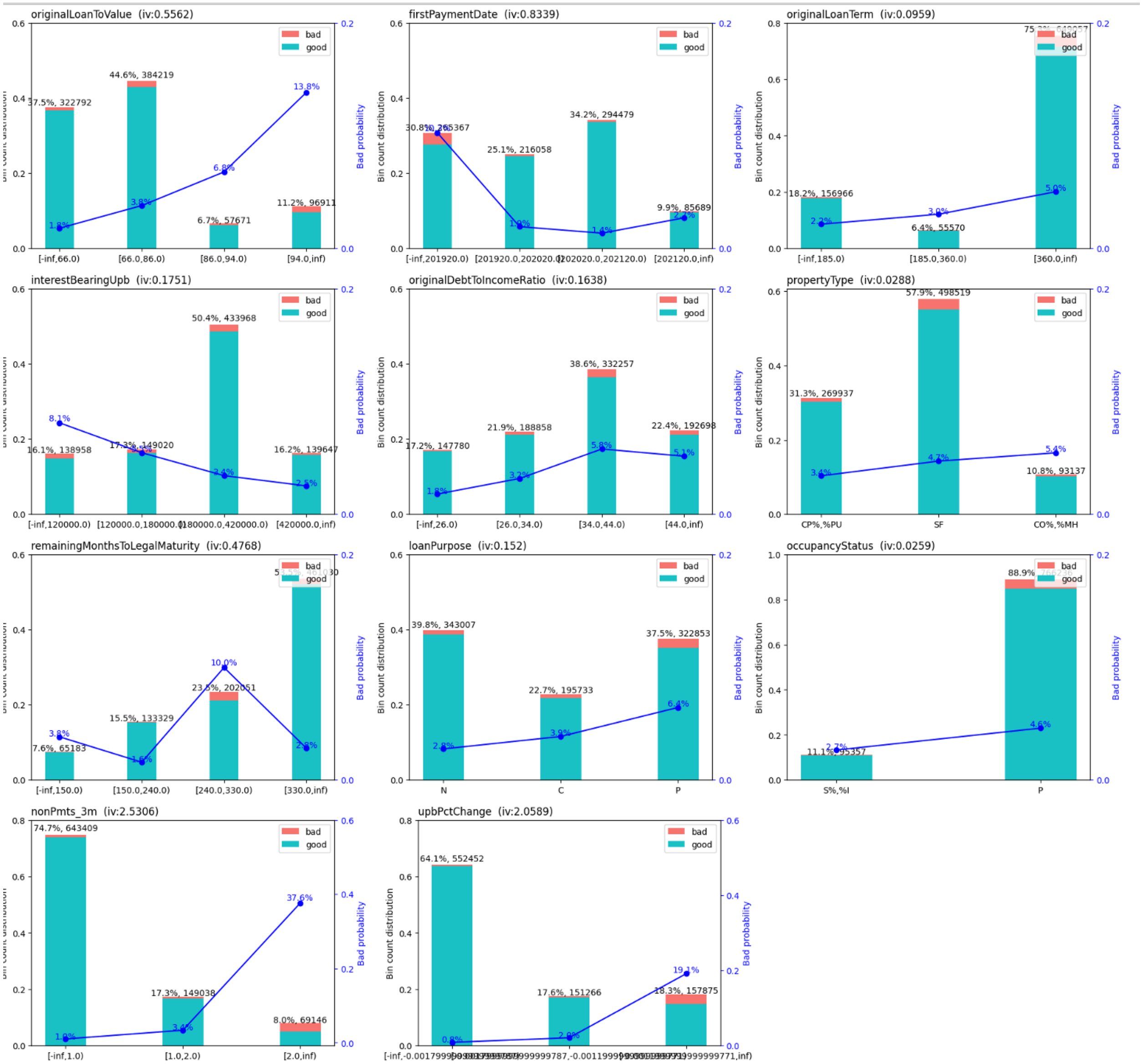

Fig 2. Weight of Evidence Binning (continued)

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

where p is the default probability. Elastic-Net mixing parameter *l1_ratio_* was 0.2 and regularization strength *C_* 21.5443.

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, 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 99.2.

Determining the best cutoff

Table 1. Cutoff and Total Expected Profit

Cutoff	Total Profit
0.52631579	60870216.9
0.47894737	60785965.1
0.57368421	60684131.6
0.62105263	60676274.6
0.66842105	60537242.2
0.38421053	60517183.3
0.43157895	60517183.3
0.71578947	60426245.3

0.76315789	60380785.8
0.33684211	60320875.6
0.81052632	60041635.1
0.85789474	59279723.2
0.28947368	59257896.5
0.90526316	58875637.6
0.24210526	58719025.8
0.95263158	57860401
1	57385489.3
0.19473684	57252875.8
0.14736842	55381122.9
0.1	49991554.7

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

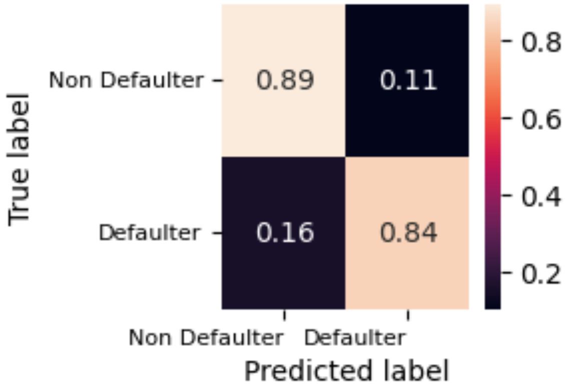

Fig 3. Confusion Matrix

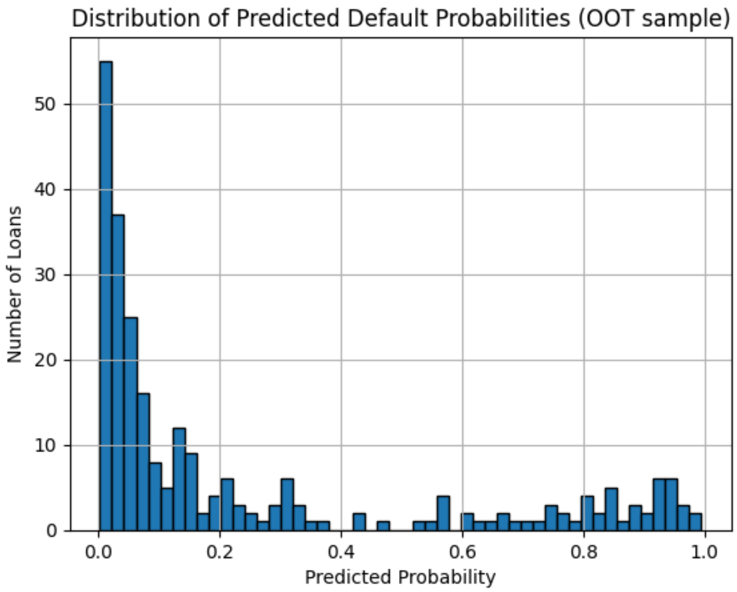

Fig 4. The Distribution of Predicted Default Probabilities

3. PD modeling

We grouped loans with similar predicted default probabilities into ratings, creating a more stable risk measurement, by approximating the ROC curve piecewise linearly using 7 line segments. As shown in Fig. 5, they followed the original ROC curve closely. The locations of the breakpoints are specific FPR values that can be mapped to probability cutoffs. Thus, we obtained 7 cutoffs or 8 bins which put a predicted probability of default into an interval/bin. 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.

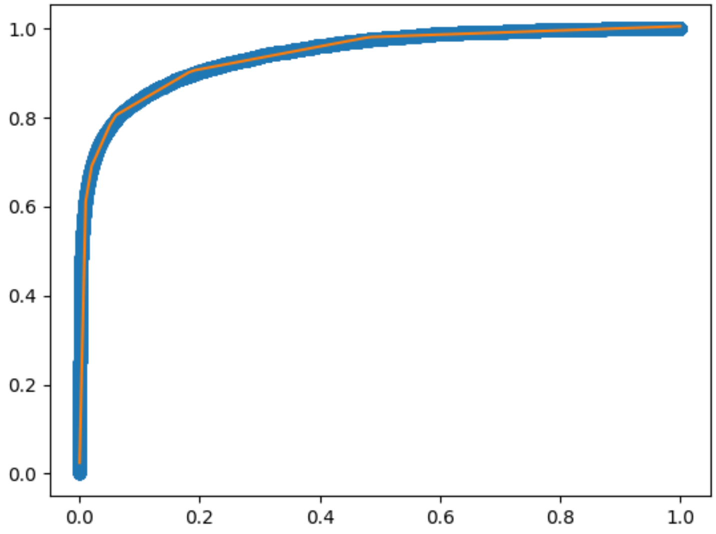

Fig 5. Piecewise linear approximation to ROC curve

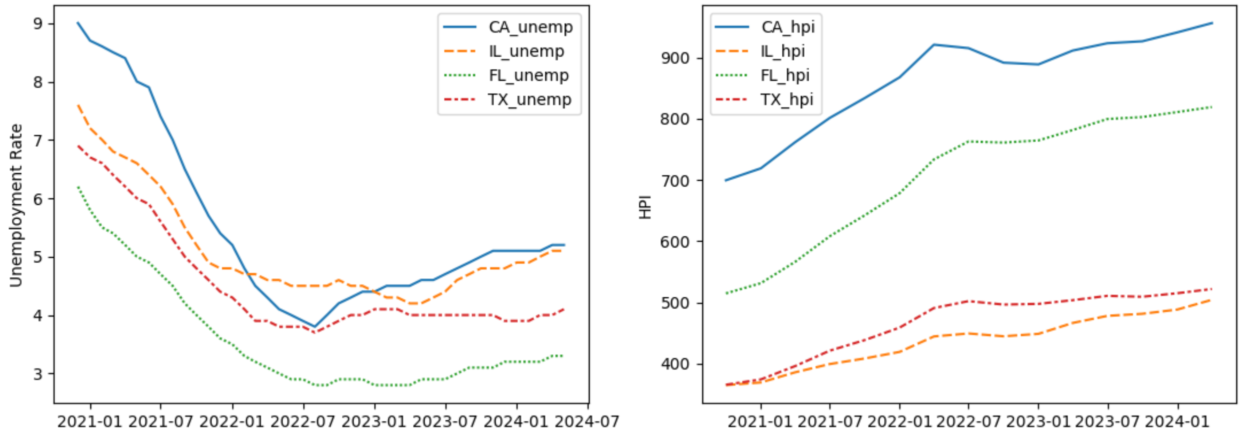

Fig 6. Macroeconomic Variables

Macroeconomic Data

We collected macroeconomic factors for each state using fredapi to model PD. We chose two economic factors: the monthly unemployment rate and the quarterly House Price Index (HPI) for the four states (CA, IL, FL, TX) in the test set. 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. We adjust its quarterly frequency by filling in the most recent available value. 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 binned into a PD group and is associated with a time index (*monthlyReportingPeriod*), a default-or-not indicator (*target*), and a *state*. For each state and each rating, at a given month, we computed the mean number of defaults as another kind of PD, which was then modeled using the ARIMAX model, which included two exogenous variables mentioned above.

There are four states and for every state there are at most eight bins, so there will be at most 32 final PD models. In fact, for some bins, due to a lack of sufficient observations, the model was not estimated. Having set up a parameter grid that contains different orders of AR, MA, Integration, and their seasonal counterparts, we selected the best model 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 [2]. 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 [3].

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, and it varies widely across the four states. Therefore, we fitted a simple SARIMA model with a smaller parameter search space to the each HPI series of the states, using the parameters that minimize AIC. We then generated forecasts for the next 12 quarters and take the average. The results were 532.73 for Texas, 970.23 for California, 524.41 for Illinois, and 836.14 for Florida.

Forecasting the long-term PD for OOT sample

We plugged in these long-term exogenous variables to the fitted model, which depended on the state and the bin to get long-term PDs for the observations in the OOT sample. The distribution is shown in Fig 7.

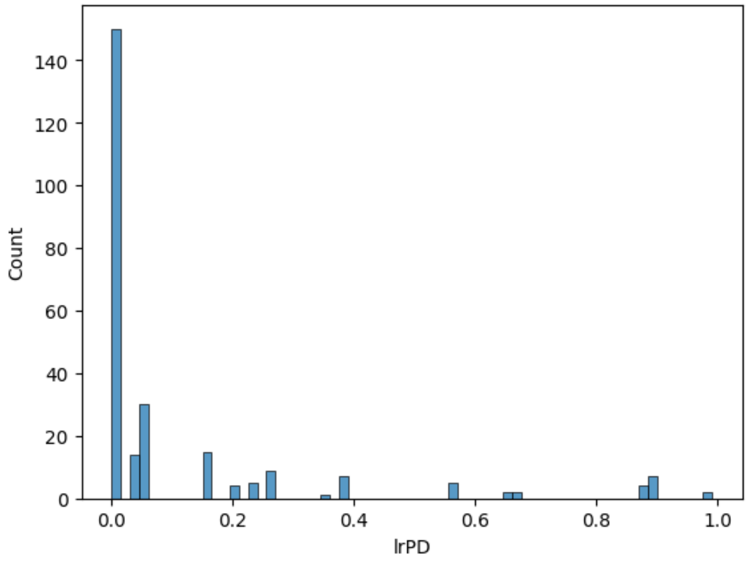

Fig 7. The Distribution of the Long-Run PD for the OOT Sample

4. LGD Modeling

Modeling LGD using XGBoost

We used XGBoosting to model Loss Given Default (LGD) because LGD is bounded between 0 and 1, is non-normally distributed, and has multiple modes. In the uncleaned training set, we computed the *LGD* defined by *actualLossCalculation* divided by *zeroBalanceRemovalUpb*, trimmed *LGD* less than 0 to 0 to indicate no loss from the default, and selected observations with non-NA *LGD* values as the sample that would be trained. This sample corresponded to 964 defaulted loans.

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 one unique value were also dropped. Missing values that had a meaning were imputed and finally, columns with missing values were dropped. Outliers were not handled due to XGBoost’s strong ability. The steps above were applied to the test set and the OOT sample consistently. There were 46 columns and no missing values in each dataset. We further used the same OneHotEncoder processing pipeline to them.

We trained the model and tuned the hyper-parameters using 3-fold cross-validation. They included the number of trees *n_estimators*, the depth of each tree *max_depth*, learning rate *learning_rate*, the percentage of observations for each tree *subsample*, and the fraction of features used per tree *colsample_bytree*, among others. Early stopping was used to prevent overfitting if the RMSE did not increase for 10 rounds. We experimented with 500 combinations of these hyperparameters using RandomizedSearchCV. The final model achieved a RMSE of 0.050 on training, and 0.068 on the test set.

We then computed the SHAP value contribution of the variables. However, due to the high dimensionality of the one-hot-encoded matrix, the results were not meaningful.

Prediction on the OOT sample

Finally, we used the learned final model to predict the LGD in the OOT sample, achieving an RMSE of 0.061. We increased predicted values from less than 0.05 to 0.05 to satisfy the Basel III LGD floor of the retail mortgages [4].

Word count: 3145

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] 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).

[3] “Noncyclical rate of unemployment,” FRED, https://fred.stlouisfed.org/series/NROU (accessed Apr. 14, 2025).

[4] Bank for International Settlements, High-level summary of basel III reforms, https://www.bis.org/bcbs/publ/d424_hlsummary.pdf (accessed Apr. 15, 2025).