Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 3, 2026

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN CREDIT RISK ASSESSMENT: A GLOBAL ANALYSIS OF PREDICTIVE MODELS IN THE BANKING SECTOR

Kamil Nazir*[1], Dr. Amer Hussain[2], Dr. Bushra Rehman[3], Aleem Rafique[4], Dawood Samuel[5]

*[1]*PhD Scholar, Department of Business Administration, Lahore Leads University, Lahore, Pakistan*
[2]Director ORIC, Assistant Professor, Lahore Leads University, Lahore, Pakistan
[3]Associate Professor, Department of Business Administration, Lahore Leads University, Lahore, Pakistan
[4,5]PhD Scholar, Department of Business Administration, Lahore Leads University, Lahore, Pakistan

*[1]ranakamilnazir@gmail.com, [2]dreamer.dba@leads.edu.pk, [3]bushra.dba@leads.edu.pk, [4]aleemrfq@gmail.com, [5]dawoodsamuel123@gmail.com

DOI: http://doi.org/10.5281/zenodo.24336785

Keywords
Credit risk assessment; Machine learning; Neural network; XGBoost; Explainable AI

Article History
Received: 05 January 2026
Accepted: 18 February 2026
Published: 04 March 2026

Copyright @Author
Corresponding Author: *
Kamil Nazir

Abstract

Credit risk assessment is a core function of banks, and the accuracy with which lenders identify borrowers who will default shapes profitability, capital adequacy, and financial stability. This study compares four predictive models for credit risk assessment: logistic regression, random forest, XGBoost, and a multilayer perceptron (MLP) neural network. The models were trained and tested on the UCI Default of Credit Card Clients dataset from Taiwan, covering 30,000 clients, of whom 6,636 (22.12 percent) defaulted on payment in the following month. All models were evaluated on an identical stratified 80/20 train-test split with class weights applied to handle imbalance, and compared using ROC AUC, precision-recall AUC, accuracy, precision, recall, F1 score, and the Kolmogorov-Smirnov statistic. The MLP neural network achieved the best overall discrimination with a ROC AUC of 0.7737, followed by XGBoost at 0.7634, logistic regression at 0.7588, and random forest at 0.7584. The MLP also led on precision-recall AUC (0.5483), recall (0.5983), F1 score (0.5285), and the Kolmogorov-Smirnov statistic (0.4142). Random forest recorded the highest accuracy (0.8145) and precision (0.6503) but the lowest recall (0.3489), missing most defaulters at the 0.5 threshold, which shows that accuracy alone is a poor guide in imbalanced credit data. Permutation importance showed that the most recent repayment status dominated the MLP predictions. The findings suggest that neural networks and gradient boosting offer modest gains over logistic regression, and that banks should judge models by ranking and recall measures rather than accuracy, while keeping explanation and fairness checks in place.

1. Introduction

Credit is the core business of banks, and credit risk, the risk that a borrower fails to meet contractual repayment obligations, is the single largest source of loss for most lenders. A loan that defaults does not only cost the bank the unpaid principal and interest; it also consumes capital, raises funding costs, and, when defaults cluster, threatens the stability of the financial system. Accurate assessment of default risk is therefore central to lending decisions, loan pricing, provisioning, and capital planning. For

decades banks have relied on credit scoring models that convert borrower characteristics into a probability of default, and those scores now govern who receives credit, on what terms, and at what price. The quality of the scoring model thus shapes both bank performance and access to finance for households and small businesses.

The traditional workhorse of credit scoring is logistic regression, valued for its transparency: each coefficient can be read as the effect of one borrower characteristic on the odds of default, and regulators and auditors can follow the reasoning behind every decision (Dumitrescu et al., 2022). Yet borrower behaviour is rarely linear. Income, repayment history, credit utilisation, and loan purpose interact in ways that a linear model cannot capture, and lender data have grown richer and more complex. These limits have pushed banks toward machine learning methods that can learn non-linear patterns from large datasets, and the research literature on the subject has expanded rapidly. A systematic review of 129 studies found tree-based ensembles and deep learning now dominating credit risk research (B. Zhang et al., 2026), while another review of 63 studies reported ensemble models as the most used methods in the field (Ayari et al., 2026).

The evidence so far favours ensemble and deep learning methods, but the advantage is modest and uneven. Gradient boosting and XGBoost have recorded very high discrimination in several studies, including a ROC AUC of 0.9714 for XGBoost on real-world loan data (X. Zhang et al., 2025), while a stacked ensemble reached an AUC of 0.761 on the German Credit Dataset, with random forest (0.749) ahead of XGBoost (0.733) in that comparison (Mathibela & Maposa, 2026). Deep learning research is growing at about 29.2 percent a year, with feedforward networks and attention architectures now prominent in fintech lending (Saputra et al., 2026). At the same time, the field faces persistent problems. Defaulters are a small minority in most credit data, often only 5 to 20 percent of cases (Saputra et al., 2026), so accuracy can mislead and imbalance handling is central to any honest comparison. There is also no agreed standard for datasets or evaluation metrics, which blocks direct comparison across studies (Ayari et al., 2026).

Accuracy is not the only test a credit model must pass. A model that cannot be explained cannot be defended to a regulator or a rejected applicant, and explainability has become a legal requirement rather than a preference. The Basel Committee has warned that machine learning models are harder to interpret than traditional models and has named explainability, governance, and accountability as areas for further work (Basel Committee on Banking Supervision, 2022, 2024). In the European Union, the AI Act classifies systems that evaluate creditworthiness as high-risk, with duties for risk management, documentation, and human oversight (European Parliament and Council of the European Union, 2024). In the United States, adverse action rules require lenders to state specific reasons when credit is refused, which complex models cannot always supply (Consumer Financial Protection Bureau, 2022; 12 CFR § 1002.9). Fairness is a related concern: screening models can discriminate on the basis of protected attributes in the training data, and formal tests now exist to detect and correct that bias (Hurlin et al., 2026). Explanation tools such as SHAP and LIME help, but they differ in kind, with SHAP offering consistent global and local attributions and LIME offering local approximations (Gramegna & Giudici, 2021).

This study addresses a gap that the reviews above make clear. Most published comparisons report results from different datasets, different samples, and different metrics, so their numbers cannot be set side by side. What is needed is a head-to-head comparison of the leading model families on a single dataset, judged by ranking measures that suit imbalanced credit data rather than by accuracy alone. This paper therefore compares logistic regression, random forest, XGBoost, and a multilayer perceptron neural network on the UCI Default of Credit Card Clients dataset, using an identical split and preprocessing for all four models. Its objectives are to measure and compare the four models on ROC AUC, precision-recall AUC, accuracy, precision, recall, F1 score, and the Kolmogorov-Smirnov statistic;

to identify which borrower features drive the best model; and to draw practical lessons for banks that must balance predictive power with explanation and fairness.

The remainder of the paper is organised as follows. The next section reviews the literature on machine learning models, explainability, and regulation in credit scoring. The methodology section describes the dataset, preprocessing, models, and evaluation measures. The results section reports the comparison, and the discussion interprets the findings for banking practice before the conclusion.

2. Literature Review

2.1 Which models work best for credit risk?

Credit risk assessment decides who receives credit and on what terms, so it sits at the centre of banking. In recent years the field has moved quickly from statistical scorecards toward machine learning, and the research literature has grown with it. Saputra et al. (2026) reported that deep learning research for fintech credit risk was growing at a compound annual growth rate of about 29.2 percent, while B. Zhang et al. (2026) screened 894 papers and selected 129 studies for a systematic review of individual credit risk assessment. Ayari et al. (2026) reviewed 63 studies published between 2018 and 2024. Across these reviews, two findings repeat. First, tree-based ensembles and deep learning dominate current research. Second, there is no single winning model, because rankings change with the dataset, the features, and the metric used.

Ensemble methods are the most used approaches in the reviewed literature (Ayari et al., 2026). Mathibela and Maposa (2026) tested a stacked ensemble of logistic regression, random forest, XGBoost, and a multilayer perceptron, with a logistic regression meta-learner, on the German Credit Dataset of 1,000 applications with a 30 percent default rate. The stacked ensemble performed best overall, with an AUC of 0.761, precision of 0.783, recall of 0.806, and an F1 score of 0.794. Notably, random forest (AUC 0.749) beat XGBoost (AUC 0.733) in that study, which challenges the common assumption that gradient boosting always ranks above random forest. X. Zhang et al. (2025) compared XGBoost, gradient boosting, random forest, and LightGBM on real-world loan data and found a different ordering: gradient boosting had the best overall classification, with accuracy of 0.8887, F1 of 0.8084, and recall of 0.8021, while XGBoost had the highest discriminatory power, with a ROC AUC of 0.9714, ahead of random forest at 0.9696 and LightGBM at 0.9683. Lenka et al. (2022) reviewed ensemble learning for imbalanced credit scoring and tested 5 base models and 14 ensemble models on the German, Australian, and Japanese credit datasets. Their best result came from CatBoost combined with genetic algorithm feature selection, with accuracy of 86.70 percent, 88.40 percent, and 86.20 percent on the three datasets, judged on accuracy, AUC, F1, Brier score, and the Kolmogorov-Smirnov statistic. They also found that resampling improved performance on imbalanced data.

Deep learning has become a major strand of the literature. Saputra et al. (2026), in a PRISMA 2020 systematic review of 30 publications, found that feedforward deep neural networks and attention architectures such as LSTM and CNN dominate, while hybrid deep-boosting frameworks such as LightGBM-Attention and GBDT-Deep FM are emerging. Ayari et al. (2026) reported that convolutional neural networks reached an AUC of 0.90 on the German dataset and 0.99 on the Australian dataset in the reviewed studies, and that the highest reported accuracy was 98.66 percent. These figures cannot be compared directly across studies, because the datasets differ in size, features, and default rates. Ayari et al. (2026) identify the absence of standardised datasets and evaluation metrics as a gap that blocks cross-study comparison, so models must be compared on one dataset rather than by comparing numbers across papers.

Not every contribution targets accuracy alone. Dumitrescu et al. (2022) proposed penalised logistic tree regression, which feeds rules extracted from short-depth decision trees into a penalised logistic regression. The method captures non-linear effects in credit data while keeping the intrinsic interpretability of logistic

regression, and it proved significantly more accurate than logistic regression and competitive with random forest on four real credit default datasets. Their work shows that the field is not only comparing black-box accuracy but also searching for models that remain explainable.

2.2 Interpretability and explainability

As models grow more complex, the ability to explain a credit decision has become as important as the decision itself. Bussmann et al. (2021) combined Shapley values with correlation networks so that artificial intelligence predictions are grouped by the similarity of their explanations. Applied to 15,000 small and medium enterprises seeking credit on peer-to-peer platforms, their method showed that risky and non-risky borrowers can be grouped by similar financial characteristics, and that these groups can be used both to explain a credit score and to predict future behaviour. Gramegna and Giudici (2021) compared SHAP and LIME directly on real Italian small and medium enterprise data. Both methods could discriminate observations into risk groups, but they differ in kind: SHAP gives consistent global and local feature attributions grounded in game theory, while LIME gives local linear approximations around single predictions. Their comparison matters for practice, because a bank must choose an explanation method that is stable enough to defend before a regulator and a customer. Bahlool et al. (2026) reviewed 43 peer-reviewed studies published between 2020 and 2025 and synthesised how performance, fairness, and explainability trade off against each other in AI credit scoring, with direct implications for AI governance and regulatory compliance in lending.

2.3 Challenges: class imbalance, fairness, and data quality

Credit data are imbalanced by nature, because most borrowers repay. Saputra et al. (2026) reported that defaulters make up only 5 to 20 percent of observations in fintech lending data, and Paz et al. (2025) found one reviewed study in which 96 percent of a real-world dataset belonged to the majority class. Abedin et al. (2023) proposed a WSMOTE-ensemble that combines weighted synthetic minority oversampling with bagging, and on 3,111 small-business records from a Chinese commercial bank it improved minority-class (default) recognition accuracy by 15.16 percent compared with competitors, while sampling methods outperformed non-sampling algorithms overall. Paz et al. (2025) added a caution: many studies claim benefits from SMOTE without reporting pre-SMOTE comparisons, so the gains are often not quantified. The choice of imbalance handling changes results enough that it must be reported and held constant across models in any fair comparison.

Fairness is a second challenge. Hurlin et al. (2026) showed that screening algorithms can discriminate between good and bad borrowers, but they can also discriminate against people sharing a protected attribute such as gender, age, or racial origin, unintentionally, through the training data or the model itself. The authors provide a formal statistical test for algorithmic fairness, a method to identify the variables responsible for unfairness, and a way to optimise the fairness-performance trade-off while keeping high forecasting accuracy. Bahlool et al. (2026) reached a similar conclusion from their review of 43 studies: performance, fairness, and explainability must be evaluated together, not one at a time. Data quality is a third challenge. B. Zhang et al. (2026) named class imbalance, reject inference, and concept drift as core data challenges in their end-to-end framework, which runs from data preprocessing and feature engineering through model training, evaluation, and deployment.

2.4 Regulatory and governance context

Regulators have moved from watching machine learning to setting expectations for it. The Basel Committee on Banking Supervision (2022) published its first stand-alone newsletter on artificial intelligence and machine learning, noting that such models can raise operational efficiency and improve risk management, but are harder to interpret than traditional models, and

carry risks around data quality, bias, cyber exposure, and third-party dependence. The Committee stressed that banks remain fully responsible for models developed by third parties, and named three areas for further work: explainability of outcomes, governance and accountability structures, and implications for bank resilience and financial stability. The Basel Committee on Banking Supervision (2024) followed with a 46-page report on the digitalisation of finance, recording bank AI and machine learning use in credit underwriting, trading, pricing, capital and liquidity planning, fraud detection, and anti-money laundering, while finding generative AI use still limited, mainly in lower-risk internal pilots.

In the European Union, the AI Act (European Parliament and Council of the European Union, 2024) classifies AI systems that evaluate the creditworthiness of natural persons or establish their credit score as high-risk under Annex III point 5(b), with requirements for risk management, data governance, technical documentation, logging, human oversight, and post-market monitoring. Standalone high-risk obligations apply from 2 December 2027, and fines reach 15 million euros or 3 percent of global annual turnover. Explainability is therefore a legal requirement, not a preference. In the United States, the Consumer Financial Protection Bureau (2022) stated in Circular 2022-03 that creditors must give applicants a statement of specific principal reasons for adverse action, and that the Equal Credit Opportunity Act and Regulation B do not permit creditors to use complex algorithms when doing so means they cannot provide the specific and accurate reasons. The circular itself was withdrawn on 12 May 2025, but the underlying Regulation B duty (12 CFR § 1002.9) was unchanged, so the duty it restated still applies. The Board of Governors of the Federal Reserve System et al. (2026) issued SR Letter 26-2, which replaced SR 11-7 as the United States model risk management guidance, keeping the three pillars of sound development, implementation and use; validation with effective challenge; and governance, policies and controls, and treating credit risk models that manage financial risk exposures as generally higher risk.

2.5 Research gap

The literature establishes three things clearly. First, ensembles and deep learning usually beat logistic regression on discrimination, but the ranking among them is unstable across datasets. Second, explainability and fairness are now required alongside accuracy, yet they are rarely evaluated together with performance on the same data. Third, the absence of standardised datasets and metrics makes cross-study comparison unreliable (Ayari et al., 2026). This paper addresses that gap by comparing four models, logistic regression, random forest, XGBoost, and a multilayer perceptron neural network, on one dataset with identical preprocessing, the same train and test split, and the same class imbalance handling, and by reporting a full set of metrics, ROC AUC, PR AUC, accuracy, precision, recall, F1, and the KS statistic, together with permutation feature importance for the best model.

3. Methodology

3.1 Dataset

The study uses the Default of Credit Card Clients dataset from the UCI Machine Learning Repository (dataset 350), covering 30,000 clients of a bank in Taiwan. The data cover client credit records from April to September 2005. The dataset has 23 raw features plus an ID column and a binary target, default payment next month, coded 1 for default and 0 for no default. Of the 30,000 clients, 6,636 (22.12 percent) defaulted and 23,364 (77.88 percent) did not. The dataset contains 0 missing cells. It is one of the most cited credit datasets in the literature, which makes results comparable with earlier work, but it covers a single market and a single year, so results generalise to other banking systems only with caution.

Table 1 summarises the key characteristics of the dataset.

Table 1. Dataset Descriptive Statistics

Characteristic	Value
Dataset	UCI Default of Credit Card Clients, Taiwan (UCI id 350)
Period covered	April to September 2005
Rows used	30,000 (all rows, no rows dropped)
Raw features	23 (plus ID column, dropped; plus binary target)
Target	Default payment next month (1 = default, 0 = no default)
Defaults	6,636 (22.12%)
Non-defaults	23,364 (77.88%)
Missing cells	0
Training set	24,000 rows (default rate 22.12%)
Test set	6,000 rows (default rate 22.12%)
Split	Stratified 80/20, seed 42

3.2 Data preprocessing

All rows were used, and no rows were dropped. The ID column was dropped because it carries no predictive information. The 14 numeric features were standardised with a standard scaler: the credit limit (LIMIT_BAL), age (AGE), the six bill statement amounts (BILL_AMT1 to BILL_AMT6), and the six previous payment amounts (PAY_AMT1 to PAY_AMT6). The 9 categorical features were one-hot encoded: sex, education, marriage, and the repayment status in each of the six months (PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6). One-hot encoding expanded these 9 features into 91 columns. The repayment status codes were kept as categorical codes from -2 to 8: -1 means paid duly, -2 means no consumption, 0 means use of revolving credit, and 1 to 8 mean payment delay in months. PAY_0 is the repayment status in September 2005, and PAY_2 to PAY_6 are the statuses from August to April 2005. Two data quality issues were kept as they were and one-hot encoded rather than recoded or removed: education contains undocumented codes 0, 5, and 6, and marriage contains the undocumented code 0. The same preprocessing was applied identically to all four models, so differences in results reflect the models, not the data preparation.

3.3 Train and test split

The data were divided with a stratified 80/20 train and test split using seed 42. The training set has 24,000 rows with a default rate of 22.12 percent, and the test set has 6,000 rows with a default rate of 22.12 percent. Stratification kept the class balance identical in both sets. The same split was used for all four models, and every metric reported in this paper comes from the test set only.

3.4 Handling class imbalance

With 22.12 percent defaults, the classes are imbalanced, so the positive class was weighted in every model. The weight was the ratio of negative to positive cases in the training set, 3.5206 (18,691 negative and 5,309 positive). Logistic regression used balanced class weights and random forest used balanced subsample class weights, XGBoost used a scale positive weight of 3.5206, and the multilayer perceptron used sample weights with the positive class weighted by 3.5206, because its classifier has no class weight parameter. The same imbalance handling principle was applied across all models, and no resampling method was used, so the comparison remains fair and the training data remain the observed data.

3.5 Models

Four models were trained, spanning the range from the traditional statistical baseline to deep learning. Logistic regression was the baseline, run with a maximum of 1,000 iterations and balanced class weights. It is the standard against which newer models are judged, and its coefficients remain directly interpretable. Random forest was run with 300 trees and balanced subsample class weights, using all available cores. It is an ensemble of decision trees that captures non-linear effects and interactions without requiring the analyst to specify them. XGBoost was run with 300 trees, a maximum depth of 6, a learning rate of 0.1, a subsample ratio of 0.8, a column sample ratio of 0.8, and a scale positive weight of 3.5206. It is a gradient boosting ensemble and is widely regarded as one of the strongest methods for tabular credit data. The multilayer perceptron neural network was run with hidden layers of 64 and 32 units, an alpha of 0.001, a maximum of 500 iterations, early stopping enabled, and the Adam solver. It is the deep learning representative in the comparison. No hyperparameter search was run. All models used reasonable default or near-default settings so that the comparison remains fair and reproducible.

3.6 Evaluation metrics

Models were compared on the test set of 6,000 clients at a decision threshold of 0.5. The primary metric was ROC AUC, which measures discrimination across all thresholds and is the standard metric for credit scoring. PR AUC was reported alongside it, because it is more informative than ROC AUC when defaults are the minority class. Accuracy, precision, recall, and the F1 score were reported at the 0.5 threshold, and the Kolmogorov-Smirnov (KS) statistic was reported as an additional measure of separation between defaulters and non-defaulters. Accuracy was not used as the leading metric, because it can mislead when the classes are imbalanced. No threshold tuning was performed, so the reported precision, recall, and F1 figures reflect the fixed 0.5 threshold, and a bank applying these models would normally tune the threshold to its own cost of false positives against false negatives.

3.7 Feature importance

Feature importance was measured for the best model using permutation importance on the test set, with five repeats and ROC AUC as the scoring metric. Each feature was shuffled in turn, and the mean drop in ROC AUC was recorded, so the ranking shows which features the model actually relies on. SHAP was installed, but permutation importance was used instead, because it is model-agnostic and directly comparable across models.

3.8 Limitations of the methodology

Four limitations should be stated plainly. First, the data come from one market and one year, Taiwan in 2005, so the results describe that setting and generalise to other banking systems only with caution. Second, no hyperparameter search was run, so the reported performance reflects reasonable default or near-default settings rather than tuned optima, and a tuned comparison could change the ranking. Third, all accuracy, precision, recall, and F1 figures use the fixed 0.5 decision threshold, and no threshold tuning was performed, although a bank would normally tune the threshold to its own cost of false positives against false negatives. Fourth, the undocumented education and marriage codes were kept as they were, so a data quality issue in the source file is carried into the models rather than corrected.

3.9 Reproducibility

The analysis was run in Python 3.12.3 with scikit-learn 1.9.1, pandas 2.3.3, numpy 2.5.4, matplotlib 3.11.2, and xgboost 3.4.1, using the script run_models.py in the project folder and the dataset file "default of credit card clients.xls" downloaded from the UCI archive. Seed 42 was fixed everywhere. The full run was performed twice with identical results, which confirms that the reported numbers are deterministic and reproducible.

4. Results

All results reported in this section come from the held-out test set of 6,000 cases, which is 20 percent of the 30,000 cases in the Taiwan credit card data. The test set contains 1,327 defaults (22.12 percent) and 4,673 non-defaults. All four models were trained on the same 24,000-case training set and evaluated on the same test set, with identical preprocessing and the same class weight for the positive class (3.5206), so the comparison is fair. Accuracy, precision, recall, and F1 are reported at a fixed decision threshold of 0.5. ROC AUC and PR AUC do not depend on a single threshold, and they are the primary measures used to rank the models in this paper.

Table 2 presents the test-set performance of the four models across seven measures. The multilayer perceptron neural network achieved the best overall result, with ROC AUC 0.7737, PR AUC 0.5483, F1 0.5285, recall 0.5983, and KS 0.4142. XGBoost was second on ROC AUC (0.7634), followed by logistic regression (0.7588) and random forest (0.7584). The ordering changes slightly on PR AUC, with the MLP again on top at 0.5483, followed by random forest at 0.5381, XGBoost at 0.5375, and logistic regression at 0.5241. Two points must be stated plainly. First, the MLP lead over XGBoost on ROC AUC is small, only 0.0103, and all four models fall within a narrow band of 0.7584 to 0.7737. Second, no single model dominates on every measure, which matches the wider literature: systematic reviews report that rankings change with the dataset and that no model wins everywhere (Ayari et al., 2026; B. Zhang et al., 2026).

Table 2. Test-Set Performance of the Four Models (n = 6,000; Threshold 0.5)

Model	ROC AUC	PR AUC	Accuracy	Precision	Recall	F1	KS
Logistic Regression	0.7588	0.5241	0.7735	0.4893	0.5531	0.5193	0.3944
Random Forest	0.7584	0.5381	0.8145	0.6503	0.3489	0.4541	0.3965
XGBoost	0.7634	0.5375	0.7673	0.4783	0.5727	0.5213	0.3992
MLP Neural Network	0.7737	0.5483	0.7638	0.4732	0.5983	0.5285	0.4142

Random forest illustrates why accuracy alone is a poor guide in credit scoring. It recorded the highest accuracy (0.8145) and the highest precision (0.6503) of the four models, yet it also recorded the lowest recall (0.3489) and the lowest F1 (0.4541). At the 0.5 threshold it correctly identified only 463 of the 1,327 defaulters and missed 864 of them. A model that misses about two thirds of the defaults it is supposed to catch is not useful to a lender, whatever its accuracy says. This is the known problem with imbalanced credit data, where defaulters are a minority of cases (Saputra et al., 2026), and it is why this paper leads with ROC AUC and PR AUC rather than accuracy. Logistic regression, the simplest model, remained competitive, with ROC AUC 0.7588 and recall 0.5531, and its transparency is an advantage that the headline numbers do not capture (Dumitrescu et al., 2022).

Table 3 reports the confusion matrices at the 0.5 threshold. The MLP correctly identified 794 defaulters and 3,789 non-defaulters, missing 533 defaulters and raising 884 false alarms. XGBoost was close, with 760 true positives, 567 missed defaults, and 829 false positives. Logistic regression identified 734 defaulters and missed 593, with 766 false positives. Random forest was the most conservative, with only 249 false positives, but it paid for that caution by missing 864 defaults. The pattern is a direct trade-off: the models that catch more defaulters also flag more good customers. Which error a bank prefers depends on its own costs, and the 0.5 threshold used here was not tuned to any cost structure. Cost-sensitive threshold tuning, as used by X. Zhang et al. (2025), is the natural next step before any of these models is deployed.

Table 3. Confusion Matrices on the Test Set

Model	True negatives	False positives	False negatives	True positives
Logistic Regression	3,907	766	593	734
Random Forest	4,424	249	864	463
XGBoost	3,844	829	567	760
MLP Neural Network	3,789	884	533	794

Note. Rows are actual outcomes, columns are predicted outcomes; order: no default, default.

Table 4 lists the top ten features by permutation importance for the best model, the MLP. Permutation importance for the best model, the MLP, shows that the most recent repayment status dominates the prediction. Using revolving credit in September 2005 (PAY_0 = 0) produced the largest drop in ROC AUC when shuffled (0.0315), followed by a two-month payment delay in September 2005 (PAY_0 = 2, 0.0276) and the amount of credit granted (LIMIT_BAL, 0.0274). The next most important feature, the amount paid in September 2005 (PAY_AMT1, 0.0080), is far behind, and all remaining features fall below 0.0070. The same pattern appears in earlier months: revolving credit use in April 2005 (PAY_6 = 0) and August 2005 (PAY_2 = 0) both appear in the top ten, together with the corresponding bill statement amounts. The reading is straightforward. How a customer behaved most recently, and whether they are already relying on revolving credit, carries far more signal than their age, education, or older payment history. This matches earlier work in which account status emerged as the strongest predictor (Mathibela & Maposa, 2026). Permutation importance was used instead of SHAP because it is model-agnostic and directly comparable across models; SHAP and LIME remain the standard tools for explaining individual decisions (Bussmann et al., 2021; Gramegna & Giudici, 2021).

Table 4. Top 10 Features by Permutation Importance, MLP Neural Network

Rank	Feature	Meaning	AUC drop
1	PAY_0 = 0	Repayment status Sep 2005: revolving credit used	0.0315
2	PAY_0 = 2	Repayment status Sep 2005: 2 months delay	0.0276
3	LIMIT_BAL	Amount of given credit (NT dollars)	0.0274
4	PAY_AMT1	Amount paid in September 2005	0.0080
5	PAY_6 = 0	Repayment status Apr 2005: revolving credit used	0.0066
6	BILL_AMT6	Bill statement amount Apr 2005	0.0066
7	BILL_AMT3	Bill statement amount Jul 2005	0.0059
8	PAY_AMT2	Amount paid in August 2005	0.0057
9	PAY_0 = -2	Repayment status Sep 2005: no consumption	0.0050
10	PAY_2 = 0	Repayment status Aug 2005: revolving credit used	0.0043

Note. Values are the mean drop in ROC AUC over five repeats.

Figures 1 and 2 show the ROC and precision-recall curves for all four models. The ROC curves sit close together, which matches the narrow range of ROC AUC values, while the precision-recall curves show the same ordering as the PR AUC column, with the MLP on top. For reference, the precision-recall baseline, which is the test default rate, is 0.2212, so all four models perform well above chance on PR AUC. Figure 3 shows the confusion matrix for the MLP, and Figure 4 shows its top 15 features by permutation importance, extending the ranking in Table 4. Training time was modest for every model: logistic regression took 1.1 seconds, the MLP 5.1 seconds, XGBoost 9.5 seconds, and random forest 74.4 seconds. Two cautions apply to any

comparison with published numbers. First, performance figures differ widely across studies because the datasets differ (Ayari et al., 2026), so the AUC values reported here should be compared with other studies only with care. Second, the full run was repeated with identical results, and the script and data are documented for reproduction, but no hyperparameter search was run, so these results describe near-default configurations rather than the best each model could achieve.

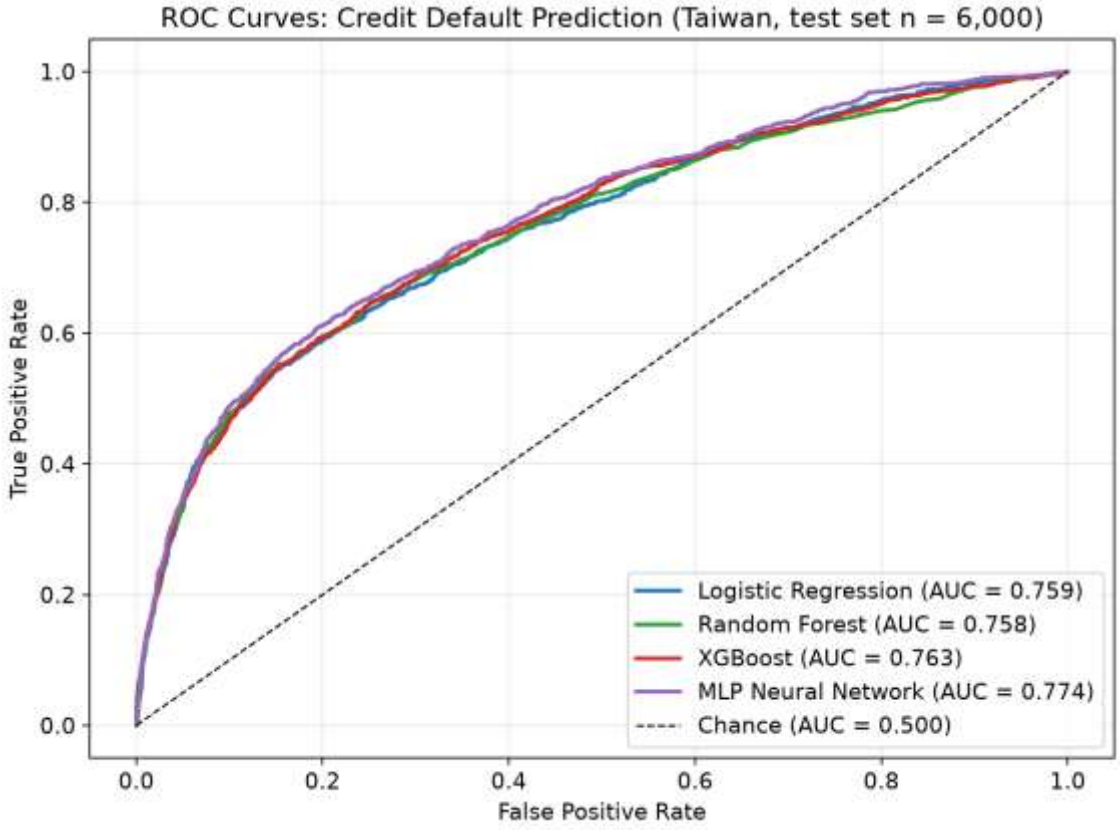


Figure 1. ROC curves for the four models

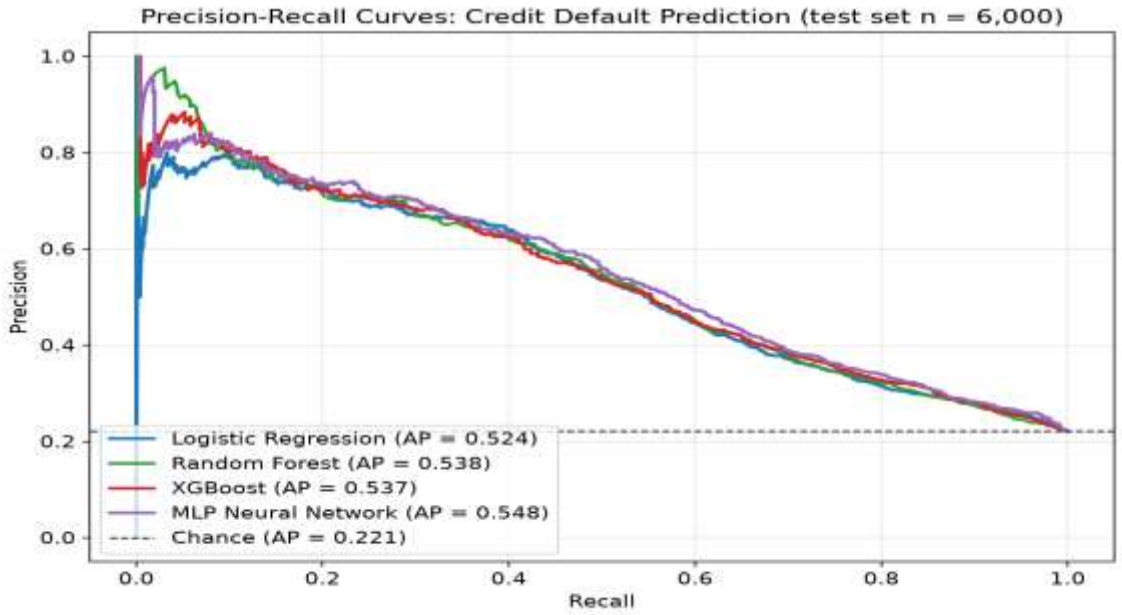


Figure 2. Precision-recall curves for the four models

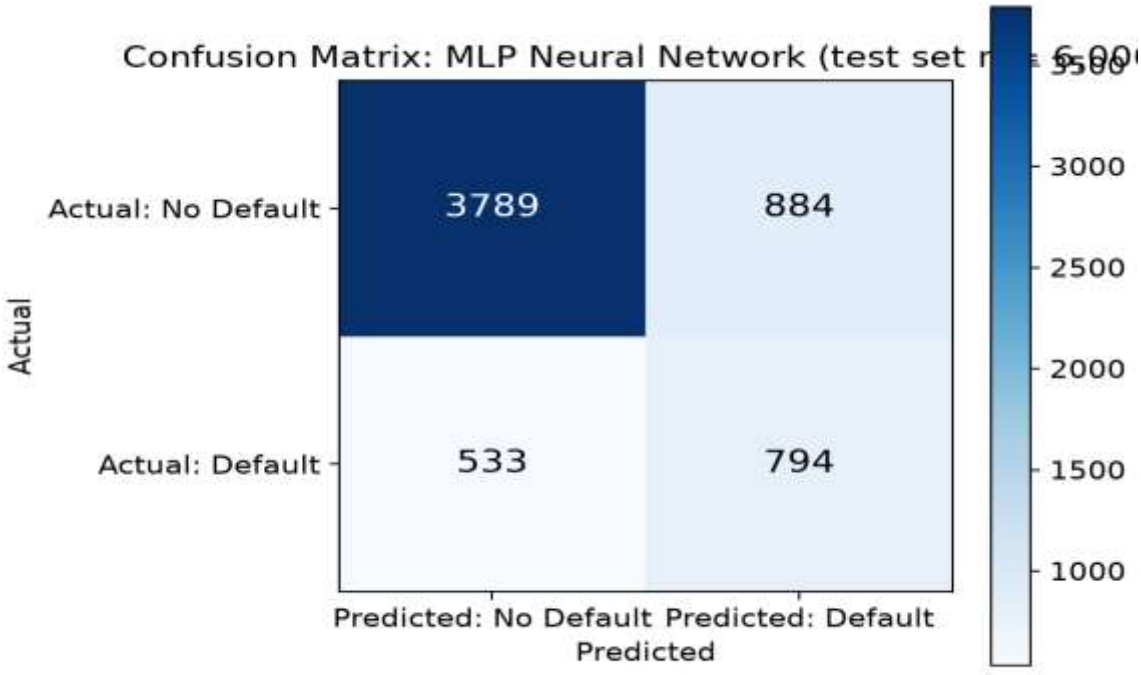


Figure 3. Confusion matrix for the MLP Neural Network (best model)

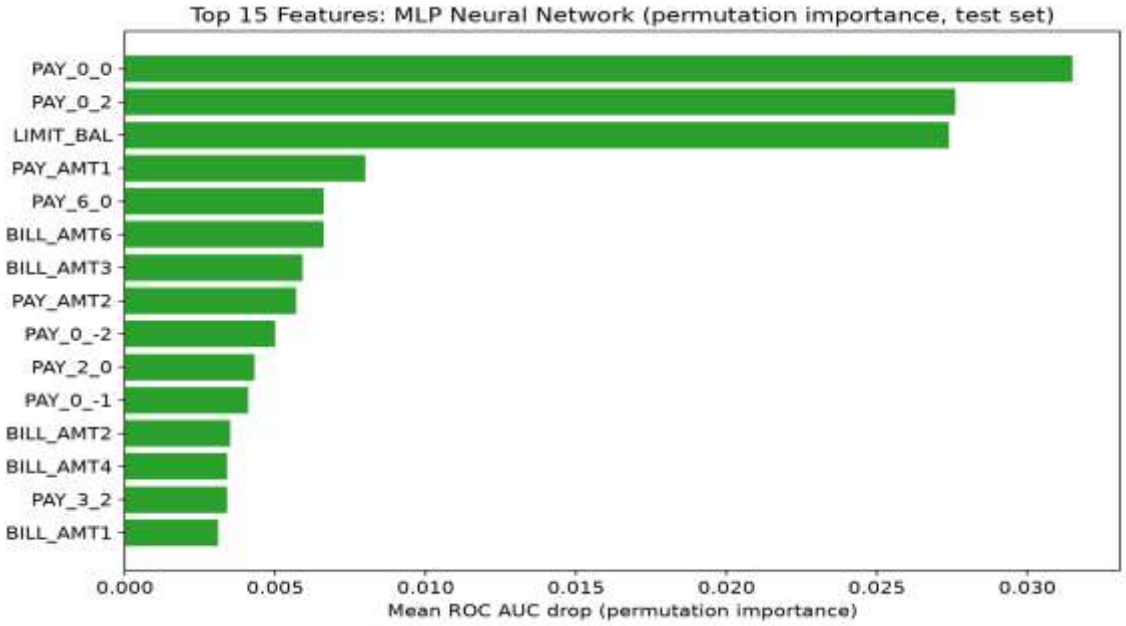


Figure 4. Top 15 features by permutation importance, MLP Neural Network

5. Discussion

For banks, the main message is that machine learning improves on logistic regression, but only by a small margin on these data. The MLP leads on five of the seven measures, yet its ROC AUC advantage over XGBoost is 0.0103, and over logistic regression it is 0.0149. A bank choosing between these models should therefore weigh more than the headline AUC. XGBoost trains quickly, has mature tooling, and is close to the best on every measure, which makes it a practical choice. The MLP needs more care: it is sensitive to scaling and tuning, and its decisions are harder to explain. Logistic regression, with ROC AUC 0.7588, is only slightly behind and is fully transparent, which still counts for a great deal in a regulated lender. This matches the wider evidence that tree-based ensembles and deep learning dominate current research, while interpretable models remain in use because regulators and auditors can follow them (B. Zhang et al., 2026; Dumitrescu et al., 2022).

The second message concerns the accuracy-recall trade-off. Random forest looked like the winner on accuracy (0.8145), but it missed 864 of 1,327 defaulters. In lending, a missed defaulter usually costs far more than a false alarm, because the false alarm costs a lost customer or a manual review while the missed default costs the loan itself. A bank that selects a model on accuracy alone would therefore choose the worst model for its real objective. The correct approach is to choose the operating threshold from the bank's own cost of each error, using cost-sensitive methods (X. Zhang et al., 2025), and to judge models on ranking measures such as ROC AUC and PR AUC, with recall and precision reported at the chosen threshold. Class imbalance handling also matters here: defaulters were only 22.12 percent of the test set, and in many real portfolios they are 5 to 20 percent (Saputra et al., 2026). The class weights used in this study are one standard approach; resampling methods such as SMOTE and its weighted variants are another, and they have improved minority-class recognition in small-business lending (Abedin et al., 2023; Paz et al., 2025).

The third message is about explainability, which is now a legal requirement rather than a preference. A bank that refuses credit must be able to state the specific reasons for the refusal, and a model that cannot supply those reasons cannot be used for the decision (Consumer Financial Protection Bureau, 2022). The circular

that restated this duty was withdrawn in 2025, but the underlying Regulation B requirement (12 CFR § 1002.9) was unchanged, so the duty still applies. In the European Union, the AI Act classifies systems that assess creditworthiness as high-risk, which brings duties for risk management, data governance, documentation, human oversight, and post-market monitoring (European Parliament and Council of the European Union, 2024). The Basel Committee has made the same point from the supervisory side: banks remain responsible for their models, including third-party models, and explainability, governance, and accountability are the areas it singled out for further work (Basel Committee on Banking Supervision, 2022, 2024). The United States supervisory guidance on model risk management, SR 26-2, keeps the same three pillars of sound development, validation with effective challenge, and governance (Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, & Federal Deposit Insurance Corporation, 2026). In practice this means that a model like the MLP in this study, however accurate, needs an explanation layer before it can be deployed: SHAP values for global and local attribution, or LIME for local explanations (Bussmann et al., 2021; Gramegna & Giudici, 2021). Performance, fairness, and explainability must be managed together, not traded off silently (Bahlool et al., 2026), and fairness should be tested formally, since models can discriminate through patterns in the training data even when no protected attribute is used (Hurlin et al., 2026).

This study has honest limitations. The data come from Taiwan and cover a single period, April to September 2005, so the results describe one market in one year and generalise to other banking systems only with caution. The data also contain undocumented codes in EDUCATION (0, 5, and 6) and MARRIAGE (0), which were kept as-is rather than recoded, so some noise remains in those features. No hyperparameter search was run, so the comparison describes near-default configurations; a tuned XGBoost or random forest could close the small gap to the MLP, and tuned ensembles or stacking have beaten single models in other studies (Mathibela & Maposa, 2026; Lenka et al., 2022). The decision threshold of 0.5 was not tuned to any cost structure. Feature importance was measured by permutation importance rather than SHAP, which gives a global ranking but no per-decision explanation. Finally, the study does not address reject inference, concept drift, or model monitoring after deployment, all of which the recent literature names as core challenges (B. Zhang et al., 2026). These limits do not weaken the central finding, which is that the four models are close and that the choice between them should turn on recall, explainability, and governance rather than on small differences in AUC.

6. Conclusion

The comparison of four predictive models on 30,000 Taiwan credit card cases found that the multilayer perceptron neural network performed best overall, with ROC AUC 0.7737, PR AUC 0.5483, F1 0.5285, and recall 0.5983, ahead of XGBoost, logistic regression, and random forest. The lead is real but small, and the models are close enough that a bank should choose on more than accuracy: random forest had the highest accuracy yet missed about two thirds of the defaulters, which shows why ranking measures and recall must lead the evaluation. The strongest signals were the most recent repayment status and the use of revolving credit, with the credit limit close behind, which matches the practical experience of lenders. This study contributes a clean, reproducible comparison on a single dataset with identical preprocessing and splits for all four models, which the literature has called for, since published performance numbers cannot be compared across different datasets (Ayari et al., 2026; B. Zhang et al., 2026). For banks, the practical conclusion is to adopt machine learning with an explanation layer and a cost-based threshold, under the model risk governance that regulators now expect. Future research should test the same comparison on multi-country and more recent data, tune the models and their thresholds, add ensemble and stacking approaches, apply SHAP explanations

and formal fairness tests, and study how models behave as data drift over time. As deep learning research in fintech lending grows rapidly (Saputra et al., 2026), the field will need shared benchmarks and standard evaluation practice before model claims can be compared with confidence. The question is no longer whether machine learning can predict default, but how to deploy it in a way that is accurate, explainable, and fair.

References

Abedin, M. Z., Chi, G., Hajek, P., & Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*, *9*(4), 3559-3579. https://doi.org/10.1007/s40747-021-00614-4

Ayari, H., Guetari, R., & Kraiem, N. (2026). Machine learning powered financial credit scoring: A systematic literature review. *Artificial Intelligence Review*, *59*, Article 13. https://doi.org/10.1007/s10462-025-11416-2

Bahlool, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, *19*(2), Article 104. https://doi.org/10.3390/jrfm19020104

Basel Committee on Banking Supervision. (2022). *Newsletter on artificial intelligence and machine learning* (BCBS Newsletter No. 27). Bank for International Settlements. https://www.bis.org/publ/bcbs_nl27.htm

Basel Committee on Banking Supervision. (2024). *Digitalisation of finance*. Bank for International Settlements.

Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, & Federal Deposit Insurance Corporation. (2026). *Supervisory guidance on model risk management* (SR Letter 26-2).

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

Consumer Financial Protection Bureau. (2022). *Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms* (87 FR 35864). https://files.consumerfinance.gov/f/documents/cfpb_2022-03_circular_2022-05.pdf

Dumitrescu, E., Hue, 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*(3), 1178-1192. https://doi.org/10.1016/j.ejor.2021.06.053

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, Article 752558. https://doi.org/10.3389/frai.2021.752558

Hurlin, C., Perignon, C., & Saurin, S. (2026). The fairness of credit scoring models. *Management Science*, *72*(1), 406-425. https://doi.org/10.1287/mnsc.2022.03888

Lenka, S. R., Bisoy, S. K., Priyadarshini, R., & Sain, M. (2022). Empirical analysis of ensemble learning for imbalanced credit scoring datasets: A systematic review. *Wireless Communications and Mobile Computing*, *2022*, Article 6584352. https://doi.org/10.1155/2022/6584352

Mathibela, M. R., & Maposa, D. (2026). Predictive modelling of credit default risk using machine learning and ensemble techniques. *Mathematical and Computational Applications*, *31*(2), Article 45. https://doi.org/10.3390/mca31020045

Paz, A., Crawford, B., Monfroy, E., Barrera-Garcia, J., Pena Fritz, A., Soto, R., Cisternas-Caneo, F., & Yanez, A. (2025). Machine learning and metaheuristics approach for individual credit risk assessment: A systematic literature review. *Biomimetics*, *10*(5), Article 326. https://doi.org/10.3390/biomimetics10050326

Saputra, M. P. A., Sukono, Riaman, Kartiwa, A., Misiran, M., & Wahid, A. J. (2026). Deep learning for credit risk prediction in fintech lending: A systematic literature review on model architectures, imbalanced data handling, and research agenda. *Journal of Risk and Financial Management*, *19*(7), Article 465. https://doi.org/10.3390/jrfm19070465

Zhang, B., Luo, J., Wu, R., Wei, J., Luo, Z., & Shen, H. (2026). Machine learning for individual credit risk assessment: A systematic literature review of state-of-the-art methods, challenges and perspectives. *Journal of Risk and Financial Management*, *19*(8), Article 607. https://doi.org/10.3390/jrfm19080607

Zhang, X., Zhang, T., Hou, L., Liu, X., Guo, Z., Tian, Y., & Liu, Y. (2025). Data-driven loan default prediction: A machine learning approach for enhancing business process management. *Systems*, *13*(7), Article 581. https://doi.org/10.3390/systems13070581