Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU37829

# *Regularization, Coefficient Stability, and Uncertainty Quantification in Logistic Regression: A Comparative Study of 30,000 Credit-Card Client Records*

**Yutian Cai**

*School of Mathematical Sciences, University of Nottingham, Nottingham, UK*
*hmyyc31@nottingham.ac.uk*

***Abstract.*** Credit-default studies often compare classifiers but give less attention to coefficient stability, categorical encoding, probability calibration, and subgroup uncertainty. This study analyzed 30,000 credit-card client records with 23 original predictors and a 22.12% default rate. Sex, education, and marital status were one-hot encoded, while continuous and repayment-status variables were standardized from the training data. Variance inflation factors were calculated for all 28 transformed predictors: only BILL_AMT1–BILL_AMT6 exceeded 10 (13.76–27.28), whereas every other predictor was below 5, locating the main multicollinearity problem in the monthly bill block. L2 logistic regression was then traced from C=0.0001 to C=100. Coefficients changed visibly after the cross-validated area under the receiver operating characteristic curve (AUC) had reached a broad plateau, separating parameter stability from predictive performance. On the same held-out test set, Random Forest achieved the strongest overall results (AUC=0.770, Brier=0.136, precision=0.653, recall=0.356, F1=0.461). Gaussian Naive Bayes had high recall (0.823) but low precision (0.280) and poor probability accuracy, whereas L2 logistic regression had the highest precision (0.692) but lower recall (0.244). Finally, Wald, Wilson, and Bayesian 95% intervals were compared for education and credit-limit groups; their differences were most visible in the 123-client education "others" group. The analysis shows why coefficient stability, discrimination, calibration, threshold-dependent classification, and subgroup uncertainty should be reported together.

***Keywords:*** logistic regression, one-hot encoding, L2 regularization, calibration, uncertainty quantification

## 1. Introduction

Credit scoring is used to assess the borrower's risk of default. Although statistical models cannot eliminate uncertainty, they can improve the consistency of evaluation and reduce dependence on manual judgment [1]. Logistic regression outputs probabilities and has interpretable coefficients. Machine learning models may be more accurate but less transparent [2, 3]. The two need to be balanced.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU37829

This study analyzed 30,000 credit card customer records in the UCI Machine Learning Repository [4, 5]. Previous studies have mostly focused on prediction accuracy [3], but this is not enough to fully reflect the reliability of the model. The correlation between predictors may cause the coefficient to be sensitive to regularization, even if the overall prediction results do not change much. Small subgroups raise a different problem: their observed default rates can fluctuate greatly.

Data preprocessing also affects interpretation, and recent credit-scoring work has emphasized transparent scorecards [6]. Variables such as gender, education and marital status are often recorded as numerical codes, and using these codes directly would impose an artificial ordering [7]. This study adopts one-hot encoding, tracks the coefficient path under different penalty strengths, and uses Random Forest importance as a nonlinear supplement. Coefficient paths and importance scores reflect observed associations, not causality.

L2 regularization can reduce the instability caused by multicollinearity by shrinking coefficients [8]. However, similar prediction results do not mean that the model coefficients are similar [9]. Model evaluation also needs to distinguish between the area under the receiver operating characteristic (ROC) curve (AUC), which measures ranking ability, and calibration, which compares predicted probabilities with observed frequencies [10, 11]. The default rate in this data set is only 22.12%. If all cases are classified as non-default, accuracy reaches 77.88% without separating the two classes, so accuracy should not be the only criterion [3].

This study aims to fill two gaps: coefficient change along the regularization path, and calibration and small-group uncertainty. Addressing them helps to distinguish stable predictions from unstable coefficients, and abnormal rates from sampling fluctuation.

The analysis proceeds in four steps. First, the logistic regression coefficient paths are traced as C (the inverse of regularization strength) increases from 0.0001 to 100. Second, cross-validated AUC and calibration are examined over the same range and interpreted alongside the coefficient changes. Third, logistic regression, Gaussian Naive Bayes and Random Forest are compared on the same test set, with Random Forest variable importance used as a supplementary reference for the logistic coefficients. Finally, Wald, Wilson and Bayesian intervals are estimated for the education groups and the credit-limit quintiles. Each step evaluates a different dimension of reliability.

## 2. Methodology

### 2.1. Data source and variables

The credit card default data set comes from a bank in Taiwan and contains the records of 30,000 clients [4, 5]. The records cover April to September 2005. Each record includes a client identifier (ID), 23 candidate predictors, and a binary indicator of whether the client defaulted in the following month. Predictors include credit limit, three demographic variables, six repayment statuses, six bill amounts and six payment amounts (Table 1). The ID is excluded because it carries no predictive information.

Table 1. Data dictionary

| Variable | Type | Description |
|---|---|---|
| ID | Identifier | Client identifier; excluded from analysis |
| LIMIT_BAL | Continuous | Amount of given credit in NT dollars |
| SEX | Categorical | 1 = male; 2 = female |
| EDUCATION | Categorical | 1 = graduate school; 2 = university; 3 = high school; 4 = others; 0/5/6 = undocumented |

DOI: 10.54254/2753-8818/2026.PU37829

Table 1. (continued)

| | | |
|---|---|---|
| MARRIAGE | Categorical | 1 = married; 2 = single; 3 = others; 0 = undocumented |
| AGE | Continuous | Age in years |
| PAY_0 | Ordinal | Repayment status in September 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| PAY_2 | Ordinal | Repayment status in August 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| PAY_3 | Ordinal | Repayment status in July 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| PAY_4 | Ordinal | Repayment status in June 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| PAY_5 | Ordinal | Repayment status in May 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| PAY_6 | Ordinal | Repayment status in April 2005; -2/-1/0 = no delay or paid duly; 1-8 = months overdue |
| BILL_AMT1 | Continuous | Bill statement amount in September 2005 (NT dollars) |
| BILL_AMT2 | Continuous | Bill statement amount in August 2005 (NT dollars) |
| BILL_AMT3 | Continuous | Bill statement amount in July 2005 (NT dollars) |
| BILL_AMT4 | Continuous | Bill statement amount in June 2005 (NT dollars) |
| BILL_AMT5 | Continuous | Bill statement amount in May 2005 (NT dollars) |
| BILL_AMT6 | Continuous | Bill statement amount in April 2005 (NT dollars) |
| PAY_AMT1 | Continuous | Previous payment amount in September 2005 (NT dollars) |
| PAY_AMT2 | Continuous | Previous payment amount in August 2005 (NT dollars) |
| PAY_AMT3 | Continuous | Previous payment amount in July 2005 (NT dollars) |
| PAY_AMT4 | Continuous | Previous payment amount in June 2005 (NT dollars) |
| PAY_AMT5 | Continuous | Previous payment amount in May 2005 (NT dollars) |
| PAY_AMT6 | Continuous | Previous payment amount in April 2005 (NT dollars) |
| default.payment.next.month | Binary outcome | 1 = default in the following month; 0 = no default |

## 2.2. Categorical encoding and descriptive grouping

SEX, EDUCATION and MARRIAGE are category labels, not equidistant variables. This study adopts one-hot encoding to avoid imposing a linear ordering [7]. Female, graduate education and married status were the reference categories, and each indicator was coded 0 or 1. After coding, 23 predictors are expanded to 28 columns. The descriptive education grouping is separate from the model encoding: code 4 is kept as “others”, and codes 0, 5 and 6 are combined into “unknown”. Keeping the smallest group separate makes it possible to examine interval uncertainty under different sample sizes.

## 2.3. Data splitting, standardization, and model fitting

The data are divided into training and test sets by an 80:20 stratified split (random seed 42). Continuous variables and repayment status variables are standardized by the mean and standard deviation of the training set. The one-hot encoded indicators are not standardized. Preprocessing is refitted in each cross-validation fold. This study does not use resampling or synthetic data generation.

DOI: 10.54254/2753-8818/2026.PU37829

First, an unpenalized logistic regression was fitted. Coefficients, standard errors, and 95% confidence intervals (CIs) were reported. L2-regularized models were then fitted over a logarithmic grid of C from 0.0001 to 100. The objective function is $L(\beta) + (1/2C)\|\beta\|^2$, where $L$ is the negative log-likelihood and the equivalent penalty weight is $\lambda = 1/C$, so a smaller C implies stronger shrinkage [8]. C was selected by five-fold cross-validation, and the coefficients at every grid point were stored to draw the shrinkage paths. Variance inflation factors (VIFs) were calculated to assess collinearity.

For client i, equation (1) gives the probability of default, and equation (2) gives its log-odds form. Under repeated sampling and correct model specification, about 95% of intervals constructed in this way would contain the true coefficient. A CI containing zero leaves the direction of the association unclear, but it does not show that the predictor is irrelevant.

$$p_i = \Pr(Y_i = 1 \mid x_i) = 1 / \{1 + \exp[-(\beta_0 + x_i^T\beta)]\} \tag{1}$$

$$log[p_i/(1 - p_i)] = \beta_0 + x_i^T\beta \tag{2}$$

where $p_i$ is the conditional probability of default for client i, $Y_i$ is the default indicator (1 = default in the following month), $x_i$ is the vector of transformed predictors, $\beta_0$ is the intercept, $\beta$ is the coefficient vector, and the superscript $T$ denotes the transpose.

The coefficient test is exploratory and not used for variable screening. The p-value is unadjusted, and it needs to be interpreted in combination with the CI, VIF and regularization path.

Gaussian Naive Bayes and Random Forest were fitted on the same split. Naive Bayes assumes conditional independence, and Random Forest can capture nonlinear effects and interactions [12]. The Random Forest uses 300 trees, min_samples_leaf=5, max_features='sqrt', and random seed 42. These settings were not extensively tuned, so the model serves as a practical benchmark.

### 2.4. Evaluation measures and interval estimation

This study uses AUC to assess the discriminative ability. The Brier score measures the mean squared error between the predicted probability and the actual result. A lower score indicates more accurate predictions [11]. The calibration curve compares the predicted risk and the observed default rate [10]. The training set default rate is used as a constant prediction and serves as the Brier score baseline. Precision, recall, and F1 score evaluate the classification performance at the threshold of 0.50.

The subgroup interval analysis uses all 30,000 records and is separate from the model fitting. Clients are grouped by education level and credit limit quintiles. Sample size, number of defaults, and default rate are reported for each group. Wald, Wilson, and Beta(1,1) Bayesian 95% intervals are calculated. The three methods yield consistent results in large sample groups. They can diverge when samples are small or proportions are close to 0 or 1 [13, 14]. For the smallest education group, a Jeffreys Beta(0.5,0.5) interval is also reported to examine prior sensitivity.

## 3. Results

### 3.1. Data profile and exploratory analysis

No missing values or duplicate records were found, so neither imputation nor row removal was required. Table 2 summarizes the retained variables.

DOI: 10.54254/2753-8818/2026.PU37829

Table 2. Descriptive statistics of the 30,000 client records

| Variable | Mean | SD | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|
| LIMIT_BAL | 167,484.32 | 129,747.66 | 10,000 | 50,000 | 140,000 | 240,000 | 1,000,000 |
| SEX | 1.60 | 0.49 | 1 | 1 | 2 | 2 | 2 |
| EDUCATION | 1.85 | 0.79 | 0 | 1 | 2 | 2 | 6 |
| MARRIAGE | 1.55 | 0.52 | 0 | 1 | 2 | 2 | 3 |
| AGE | 35.49 | 9.22 | 21 | 28 | 34 | 41 | 79 |
| PAY_0 | -0.02 | 1.12 | -2 | -1 | 0 | 0 | 8 |
| PAY_2 | -0.13 | 1.20 | -2 | -1 | 0 | 0 | 8 |
| PAY_3 | -0.17 | 1.20 | -2 | -1 | 0 | 0 | 8 |
| PAY_4 | -0.22 | 1.17 | -2 | -1 | 0 | 0 | 8 |
| PAY_5 | -0.27 | 1.13 | -2 | -1 | 0 | 0 | 8 |
| PAY_6 | -0.29 | 1.15 | -2 | -1 | 0 | 0 | 8 |
| BILL_AMT1 | 51,223.33 | 73,635.86 | -165,580 | 3,558.75 | 22,381.5 | 67,091 | 964,511 |
| BILL_AMT2 | 49,179.08 | 71,173.77 | -69,777 | 2,984.75 | 21,200 | 64,006.25 | 983,931 |
| BILL_AMT3 | 47,013.15 | 69,349.39 | -157,264 | 2,666.25 | 20,088.5 | 60,164.75 | 1,664,089 |
| BILL_AMT4 | 43,262.95 | 64,332.86 | -170,000 | 2,326.75 | 19,052 | 54,506 | 891,586 |
| BILL_AMT5 | 40,311.40 | 60,797.16 | -81,334 | 1,763 | 18,104.5 | 50,190.5 | 927,171 |
| BILL_AMT6 | 38,871.76 | 59,554.11 | -339,603 | 1,256 | 17,071 | 49,198.25 | 961,664 |
| PAY_AMT1 | 5,663.58 | 16,563.28 | 0 | 1,000 | 2,100 | 5,006 | 873,552 |
| PAY_AMT2 | 5,921.16 | 23,040.87 | 0 | 833 | 2,009 | 5,000 | 1,684,259 |
| PAY_AMT3 | 5,225.68 | 17,606.96 | 0 | 390 | 1,800 | 4,505 | 896,040 |
| PAY_AMT4 | 4,826.08 | 15,666.16 | 0 | 296 | 1,500 | 4,013.25 | 621,000 |
| PAY_AMT5 | 4,799.39 | 15,278.31 | 0 | 252.5 | 1,500 | 4,031.5 | 426,529 |
| PAY_AMT6 | 5,215.50 | 17,777.47 | 0 | 117.75 | 1,500 | 4,000 | 528,666 |
| default.payment.next.month | 0.22 | 0.42 | 0 | 0 | 0 | 0 | 1 |

The sample size is 30,000, with mean and median credit limits of NT$167,484 and NT$140,000. The average age is 35.49 and the default rate is 22.12% (6,636 cases). The summary values for SEX, EDUCATION and MARRIAGE describe category codes rather than quantities. Bill and payment amounts are right-skewed, with maxima far above the upper quartiles, and standardization does not remove this skewness.

Figure 1 shows strong Pearson correlations among BILL_AMT1–BILL_AMT6: 0.951 for BILL_AMT1/BILL_AMT2, 0.946 for BILL_AMT5/BILL_AMT6 and 0.803 even for BILL_AMT1/BILL_AMT6. This substantial information overlap, alongside weaker positive correlations among repayment statuses, motivated the VIF and regularization analyses.

DOI: 10.54254/2753-8818/2026.PU37829

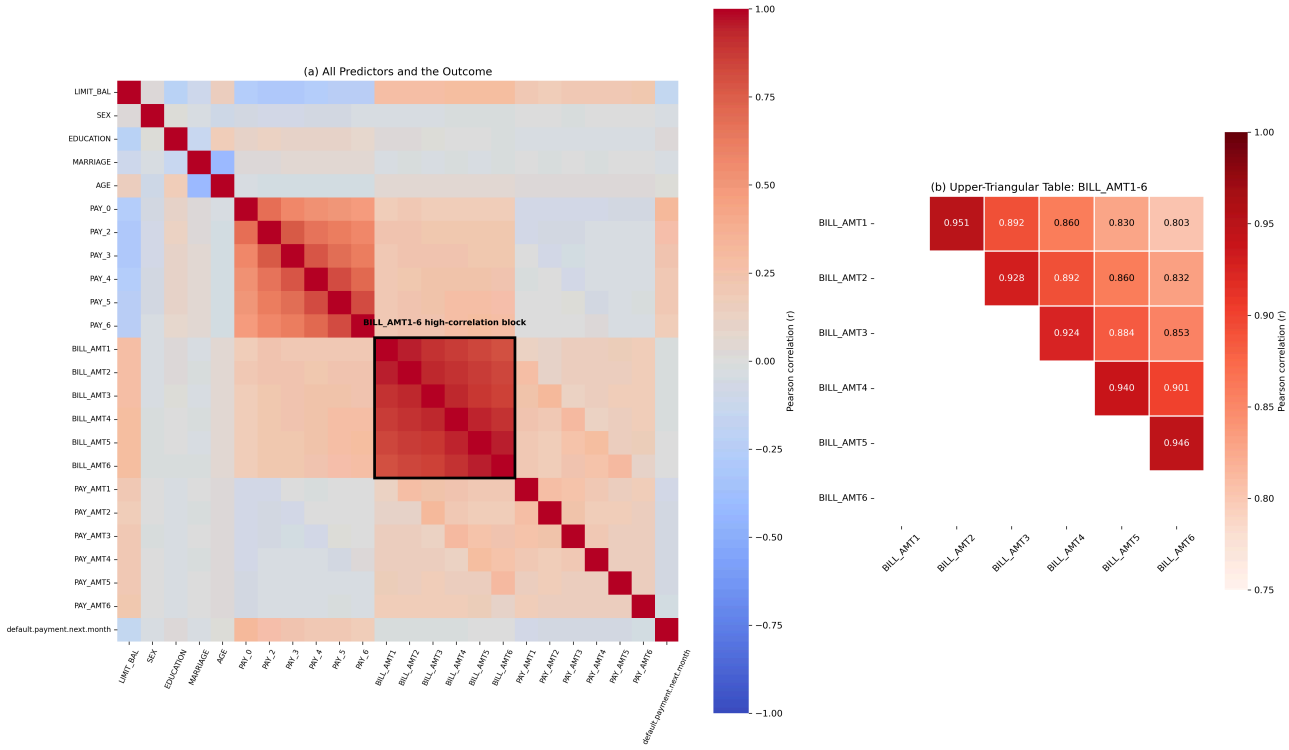



Figure 1. Correlation structure and the high-correlation bill-amount block (data from: [5])

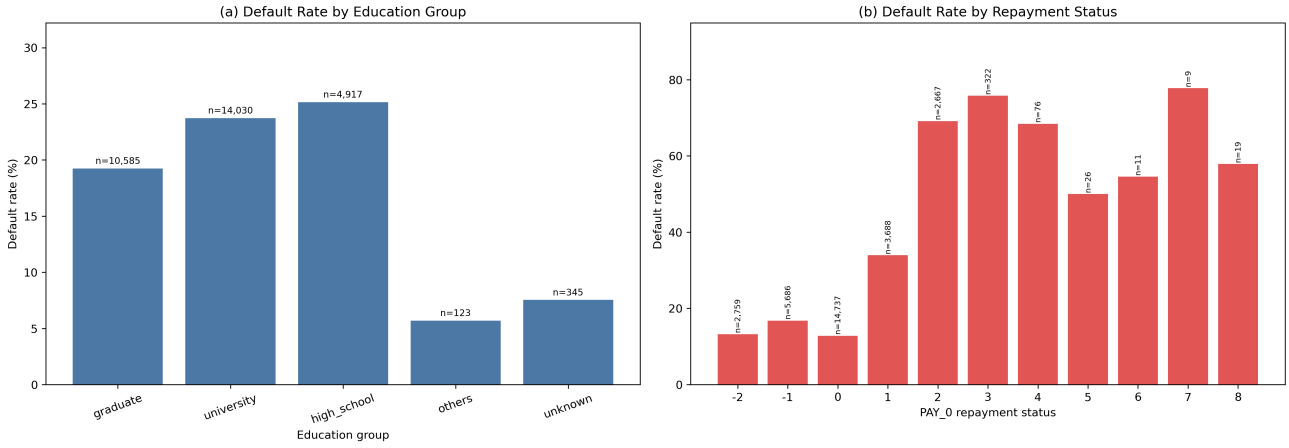



Figure 2. Default rates by education and initial repayment status (data from: [5])

In Figure 2a, the high-school group has the highest default rate (25.16%; 1,237/4,917), followed by the university (23.73%; 3,330/14,030) and graduate (19.23%; 2,036/10,585) groups. The “others” and “unknown” groups show rates of 5.69% (7/123) and 7.54% (26/345). Figure 2b shows a clear upward trend in default rates with longer overdue periods, but small sample sizes in the most overdue categories warrant caution in interpretation.

## 3.2. Baseline logistic regression

Baseline logistic regression was fitted to the preprocessed training set. Numerical coefficients represent associations per one-standard-deviation increase; indicator coefficients compare each

DOI: 10.54254/2753-8818/2026.PU37829

category with its reference group.

Table 3. Baseline logistic-regression coefficients after one-hot encoding

| Predictor | Coefficient | Standard error | 95% CI lower | 95% CI upper | p-value |
|---|---|---|---|---|---|
| const | -1.3294 | 0.0409 | -1.4096 | -1.2492 | <0.0001 |
| LIMIT_BAL | -0.1259 | 0.0231 | -0.1712 | -0.0806 | <0.0001 |
| AGE | 0.0551 | 0.0192 | 0.0175 | 0.0928 | 0.0041 |
| PAY_0 | 0.6607 | 0.0223 | 0.6170 | 0.7043 | 0.0000 |
| PAY_2 | 0.0922 | 0.0270 | 0.0393 | 0.1450 | 0.0006 |
| PAY_3 | 0.0955 | 0.0300 | 0.0366 | 0.1544 | 0.0015 |
| PAY_4 | 0.0601 | 0.0323 | -0.0032 | 0.1235 | 0.0627 |
| PAY_5 | -0.0006 | 0.0339 | -0.0670 | 0.0658 | 0.9851 |
| PAY_6 | 0.0154 | 0.0282 | -0.0399 | 0.0708 | 0.5851 |
| BILL_AMT1 | -0.3231 | 0.0891 | -0.4978 | -0.1484 | 0.0003 |
| BILL_AMT2 | 0.1054 | 0.1178 | -0.1256 | 0.3363 | 0.3712 |
| BILL_AMT3 | 0.0330 | 0.1057 | -0.1741 | 0.2401 | 0.7551 |
| BILL_AMT4 | 0.0505 | 0.0965 | -0.1387 | 0.2396 | 0.6009 |
| BILL_AMT5 | 0.0437 | 0.0994 | -0.1512 | 0.2386 | 0.6602 |
| BILL_AMT6 | -0.0205 | 0.0776 | -0.1726 | 0.1316 | 0.7917 |
| PAY_AMT1 | -0.1999 | 0.0409 | -0.2802 | -0.1197 | <0.0001 |
| PAY_AMT2 | -0.1786 | 0.0465 | -0.2698 | -0.0874 | 0.0001 |
| PAY_AMT3 | -0.0335 | 0.0327 | -0.0975 | 0.0305 | 0.3047 |
| PAY_AMT4 | -0.0385 | 0.0290 | -0.0954 | 0.0183 | 0.1841 |
| PAY_AMT5 | -0.0185 | 0.0286 | -0.0745 | 0.0375 | 0.5176 |
| PAY_AMT6 | -0.0293 | 0.0255 | -0.0792 | 0.0206 | 0.2503 |
| SEX_male | 0.1194 | 0.0344 | 0.0520 | 0.1869 | 0.0005 |
| EDUCATION_high_school | -0.1599 | 0.0536 | -0.2650 | -0.0548 | 0.0029 |
| EDUCATION_others | -1.0377 | 0.4289 | -1.8784 | -0.1970 | 0.0155 |
| EDUCATION_university | -0.0987 | 0.0398 | -0.1768 | -0.0206 | 0.0132 |
| EDUCATION_unknown | -1.2156 | 0.2328 | -1.6719 | -0.7593 | 0.0000 |
| MARRIAGE_others | 0.0196 | 0.1514 | -0.2771 | 0.3163 | 0.8968 |
| MARRIAGE_single | -0.1794 | 0.0389 | -0.2556 | -0.1032 | <0.0001 |
| MARRIAGE_unknown | -1.1927 | 0.5242 | -2.2201 | -0.1653 | 0.0229 |

Table 3 reports coefficients, standard errors, 95% confidence intervals and p-values. Among standardized numerical predictors, PAY_0 had the largest positive coefficient, $\beta = 0.661$ (95% CI 0.617–0.704), and BILL_AMT1 the most negative, $\beta = -0.323$ (95% CI -0.498 to -0.148). PAY_AMT1, PAY_AMT2 and LIMIT_BAL were also negative (-0.200, -0.179 and -0.126). The male coefficient relative to females was 0.119 (95% CI 0.052–0.187). Relative to graduate education, high school, others, university and unknown had coefficients of -0.160, -1.038, -0.099 and -1.216. These are adjusted associations, not causal effects.

DOI: 10.54254/2753-8818/2026.PU37829

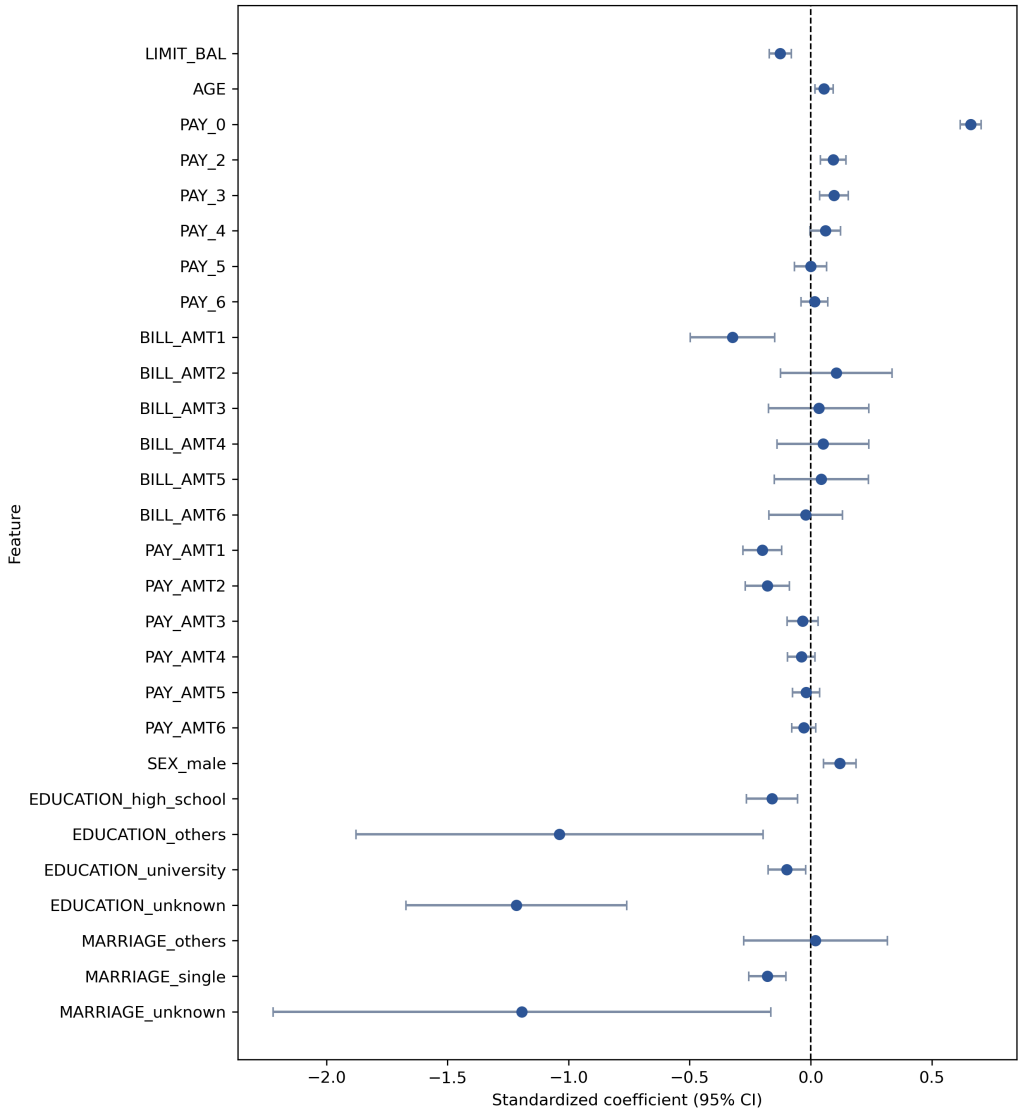



Figure 3. Baseline logistic-regression coefficients after one-hot encoding (data from: [5])

Figure 3 visualizes the uncertainty in Table 3. Intervals included zero for PAY_4–PAY_6, BILL_AMT2–BILL_AMT6, PAY_AMT3–PAY_AMT6 and MARRIAGE_others. Bill-amount intervals were generally wide (e.g., BILL_AMT2: -0.126 to 0.336), as were those for the small EDUCATION_others, EDUCATION_unknown and MARRIAGE_unknown categories. This reflects uncertainty in individual adjusted coefficients, not evidence that billing history or rare categories are unimportant.

Table 4 locates the strongest multicollinearity in the monthly bill block: only BILL_AMT1–BILL_AMT6 had VIFs above 10 (13.76–27.28). All other predictors were below 5, with PAY_5 highest at 4.67. Individual bill coefficients therefore require caution and examination across L2 penalties.

Table 4. Variance inflation factors for all transformed predictors

| Predictor | VIF |
|---|---|
| LIMIT_BAL | 1.52 |

DOI: 10.54254/2753-8818/2026.PU37829

Table 4. (continued)

| | |
|---|---|
| AGE | 1.36 |
| PAY_0 | 1.90 |
| PAY_2 | 3.15 |
| PAY_3 | 3.63 |
| PAY_4 | 4.23 |
| PAY_5 | 4.67 |
| PAY_6 | 3.21 |
| BILL_AMT1 | 13.76 |
| BILL_AMT2 | 27.28 |
| BILL_AMT3 | 24.74 |
| BILL_AMT4 | 20.74 |
| BILL_AMT5 | 24.20 |
| BILL_AMT6 | 14.86 |
| PAY_AMT1 | 1.69 |
| PAY_AMT2 | 2.15 |
| PAY_AMT3 | 1.80 |
| PAY_AMT4 | 1.60 |
| PAY_AMT5 | 1.66 |
| PAY_AMT6 | 1.18 |
| SEX_male | 1.51 |
| EDUCATION_high_school | 1.28 |
| EDUCATION_others | 1.01 |
| EDUCATION_university | 1.47 |
| EDUCATION_unknown | 1.02 |
| MARRIAGE_others | 1.03 |
| MARRIAGE_single | 1.89 |
| MARRIAGE_unknown | 1.01 |

### 3.3. Coefficient shrinkage and regularization strength

Figure 4 shows the trajectory of 28 coefficients with the C value (0.0001–100). The 8 coefficients with the largest absolute value are highlighted at C=100, and the rest are gray. PAY_0 is always the maximum positive coefficient. The coefficients of EDUCATION_others, EDUCATION_unknown and MARRIAGE_unknown change markedly as regularization weakens. Their larger coefficients are related to small samples and broad confidence intervals. The coefficients of the bill-related variables diverge as C increases, indicating that the size of the coefficient is not equivalent to stability.

DOI: 10.54254/2753-8818/2026.PU37829

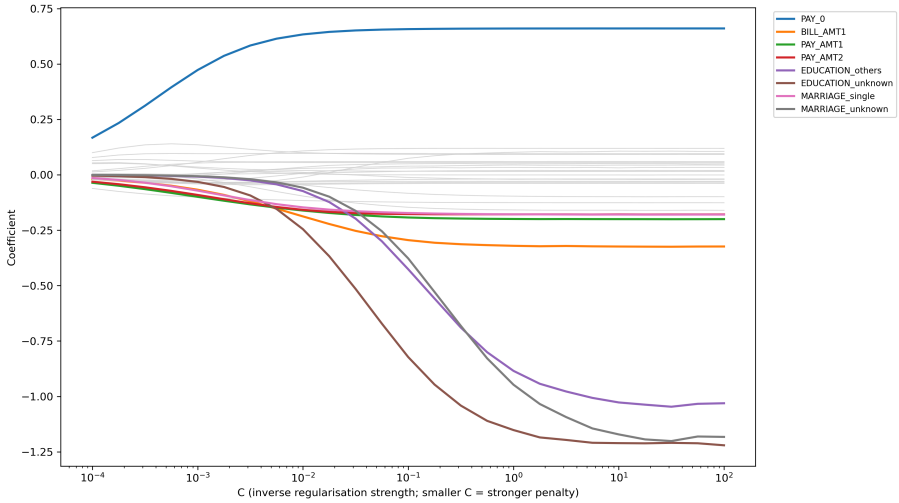



Figure 4. L2 coefficient shrinkage paths after one-hot encoding (data from: [5])

Strong regularization (C<0.01) shrinks most coefficients toward zero. As C increases, some bill-amount and rare-category coefficients keep changing even after AUC reaches a plateau. Therefore, C=0.01 marks a point on the performance plateau, not a general threshold for coefficient stability.

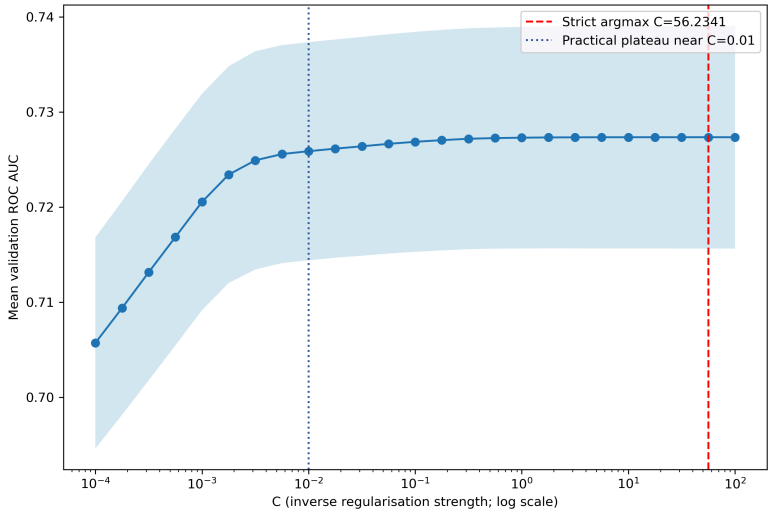



Figure 5. Training-set cross-validated AUC across regularization strengths (data from: [5])

Discriminative ability changes much less than the coefficients do (Figure 5). The mean cross-validated AUC rises from 0.70572 at C=0.0001 to 0.72588 at C=0.01 and then stabilizes. The grid-search optimum, C=56.23 (AUC=0.72734825), is only 0.00147183 higher than the value at C=0.01. The increase is smaller than the variation between the folds, indicating that the performance improvement is limited. Most of the discriminative ability is already achieved around C=0.01. The one-standard-error rule would favor a smaller C, but the strict grid-search result is retained for reproducibility.

DOI: 10.54254/2753-8818/2026.PU37829

### 3.4. Calibration and ROC performance

Compared with C=0.01 and C=56.23, the calibration curve at C=0.0001 has a greater deviation from the diagonal (Figure 6). Although the coefficients have changed, the curves of the latter two are closer. The AUC of the selected model on the test set is 0.710 (Figure 7), which is lower than the mean of cross-validation. This difference may stem from the process of data division and model selection, because the data used in each estimate are not the same.

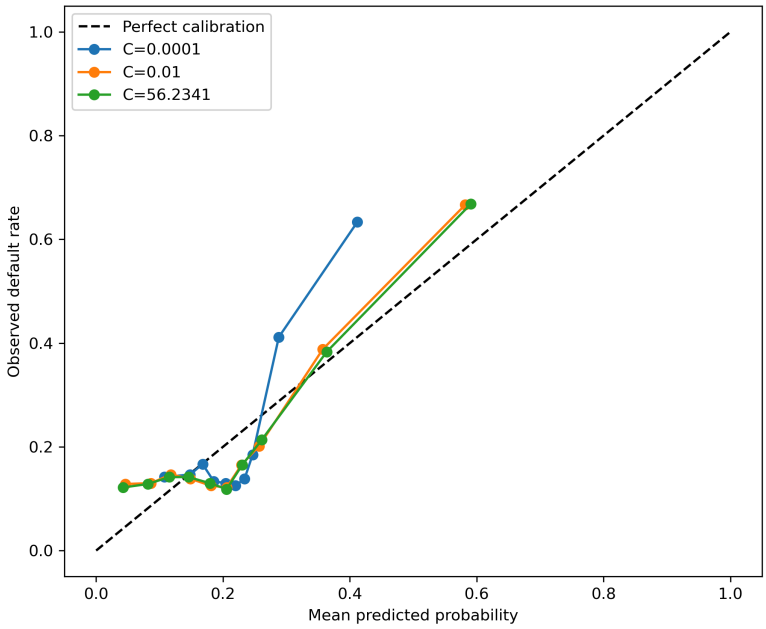



Figure 6. L2 logistic calibration across selected C values (data from: [5])

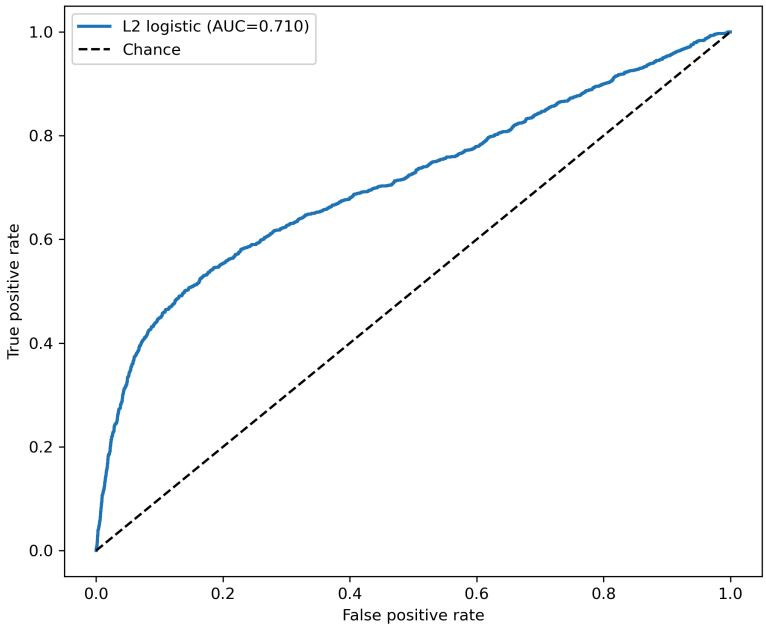



Figure 7. ROC curve for the selected one-hot L2 logistic model (data from: [5])

DOI: 10.54254/2753-8818/2026.PU37829

### 3.5. Comparison of three predictive models

Table 5 compares the three fitted models and the constant-probability benchmark on the same test set, using AUC, Brier score and classification metrics at the 0.50 threshold.

Table 5. Test-set discrimination, probability accuracy, and threshold-dependent classification metrics for the three fitted models and null benchmark

| Model | AUC | Brier score | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Random Forest | 0.770 | 0.136 | 0.653 | 0.356 | 0.461 |
| Gaussian Naive Bayes | 0.723 | 0.345 | 0.280 | 0.823 | 0.418 |
| L2 Logistic Regression | 0.710 | 0.147 | 0.692 | 0.244 | 0.361 |
| Null model | 0.500 | 0.172 | n/a | 0.000 | 0.000 |

Random Forest performs best in AUC (0.770), Brier score (0.136), and F1 score (0.461). Gaussian Naive Bayes has the highest recall (0.823) but lower precision (0.280), and its Brier score (0.345) is worse than that of the null model (0.172). Logistic regression has the highest precision (0.692) but a recall of only 0.244. The null model predicts all samples as “non-default” because its predicted probability (0.221) is below the 0.50 threshold. Its recall and F1 score are therefore 0, and its precision is undefined and reported as n/a. The precision, recall and F1 rankings depend on this threshold.

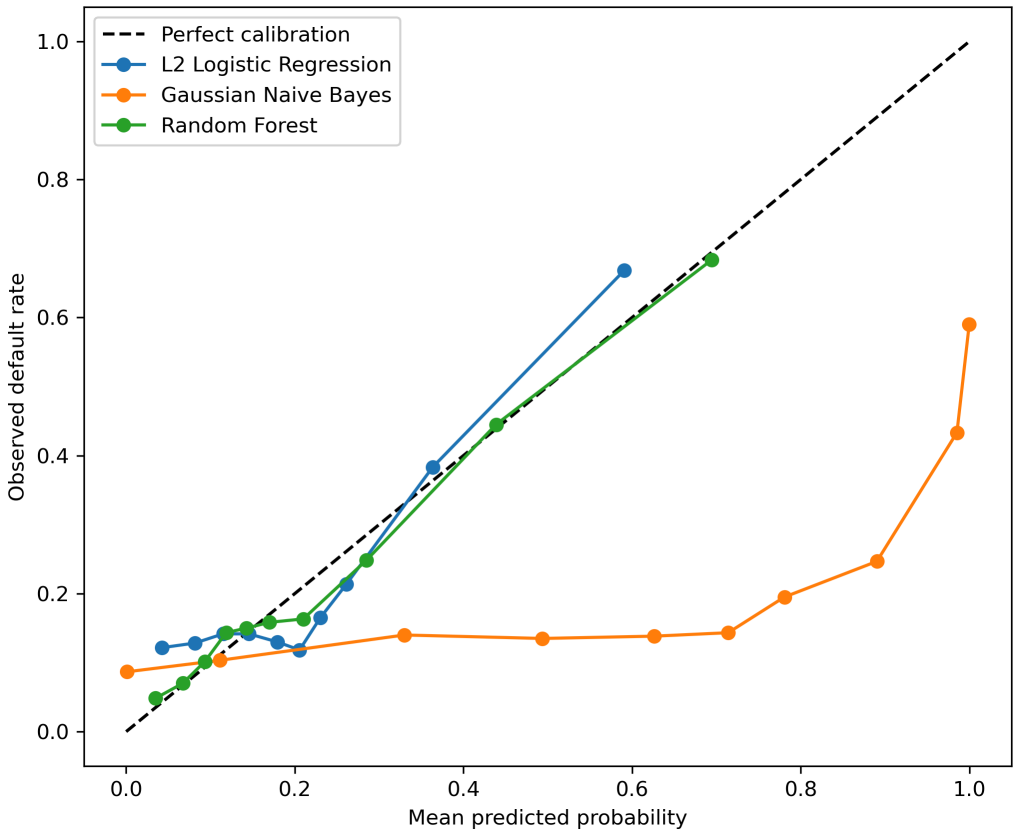



Figure 8. Calibration curves for the three predictive models (data from: [5])

The Random Forest calibration curve is closest to the diagonal across most risk intervals (Figure 8). Logistic regression deviates more in the high-risk range, while Gaussian Naive Bayes is clearly

DOI: 10.54254/2753-8818/2026.PU37829

miscalibrated. These patterns are consistent with the Brier score ranking. If Gaussian Naive Bayes is used for probability prediction, it should be recalibrated, or its use should be limited to ranking tasks.

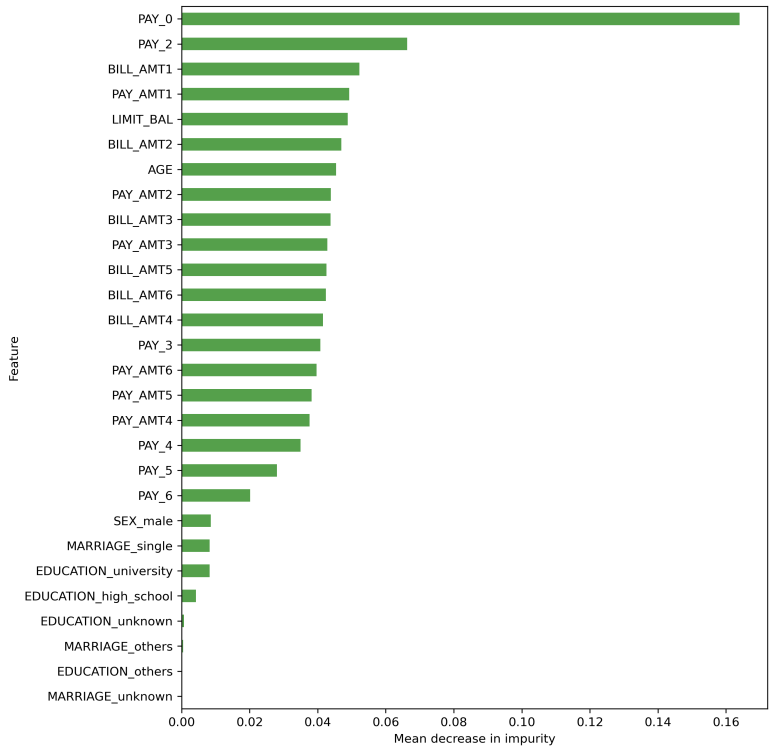



Figure 9. Random Forest feature importance after one-hot encoding (data from: [5])

Random Forest importance ranks PAY_0 first (0.164) and PAY_2 second (0.066), followed by several bill, payment, credit-limit and age predictors (Figure 9). Both models identify recent repayment status as the main predictor. Signed logistic regression coefficients are not directly comparable to non-negative Random Forest importance values. In addition, correlated predictors may share importance, and Random Forest can capture nonlinear effects and interactions between variables [12].

## 3.6. Uncertainty in grouped default rates

Table 6. Grouped default rates with Wald, Wilson, and Bayesian 95% intervals, Bayesian limits from a Beta(1,1) prior except in the sensitivity row

| Group | n | Defaults | Rate (%) | Wald 95% CI (%) | Wilson 95% CI (%) | Bayesian 95% CI (%) |
|---|---|---|---|---|---|---|
| Education: graduate | 10,585 | 2,036 | 19.23 | 18.48–19.99 | 18.50–20.00 | 18.50–20.00 |
| Education: university | 14,030 | 3,330 | 23.73 | 23.03–24.44 | 23.04–24.45 | 23.04–24.45 |
| Education: high school | 4,917 | 1,237 | 25.16 | 23.94–26.37 | 23.96–26.39 | 23.96–26.39 |
| Education: others | 123 | 7 | 5.69 | 1.60–9.79 | 2.78–11.28 | 2.83–11.29 |
| Education: unknown | 345 | 26 | 7.54 | 4.75–10.32 | 5.19–10.81 | 5.21–10.82 |
| Credit limit: Q1 | 7,676 | 2,440 | 31.79 | 30.75–32.83 | 30.76–32.84 | 30.76–32.84 |
| Credit limit: Q2 | 4,822 | 1,244 | 25.80 | 24.56–27.03 | 24.58–27.05 | 24.58–27.05 |
| Credit limit: Q3 | 6,123 | 1,216 | 19.86 | 18.86–20.86 | 18.88–20.88 | 18.88–20.88 |

DOI: 10.54254/2753-8818/2026.PU37829

Table 6. (continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Credit limit: Q4 | 5,421 | 914 | 16.86 | 15.86–17.86 | 15.89–17.88 | 15.89–17.88 |
| Credit limit: Q5 | 5,958 | 822 | 13.80 | 12.92–14.67 | 12.94–14.70 | 12.94–14.70 |
| Sensitivity: others, Jeffreys prior | 123 | 7 | 5.69 | n/a | n/a | 2.58–10.85 |

In the three large education groups, the three intervals almost coincide (Table 6). For the graduate group, the Wald interval is 18.48%–19.99%, and the Wilson and Beta(1,1) intervals are both 18.50%–20.00%. The intervals for the "others" group (123 clients, 7 defaults) are 1.60%–9.79% (Wald), 2.78%–11.28% (Wilson) and 2.83%–11.29% (Bayesian) respectively. The Wald interval has the lowest lower limit, consistent with its known coverage problems in small samples [13, 14]. Wilson and Jeffreys intervals are more reliable alternatives [13, 15].

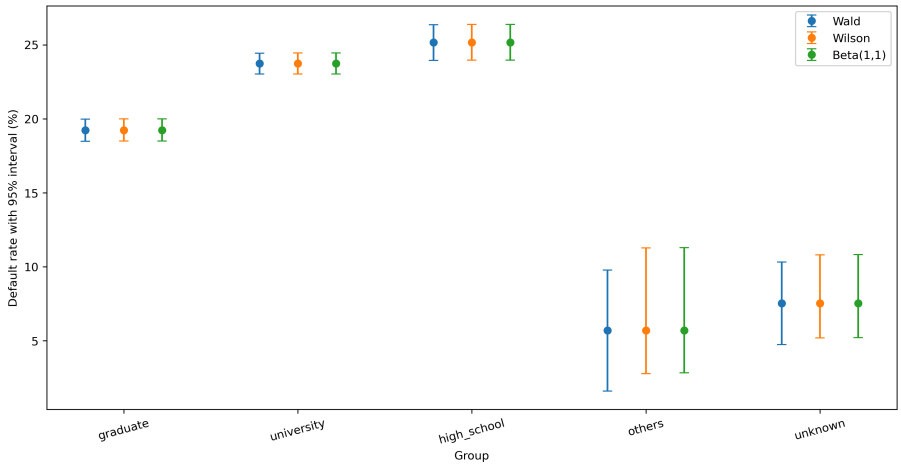



Figure 10. Default-rate intervals by education group (data from: [5])

Figure 10 shows near-complete overlap among methods for the three large education groups, but wider intervals for "others" (n=123) and "unknown" (n=345). Combining these groups would obscure the smallest category and the clearest contrast between interval methods.

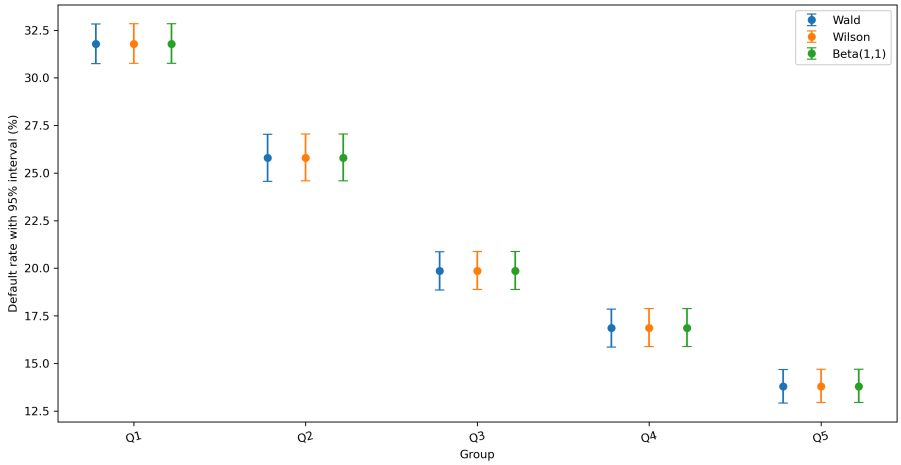



Figure 11. Default-rate intervals by credit-limit quintile (data from: [5])

The default rate decreases steadily from 31.79% in the lowest credit-limit quintile (Q1) to 13.80% in the highest quintile (Q5). Because all five quintile groups are large, the Wald, Wilson, and

DOI: 10.54254/2753-8818/2026.PU37829

Bayesian intervals almost completely overlap (Figure 11). For example, the Wilson interval is 30.76%–32.84% for Q1 and 12.94%–14.70% for Q5.

When using the Jeffreys Beta(0.5,0.5) prior, the interval for the "others" group becomes 2.58%–10.85%. This variation (0.25 and 0.44 percentage points) is smaller than the difference between the lower limits of the Wald and Wilson intervals. This indicates that the limited sample size has a more substantial effect than the choice of weak priors.

## 4. Discussion and limitations

### 4.1. Discussion

The coefficient path and performance curve reflect different aspects of stability. The VIFs of BILL_AMT1 to BILL_AMT6 exceed 10, indicating substantial overlap among these variables. This helps explain why their coefficients fluctuate or even change sign as the penalty weakens. PAY_0 remains positive and stable under most C values. Although the coefficients continue to change, the AUC at C=0.01 is only 0.00147 lower than at the selected value, indicating that discriminative ability stabilizes before individual coefficients do.

Random Forest has the best overall test-set performance. Gaussian Naive Bayes gains recall at the expense of precision and probability calibration. Its Brier score is even worse than that of the null model. Although the bill variables are highly correlated, this alone does not prove a violation of the conditional independence assumption within default classes. Both Random Forest and logistic regression support the predictive relevance of PAY_0, but neither establishes causality.

For groups with larger sample sizes in education and credit limit, the results of the three intervals are close. The education "others" group (123 clients, 7 defaults) has wide intervals, the choice of method matters more, and the Wald lower limit is noticeably lower. Changing the prior from Beta(1,1) to Jeffreys Beta(0.5,0.5) has little impact. Therefore, the uncertainty mainly reflects the limited information in this subgroup.

The education-group results do not constitute a fairness audit. Estimates for the smallest education groups are imprecise. A comprehensive audit would require group-specific error measures and actual approval decisions, which this study does not include. This analysis shows how estimation uncertainty changes with subgroup size.

### 4.2. Limitations

These records come from a bank in Taiwan (2005). Due to the differences in supervision, rules and customer composition, the applicability of results is restricted. This study is only a case study, not a generalizable model.

One-hot encoding avoids misinterpreting categorical codes, but the coefficients depend on the reference groups (female, graduate school, married). Small categories (education "others", marriage "unknown") have wide confidence intervals and are sensitive to weak regularization. The coefficients reflect conditional associations, not causal effects.

The analysis uses a single 80:20 split, with preprocessing refitted within each cross-validation fold. Future work could use nested cross-validation or temporal validation [9]. Impurity-based Random Forest importance can be spread across correlated predictors, and permutation importance or Shapley values could test the stability of the ranking [6].

Only L2 regularization is used in this study. L1 or elastic-net penalties could be used to examine sparsity. The 0.50 threshold does not account for misclassification costs or credit policy. Future work

DOI: 10.54254/2753-8818/2026.PU37829

should specify these costs, use decision curves, and select the threshold on validation data.

The baseline p-values are exploratory only; they are not adjusted for multiple testing and are not used for variable selection. Some small p-values may be due to chance. The p-values should be interpreted together with the CIs, VIFs and coefficient paths rather than through significance thresholds alone.

The education and credit-limit group comparisons are not adjusted for other covariates. The differences between groups are therefore descriptive only. The Bayesian sensitivity check considers only two weak priors. Future work could use a hierarchical binomial model to partially pool small groups while reporting the uncertainty of each subgroup.

## 5. Conclusion

Credit default analysis should not rely solely on AUC. Feature coding, multicollinearity, coefficient stability, calibration, classification performance and subgroup uncertainty reflect different aspects of model reliability. A comprehensive examination of these factors helps distinguish robust signals from coefficients affected by correlated predictors. It also clarifies differences between ranking ability, probability accuracy, and practicality.

The study examined credit-default modeling through four connected views. First, one-hot encoding removed an unjustified numeric ordering from sex, education, and marital status, and the complete L2 path showed which transformed logistic coefficients persisted as the penalty changed. PAY_0 remained the dominant positive signal, whereas the correlated bill coefficients and several rare-category contrasts were less stable. Second, mean cross-validated AUC had already reached 0.72588 at C=0.01 and increased by only 0.00147 at the grid-search optimum C=56.23, while calibration changed little once the penalty was moderate. Coefficient stability and predictive quality were therefore related but not interchangeable.

Third, Random Forest had the best AUC, Brier score, and F1 score in the three-model comparison. Gaussian Naive Bayes ranked cases slightly better than L2 logistic regression but produced substantially worse probabilities, and Random Forest and logistic regression agreed that recent repayment status was central. Fourth, the Wald, Wilson, and Beta(1,1) limits were almost identical in the three large education groups and in every credit-limit quintile. They separated for the 123-client “others” group, where the Wald interval was both lower and narrower than the other two. Replacing the uniform prior with the Jeffreys prior moved the Bayesian limits by less than the gap between the Wald and Wilson lower limits, so the main source of uncertainty in that group was its size rather than the choice between two weak priors.

The combination of explainable regularization, nonlinear benchmark, null model and interval estimation can provide a basis for model validation. It can also support threshold selection and review. Although the framework does not directly determine the credit policy, it can ensure that the predictive performance and uncertainty remain visible throughout the decision-making process.

## References

[1] Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science, 8*, 180–201. https://doi.org/10.1016/j.jfds.2022.07.002

[2] Ayari, H., Guetari, R., & Kraïem, N. (2025). Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review, 59*(1), 13. https://doi.org/10.1007/s10462-025-11416-2

[3] Bhandary, R., & Ghosh, B. K. (2025). Credit card default prediction: An empirical analysis on predictive performance using statistical and machine learning methods. *Journal of Risk and Financial Management, 18*(1), 23. https://doi.org/10.3390/jrfm18010023

DOI: 10.54254/2753-8818/2026.PU37829

[4] Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020
[5] Yeh, I.-C. (2009). Default of credit card clients [Data set]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C55S3H
[6] Hlongwane, R., Ramabao, K., & Mongwe, W. (2024). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLOS ONE, 19*(8), e0308718. https://doi.org/10.1371/journal.pone.0308718
[7] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. (2024). Effective methods of categorical data encoding for artificial intelligence algorithms. *Mathematics, 12*(16), 2553. https://doi.org/10.3390/math12162553
[8] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634
[9] Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal, 65*(8), 2200302. https://doi.org/10.1002/bimj.202200302
[10] Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., & Ziegler, A. (2023). Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine, 42*(29), 5451–5478. https://doi.org/10.1002/sim.9921
[11] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
[12] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[13] Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101–133. https://doi.org/10.1214/ss/1009213286
[14] McGrath, O., & Burke, K. (2024). Binomial confidence intervals for rare events: Importance of defining margin of error relative to magnitude of proportion. *The American Statistician, 78*(4), 437–449. https://doi.org/10.1080/00031305.2024.2350445
[15] Zhou, X.-H., Li, C.-M., & Yang, Z. (2008). Improving interval estimation of binomial proportions. *Philosophical Transactions of the Royal Society A, 366*(1874), 2405–2418. https://doi.org/10.1098/rsta.2008.0037