Journal of Financial Risk Management, 2026, 15(3), 237-258
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

# Comparative Analytics of Machine Learning and Traditional Models in Mortgage Credit Risk: Evidence from Freddie Mac Data

**Mammy F. Sanyang, Johnathan Mun**

School of Business, San Francisco Bay University, Fremont, CL, USA
Email: johnathan.mun@sfbu.edu

**How to cite this paper:** Sanyang, M. F., & Mun, J. (2026). Comparative Analytics of Machine Learning and Traditional Models in Mortgage Credit Risk: Evidence from Freddie Mac Data. *Journal of Financial Risk Management, 15,* 237-258.
https://doi.org/10.4236/jfrm.2026.153014

**Received:** June 18, 2026
**Accepted:** August 9, 2026
**Published:** August 12, 2026



## Abstract

This paper compares the performance of traditional models and machine learning (ML) models for mortgage credit risk prediction using the Freddie Mac Single-Family Loan-Level Dataset. A sample of 100,000 loans from 2022 and 2023 was used to train and test logistic regression models (with LASSO, Ridge, and Elastic Net penalties), a generalized additive model (GAM), and three boosting ML models (XGBoost, CatBoost, and LightGBM). All models used the same training and test data splits, ensuring a fair comparison. The models were evaluated using ROC-AUC, Precision-Recall metrics, Brier scores, F1 scores, and calibration measures. Among the logistic regression models tested, engineered logistic regression had the highest ROC-AUC, 0.7956. The GAM had the highest F1-score of 0.1542, with the lowest ECE (0.000703) and Brier score (0.02283) among all models tested. The CatBoost model had the highest ROC-AUC among all models tested at 0.7988, but also the worst calibration (ECE = 0.3386, Brier = 0.1835). Furthermore, the traditional logistic regression models outperform the ML boosting models in Brier score, with values of 0.0228 - 0.0229 compared to 0.1797 - 0.1853 for the boosting models. Finally, the PSI interest rate of 3.121 for the period from 2022 to 2023 shows the need to recalibrate these models over time. Overall, the results suggest that there may be little benefit to using ML models for credit risk prediction tasks. However, CatBoost may be useful for creating early warning systems for mortgage loan defaults. At the same time, GAMs may be useful for estimating losses on mortgage loans in default and reporting that information to regulatory authorities.

## Keywords

Machine Learning, Credit Risk, Random Forest, Logistic Regression,

Calibration, Mortgage Risk, CatBoost, XGBoost, LightGBM, Ridge, Elastic Net, LASSO

## 1. Introduction

Given the importance of credit risk management in the banking industry, particularly in mortgage lending, there have been attempts to use methods beyond standard logistic regression to predict the likelihood of loan default better (Lessmann et al., 2015). Despite the number of machine learning (ML) models proposed as alternatives to the standard logistic regression model for predicting credit risk, one of the major unanswered questions in the field is whether these models offer any performance improvements for banks in real-world banking situations. Factors that influence the use of these models within banks include model calibration, robustness, and regulatory requirements for their implementation (Shi et al., 2022).

The idea of Basel requirements on computing the probability of default (PD) is to obtain the expected loss at default on a loan using the following equation:

$$EL_i = PD_i \times LGD_i \times EAD_i$$

where $PD_i$ = probability of default, $LGD_i$ = loss given default, and $EAD_i$ = exposure at default.

For a whole portfolio, one would use

$$EL_{portfolio} = \sum_{i=1}^{n} PD_i \times LGD_i \times EAD_i$$

Since the current research focuses on estimations, the models tested in this paper will estimate the first and most important component of expected credit loss, the probability of default. Thus, calibration and accuracy are important aspects of these models. A model may have a good area under the curve (AUC) measure, but if the probabilities are not well calibrated, the model will yield poor estimates of expected credit loss.

The logistic regression and (ML) models will be compared using Freddie Mac loan-level data (Freddie Mac, 2023). The data are from 2022 and 2023 and are used to evaluate the calibration, accuracy, and robustness of the models outside of the training data.

This research also uses the Brier score for error measurement, defined as:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - Y_i\right)^2$$

where $\hat{p}_i$ = predicted probability of default and $Y_i$ = actual default indicator.

A lower Brier score indicates that the model's predicted probabilities are more accurate. This is critical for identifying whether logistic regression performs much better than other ML methods. The ML models listed in **Table 1** will be tested in the current research project.

Table 1. Model runs comparing Logistic Regression (LR) and ML.

| # | Model | Family |
|---|---|---|
| 1 | Full Logistic Regression (all variables) | Traditional LR |
| 2 | Significant Logistic Regression ($p \leq 0.10$) | Traditional LR |
| 3 | Full LN Logistic Regression | Log-Transformed LR |
| 4 | Significant LN Logistic Regression | Log-Transformed LR |
| 5 | Engineered Logistic Regression (interactions) | Engineered LR |
| 6 | Ridge Logistic Regression | Regularized LR |
| 7 | LASSO Logistic Regression | Regularized LR |
| 8 | Elastic Net Logistic Regression | Regularized LR |
| 9 | Generalized Additive Model (GAM) | Nonlinear LR |
| 10 | XGBoost | Gradient Boosting ML |
| 11 | LightGBM | Gradient Boosting ML |
| 12 | CatBoost | Gradient Boosting ML |

### 1.1. Research Questions

1) Can machine learning models (XGBoost, LightGBM, CatBoost) outperform traditional logistic regression (Ridge, LASSO, Elastic Net) and a generalized additive model (GAM) in terms of predicting mortgage default risk?

2) Do machine learning models have an advantage over the logistic regression model if both models are trained and tested on the same data?

3) How do the twelve models compare in terms of their calibrated models (ECE, Brier score), and what are the effects on estimating regulatory loss?

4) Given the comparison between the two models, which one should a bank utilize?

### 1.2. Objectives

1) To test and compare all twelve models of machine learning and traditional logistic regression.

2) Additionally, to calculate and calibrate each of these models, including calculating the expected calibration error and Brier score, and to understand the impact of the Basel requirements for expected loss calculations.

3) Furthermore, to determine whether models trained on the 2022 loans dataset would remain accurate with loans in 2023, with its drastic changes in interest rates.

4) Finally, to provide recommendations regarding which of these models would be best to use in the management of the bank's credit risks.

## 2. Literature Review

The management of credit risk remains one of the most important areas within

the banking industry. The failure to properly assess the likelihood that borrowers will default on their loans has, on many occasions, led to banks suffering losses. The 2008 financial crisis exposed the consequences of poor management of credit risk in the mortgage sector (Addo et al., 2018). This is one of the reasons that research on artificial intelligence (AI) is done.

However, a higher AUC score is not the only consideration for banks. Other considerations may include whether the models are understandable to bank auditors and regulators, and whether the models stand up to challenges from loan borrowers and regulators, that is, they create an audit trail that is defensible and robust.

While a great deal of literature has been published on credit risk and artificial intelligence, it has not been distributed evenly across the topic. The literature does provide evidence that ML models outperform traditional models in certain instances. However, research has been unable to determine the impact of this performance on banks subject to strict model-explainability regulations.

### 2.1. The Limits of Traditional Credit Risk Models

Although logistic regression is not the most effective algorithm for credit risk assessment, it is used due to regulatory approval and the institutional understanding it provides. Banks that use probability estimates from the logistic regression model also use these probabilities in accordance with regulatory systems such as the Basel Accords and the Federal Reserve's SR 11-7 model risk-management instructions (Bücker et al., 2022). Thus, logistic regression has been the option of least resistance for the banks that use it, given the high probability of regulatory noncompliance if they had implemented another risk assessment model.

However, the interpretability of logistic regression comes at a cost to the model's predictive performance. Because logistic regression assumes a linear relationship between the features and the outcome, it cannot, by construction, capture nonlinear interactions between features and the likelihood of loan default in its standard form. Dumitrescu et al. (2022) found that their logistic regression model, with additional terms from decision trees to account for nonlinear interactions, achieved higher accuracy across several datasets.

Thus, the findings indicate that the logistic regression model lacks sufficient information to predict the likelihood that a borrower will default on their loan. Furthermore, another drawback of logistic regression is that it is difficult to provide an objective evaluation of the model's effectiveness, due to a tendency in the literature to compare traditional loan risk assessment models with machine learning-based models using datasets on which logistic regression models had previously been applied to make lending decisions. The findings of such literature, therefore, may be influenced by the feedback mechanism the model exhibited towards its developers and lenders. This point must be kept in mind when reviewing the literature that intends to make such comparisons.

## 2.2. AI and Machine Learning at Credit Risk

### *The Broader Consensus and Its Limits*

Shi et al. (2022), after reviewing 76 research articles over eight years, found that deep learning models outperformed both traditional models and statistical algorithms for the same tasks and benchmarks. However, there were several problems in existing literature on this topic, including data imbalance, a lack of model transparency, and poor use of datasets. These findings were replicated by Ayari et al. (2025), who reviewed 63 articles published between 2018 and 2024 and discovered findings similar to those of Shi et al. Furthermore, XGBoost emerged as the highest-performing machine-learning algorithm in these articles, with many researchers in the field beginning to use hybrid and ensemble models in their natural language processing approaches.

However, there were some issues with the research study designs performed by these researchers. For example, Seitshiro & Govender (2024) reviewed the commercial banking credit registry and found that using a standard logistic regression without a transformation outperformed complex ML models. Furthermore, when using the weight-of-evidence transformation, which is often performed as part of model pre-processing, the results did not improve as much as expected. This outcome, however, challenges the findings of other researchers, as it shows that the performance of logistic regression and ML algorithms is entirely dependent on the data and preprocessing steps used. Thus, research suggesting that ML algorithms outperform logistic regression models should be interpreted with caution.

### *Model Performance and the Metrics Problem*

Another issue with prior research on the performance of logistic regression and other ML algorithms is that most studies use accuracy as their primary performance metric. Because the data on credit defaults is inherently imbalanced, most loans in the dataset are non-default loans. A model that always predicts non-default loans will achieve high accuracy without having learned much about the loans. Among the most notable studies on this topic is that of Chang et al. (2024), who used a more comprehensive and accurate set of metrics to assess the performance of XGBoost, LightGBM, AdaBoost, and feedforward neural network models for predicting credit card default. In the cited study, the XGBoost approach achieved an impressive 99.4% prediction accuracy. In other studies, researchers have found different results for the same XGBoost models. For instance, Hossain et al. (2025) found that XGBoost models achieved only 88.7% accuracy on a banking dataset, with AUCs of 80.3% and 91.3%. However, the authors used the same accuracy and AUC scores to determine that the performance of the machine-learning algorithms was not as high as suggested by other studies on the topic. Thus, these differing accuracy results highlight the potential for studies that rely solely on accuracy metrics to draw misleading conclusions about the performance of logistic regression and ML models.

### *XGBoost*

XGBoost is usually a top performer for credit risk prediction because it captures

nonlinear relationships and interactions and handles structured tabular data very well. The general boosting form is

$$\hat{y}_i = \sum_{m=1}^{M} f_m(X_i)$$

where each $f_m$ is a decision tree added sequentially.

The objective function is

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m)$$

where $l(y_i, \hat{y}_i)$ is the prediction loss and $\Omega(f_m)$ is the regularization penalty that controls the tree complexity.

***LightGBM***

LightGBM, built for speed and scalability, is another gradient-boosting method that is efficient on large datasets. It is especially useful when the dataset has categorical or high-dimensional features.

***CatBoost***

The mortgage dataset includes categorical variables such as Loan Purpose, Occupancy Status, Property State, and First Time Home Buyer. CatBoost handles categorical variables well without requiring as much manual encoding. This approach may outperform random forest and logistic regression when categorical features are important.

## 2.3. Research Gaps and the Contribution of This Study

There are three main conclusions from the literature. First, while there is an acknowledged technological advantage to AI credit risk models, the reliability of their results depends on the methods used to build and evaluate them. Second, while there are an increasing number of articles that use large datasets of mortgage loans, such as Freddie Mac's dataset, to evaluate mortgage credit risk, these studies are not as thorough as those focused on general credit risk. Third, and most important, is that there is a gap between the technological findings on these models' performance and the considerations relevant to their deployment within a banking organization. This study is intended to fill these gaps in the literature.

Overall, this study aims to provide a more reliable evaluation of the performance of advanced AI-based models compared with traditional logistic regression for credit risk, using Freddie Mac's mortgage loan dataset. By including considerations of explainability and governance requirements for AI models, this study also addresses a gap in the literature beyond performance evaluations. Thus, this study contributes to existing literature not only through its technical application of information on AI models and credit risk but also by providing recommendations on their applicability to banking organizations.

Due to computational constraints, this study was limited to evaluating loans originated in the 2022 and 2023 calendar years. Thus, the analysis did not account for the long-term cycles of the loans or for differences in loan behavior across

economic conditions and interest rates. While the time frame selected for this study was not representative of past periods of economic stress, it reflected the economic environment following the COVID-19 pandemic. The findings of this study, then, are specific to this economic environment, and their application elsewhere requires judgment regarding how much the economic conditions have shifted. Future studies could expand the longitudinal time frame for which Freddie Mac's loan data was used to evaluate AI-based credit risk models.

## 3. Methodology and Analysis

### 3.1. Data Source and Sample Construction

Table 2 provides an overview of the data used, including the loan-level mortgage data from the Freddie Mac Single-Family Loan Dataset. The final dataset contains approximately 100,000 observations, evenly split between loans from the 2022 and 2023 loan cohorts. Each loan record includes variables related to the borrower's credit score, loan-to-value ratio (LTV), debt-to-income ratio (DTI), unpaid principal balance (UPB), interest rate, occupancy of the mortgage loan, purpose of the mortgage loan, the number of borrowers of the mortgage loan, and the state of the property that is being mortgaged. The variable of interest is the default status of the mortgages, represented as a binary variable that takes the value 1 if the mortgage has defaulted (missed payments for 90 or more days) and 0 otherwise. Thus, the dataset's features relate to borrowers' credit quality, their income relative to their mortgage payments, mortgage segmentation, and borrowers' leverage. Furthermore, there were no missing values for any variable in the final dataset. The default rate for mortgages in the dataset is approximately 2.43%, with 2022 mortgages at 3.02% and 2023 mortgages at 1.85%.

Table 2. Data overview.

| Metric | Value |
|---|---|
| Total Observation | 100,000 |
| 2022 Observation | 50,000 |
| 2023 Observation | 50,000 |
| Total Variables | 12 |
| Missing Variables | 0 |
| Overall Default | 2.43% |
| 2022 Default | 3.02% |
| 2023 Default | 1.85% |
| Train set (70/30 stratified split) | 70,000 rows - 1704 defaults (2.43%) |
| Test set (70/30 stratified split) | 30,000 rows - 730 defaults (2.43%) |
| Random State | 42 |

## 3.2. Modeling Framework, Evaluation Metrics, and Statistical Significance

### 3.2.1. Validation Design

All of the models were trained on the same 70,000-row training dataset and then tested on the same 30,000-row dataset; the split of approximately 70% training and 30% testing data is stratified to reflect the default rate of the full 100,000-row dataset. The default rate is the same for both the training and testing datasets: 2.43% of each dataset defaults on its loan. There are 1704 defaults in the training dataset and 730 defaults in the test dataset. Therefore, the comparison between the two models is unbiased. Furthermore, an assessment of out-of-time data (2022 versus 2023 loans) was also performed for the logistic regression and random forest models.

### 3.2.2. Coefficient Estimation and Data Preprocessing

Using encoded categorical variables (Occupancy Status, First Time Homebuyer Flag, Loan Purpose, and Property State), one category from each variable was removed to reduce multicollinearity effects. All continuous variables were then preprocessed by standardizing them to have a mean of zero and a variance of one, prior to fitting the models. For these reasons, the computed probabilities from the gradient boosting machine models (XGBoost, LightGBM, CatBoost) differ from those of the logistic regression model. As a result, the Brier scores from the gradient boosting machines are not directly comparable to those from the logistic regression model. Consequently, the loss estimates from the ML models must be calibrated to reflect the true probabilities before using them to predict loan defaults.

### 3.2.3. Logistic Regression Model Sequence (Models 1 - 5)

Five logistic regression models were fitted to the data in sequence. Model 1 included all independent variables in their raw, standardized form for the logistic regression. Model 2 included only variables statistically significant at $p \le 0.10$ in Model 1. Model 3 was a model in which the continuous variables were transformed using natural logs. Model 4 included only the variables that were statistically significant in Model 3. Finally, Model 5 included interaction terms between the continuous variables, formed as combinations of variables that yielded the maximum pseudo-R-squared for the model or the lowest AIC score.

***The Core Credit-Risk Modeling Framework***

The dependent variable is a binary default indicator:

$$Y_i = \begin{cases} 1, & \text{if loan } i \text{ becomes } 90+\text{ days delinquent} \\ 0, & \text{otherwise} \end{cases}$$

Let each loan have features

$$X_i = \left(CreditScore_i, LTV_i, DTI_i, UPB_i, Rate_i, Occupancy_i, Purpose_i, State_i, Borrowers_i\right)$$

The goal is to estimate

$$PD_i = P\left(Y_i = 1 \mid X_i\right)$$

where $PD_i$ is the predicted probability of default.

***Logistic Regression***

Logistic regression is the typical traditional benchmark, where we define it as

$$P(Y_i = 1 \mid X_i) = p_i$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})}}$$

Equivalently, the log-odds ratio is

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

Logistic regression assumes a linear relationship between the predictors and the log-odds of default, which makes it transparent and easy to understand. However, it may miss nonlinear patterns in mortgage risk. A maximum likelihood function is used to solve for $p_i$ such that

$$L(\beta) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}$$

Moreover, the log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \left[ Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i) \right]$$

***Modifications of Logistic Regression Models***

We also tested other methods of regularized logistic regression, such as Ridge, LASSO, and Elastic Net, as these methods maintain interpretability while offering greater coefficient stability.

***Ridge Logistic Regression***

$$\min_{\beta} \left[ -\ell(\beta) + \lambda \sum_{j=1}^{k} \beta_j^2 \right]$$

Ridge-based regressions are useful when the predictors are highly correlated with each other, for example, interest rate, loan purpose, and refinance activity.

***LASSO Logistic Regression***

$$\min_{\beta} \left[ -\ell(\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \right]$$

The LASSO approach performs variable selection by shrinking the coefficients of less important variables to zero.

***Elastic Net Logistic Regression***

$$\min_{\beta} \left[ -\ell(\beta) + \lambda \left( \alpha \sum_{j=1}^{k} |\beta_j| + (1 - \alpha) \sum_{j=1}^{k} \beta_j^2 \right) \right]$$

Elastic Net combines the methods of Ridge and LASSO. Elastic Net is another model that could be applied to this problem, as its results would be explainable to regulators and would meet Basel banking requirements.

***ROC-AUC and Precision-Recall***

We use ROC-AUC calculations as a measurement of accuracy. Accuracy itself can be calculated using the following equation, where the variables represent the number of true positives ( *TP*), false positives (*FP*), true negatives ( *TN*), and false negatives (*FN*):

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve plots

$$TPR(\tau) \text{ against } FPR(\tau)$$

across different thresholds $\tau$.

The Precision-Recall curve plots

$$Precision(\tau) \text{ against } Recall(\tau)$$

This is especially important because mortgage default is rare. Precision-Recall AUC may be more informative than ROC-AUC when the default rate is only 2% - 3%.

***Population Stability Index***

The Population Stability Index (PSI) splits the distribution of a feature $X$, into bins. Let

$$A_j = \text{proportion of observations in bin } j \text{ in the training year}$$

$$E_j = \text{proportion of observations in bin } j \text{ in the validation year}$$

Then

$$PSI = \sum_{j=1}^{J} (E_j - A_j) \ln\left(\frac{E_j}{A_j}\right)$$

Typical interpretations include a little shift where

$$PSI < 0.10$$

a moderate shift where

$$0.10 \le PSI < 0.25$$

and a major shift where

$$PSI \ge 0.25$$

Our analysis indicates an interest-rate PSI above 3.0, which is a very large distribution shift.

***Generalized Additive Model (GAM) on Logistic Regression***

To further improve the competitiveness of our standard logistic regression with

ML, we also tested interactions among the independent variables. For example,

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1 CS_i+\beta_2 LTV_i+\beta_3 DTI_i+\beta_4 R_i+\beta_5\left(CS_i\times LTV_i\right)+\beta_6\left(LTV_i\times DTI_i\right)+\varepsilon_i$$

Other useful interactions that were tested include

$$Credit\ Score\left(CS\right)\times LTV$$

$$LTV\times DTI$$

$$Interest\ Rate\times Loan\ Purpose$$

$$First\ Time\ Homebuyer\times LTV$$

$$Occupancy\ Status\times Loan\ Purpose$$

This interaction helps to address the linearity limitation of logistic regression. Furthermore, rather than testing the effects of the variables in their current forms, we also tested the spline-transformed versions of each of those variables:

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+f_1\left(CS_i\right)+f_2\left(LTV_i\right)+f_3\left(DTI_i\right)+f_4\left(Rate_i\right)+\varepsilon_i$$

where each $f_j\left(\cdot\right)$ is a nonlinear spline function. The advantage of using GAMs is that they can capture nonlinear relationships among variables while still providing interpretable results for study readers.

The diagrams compare logistic regression against random forest models. The dashed line represents the performance of random classification. The shapes of the curves are similar, indicating temporal stability of the models, and the differences in AUC values reflect the models' relative performance in discriminating between the classes.

Performance was measured using ROC-AUC, Precision-Recall AUC, F1 score, and Brier score. Calibration was also evaluated using the expected calibration error and the calibration slope. The statistical significance of the difference in ROC-AUC was also tested.

For ML calibration, XGBoost, LightGBM, and CatBoost use class-weighted training, which affects their raw probability outputs. Their ECE values (0.34 - 0.35) and Brier values (0.8 - 0.19) indicate that they are not comparable to the calibration measures of logistic regression without post hoc calibration.

#### 3.2.4. Advanced Machine Learning Models (Models 10 - 12)

All advanced ML models (XGBoost, LightGBM, and CatBoost, which support categorical features) are gradient boosting ensemble methods. All were trained on the same 70,000-row training set and evaluated on the same 30,000-row test set, as well as all logistic regression models, and were tested at a threshold of 0.10.

## 4. Analytical Results

### 4.1. Exploratory Data Analysis

Mortgage default was relatively uncommon in the dataset, with only 2.43% of the

loans defaulting (Table 3). Credit score was the strongest factor predicting default, followed by loan-to-value (LTV) and debt-to-income (DTI) ratio (Table 4). The number of borrowers, property state, loan purpose, and the year the loan was obtained also influenced the default risk, indicating that mortgage risk performance varies with loan features, location, borrower characteristics, and the year the loan was obtained.

Table 3. Distribution of dependent variables (Default).

| Default Status | Count | Percentage |
|---|---|---|
| Non-Default (0) | 97,566 | 97.57% |
| Default (1) | 2434 | 2.43% |
| Total | 100,000 | 100% |

Table 4. Top variables associated with default.

| Variable | Association Statistic | Method |
|---|---|---|
| Credit Score | 0.1495 | Absolute Pearson Correlation |
| Loan Purpose | 0.0581 | Cramér's V |
| Number of Borrowers | 0.0468 | Absolute Pearson Correlation |
| Property State | 0.0433 | Cramér's V |
| Year | 0.0380 | Absolute Pearson Correlation |
| Original DTI | 0.0360 | Absolute Pearson Correlation |
| Occupancy Status | 0.0304 | Cramér's V |
| Original LTV | 0.0222 | Absolute Pearson Correlation |

## 4.2. Analysis and Results of Logistic Regression (Models 1 - 5)

All the logistic regression models were sequentially estimated on 70,000-row training sets and 30,000-row test sets. As shown in Tables 5-9, credit scores had the largest negative coefficient among all variables, indicating that higher credit quality reduces the probability of default. The year the loan was obtained, DTI, LTV, loan purpose, number of borrowers, interest rate, and UPB also provide meaningful signals.

Table 5. Logistic regression of all variables (Model 1).

| Variable | Coefficient | SE | Z-score | *p*-value | Sig ($p \leq 0.10$) |
|---|---|---|---|---|---|
| Constant | −4.537 | 0.632 | −7.178 | 0.000 | Yes |
| Credit Score Z | −0.924 | 0.023 | −40.742 | 0.000 | Yes |
| Original LTV Z | 0.480 | 0.033 | 14.368 | 0.000 | Yes |
| Original DTI Z | 0.060 | 0.012 | 5.013 | 0.000 | Yes |

**Continued**

| | | | | | |
|---|---|---|---|---|---|
| Original UPB Z | 0.138 | 0.026 | 5.228 | 0.000 | Yes |
| Original Interest Rate Z | 0.083 | 0.028 | 2.955 | 0.003 | Yes |
| Number of Borrowers Z | −0.451 | 0.024 | −18.573 | 0.000 | Yes |
| First Time Homebuyer Flag Y | −0.093 | 0.055 | −1.690 | 0.091 | Yes |
| Occupancy Status P | 0.529 | 0.105 | 5.027 | 0.000 | Yes |
| Loan Purpose N | −0.520 | 0.106 | −4.891 | 0.000 | Yes |
| Loan Purpose P | −0.494 | 0.066 | −7.516 | 0.000 | Yes |
| Year | −0.470 | 0.055 | −8.577 | 0.000 | Yes |
| Property State WA | 0.064 | 0.643 | 0.100 | 0.920 | No |
| Property State WI | −0.119 | 0.655 | −0.182 | 0.856 | No |
| Property State WY | −0.143 | 0.953 | −0.150 | 0.881 | No |

**Table 6.** Logistic regression of significant variables only $p \leq 0.10$ (Model 2).

| Variable | Coefficient | SE | Z-score | *p*-value | CI Lower | CI Upper |
|---|---|---|---|---|---|---|
| Constant | −4.121 | 0.112 | −36.720 | 0.000 | −4.341 | −3.901 |
| Credit Score Z | −0.925 | 0.023 | −41.090 | 0.000 | −0.969 | −0.881 |
| Original LTV Z | 0.439 | 0.032 | 13.532 | 0.000 | 0.376 | 0.503 |
| Original DTI Z | 0.064 | 0.012 | 5.122 | 0.000 | 0.039 | 0.088 |
| Original UPB Z | 0.200 | 0.023 | 8.740 | 0.000 | 0.155 | 0.244 |
| Original Interest Rate Z | 0.083 | 0.028 | 3.004 | 0.003 | 0.029 | 0.137 |
| Number of Borrowers Z | −0.451 | 0.024 | −18.425 | 0.000 | −0.499 | −0.403 |
| First Time Homebuyer Flag Y | −0.093 | 0.055 | −1.689 | 0.091 | −0.200 | 0.015 |
| Occupancy Status P | 0.529 | 0.105 | 5.028 | 0.000 | −0.200 | 0.735 |
| Loan Purpose N | −0.520 | 0.106 | −4.891 | 0.000 | −0.729 | −0.312 |
| Loan Purpose P | −0.494 | 0.066 | −7.517 | 0.000 | −0.622 | −0.365 |
| Year | −0.467 | 0.055 | −8.536 | 0.000 | −0.574 | −0.360 |
| Property State GU | 2.392 | 1.119 | 2.138 | 0.033 | 0.199 | 4.585 |
| Property State HI | 1.115 | 0.313 | 3.564 | 0.000 | 0.502 | 1.727 |

**Table 7.** LN transformed all variables (Model 3).

| Variable | Coefficient | SE | Z-score | *p*-value | Sig ($p \leq 0.10$) |
|---|---|---|---|---|---|
| Constant | −1631 | 841.016 | −0.002 | 0.998 | NO |
| ln(Credit Score) | −14.147 | 0.348 | −40.696 | 0.000 | Yes |
| ln (Original LTV) | 1.516 | 0.113 | 13.385 | 0.000 | Yes |

**Continued**

| | | | | | |
|---|---|---|---|---|---|
| ln(Original DTI) | 0.876 | 0.098 | 8.979 | 0.000 | Yes |
| ln(Original UPB) | 0.089 | 0.048 | 1.860 | 0.063 | Yes |
| Original Interest Rate_ | 0.137 | 0.114 | 1.200 | 0.230 | NO |
| Number of Borrowers_ | −2.048 | 0.113 | −18.114 | 0.000 | Yes |
| Occupancy Status P | 0.668 | 0.148 | 4.527 | 0.000 | Yes |
| Loan Purpose N | −0.802 | 0.153 | −5.253 | 0.000 | Yes |
| Loan Purpose P | −0.694 | 0.084 | −8.234 | 0.000 | Yes |
| Year | −0.460 | 0.056 | −8.256 | 0.000 | Yes |
| Property State WA | 0.065 | 0.930 | 0.070 | 0.944 | No |
| Property State WI | −0.279 | 0.947 | −0.294 | 0.769 | No |

**Table 8.** LN transformed significant variables (Model 4).

| Variable | Coefficient | SE | Z-score | *p*-value | CI Lower | CI Upper |
|---|---|---|---|---|---|---|
| Constant | 80.067 | 2.249 | 35.601 | 0.000 | 75.659 | 84.475 |
| ln(Credit Score) | −14.195 | 0.345 | −41.193 | 0.000 | −14.870 | −13.519 |
| ln(Original LTV) | 1.327 | 0.107 | 12.422 | 0.000 | 1.117 | 1.536 |
| ln(Original DTI) | 0.932 | 0.097 | 9.631 | 0.000 | 0.742 | 1.121 |
| ln(Original UPB) | 0.222 | 0.041 | 5.376 | 0.000 | 0.141 | 0.302 |
| Original Interest Rate | −0.271 | 0.109 | −2.500 | 0.012 | −0.484 | −0.059 |
| Number of Borrowers_ | −2.047 | 0.114 | −17.919 | 0.000 | −2.271 | −1.823 |
| Occupancy Status P | 0.681 | 0.148 | 4.611 | 0.000 | 0.392 | 0.971 |
| Loan Purpose N | −0.826 | 0.153 | −5.394 | 0.000 | −1.126 | −0.526 |
| Loan Purpose P | −0.706 | 0.085 | −8.262 | 0.000 | −0.873 | −0.538 |
| Year | −0.463 | 0.056 | −8.298 | 0.000 | −0.572 | −0.353 |
| Property State GU | 3.350 | 1.772 | 1.890 | 0.059 | −0.123 | 6.822 |
| Property State HI | 1.766 | 0.448 | 3.938 | 0.000 | 0.887 | 2.645 |

**Table 9.** "Engineered" logistic regression with interactions (Model 5).

| Variable | Coefficient | SE | Z-score | *p*-value | Sig ($p \leq 0.10$) |
|---|---|---|---|---|---|
| Constant | −4.346 | 0.033 | −130.845 | 0.000 | Yes |
| ln(Credit Score) | −0.931 | 0.095 | −9.845 | 0.000 | Yes |
| ln (Original LTV) | −0.245 | 0.109 | −2.238 | 0.025 | Yes |
| ln(Original DTI) | 0.333 | 0.043 | 7.805 | 0.000 | Yes |
| ln(Original UPB) | 0.139 | 0.024 | 5.673 | 0.000 | Yes |

**Continued**

| | | | | | |
|---|---|---|---|---|---|
| Original Interest Rate_ | 0.554 | 0.199 | 2.782 | 0.005 | Yes |
| Number of Borrowers_ | −0.433 | 0.025 | −17.609 | 0.000 | Yes |
| Credit Score × Original LTV | 0.473 | 0.152 | 3.109 | 0.002 | Yes |
| Original DTI/Credit Score | −0.309 | 0.119 | −2.589 | 0.010 | Yes |
| Original LTV × Original DTI | 0.356 | 0.137 | 2.594 | 0.009 | Yes |
| Original Interest Rate$^2$ | −0.022 | 0.167 | −0.132 | 0.895 | NO |
| Occupancy Status P | 0.128 | 0.032 | 4.039 | 0.000 | Yes |
| Loan Purpose N | −0.110 | −0.110 | −4.236 | 0.000 | Yes |
| Loan Purpose P | −0.221 | −0.221 | −7.706 | 0.000 | Yes |
| Year | −0.221 | −0.221 | −7.936 | 0.000 | Yes |

Table 10 shows that Model 5, the "Engineered" model, has the best AIC and McFadden $R^2$ statistics, indicating that it has the most concise specification for the dataset. GAM (Model 9) has the best ROC-AUC statistics and the lowest Brier score. The models that utilize log-transformations of the variables have a lower AIC than the models that use the raw variables, but have similar hold-out AUC scores, indicating that the features engineered for Model 5 provide more value than the log-transformations alone.

**Table 10.** Goodness of fit results on a 70/30 split (Models 1 - 9).

| Model | McFadden $R^2$ | AIC | BIC | Log-Like | Holdout AUC | Brier |
|---|---|---|---|---|---|---|
| Model 1 (Full LR) | 0.1200 | 6175 | 6723 | −3082.6 | 0.7895 | 0.0229 |
| Model 2 (Sig LR) | 0.1202 | 6069 | 6186 | −3028.9 | 0.7890 | 0.0229 |
| Model 3 (LN Full) | 0.1212 | 6167 | 6715 | −3077.5 | 0.7927 | 0.0229 |
| Model 4 (LN Sig) | 0.1172 | 6086 | 6186 | −3037.2 | 0.7870 | 0.0229 |
| Model 5 (Engineered) | 0.1248 | 6047 | 6205 | −3016.9 | 0.7956 | 0.0229 |
| Model 6 (Ridge) | 0.1242 | 6048 | 6189 | −3017.2 | 0.7947 | 0.0229 |
| Model 7 (LASSO) | 0.1241 | 6048 | 6190 | −3017.4 | 0.7946 | 0.0229 |
| Model 8 (Elastic Net) | 0.1241 | 6048 | 6190 | −3017.4 | 0.7946 | 0.0229 |
| Model 9 (GAM) | 0.1285 | 6099 | 6572 | −3042.5 | 0.7971 | 0.0228 |

The PSI analysis (Figure 1) helps to show the extent of the change in the distribution of each of these features between the training and validation years. The interest rate feature had a significant shift in its distribution between the two validation years (PSI > 0.25). Additionally, the distribution of loan purposes showed a significant shift (PSI = 0.171). The change in the purpose of these loans is therefore the cause of the difference in the model's performance between these two validation years.

| Features | Types | PSI | Interpretation |
|---|---|---|---|
| Original Interest Rate | Numerical | 3.121 | Major shift-PSI > 0.25 |
| Loan Purpose | Categorical | 0.171 | Moderate shift |
| First Time Homebuyer Flag | Categorical | 0.045 | Minor shift |
| Original DTI | Numerical | 0.043 | Minor shift |
| Original LTV | Numerical | 0.039 | Minor shift |
| Credit Score | Numerical | 0.015 | Stable |
| Property State | Categorical | 0.010 | Stable |
| Original UPB | Numerical | 0.007 | Stable |
| Occupancy Status | Categorical | 0.002 | Stable |
| Number of Borrowers | Numerical | 0.001 | Stable |

Figure 1. Distribution shift between 2022 and 2023.

### 4.3. Precision, Recall, and Threshold Optimization Analysis

At a threshold of 0.10, the logistic regression models will have fewer false positive values than the ML models. The reason is that the probabilities predicted by the logistic regression models are well calibrated. Model 2 reported that 1011 loans were potentially in default and identified 150 defaults. However, the XGBoost model caught 727 defaults using the same threshold value across 25,961 loans. This performance difference would not be acceptable without adjusting the model's threshold.

Table 11 compares the 12 models that were run. The key findings in the analysis include the following:

Table 11. All 12 models using the same 30,000-row test set.

| Model | AUC | Accuracy | Precision | Recall | F1 | Brier | TN | FP | FN | TP | ECE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1: Full LR | 0.7895 | 94.46% | 0.1217 | 0.2055 | 0.152 | 0.022 | 28,187 | 1083 | 580 | 150 | 0.001 |
| 2: Sig LR | 0.7890 | 94.70% | 0.1292 | 0.2055 | 0.158 | 0.022 | 28,259 | 1011 | 580 | 150 | 0.001 |
| 3: LN Full | 0.7927 | 94.92% | 0.1283 | 0.1877 | 0.152 | 0.022 | 28,339 | 931 | 593 | 137 | 0.002 |
| 4: LN Sig | 0.7870 | 94.84% | 0.1256 | 0.1877 | 0.150 | 0.022 | 28,316 | 954 | 593 | 137 | 0.001 |
| 5: Engineered | 0.7956 | 94.73% | 0.1250 | 0.1945 | 0.152 | 0.022 | 28,276 | 994 | 588 | 142 | 0.001 |
| 6: Ridge | 0.7947 | 94.71% | 0.1225 | 0.1904 | 0.149 | 0.022 | 28,274 | 996 | 591 | 139 | 0.001 |
| 7: LASSO | 0.7946 | 94.73% | 0.1230 | 0.1904 | 0.149 | 0.022 | 28,279 | 991 | 591 | 139 | 0.001 |
| 8: Elastic Net | 0.7946 | 94.73% | 0.1230 | 0.1904 | 0.149 | 0.022 | 28,279 | 991 | 591 | 139 | 0.001 |
| 9: GAM | 0.7971 | 94.52% | 0.1235 | 0.2055 | 0.154 | 0.022 | 28,205 | 1065 | 580 | 150 | 0.001 |
| 10: XGBoost | 0.7967 | 13.45% | 0.0272 | 0.9959 | 0.053 | 0.185 | 3309 | 25,961 | 3 | 727 | 0.346 |
| 11: LightGBM | 0.7965 | 16.30% | 0.0281 | 0.9945 | 0.054 | 0.179 | 4164 | 25,106 | 4 | 726 | 0.335 |
| 12: CatBoost | 0.7988 | 17.80% | 0.0285 | 0.9918 | 0.055 | 0.183 | 4617 | 24,653 | 6 | 724 | 0.338 |

- Engineered LR (Model 5) has the lowest AIC and the highest ROC-AUC (0.7956) among LR models.
- GAM (Model 9) has the lowest ECE (0.3386), but its class-weighted probability outputs are not calibrated.
- Model 2 has the highest F1 score among all LR models (0.1586). All LR models are better calibrated than all gradient-boosting ML models.
- The ECE and Brier gap is approximately 20 - 500× and 8× (0.023 vs. 0.18), respectively. This means that if a bank uses CatBoost to calculate loss (EL = PD × LGD × EAD) using raw model probabilities, its loss estimation will be wildly inaccurate, which is why using post hoc calibration before any ML model is used is mandatory.

Figure 2 shows that Credit Score is by far the most important variable across all ML models, consistent with the logistic regression results. Original DTI, LTV, original UPB, and the year the loan was obtained are the next critical variables. The consistency we obtained in the variable importance rankings across all twelve models provides relatively strong evidence that these are the correct underlying economic drivers of mortgage default risk in this dataset.

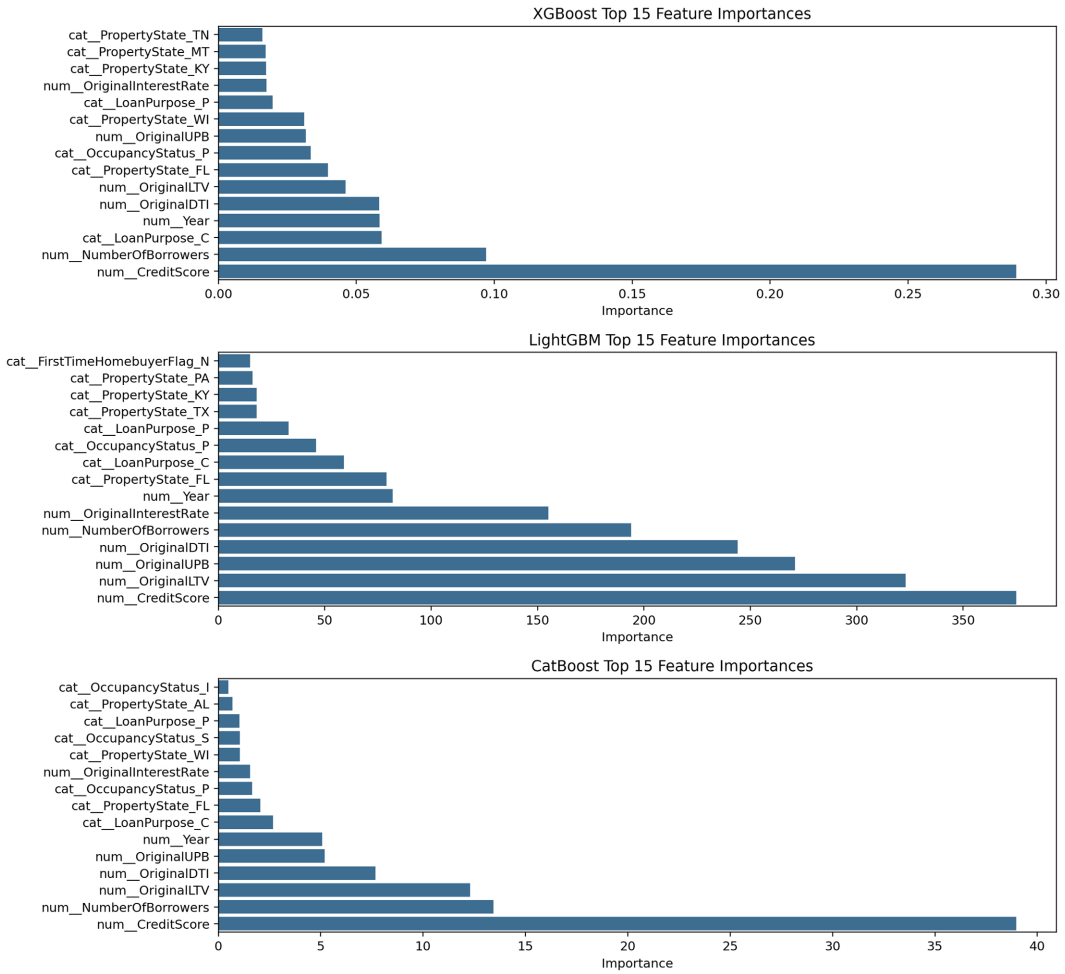


Figure 2. Machine models (model 10 - 12).

Figure 3 illustrates confusion matrices for XGBoost, LightGBM, and CatBoost at a threshold of 0.10. It shows that XGBoost detects 727 of 730 defaults (99.6% recall) but reports 25,961 false positives. LightGBM detects 726 defaults (99.5% recall), with 25,106 false positives. CatBoost detects 724 defaults (99.2% recall) and 24,653 false positives. These ML models are effective early-warning tools if the false-positive burden can be minimized; however, their operational threshold should be adjusted based on the institution's proportional cost of missed defaults versus false alarms.

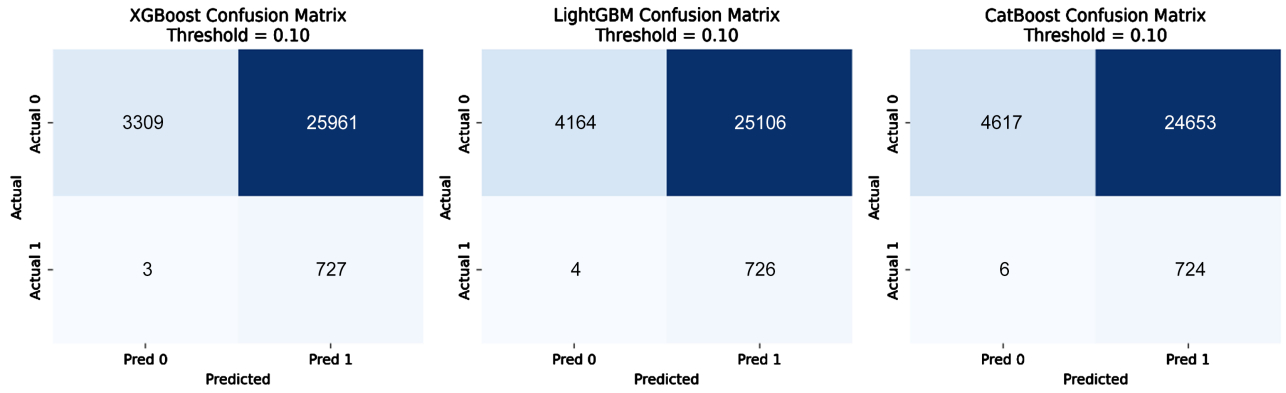


Figure 3. Confusion matrices for XGBoost, LightGBM, and CatBoost at threshold 0.10.

All the models performed similarly in risk ranking, with CatBoost leading, which is why it is the recommended ML model for risk ranking applications (Figure 4).

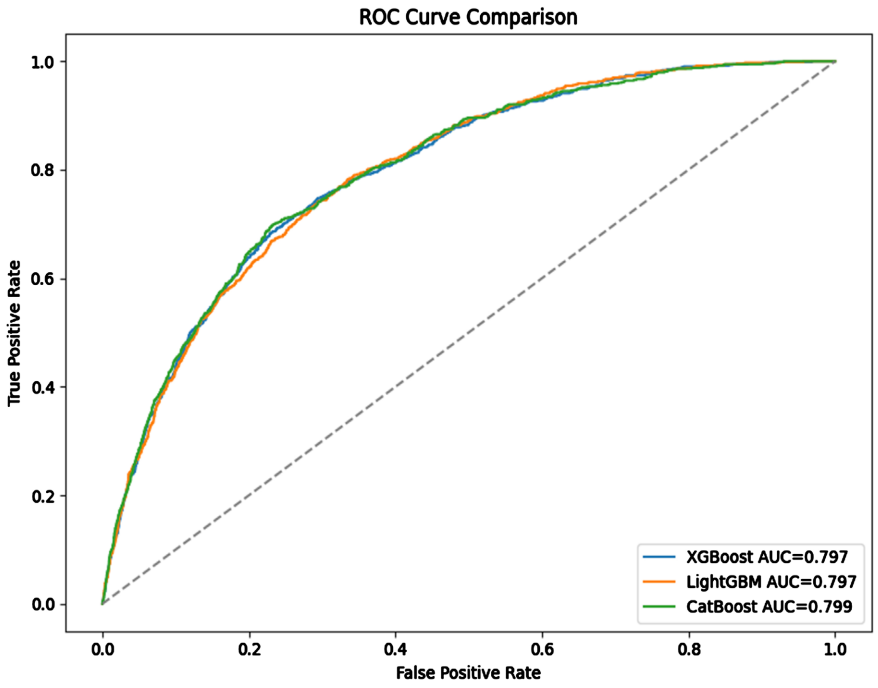


Figure 4. ROC curve comparison for XGBoost, LightGBM, and CatBoost.

## 5. Summary and Conclusions

A 30,000-row test set was evaluated on a 70/30 split with all twelve models. The findings in this research suggest that Credit Score is the strongest predictor of mortgage default rates across all 12 models tested. Number of borrowers, original LTV, original DTI, loan purpose, original interest, and the year the loan was obtained were the other key predictors.

The findings are similar and relatively consistent across all family models tested. Engineered Logistic Regression (Model 5) has the best AIC (604,799) and McFadden $R^2$ (0.1248) among the logistic regression family models. GAM (Model 9) has the best ROC-AUC (0.7971), best ECE (0.000703), and best Brier score (0.02283) compared with the rest of the logistic family models, showing that non-linear spline effects do capture additional structure beyond what linear specification can represent. Model 2, the significant logistic regression, having the highest F1 Score (0.1586) among traditional models, is the most interpretable and governance-friendly specification.

CatBoost (Model 12), among all the ML models, has the highest overall ROC-AUC (0.7988), confirming it as the best model for risk ranking and early warning surveillance. However, it is important to note that its raw probability output was not calibrated with its ECE (0.3386) and Brier (0.1835), and that it requires post hoc calibration before use in expected loss estimation.

**Key Conclusions**

1) All logistic regression models are better calibrated on raw probability outputs than all of the ML models. The ECE between these two models is within 200× to 500× of each other, and the Brier score is within approximately 8× of each other. This affects the regulatory requirements for estimating expected loss from these models.

2) The gap between the performance of logistic regression and ML models is much smaller than that often reported in the literature in the context of evaluating the models on the same data. The best of each model shows a difference in AUC of only 0.0017 between the logistic regression model (GAM, AUC = 0.7971) and the ML model (CatBoost, AUC = 0.7988).

3) Two models are recommended for deployment: CatBoost for early risk ranking and early warning surveillance, and GAM for expected loss estimation and regulatory reporting.

4) Elastic Net has a similar result to LASSO on the dataset, confirming that for simple mortgage credit risk data with a little multicollinearity, the additional Elastic Net modeling does not lead to an additional meaningful value.

5) The interest rate PSI between 2022 and 2023 of 3.121 indicates a huge shift in the distribution of interest rates. Each of the 12 models needs to be recalibrated annually or whenever the distribution of interest rates shifts.

## Limitations and Future Research

The study was conducted only on loans from 2022 to 2023, which reflects the post-

COVID-19 pandemic interest-rate era and cannot be generalized to stress periods or longer loan cycles. It is recommended that future studies include the following considerations:

- Extend the dataset to incorporate longitudinal interest rate cycles.
- Apply some survivorship analysis to the data, like the Cox proportional hazards model, to model time to default.
- Use post-hoc calibrations, like the Platt scaling with isotonic regression, and standardize the beta with calibrations for the ML models.
- Determine whether the GAM's strong calibration performance persists on out-of-sample vintages after 2023.

## Recommended Additional Machine Learning Methods

### Cost-Sensitive Random Forest or XGBoost

Instead of resampling, you can penalize mistakes on default loans more heavily. For weighted binary cross-entropy,

$$Loss = -\sum_{i=1}^{n}\left[ w_1 Y_i \log\left(\hat{p}_i\right) + w_0\left(1-Y_i\right)\log\left(1-\hat{p}_i\right)\right]$$

where $w_1 > w_0$ because default cases are rarer and more costly to miss.

### Neural Network/Multilayer Perceptron

A simple neural network can be written as

$$h_i = g\left(W_1 X_i + b_1\right)$$

$$\hat{p}_i = \sigma\left(W_2 h_i + b_2\right)$$

where $g(\cdot)$ is an activation function and $\sigma(\cdot)$ is the logistic sigmoid function. A deep learning comparison is also applied to various ML models, but it may not outperform gradient boosting on tabular mortgage data.

### Survival Analysis Model

This would be a major upgrade because mortgage default is naturally a time-to-event problem, not just a binary classification problem. Instead of asking whether the borrower defaults, survival analysis asks.

$$P\left(T_i \le t \mid X_i\right)$$

where $T_i$ is the time until default.

A Cox proportional hazards model is

$$h\left(t \mid X_i\right) = h_0\left(t\right)\exp\left(\beta^{\mathrm{T}} X_i\right)$$

where: $h\left(t \mid X_i\right)$ is the hazard of default at time $t$, and $h_0\left(t\right)$ is the baseline hazard. This is important because a loan that defaults after 3 months and one that defaults after 30 months are different, but binary classification treats them the same.

### Model Calibration Methods

Our initial finding is that random forest has poor calibration; hence, we added

three calibration correction methods.

**1) Platt scaling**

$$P(Y=1|\hat{s})=\frac{1}{1+e^{-(a\hat{s}+b)}}$$

where $\hat{s}$ is the raw model score.

**2) Isotonic regression**

$$\hat{p}_{calibrated}=f(\hat{p})$$

where $f$ is a nondecreasing function.

**3) Beta calibration**

$$\text{logit}(p)=a\log(\hat{p})-b\log(1-\hat{p})+c$$

Beta calibration can be useful when predicted probabilities are skewed. We were also able to compare the random forest results before and after calibration.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks, 6,* Article 38. https://doi.org/10.3390/risks6020038

Ayari, H., Guetari, P. R., & Kraïem, P. N. (2025). Machine Learning Powered Financial Credit Scoring: A Systematic Literature Review. *Artificial Intelligence Review, 59,* Article No. 13.

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, Auditability, and Explainability of Machine Learning Models in Credit Scoring. *Journal of the Operational Research Society, 73,* 70-90. https://doi.org/10.1080/01605682.2021.1922098

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks, 12,* Article 174. https://doi.org/10.3390/risks12110174

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects. *European Journal of Operational Research, 297,* 1178-1192. https://doi.org/10.1016/j.ejor.2021.06.053

Freddie Mac (2023). Single-Family Loan-Level Dataset. https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

Hossain, S., Sajal, A., Jamee, S. S., Tisha, S. A., Siddique, M. T., Obaid, M. O. et al. (2025). Comparative Analysis of Machine Learning Models for Credit Risk Prediction in Banking Systems. *The American Journal of Engineering and Technology, 7,* 22-33. https://doi.org/10.37547/tajet/volume07issue04-04

Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research, 247,* 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

Seitshiro, M. B., & Govender, S. (2024). Credit Risk Prediction with and without Weights of Evidence Using Quantitative Learning Models. *Cogent Economics & Finance, 12,* Ar-

ticle 2338971. https://doi.org/10.1080/23322039.2024.2338971

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine Learning-Driven Credit Risk: A Systemic Review. *Neural Computing and Applications, 34,* 14327-14339. https://doi.org/10.1007/s00521-022-07472-2