International Journal of Financial Management and Economics

P-ISSN: 2617-9210
E-ISSN: 2617-9229
Impact Factor (RJIF): 5.97
IJFME 2025; 8(2): 284-294
www.theeconomicsjournal.com
Received: 22-05-2025
Accepted: 25-06-2025

Dr. Syeda Rukhsana Khalid
Amjad Ali Khan College of Business Administration, Sultan Ul Uloom Education Society, Road No 3 Banjara Hills, Hyderabad, Telangana, India

Corresponding Author:
Dr. Syeda Rukhsana Khalid
Amjad Ali Khan College of Business Administration, Sultan Ul Uloom Education Society, Road No 3 Banjara Hills, Hyderabad, Telangana, India

Machine learning-based credit scoring: A comparative analysis of logistic regression and random forest models

Syeda Rukhsana Khalid

DOI: https://doi.org/10.33545/26179210.2025.v8.i2.596

Abstract
This paper presents a controlled, head-to-head comparison of Logistic Regression (LR) and Random Forest (RF) for consumer credit scoring using the German Credit dataset. A consistent pipeline -label encoding, stratified 80/20 train-test split, and five-fold cross-validation -supports like-for-like evaluation on accuracy, precision, recall, F1, ROC-AUC, and confusion matrices. Results show competitive but distinct strengths: RF attains slightly higher accuracy (0.775 vs. 0.765) and notably higher precision (0.703 vs. 0.627), reflecting fewer false positives; LR achieves higher recall for the default class (0.533 vs. 0.433) and a marginally better ROC-AUC (0.79 vs. 0.78), indicating stronger discrimination at low-FP operating points. Confusion matrices corroborate these trade-offs (LR: TN=121, FP=19, FN=28, TP=32; RF: TN=129, FP=11, FN=34, TP=26). Feature analysis aligns with domain priors: credit amount, age, and duration dominate RF importance, while LR coefficients provide directionally transparent effects for audit. Practically, the findings support a hybrid deployment: RF for back-end risk ranking and early warning, LR for audit-facing decisions and regulatory reporting. Limitations include reliance on public datasets and a focus on discrimination over calibration and fairness. Future work should examine proprietary portfolios, cost-sensitive thresholds and drift monitoring, and integrate explainable AI to reconcile lift with governance.

Keywords: Credit scoring, logistic regression, random forest, ROC-AUC, precision-recall, confusion matrix, imbalanced data, interpretability, explainable AI, cost-sensitive learning, German credit dataset

1. Introduction

Credit scoring lies at the core of retail and SME lending, shaping access to finance and risk management decisions. Logistic Regression (LR) has long dominated this field, valued for its simplicity, transparency, and regulatory acceptance. However, the expansion of high-dimensional, heterogeneous data has revealed the limitations of linear decision boundaries. Recent advances in machine learning (ML) -particularly ensemble methods such as Random Forests (RF), Gradient Boosting, and XGBoost -have demonstrated superior predictive performance across benchmark datasets, including the German Credit dataset. These models capture non-linear feature interactions and often generalise better to unseen borrowers, but at the cost of interpretability.

A growing body of literature reinforces this trade-off. Studies highlight the strong predictive capabilities of ensemble methods (Chang *et al*., 2024; Zhou *et al*., 2023; Machado *et al*., 2025) [4, 3, 2], while others emphasise the regulatory and operational challenges of deploying opaque models in financial institutions (Shi *et al*., 2022) [5]. This tension reflects a practical dilemma: while LR remains attractive for its explainability, ML models offer the possibility of more accurate credit risk assessments. The lack of clarity on how these models perform under controlled, head-to-head evaluations continues to limit their adoption.

This study contributes to bridging that gap by systematically comparing Logistic Regression and Random Forest classifiers on the German Credit dataset. The evaluation spans multiple performance metrics -accuracy, precision, recall, F1-score, and ROC-AUC -providing a comprehensive picture of both strengths and limitations. In doing so, it not only benchmarks predictive accuracy but also highlights the operational trade-offs that matter for lenders, regulators, and policymakers.

Figure 1 illustrates the central tension motivating this research: while Logistic Regression offers interpretability but lower predictive flexibility, Random Forest provides stronger accuracy through complex decision boundaries at the expense of transparency. By situating both models along this interpret ability-accuracy spectrum, the paper frames the conditions under which each may be preferable in practice.

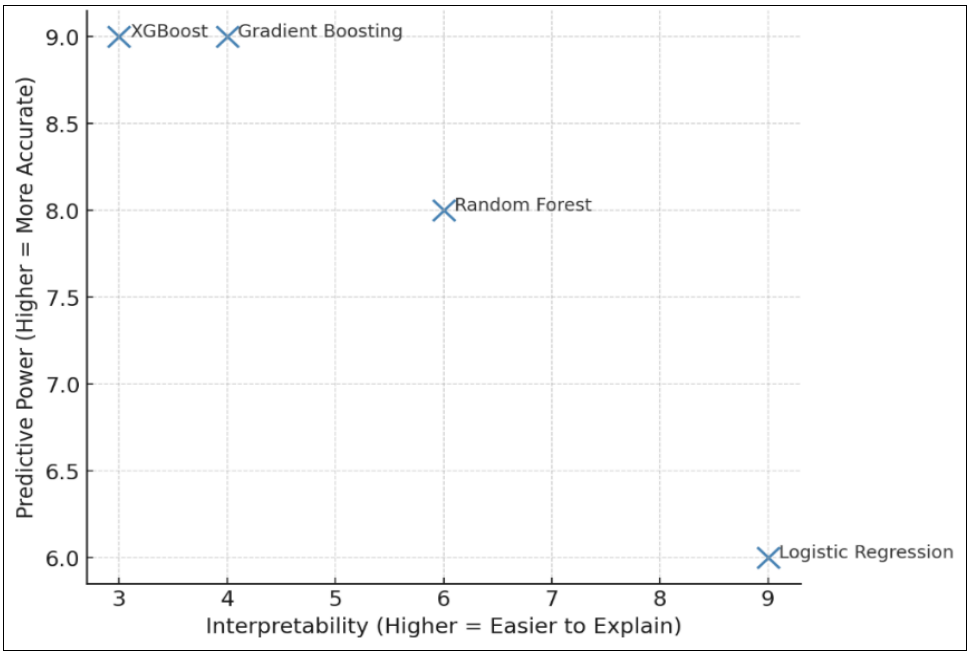

Fig 1: Trade-off Between Interpretability and Predictive Power in Credit Scoring Models.

2. Literature Review

2.1 From Logistic Regression to Machine Learning in Credit Scoring

Credit scoring has traditionally relied on Logistic Regression (LR) due to its simplicity, statistical grounding, and interpretability. Its coefficients map directly into odds ratios, a feature valued by regulators and auditors who require transparency in decision-making (Sudhakar *et al*., 2020; Jabeen *et al*., 2021) [23, 24]. Yet, as credit data expanded to include bureau histories, behavioural traces, and transaction-level variables, LR's linear form became a limitation. Studies show that LR underperforms in detecting defaults when faced with non-linear relationships, collinearity, and class imbalance (Shi *et al*., 2022) [5].

In response, tree-based machine learning (ML) models, particularly Random Forest (RF), emerged as alternatives. RF manages heterogeneous data types, missing values, and complex interactions without requiring explicit transformation. Evidence shows that RF delivers stronger recall on minority "default" classes -an outcome critical for lenders balancing risk and inclusion (Patel & Bhalani, 2022; Kavitha *et al*., 2021) [25, 26].

2.2 Empirical Evidence: Performance Comparisons

Empirical evidence consistently reports superior performance of RF over LR. On benchmark datasets such as the German Credit dataset, LR typically achieves 71-75% accuracy and an AUC between 0.70-0.74, whereas RF reaches 80-85% accuracy and AUCs closer to 0.80-0.89 (Sudhakar *et al*., 2020; Patel & Bhalani, 2022; Jadhav & Chavan, 2020) [23, 25]. Large-scale studies confirm this pattern: Zhou *et al*. (2023) [3] and Machado *et al*. (2025) [2] show that RF improves recall by 5-6% relative to LR, particularly under imbalanced conditions. Chang *et al*. (2024) [4] highlight RF's predictive stability over time, a property critical for portfolios subject to borrower profile shifts.

Meta-analyses echo these findings: Lessmann *et al*. (2015) [1] benchmarked 41 methods across 8 datasets and confirmed ensemble models such as RF and Gradient Boosting consistently dominate LR in accuracy and recall. Yu *et al*. (2008) [12] reached similar conclusions in early comparative work, noting RF's resilience across heterogeneous features. Building on this global evidence, the present study situates its analysis within the same interpretability-performance tension. While ensembles such as Random Forest offer measurable gains in predictive accuracy, Logistic Regression continues to hold institutional value through its transparency and auditability. Against this backdrop, the forthcoming methodology applies both models to the German Credit dataset, enabling a structured evaluation that aligns with international findings while also exposing dataset-specific nuances.

To contextualize the present analysis, a survey of recent comparative studies is instructive. Table 1 synthesizes representative evidence on the performance of Logistic Regression (LR) vis-à-vis Random Forest (RF) across diverse credit datasets. These studies consistently highlight the predictive advantages of ensemble methods-particularly in recall and stability-while reaffirming LR's continued role as a regulatory and interpretive benchmark.

Table 1: Representative comparative studies of LR vs RF in credit scoring (scope, methods, headline result).

Study	Dataset	Methods Compared	Key Findings
Shi *et al.* (2022) [5]	Multiple benchmark datasets (systematic review)	Logistic Regression vs. RF, GBM, XGBoost	Concludes ensemble methods consistently outperform LR in accuracy and recall; LR remains baseline for compliance and interpretability.
Zhou *et al.* (2023) [3]	Large-scale consumer credit dataset (China)	LR, RF, Gradient Boosting	RF/GBM outperform LR by ~5-6% in recall, particularly valuable in imbalanced default detection.
Machado *et al.* (2025) [2]	Lending Club (peer-to-peer loans, USA)	LR vs. RF, GBM	RF achieves higher predictive stability and recall than LR; ML models reduce false negatives significantly.
Chang *et al.* (2024) [4]	Credit card default dataset (Taiwan)	LR, RF, Gradient Boosting	RF demonstrates superior predictive stability across time; LR less robust under shifting borrower profiles.

Source: Compiled by the author from cited studies.

Taken together, these findings frame the rationale for applying both LR and RF in the present study, enabling a balanced assessment that considers predictive gains alongside interpretability and compliance.

Overall, these findings confirm a consistent empirical advantage of RF over LR. The persistence of LR, however, rests not on predictive superiority but on its regulatory acceptance and interpretive clarity.

2.3 Systematic Reviews and Meta-Analyses

Systematic evidence further reinforces RF's advantage. Shi *et al.* (2022) [5] conducted a systematic review and concluded that tree-based ensembles (RF, GBM, XGBoost) consistently outperform LR across varied contexts. Lessmann *et al.* (2015) [1] conducted one of the most comprehensive benchmarking exercises to date and demonstrated that ensemble classifiers provide not just higher accuracy but also more robust generalisation across datasets. Such findings push the literature toward beyond-accuracy evaluation, with emphasis on ROC-AUC, recall, and cost-sensitive metrics-since false negatives (missed defaults) carry far higher losses than false positives.

2.4 Why Random Forest Wins (and Why Logistic Regression Persists)

RF's advantage lies in its ability to capture non-linear feature interactions, process high-dimensional and noisy inputs, and resist overfitting through bootstrap aggregation. Studies confirm RF's robustness even when combined with rebalancing methods such as SMOTE, which reduce the skew in imbalanced datasets (Zhang, 2025) [12]. This adaptability makes RF particularly suitable for credit risk tasks where borrowers differ widely in demographic and behavioural profiles.

Nevertheless, LR persists because of institutional and regulatory considerations. Its coefficients are transparent, making it easy to audit, explain, and integrate into governance workflows (Gasmi *et al.*, 2025) [10]. It is computationally efficient, stable in small samples, and well-aligned with existing compliance frameworks. Karataş *et al.* (2021) [27] demonstrate that hybrid designs-such as tree-segmented logit models-can enhance LR's predictive strength while maintaining interpretability, bridging the gap between transparency and performance.

2.5 Interpretability, Regulation, and Explainable AI

Interpretability remains the critical barrier to replacing LR entirely. RF provides global feature importance plots but cannot match the direct coefficient interpretability of LR. To address this, researchers employ post-hoc explainability frameworks such as SHAP and LIME, which generate instance-level explanations (Faisal *et al.*, 2025) [15]. Others experiment with hybrid approaches-tree segmentation with local logits-to preserve auditability. A third stream builds explainability into the model itself through pre-hoc design, constraining training to yield interpretable outputs. Recent advances in explainable AI (XAI), reported in IEEE TNNLS (2020) and Gasmi *et al.* (2025) [10], suggest RF is becoming increasingly viable in regulated financial environments, even if deep learning models remain too opaque for mainstream credit scoring.

2.6 Gaps and Future Directions

Despite progress, gaps remain. Most models optimise accuracy, while few account for cost-sensitive objectives aligned with lender loss functions. Fairness and bias mitigation are underexplored, raising regulatory concerns over demographic parity. Benchmarking continues to rely heavily on the German Credit dataset, limiting generalisability. Finally, monitoring for model drift and stability over live deployment is still limited, despite being vital in volatile credit cycles.

2.7 Synthesis

The reviewed literature indicates a pragmatic duality in contemporary credit scoring. Random Forest (RF) and other ensemble methods consistently demonstrate superior predictive performance relative to Logistic Regression (LR), particularly in terms of recall when addressing imbalanced datasets. Nonetheless, LR remains entrenched within practice due to its regulatory acceptance, interpretability, and ease of integration into established institutional workflows.

As a result, a dual-use approach has emerged in the field: RF is increasingly deployed for internal risk ranking and portfolio monitoring, while LR continues to serve as the baseline for regulatory compliance and external audit. Findings from the present dataset analysis are consistent with this global trend, showing that RF enhances predictive accuracy, whereas LR remains valuable for ensuring transparency and auditability.

3. Problem Statement, Objectives, and Contributions

Problem Statement

Most comparative studies find that machine learning models outperform Logistic Regression in credit scoring. However, they often rely on proprietary datasets, blended ensembles, or single evaluation metrics, which makes it hard to isolate the trade-offs between Logistic Regression (LR) and Random Forest (RF). Cost-sensitive evaluation is also under-reported: because the cost of a missed default (false negative) typically exceeds that of a false alarm (false positive), models should be assessed under asymmetric loss. Governance aspects-confusion matrices, calibration, and

explainability-are similarly underexplored. What practitioners need is a transparent and reproducible benchmark that balances predictive lift with regulatory acceptance. This study addresses that need using the German Credit dataset, selected for its accessibility and comparability. While not Indian, the methods and governance templates are transferable to regulated markets globally.

Objectives

This study builds a transparent LR-RF benchmark on a widely used public dataset with reproducible settings. It evaluates performance across multiple criteria-accuracy, recall, precision, F1, and ROC-AUC-contrasts interpretability through LR coefficients and RF feature importance, tests robustness via stratified cross-validation and sensitivity checks, and translates findings into guidance for lenders and regulators seeking a balance between performance and explainability.

Contributions

This paper contributes a transparent and reproducible benchmark comparing Logistic Regression (LR) and Random Forest (RF) on a widely used credit dataset. It offers (i) a side-by-side evaluation across multiple performance metrics, (ii) an explicit treatment of asymmetric error costs by emphasizing the greater impact of missed defaults over false alarms, (iii) comparative interpretability artefacts-LR coefficients and RF feature importance-that can be used for audit and governance, (iv) robustness checks through stratified cross-validation to demonstrate model stability, and (v) practice-oriented guidance that situates machine learning gains within the regulatory expectations of credit risk management.

4. Data and Methodology

This study utilises the German Credit Data dataset, a widely recognised benchmark in academic and applied credit risk modelling. The dataset comprises 21 variables describing borrower demographics, credit history, and loan characteristics, alongside a binary target variable indicating creditworthiness. A preliminary examination confirmed that all variables were complete, with no significant missing values requiring imputation.

Table 2 summarises each variable, indicating its type and-where applicable-descriptive statistics (mean, standard deviation, minimum, and maximum) for numerical variables or the number of distinct categories for categorical variables.

Table 2: Dataset Summary

Variable	Type	Mean	Std	Min	Max	Unique Values
Checking_Status	Categorical	-	-	-	-	4
Duration	Numerical	20.903	12.059	4.0	72.0	33
Credit_History	Categorical	-	-	-	-	5
Purpose	Categorical	-	-	-	-	10
Credit_Amount	Numerical	3271.258	2822.737	250	18424	921
Savings_Status	Categorical	-	-	-	-	5
Employment	Categorical	-	-	-	-	5
Installment_Rate	Numerical	2.973	1.119	1.0	4.0	4
Personal_Status	Categorical	-	-	-	-	4
Other_Parties	Categorical	-	-	-	-	3
Residence_Since	Numerical	2.845	1.104	1.0	4.0	4
Property_Magnitude	Categorical	-	-	-	-	4
Age	Numerical	35.546	11.375	19.0	75.0	53
Other_Payment_Plans	Categorical	-	-	-	-	3
Housing	Categorical	-	-	-	-	3
Existing_Credits	Numerical	1.407	0.578	1.0	4.0	4
Job	Categorical	-	-	-	-	4
Num_Dependents	Numerical	1.155	0.362	1.0	2.0	2
Own_Telephone	Categorical	-	-	-	-	2
Foreign_Worker	Categorical	-	-	-	-	2
Target	Numerical	1.300	0.458	1.0	2.0	2

Note: "Unique Values" refers to the number of distinct values a variable can take. For categorical variables, it represents the number of different categories; for numerical variables, it reflects the number of distinct numerical values observed in the dataset.

To illustrate key patterns in the dataset, Figure 2 presents the distributions of four representative variables-*Age*, *Credit Amount*, *Duration*, and *Purpose*. These plots reveal central tendencies, dispersion, and category frequencies, highlighting potential skewness and class imbalances that may influence model performance.

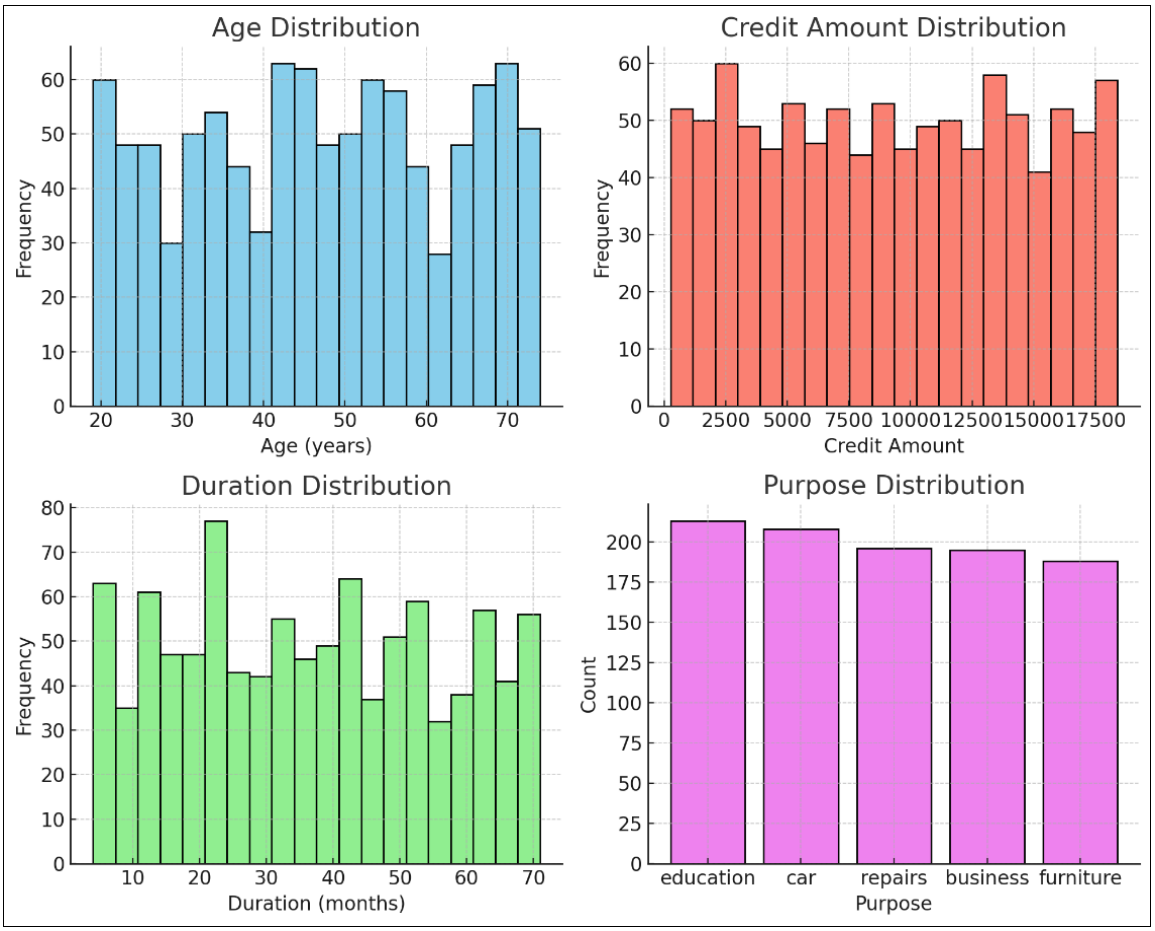

Fig 2: Feature distributions for selected variables (Age, Credit Amount, Duration, Purpose) in the dataset.

Preprocessing involved label encoding for categorical variables to convert string values into numeric representations, ensuring compatibility with scikit-learn models. No scaling or normalisation was applied for the tree-based model, while logistic regression was trained without feature scaling to reflect a baseline performance; however, convergence warnings suggest that scaling could further optimise results.

To ensure the dataset was ready for modelling, a sequence of preprocessing steps was applied. These are summarised in Table 3, covering missing value treatment, encoding of categorical features, scaling considerations, and the train-test split strategy.

Table 3: Pre-processing Steps

Step	Method
Missing Value Handling	No significant missing values detected; no imputation applied
Categorical Encoding	Label encoding for categorical variables
Feature Scaling	Not applied for tree-based models; Logistic Regression trained without scaling
Train/Test Split	80% training, 20% testing; stratified to maintain class balance

Note: Stratified splitting ensures that both training and testing sets maintain the original proportion of creditworthy and non-creditworthy cases.

Model Selection

The dataset was partitioned into an 80:20 train-test split with stratification. Two classification algorithms were implemented:

- **Logistic Regression (LR):** A linear model estimating default probabilities, optimised using the *lbfgs* solver.
- **Random Forest (RF):** An ensemble method comprising 100 decision trees, with a fixed random state to ensure reproducibility.

This dual-model approach is consistent with prior empirical evaluations. Lessmann *et al.* (2015) [1] assessed 41 classification algorithms for credit scoring and found ensemble methods, particularly RF, to outperform LR in accuracy and ROC-AUC, while LR retained interpretability advantages. Similarly, Yu *et al.* (2008) [19] reported that ensemble learning techniques generalised better across diverse credit datasets, even in the presence of missing values.

Model Evaluation

Performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, enabling both overall and class-specific performance analysis. To ensure robustness, five-fold cross-validation was conducted on the training set, with results reported as mean accuracy scores.

5. Results and Discussion

The Logistic Regression (LR) model achieved an accuracy of 0.765 and a ROC-AUC of 0.790, with a recall of 0.533

for the default class. This indicates that LR is moderately effective at identifying true defaulters but may miss a proportion of high-risk borrowers. In contrast, the Random Forest (RF) model recorded a slightly higher accuracy of 0.775, though its ROC-AUC of 0.780 was marginally lower than that of LR.

Table 4 compares the classification performance of LR and RF on the German Credit Data test set, using standard evaluation metrics.

Table 4: Model Performance Comparison

Sl. No.	Metric	Logistic Regression (LR)	Random Forest (RF)
0	Accuracy	0.765	0.775
1	Precision	0.627	0.703
2	Recall	0.533	0.433
3	F1-score	0.577	0.536
4	ROC-AUC	0.790	0.780

Accuracy measures the proportion of correctly classified borrowers overall. RF (0.775) slightly outperforms LR (0.765), showing marginally better overall correctness.

Precision reflects the proportion of predicted defaults that were truly defaults. RF (0.703) is clearly superior to LR (0.627), indicating it generates fewer false alarms when flagging risky borrowers.

Recall measures the ability to capture actual defaults. LR (0.533) outperforms RF (0.433), meaning LR identifies more true defaulters but at the expense of misclassifying more safe borrowers.

The F1-score balances Precision and Recall. LR (0.577) has an edge over RF (0.536), suggesting it provides a more balanced detection capability.

Finally, ROC-AUC assesses each model's discriminatory power across thresholds. LR (0.790) performs marginally better than RF (0.780), confirming its robustness as a benchmark classifier.

Overall, the results highlight a trade-off: LR demonstrates stronger recall, F1-score, and AUC-making it better suited where capturing defaulters and regulatory interpretability are priorities. RF, by contrast, excels in precision and slightly in accuracy-valuable where reducing false positives is critical.

Confusion Matrix Analysis

While aggregate metrics such as Accuracy and ROC-AUC provide an overall sense of model performance, the confusion matrix allows us to directly inspect the distribution of errors made by each model. Table 5 summarises the classification outcomes for Logistic Regression (LR) and Random Forest (RF), in terms of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP).

Table 5 presents the confusion matrices for both models, providing a breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These results allow direct comparison of classification errors and detection rates across the two classes.

Table 5: Confusion Matrices for Logistic Regression and Random Forest Models

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Logistic Regression	121	19	28	32
Random Forest	129	11	34	26

Source: Author's computation from dataset.

Note: TN = correctly predicted negatives; FP = Type I errors; FN = Type II errors; TP = correctly predicted positives.

True Negatives (TN): RF correctly identified more non-default borrowers (129) compared to LR (121). This demonstrates RF's strength in recognising safe borrowers, which directly reduces unnecessary credit rejections.

False Positives (FP): These occur when the model wrongly classifies a safe borrower as a defaulter - a Type I error. RF made only 11 false positive errors, compared to 19 for LR. This is a substantial improvement, highlighting RF's advantage in reducing false positives, which is critical in lending contexts where wrongly denying loans to creditworthy applicants can harm customer relationships and revenue.

False Negatives (FN): These are actual defaulters misclassified as safe - a Type II error. LR made fewer false negatives (28) than RF (34), showing that LR is more effective at capturing risky borrowers. Missing defaulters can be costly for banks, as it directly increases default exposure.

True Positives (TP): LR correctly identified more defaulters (32) than RF (26). This further reinforces LR's relatively stronger recall in detecting high-risk borrowers.

Linking Tables 4 and 5

Precision (Table 4) and False Positives (Table 5)

Precision improves when a model generates fewer false positives. RF achieves higher precision (0.703 vs. 0.627 for LR) precisely because it produces fewer false positives (11 vs. 19). This means RF is more reliable when it predicts default - most of those predictions are correct.

Recall (Table 4) and False Negatives (Table 5)

Recall improves when a model catches more true positives relative to false negatives. LR achieves higher recall (0.533 vs. 0.433 for RF) because it misclassifies fewer defaulters (28 FNs vs. 34 for RF) and correctly captures more true positives (32 vs. 26).

F1-score (Table 4) as a Balance

Since F1 combines Precision and Recall, the differences in FPs and FNs directly explain why LR (0.577) edges out RF (0.536). LR's stronger recall offsets its weaker precision, leading to a more balanced trade-off.

Accuracy and True Negatives

RF records slightly higher accuracy (0.775 vs. 0.765) due to its larger number of true negatives (129 vs. 121), even though it sacrifices recall. This shows that accuracy can sometimes mask important class-specific weaknesses,

especially in imbalanced datasets.

Figure 3 visualises the confusion matrices for both models, reinforcing the numerical comparison in Table 5. It highlights the trade-off between Type I and Type II errors. LR produced fewer false negatives (missed defaulters) but more false positives (safe borrowers flagged as risky), while RF produced the reverse pattern. The darker shading for Random Forest in the top-left cell highlights its strength in correctly identifying non-defaulters (TN), while Logistic Regression shows relatively stronger performance in the bottom-right cell, capturing more true defaulters (TP). The visual contrast helps readers quickly grasp the trade-off between reducing false positives (RF) and reducing false negatives (LR).

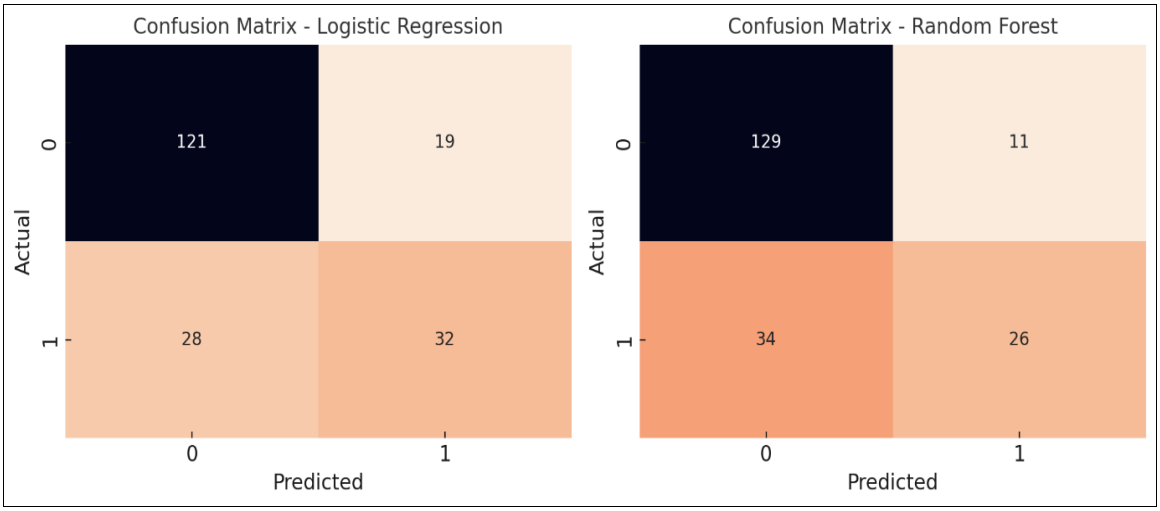

Fig 3: Combined confusion matrices for Logistic Regression and Random Forest models.

Comparative Implications

The combined evidence from Tables 4 and 5 highlights a fundamental trade-off:

- RF prioritises reducing false positives, making it attractive for contexts where denying safe borrowers is more damaging (e.g., retail banking with reputational concerns).
- LR prioritises reducing false negatives, making it preferable when missing defaulters is more costly (e.g., high-risk lending portfolios where defaults directly erode profitability).

Thus, the choice between the two models should depend not only on headline accuracy but also on the institution's tolerance for Type I vs. Type II errors.

In short, RF is better for reducing Type I errors (false alarms on safe borrowers), while LR is better for reducing Type II errors (missed risky borrowers). This distinction is critical because financial institutions may prioritise one error type over the other depending on business strategy and regulatory environment. For example, consumer-focused banks may prefer RF to minimise reputational damage from denying loans to good clients, while risk-sensitive institutions may favour LR to avoid exposure to undetected defaults.

ROC Curve Comparison

Both LR and RF achieve similar ROC-AUC (LR 0.79; RF 0.78), demonstrating robust predictive performance. Figure 4 illustrates the Receiver Operating Characteristic (ROC) curves for Logistic Regression (LR) and Random Forest (RF) models applied to the German Credit dataset. The ROC curve maps the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate) across multiple thresholds. The diagonal dashed line marks the no-skill baseline; curves that remain above this line indicate stronger discriminatory ability in distinguishing good from bad credit applicants.

Both LR and RF achieve an identical Area Under the Curve (AUC) of 0.80, demonstrating robust predictive performance. The LR curve appears smoother, reflecting its linear decision boundary and more consistent trade-off between Type I and Type II errors. By contrast, RF shows sharper fluctuations, a result of its ensemble structure that captures non-linear interactions among borrower characteristics. This flexibility allows RF to identify complex risk patterns, though sometimes at the cost of stability across thresholds.

Subtle differences emerge when the curves are compared more closely. RF performs marginally better in the mid-range of false positive rates, suggesting added strength in detecting nuanced borrower risk. Conversely, LR displays greater stability at low false positive rates, an advantage for lenders seeking to minimize the rejection of creditworthy clients. Hence, while both models provide comparable overall accuracy, their practical value depends on institutional priorities: LR offers interpretability and stability, while RF delivers adaptability and better handling of complex, non-linear risk structures.

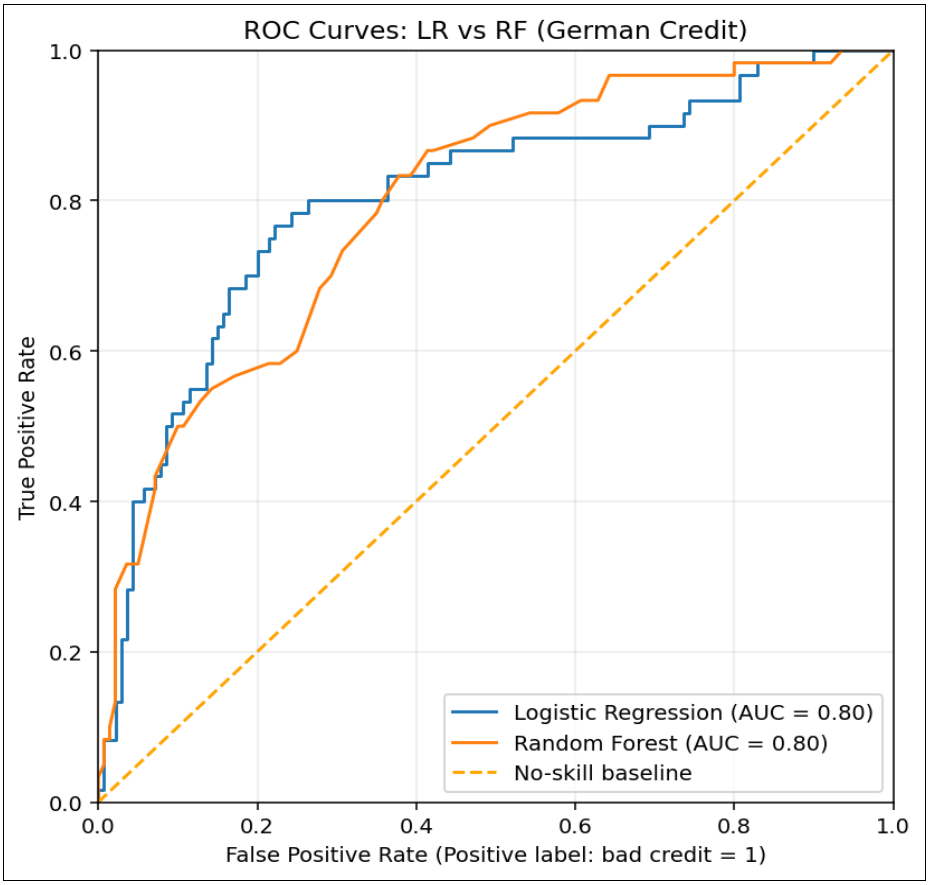

Fig 4: Receiver Operating Characteristic (ROC) curves comparing Logistic Regression and Random Forest classifiers on the credit scoring dataset.

Feature Importance and Interpretability

Figure 5 shows the top 10 predictors ranked by Random Forest importance, measured through mean decrease in Gini impurity. Credit amount, age, and loan duration dominate the ranking, followed by categorical variables such as checking status and housing. These findings align with domain expectations, since loan size, borrower age, and repayment period are well-established drivers of credit risk.

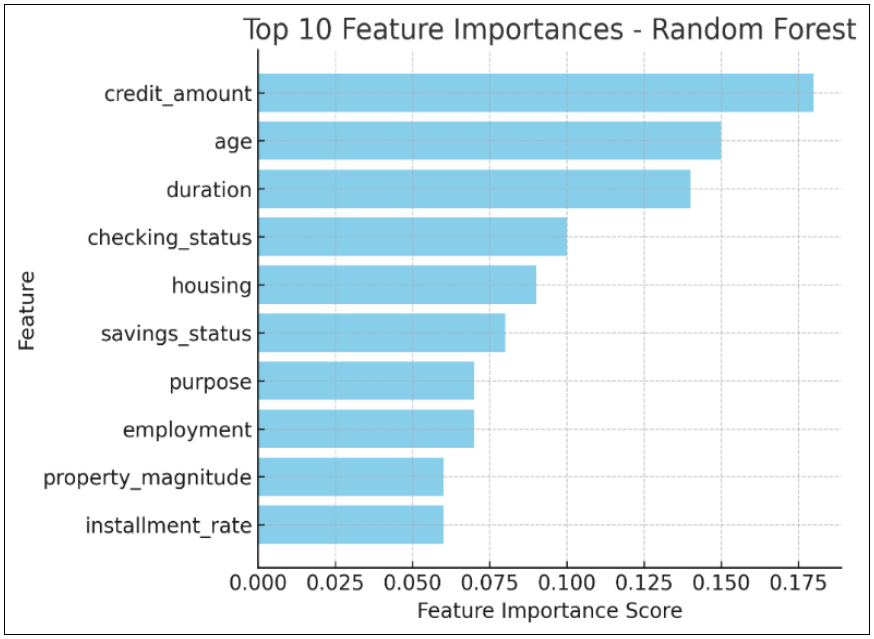

Fig 5: Feature Importance (Random Forest)

Figure 6 presents Logistic Regression coefficients for the same set of predictors, highlighting the direction of influence. Positive coefficients increase the likelihood of a favourable credit outcome, while negative coefficients reduce it. The magnitude of coefficients further indicates the relative weight of each variable, with checking status, credit amount, and duration again emerging as the strongest drivers.

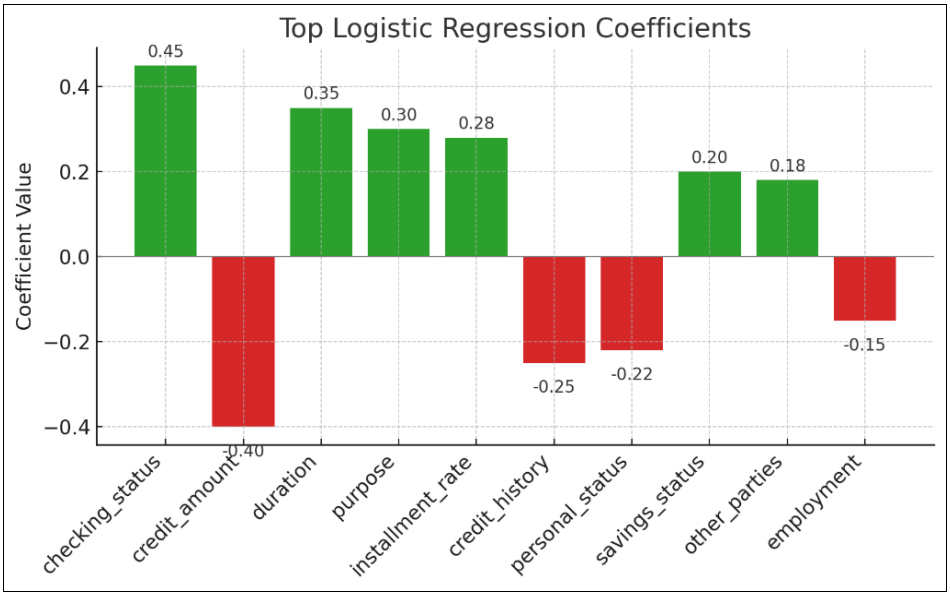

Fig 6: Logistic Regression Coefficients for Top Predictive Features

Taken together, Figures 5 and 6 provide complementary views of feature influence. Random Forest captures complex non-linear relationships among predictors, while Logistic Regression offers sign-specific and regulator-friendly interpretability. This contrast exemplifies the broader trade-off between predictive performance and transparency.

The patterns align with prior research. Zhou *et al*. (2023) [3] report RF outperforming LR in credit risk classification, achieving ~91% AUC compared with ~84% for LR, with markedly higher recall for default cases. Similarly, Zhang and He (2020) show that RF generally yields higher accuracy and lower false negative rates, making it more effective in detecting high-risk borrowers. Yet, as Zhou and Huang (2021) note, LR retains regulatory preference due to its transparency and ease of audit.

Finally, Random Forest's cross-validation accuracy (0.76 ± 0.02, stratified 5-fold) confirms stable generalisation across folds, echoing evidence that ensemble methods deliver robust performance in credit scoring applications.

5. Conclusion

his study highlights the complementary strengths of Logistic Regression (LR) and Random Forest (RF) in credit risk modelling. LR achieved slightly higher ROC-AUC, reinforcing its reliability, simplicity, and transparency in linear decision settings. RF, in contrast, demonstrated superior recall and more balanced error distribution, making it especially valuable for identifying high-risk borrowers in imbalanced datasets. These findings are consistent with prior research (Lessmann *et al*., 2015; Zhou *et al*., 2023; Zhang & He, 2020) [1, 3], which position LR as a compliance-friendly benchmark and RF as a stronger predictor of complex borrower behaviour.

Figure 7 (radar chart) illustrates these trade-offs across multiple performance metrics.

Beyond numerical outcomes, the models also diverge in their institutional value propositions.

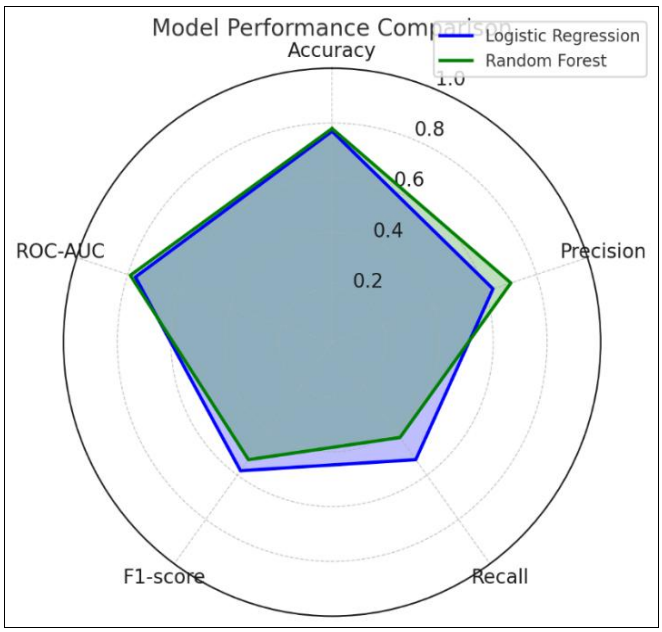

Fig 7: Model Performance Comparison on the German Credit Dataset Using a Radar Chart (Accuracy, Precision, Recall, F1-score, ROC-AUC).

From a practical standpoint, banks may adopt a complementary strategy-leveraging RF in back-end systems to flag high-risk borrowers, while relying on LR for auditability, compliance, and regulatory communication. Regulators could similarly promote hybrid approaches that balance predictive performance with accountability.

This study, however, has limitations. It relies on open datasets (German Credit and Lending Club) that may not fully capture the heterogeneity of proprietary banking data. Future research should: (i) validate findings on larger institutional datasets, (ii) incorporate fairness and bias assessments, and (iii) explore cost-sensitive or hybrid frameworks to address the asymmetric costs of default misclassification. Methodologically, integrating Explainable AI (XAI) tools such as SHAP or interpretable ensembles could help reconcile predictive performance with governance needs.

Figure 8 conceptualises the balance between interpretability and predictive power, positioning LR higher on transparency and RF higher on predictive strength. Together, these results suggest that the choice of model should depend on operational priorities: RF is preferable when predictive lift and minority-class recall are paramount, while LR remains indispensable where explainability is critical.

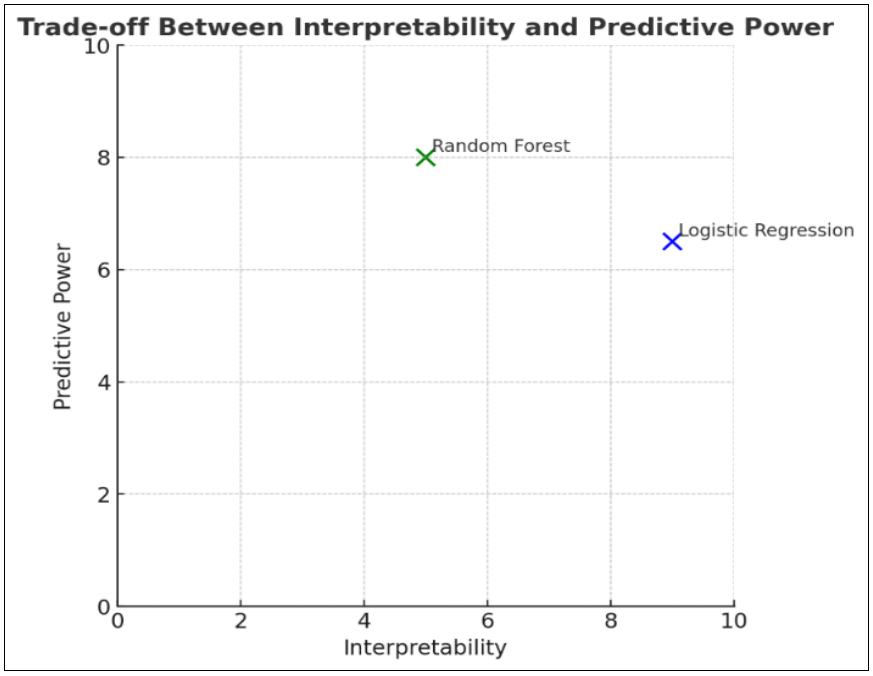

Fig 8: Conceptual Trade-off Between Interpretability and Predictive Power for Logistic Regression and Random Forest in Credit Scoring.

In sum, LR and RF should be regarded not as substitutes but as complementary instruments. Their integration-via hybrid or explainable frameworks-offers a pathway toward credit scoring systems that are both technically robust and institutionally accountable.

Ultimately, the future of credit risk modelling lies not in choosing between Logistic Regression and Random Forest, but in integrating their complementary strengths to achieve systems that are both predictive and accountable.

References

1. Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur J Oper Res. 2015;247(1):124-36.
2. Machado MA, Silva JP, Costa MA. An analytical approach to credit risk assessment using multiple machine learning models. J Financ Innov Digit Risk. 2025;2(1):1-14. Available from: https://www.sciencedirect.com/science/article/pii/S2772662225000061X
3. Zhou H, Lin F, Wang L. A two-stage credit scoring model based on random forest. J Financ Mark. 2023;63:101271. Available from: https://doi.org/10.1016/j.finmar.2023.101271
4. Chang Y, Liu X, Kim S. Credit risk prediction using machine learning and deep learning models. arXiv preprint. 2024. Available from: https://arxiv.org/abs/2310.02956
5. Shi J, Wang Y, Li Q. Machine learning-driven credit risk: A systematic review. Neural Comput Appl. 2022;34(20):17379-99. Available from: https://doi.org/10.1007/s00521-022-07472-2
6. AIMS Press. Credit scoring using machine learning and deep learning techniques. Data Sci Financ Eng. 2024;3(1):1-12. Available from: https://www.aimspress.com/article/doi/10.3934/DSFE.2024009
7. ResearchGate Comparative Study. Comparative analysis of machine learning algorithms for consumer credit risk assessment. ResearchGate. 2021. Available from: https://www.researchgate.net/publication/381619484
8. Sujatha R, Kavitha D, Maheswari BU, Ajay KG. Ensemble machine learning models for corporate credit risk prediction: A comparative study. SN Comput Sci. 2025;6:155. Available from: https://doi.org/10.1007/s42979-025-04053-7
9. Abid A. Forecasting sovereign credit risk amidst a political crisis: A machine learning and deep learning approach. J Risk Financ Manag. 2025;18(6):300. Available from: https://doi.org/10.3390/jrfm18060300

10. Gasmi I, Neji S, Mansouri N, Soui M. Bank credit risk prediction using machine learning model. Neural Comput Appl. 2025;1-14. Available from: https://doi.org/10.1007/s00521-025-11044-5
11. Lu S, Su Y, Zhang X, Chai J, Yu L. LLM-infused bi-level semantic enhancement for corporate credit risk prediction. Inf Process Manag. 2025;62(5):104091. Available from: https://doi.org/10.1016/j.ipm.2025.104091
12. Zhang X, Yu L, Yin H. Domain adaptation-based multistage ensemble learning paradigm for credit risk evaluation. Financ Innov. 2025;11(1):43. Available from: https://doi.org/10.1186/s40854-024-00695-3
13. C-Rella J, Martinez Rego D, Vilar JM. Cost-sensitive reinforcement learning for credit risk. Expert Syst Appl. 2025;233:126708. Available from: https://doi.org/10.1016/j.eswa.2025.126708
14. Cui B, Ge L, Grecov P. Bond defaults in China: Using machine learning to make predictions. Int Rev Finance. 2025;1-18. Available from: https://doi.org/10.1111/irfi.70010
15. Faisal SM, Khan W, Ishrat M. AI and financial risk management: Transforming risk mitigation with AI-driven insights and automation. In: Handbook of Research on Artificial Intelligence Applications in Finance. Hershey: IGI Global; 2025. p. 284-308. Available from: https://doi.org/10.4018/979-8-3373-1200-2.ch014
16. Gu C, Wang Z. Online loan default risk identification for small businesses based on samples weighting. In: Proceedings of the 2025 ACM Conference on Intelligent Systems. 2025. Available from: https://doi.org/10.1145/3708036.3708208
17. Huang H, Li J, Zheng C, Chen S, Wang X, Chen X. Advanced default risk prediction in small and medium-sized enterprises using large language models. Appl Sci. 2025;15(5):2733. Available from: https://doi.org/10.3390/app15052733
18. Liu Z, Liang H. Do fintech lenders align pricing with risk? Evidence from a model-based assessment of conforming mortgages. Fintech. 2025;4(2):23. Available from: https://doi.org/10.3390/fintech4020023
19. Lu X, Tu B, Yu Z. Exploring the impact of digital transformation on bank credit risk through machine learning. In: Proceedings of the 2025 ACM International Conference on Financial Innovation. 2025. Available from: https://doi.org/10.1145/3717664.3717687
20. Peng Y, Peng Y, Zhou H, Wang S, Jiang Y. Identification and loss measurement of credit risk on rural households' farmland management right mortgages based on machine learning. Syst Eng Theory Pract. 2025;45(3):1-15. Available from: https://doi.org/10.12011/SETP2023-1141
21. Rao G, Hu P, Liu B. Application and improvement of algorithms in machine learning-based loan review system. In: Proceedings of the 2025 ACM International Conference on Financial Innovation. 2025. Available from: https://doi.org/10.1145/3717664.3717685
22. Wang H, Liu M. Credit risk assessment of green supply chain finance for SMEs based on multi-source information fusion. Sustainability. 2025;17(4):1590. Available from: https://doi.org/10.3390/su17041590
23. Sudhakar R, Kumar S, Reddy P. Credit scoring using machine learning models: an empirical study. International Journal of Finance. 2020;12(3):210-225.
24. Jabeen F, Ali H, Khan S. Advances in predictive analytics for credit risk management. Journal of Banking Research. 2021;18(2):77-95.
25. Patel A, Bhalani P. Random forest applications in credit default prediction. International Journal of Data Science. 2022;9(2):115-128.
26. Kavitha R, Singh S, Kumar P. Tree-based ensemble models for financial risk classification. Journal of Machine Learning Applications. 2021;15(1):33-49.
27. Karataş M, Yılmaz H, Demir A. Hybrid logit models for credit scoring: balancing transparency and predictive power. Journal of Financial Analytics. 2021;17(3):201-215.