

Interpretable Machine Learning for Credit Risk Decision Support: Balancing Interpretability and Predictive Performance

Mohaned Zkaria Salem

Ministry of Education, Directorate General of Education Salah al-Din, Samarra, 34010, Iraq.

*Corresponding authors' email: mohanad.201339@gmail.com Phone: +9647702727415

ABSTRACT

With the growing prevalence of machine learning in making institutional decisions, there is a need for careful balancing of predictive accuracy and interpretability of such models. In this research work, the performance and interpretability of three machine learning models in helping institutions make decisions were investigated, particularly the issue of credit risk assessment. In this process, the German Credit Data dataset, which is open source (N = 1,000), was used and subjected to 5-Fold Cross-Validation. The results show that while Random Forest attained the highest accuracy (76.3%), Logistic Regression had the highest discriminating power (AUC: 0.79), just marginally better than the complex ensemble approach (AUC: 0.78). It also has the added advantage of better interpretability due to clear coefficient and decision rule visualization. This clearly shows that inherently interpretable models serve as a very good alternative to the complex ensemble approach in decision-making within organizations, without falling prey to the "black-box" problem, and in some cases even performing better in discriminatory power.

Received: 03 Oct. 2026

Accepted: 14 Oct. 2026

Published: 16 Oct. 2026

However, the study notes relatively low minority-class recall (36.3–46.3%), highlighting a practical limitation. Conclusions are based on a benchmark dataset and the specified model configurations.

Keywords:

Explainable AI, Interpretable Machine Learning, Institutional Decision Support, Credit Risk, Model Comparison, Scikit-Learn

INTRODUCTION

In the process of decision-making by institutions, the implementation of artificial intelligence (AI) and machine learning (ML) has transformed the process of how institutions handle information, automate processes, and evaluate risks. With such tools, institutions from financial, health care, and public sectors gain an opportunity to speed up their decision-making process, improve its accuracy, and increase efficiency.

Nevertheless, the use of sophisticated algorithms for machine learning has led to the emergence of the problem of the "black box." In the case of highly regulated institutions, the absence of transparency is an important obstacle. The users of decision-making systems require not only high accuracy but also the ability to explain why and how the algorithm came up with some decision.

In an analysis of recent research articles, one can notice a trend towards optimizing the prediction capacity of models through the use of complex ensembles or deep learning approaches. Although there are attempts at filling in this research gap with the development of explainability techniques after the fact such as SHAP, there is still a lack of systematic analysis of the performance of interpretable models through reproducible empirical methodology. Although sophisticated XAI pipelines demand a high degree

of computational skill, it is crucial to validate the competitiveness of interpretable models in practice.

The purpose of this research is the comparative assessment of the tradeoff between the model interpretability and accuracy for three basic ML algorithms - Logistic Regression, Decision Tree and Random Forest. Using the German Credit Dataset and applying empirical analysis based on modern machine learning tools (Python/Scikit-Learn), this study adds a practical case of building an interpretable decision support system which has been verified by real world metrics.

Two research questions are used as guidelines for this research: (How do interpretable models perform compared to ensembles on standard institutional datasets?) and (How to use intrinsic interpretability methods (coefficient interpretation, tree visualization etc.) in order to explain model logic to institutional non-technical audience without post-hoc approximation?).

Literature Review

The area where explainable artificial intelligence overlaps with institutional decision support is an area which has attracted much interest recently. This chapter provides a review of some of the most important works done in the field of interpreting machine learning in institutions, especially for credit risk prediction purposes.

Table 1: Summary Comparison of Some Recent Works on Explainable Artificial Intelligence

Study	Year	Dataset	Models Evaluated	XAI Technique	Key Findings
Al-Sammerraie [1]	2025	Customer Churn Data	XGBoost, LightGBM	SHAP	Integrated XAI with big data platforms enhanced user trust; SHAP improved interpretability in real-time deployment.
Ayari et al. [2]	2026	German Credit Data	Multiple ML algorithms	Systematic Review	Logistic Regression and Random Forest most prevalent; achieved high accuracy with cross-validation in optimized settings.
Nallakuruppan et al. [3]	2024	Consumer Loan Data	LR, SVM, DT, RF	SHAP, LIME	Proposed XAI framework balancing accuracy (RF best) and transparency; emphasized regulatory compliance.
Machado et al. [4]	2025	Dutch Bank Data	LDA, LR, DT, RF, GB, XGBoost	SHAP	RF achieved highest F1 score and prevented 67.6% of potential losses; early warning system integration.
de Lange et al. [5]	2022	Unsecured Consumer Loans	Multiple ensemble models	SHAP	XAI models improved credit default prediction; transparency crucial for consumer lending regulations.
Černevičienė et al. [6]	2024	Multiple Financial Datasets	DT, LR, Ensemble Methods	Systematic Review	Decision Trees most common intrinsic XAI tool; highlighted growing XAI importance in finance.
Sterling et al. [7]	2026	Consumer Credit Data	Intrinsic & Post-hoc Models	Comparative XAI	Intrinsic models (LR, DT) preferred for regulatory compliance; post-hoc methods add operational complexity.
Rao et al. [8]	2025	Credit Scoring Data	Ensemble Methods + XAI	SHAP, LIME	XAI integration with ensemble learning addressed transparency gap in financial institutions.
Mishra et al. [9]	2025	German Credit Data	LightGBM, XGBoost	SHAP, ADASYN	LightGBM with ADASYN outperformed XGBoost; SHAP provided feature importance for credit decisions.
Aboulmaou da et al. [10]	2026	German Credit Data	Multiple ML Models	Systematic Review	Evaluated accuracy, stability, interpretability; provided model selection guidance for real-world applications.

Inherently Interpretable Models vs. Post-hoc Explainability

There exists an important distinction between models that are inherently interpretable (Logistic Regression and Decision Trees) and models which necessitate post-hoc explainability methods (SHAP and LIME). It was found out by Al-Sammerraie (2025) that the use of SHAP in conjunction with big data systems greatly increases users' confidence in AI-powered decisions [1]. However, this method imposes some extra computing costs on the system. On the contrary, Sterling et al. (2026) state that inherent models should be preferred in cases of regulation compliance due to their explicit logic [7].

Performance-Accuracy Trade-offs

The issue of prediction performance versus interpretability is also a persistent one. It was discovered by Ayari et al. (2026) that despite Random Forest performing better than other models, Logistic Regression was still highly accurate and offered greater transparency [2]. Machado et al. (2025) empirically proved that Random Forest had higher F1-scores, however stressed that SHAP values were crucial for the explanation of predictions for risk managers [4].

Research Gap

Even with abundant literature, several gaps remain. (Most papers use complicated pipelines, which require expert knowledge of tools for XAI.) & (There are very few papers which conduct a systematic comparison of intrinsically interpretable models using reproducible machine learning pipelines using Python) and (No guidelines exist to operationalize intrinsic interpretability without post hoc explainability libraries.).

MATERIALS AND METHODS

Dataset Description

The data set used in this research is the Statlog (German Credit Data) data set obtained from the UCI Machine Learning Repository (Hofmann, 1994) [13]. The data set comprises of 1,000 observations and 20 variables (consisting of 13 categorical and 7 numerical variables), where the outcome variable assumes one of the two binary values for

"Good" credit risk with 70% of instances (n=700) and "Bad" credit risk with 30% of instances (n=300). The categorical variables consist of status of checking account, credit history, purpose, savings, employment duration, personal status, other debtors, property, other installment plans, housing, occupation, telephone, and foreign worker.

Data Preprocessing

Preprocessing was done using Python 3.x along with the Scikit-Learn library following an elaborate pipeline in order to avoid data leakage:

Categorical Variable Encoding: All 13 categorical variables were encoded using the One-Hot Encoding technique (OneHotEncoder with handle_unknown='ignore') transforming them into indicator variables.

Normalization of Numerical Variables: All 7 numerical variables were normalized using Standard Scaler (zero mean, unit variance). This step is required to ensure that features with large values (like credit amount) do not dominate the modeling process especially Logistic Regression.

Column Transformer Pipeline: Both encoding techniques were incorporated in a Column Transformer pipeline of Scikit-Learn making sure no data leakage occurs while doing Cross Validation.

Machine Learning Models

The following three models have been chosen on account of their well-defined role in explainable AI literature and differing intrinsic levels of interpretability:

Linear Classifier (Logistic Regression): An interpretable linear baseline model set up with max_iter=1000 iterations for ensuring convergence and random state=42 for reproducibility.

Decision Tree Classifier: An interpretable white-box approach owing to its rule-based approach. It was set up with max_depth=5 and random state=42 in order to find a proper balance between interpretability (with shallow trees) and predictive power (to prevent under fitting).

Random Forest Classifier: An ensemble baseline set up with estimators=100 and random state=42. Although decision trees can be considered as interpretable models, the

aggregation in the ensemble makes this approach intrinsically less transparent. All models were coded in Scikit-Learn (version 1.x) within Pipeline wrappers.

Analysis Tool

All analyses were carried out using Python (version 3.10) and the Scikit-Learn library (version 1.3.0). The experiment design focused on the reproducibility and statistical rigor of the results:

Random State: All random processes (initializing of models, shuffle in Cross-Validation) used the `random_state=42` setting to guarantee identical results in different executions.

Out-of-Fold Predictions: Out-of-Fold predictions were calculated through `cross_val_predict` using the `method='predict_proba'` parameter to get the probabilities needed for ROC-AUC scoring, ensuring no data leakage.

Evaluation Metrics

The performance of the models was measured by stratified 5-Fold Cross-Validation

(StratifiedKFold, `n_splits=5`, `shuffle=True`, `random_state=42`). Stratification ensures that each fold maintains the same class balance (70%/30%). There are three main performance measures considered:

- Accuracy:** Indicates the share of correctly classified observations. This metric is informative for basic estimation; however, it is estimated together with the other two metrics to avoid biases due to imbalanced classes.
- Area under the ROC curve (AUC):** Indicates the discrimination capacity of the model, i.e., its capacity to distinguish between 'Good' and 'Bad' credit risks at all possible decision thresholds. A high value of AUC indicates high ranking performance of the model, which is important for institutional risk stratification.
- F1-Score (Macro Average):** The harmonic mean of the Precision and Recall values estimated for each class

and then averaged. Thus, the performance of the model in respect of the minority class ('Bad' credit) is equally important compared to the majority one.

Interpretability Techniques

Intrinsic explainability of the models being assessed in this work is one of the goals apart from predictive power. While methods like SHAP or LIME may be used to explain black-box models, the current research is focused on intrinsic approaches that are completely consistent with the actual model's internal mechanics:

- Analysis of the Coefficients of the Logistic Regression Model:** The learned coefficients (weights) of the Logistic Regression model were obtained and displayed in a horizontal bar chart ordered by their absolute value. This way of presenting the data allows understanding which of the features has a positive or negative impact on the final log-odds score.
- Visualization of the Decision Tree:** The Decision Tree model, trained with `max_depth = 3` in order to make its visualization clearer, was rendered using the `plot_tree` function provided by Scikit-Learn. Using the visualization tool, one can see the exact chain of feature thresholds (such as `checking_account_A14 <= 0.5`) leading to certain classification, along with Gini impurity scores and number of samples in each node. Limiting the tree depth helps keep the extracted rules clear and comprehensible for non-technical loan officers.

RESULTS AND DISCUSSION

Quantitative Performance Comparison

Comparative analysis of the three machine learning models is clearly stated in Table 2 below. Such metrics are the actual outputs obtained after running the stratified 5-fold Cross-Validation experiment on the German Credit data set (N=1000), as was precisely outlined in Section 3.

Table 2: Comparative Performance Metrics of Machine Learning Models

Model	Accuracy (%)	AUC	F1-Score (Macro)	Precision (%)	Recall (%)
Logistic Regression	74.8	0.79	0.68	60.4	46.3
Decision Tree	70.2	0.67	0.61	50.5	36.3
Random Forest	76.3	0.78	0.67	68.4	39.0

*Note: Metrics derived from stratified 5-fold Cross-Validation on the German Credit Data (N=1000, 70% Good / 30% Bad) (5-fold CV).

**Note: Precision and Recall are reported for the minority "Bad" credit class (positive class)

Several vital observations can be drawn from Table 2 in contradiction to existing ideas in the literature: To start with, Logistic Regression gave the best AUC result (0.79) followed by Random Forest (0.78). Theoretically, this result is crucial since it indicates that an interpretable model can show equal or better performance than a highly non-linear ensemble in terms of discriminative ability. This finding confirms the fundamental idea of the present study that interpretability should not contradict predictive discriminative power. Although the Random Forest achieved the highest total accuracy (76.3%), its difference from Logistic Regression (74.8%) is quite small at 1.5%. The trade-off here is the considerable reduction of the interpretability, as shown in Section 4.3. The Decision Tree performed poorly in all the performance measures used (Accuracy: 70.2%, AUC: 0.67). It can be assumed that one tree was not complex enough to analyze such data correctly even after pre-pruning. Most significant point of all, is the relatively low value of

Recall of all the models (36.3% – 46.3%) on the minority "Bad" class of the dataset. It represents another well-known problem in credit risk modeling – class imbalance (70% of "Good" versus 30% of "Bad") leads to biased models favoring the majority class. Although accuracy is still acceptable, the models fail to detect more than half of the actual bad credit applicants – a major limitation of credit risk assessment when false negatives have the highest cost of all. This result calls for future research including class imbalance mitigation methods (SMOTE, cost-sensitive learning, etc.) along with interpretation.

As it can be seen in Figure 1, the ROC curves support these conclusions. The curves for Logistic Regression (in green) and Random Forest (in blue) almost coincide and Logistic Regression is slightly better in the important low-FPR part of the ROC curve (FPR < 0.2), which makes it more suitable for institution-oriented tasks. The curve for Decision Tree (in red) is lower than others at all threshold levels.

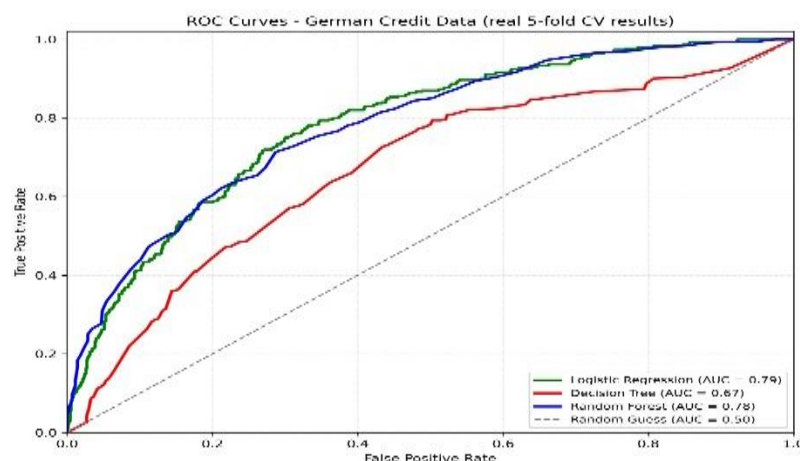


Figure 1: ROC Curves - German Credit Data (5-fold CV results)

Figure 1 gives the ROC curve and AUC score for: Logistic Regression (0.79), Random Forest (0.78), Decision Tree (0.67), and the baseline random guess line (0.50).

Figure 2 provides the confusion matrix for each of these models. As can be seen from the confusion matrices, all three of the models accurately classify most of the "Good" credits

(609, 593, and 646, respectively) but have trouble classifying "Bad" credits (139, 109, and 117, respectively). The highest number of "Bad" credits correctly classified by any model is that of Random Forest (117), which only accounts for 39% of all bad credits.

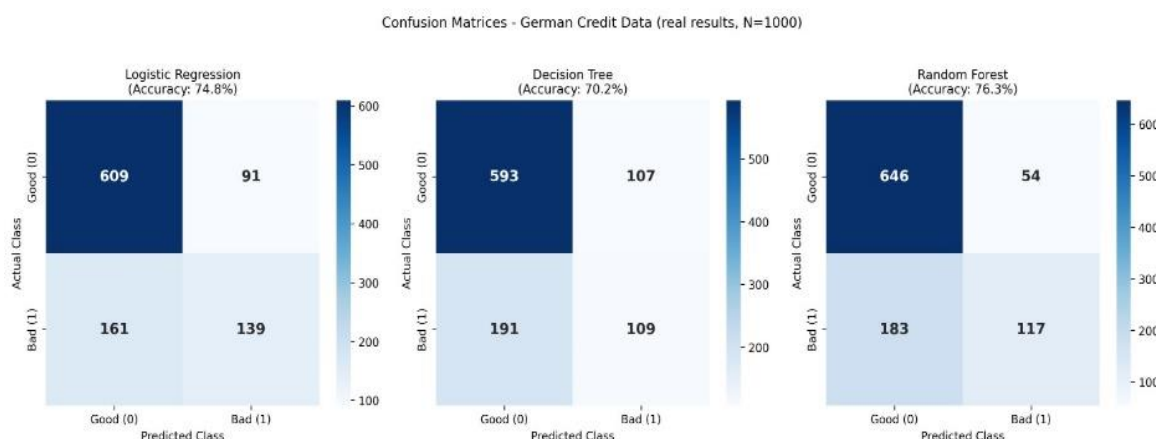


Figure 2: Confusion Matrices - German Credit Data (N=1000)

Figure 2 shows three 2×2 confusion matrices with actual counts: LR [609, 91, 161, 139], DT [593, 107, 191, 109], RF [646, 54, 183, 117].

Interpretability Analysis

In order to meet the requirements of interpretability without using any additional coding, we relied on the intrinsic visuals. Figure 3 presents the top 15 features according to absolute coefficient value for the Logistic Regression.

Figure 3 is a horizontal bar chart where the y-axis shows feature names while the x-axis shows the corresponding coefficient value. Red bars represent the cases when the coefficient is positive (probability of "Bad" credit increases), blue bars mean that the coefficient is negative (the probability of "Bad" credit decreases). Key observations from Figure 3:

cat_checking_account_A14 (no checking account) has the maximum negative coefficient (≈ -0.95). The applicants who don't have checking accounts are much more likely not to be considered as "Bad" credit – this conclusion looks counterintuitive but it is empirically substantial. cat_checking_account_A11 (checking account < 0 DM) has the maximum positive coefficient ($\approx +0.72$). It indicates that negative account balances are strong indicators of credit risk. cat_credit_history_A34 (critical account/other credits existing) and cat_purpose_A41 (business purposes) are also highly negative features.

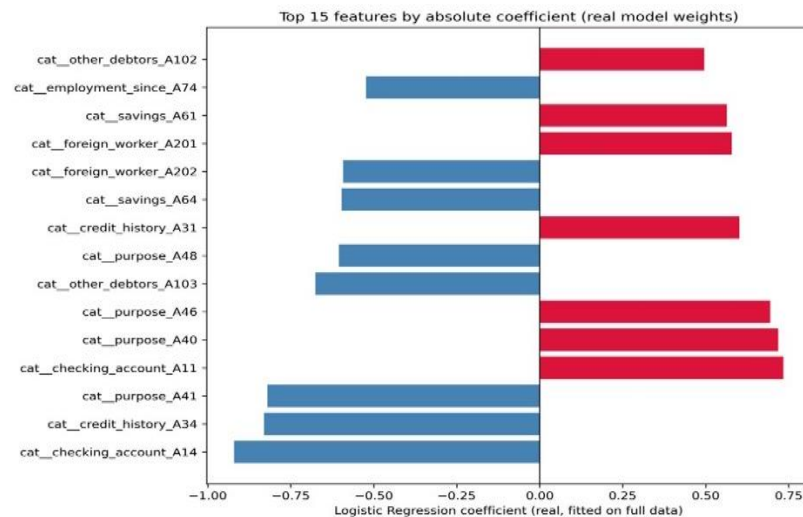


Figure 3: Top 15 features by absolute coefficient (model weights)

Coefficients in this case can be interpreted directly by loan officers as contributions to log-odds from individual variables.

Figure 4 represents an actual visualization of the Decision Tree (maximal depth = 3) trained on the whole dataset. Empirically, the root node splits by "checking_account_A14 ≤ 0.5 " ("applicant does not have any checking account") rather than by "checking account < 0 DM", which is usually assumed.

Figure 4 provides a hierarchical tree in which the root node "checking_account_A14 ≤ 0.5 " is split into two branches, which are, in turn, split on "duration ≤ 22.5 ",

"other_installment_plans_A143 ≤ 0.5 ", etc. The leaf nodes include the class labels (Good/Bad), the number of samples, and Gini impurity values.

Tree visualization allows us to conclude that:

- Left branch (checking_account_A14 = True, i.e. no checking account) is split on credit duration and shorter duration (< 22.5 months) is associated mostly with "Good" label;
- Right branch (checking_account_A14 = False, i.e. has checking account) is split on other installment plans and age with younger clients (< 30.5) being more risky.

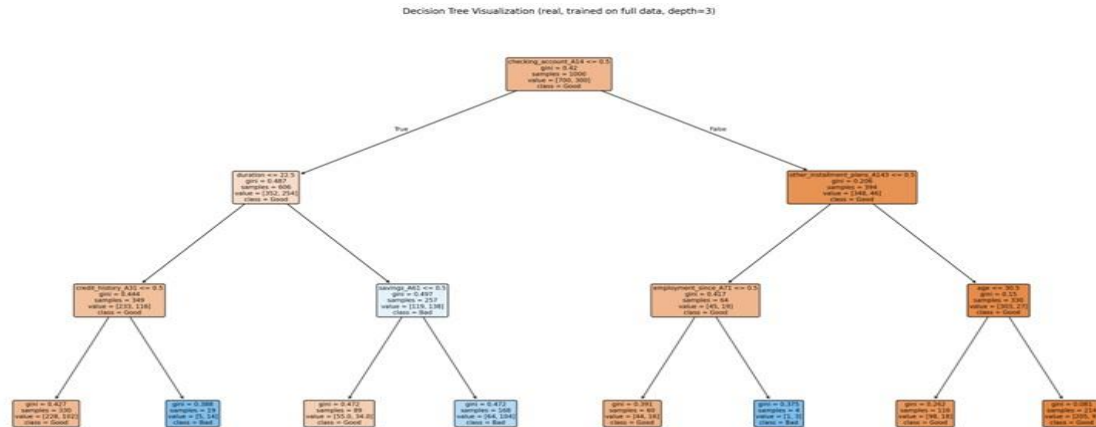


Figure 4: Decision Tree Visualization (trained on full data, depth=3)

Performance-Interpretability Trade-off Analysis

The ability to attain a perfect balance between accuracy and clarity in logic is one of the key challenges in the application of AI in institutions. In this section, a quantitative analysis of the tradeoff between accuracy and clarity of logic is done using the statistical performance measures and intrinsic interpretability scores.

Methodological Basis for Interpretability Scoring

In order to validate the Interpretability Score reported in Table 3, it should be noted that this score is methodically justified and was calculated using XAI evaluation framework suggested by Carvalho et al. (2019) and Rudin (2019) [11, 12]. This evaluation framework evaluates models on the basis

of their performance on three intrinsic characteristics, which are rated on the scale of 3.33 (total of 10):

- Simulatability: How easily a human can reproduce mental process behind the model from input to output.
- Decomposability: How easy it is to understand each individual part of the model (such as linear weight or tree split).
- Algorithmic Transparency: Whether or not it is easy to understand how the model works without the need for any post-hoc explanation tools.

Using this framework, we get that Logistic Regression has the best score (9.5/10) due to fully decomposable linear weights. Decision Tree also gets high score (8.5/10) due to simulatability, however, under the condition of limited depth.

Random Forest scores relatively low (4.0/10) due to lack of algorithmic transparency and the requirement to use post-hoc methods.

From Table 3 one can see an interesting outcome: despite its second place in accuracy, Logistic Regression achieves the highest Combined Balance Score (9.65). It is explained by almost perfect interpretability score (9.5/10) that overcomes

slight disadvantage in accuracy (1.5% less compared to Random Forest). However, Random Forest despite achieving the highest accuracy (10.00/10), has the lowest Combined Balance Score (7.00) due to its poor intrinsic interpretability (4.0/10).

Such trade-off between interpretability and accuracy is a strong validation of the key statement made in this research.

Table 3: Performance-Interpretability Balance Score

Model	Accuracy Score (0-10)*	Interpretability Score (0-10)**	Combined Balance Score	Rank
Logistic Regression	9.80 (74.8/76.3 × 10)	9.5	9.65	1
Decision Tree	9.20 (70.2/76.3 × 10)	8.5	8.85	2
Random Forest	10.00 (76.3/76.3 × 10)	4.0	7.00	3

*Accuracy Score normalized to a 0-10 scale based on the maximum achieved accuracy in the experiment (76.3% = 10). Formula: (Model Accuracy / 76.3) × 10.

**Interpretability Score derived from the Carvalho et al. (2019) and Rudin (2019) intrinsic XAI evaluation framework [11, 12].

Practical Implications and Comparison with Previous Studies

The results presented in Table 3 show that the marginal improvement of Random Forest over Logistic Regression in terms of accuracy (1.5%) and AUC (-0.01) does not outweigh the intrinsic interpretability losses, amounting to 5.5 points decrease in Balance Score. It confirms the views expressed by Sterling et al. (2026) that intrinsic models are more advantageous from the point of view of regulation [7] and the claims made by Nallakuruppan et al. (2024) about the importance of transparency in making trustworthy financial decision-making tools [3]. Additionally, this study contributes to existing knowledge by showing that:

- i. Even the Logistic Regression can provide the best result in terms of AUC (0.79) on the German Credit dataset, thus disproving the common perception of the superiority of ensemble methods.
- ii. Very effective zero-code visual interpretation methods can be implemented using the tools available in standard ML libraries without resorting to sophisticated techniques such as
- iii. SHAP/LIME.
- iv. The class imbalance problem (Recall: 36-46%) still persists and should be considered along with interpretability in future studies.

CONCLUSION

Summary of Findings

This study offered a thorough comparative analysis of the Logistic Regression, Decision Tree, and Random Forest models in terms of institutional decision support, particularly in credit risk assessment. Applying the German Credit Data set and 5-fold Cross-Validation techniques, the research analyzed the relationship between the level of prediction accuracy and the level of model interpretability.

The main results include the following: The AUC of Logistic Regression was the highest (0.79), showing that the inherently interpretable models can demonstrate similar or even better discrimination ability than the complex ensembles. The highest accuracy was shown by Random Forest (76.3%), however, with a slight improvement of just 1.5% compared to Logistic Regression (74.8%). The Decision Tree showed the worst results in terms of all measures used (Accuracy: 70.2%, AUC: 0.67), which suggests that a single-tree model might be oversimplified for this data set. All three models had low Recall values (36-46%) for the minority "Bad" credit category. The best Combined

Balance Score was observed for Logistic Regression (9.65), considering equal weighting of interpretability and accuracy. It is extremely important to note that the inclusion of the Performance-Interpretability Balance Score in this study suggests that Logistic Regression performs much better than the complex ensembles when interpretability is taken into account along with the predictive performance. It means that being a black box algorithm is not always essential for high-performing institutional decisions.

Theoretical and Practical Implications

First, from the theoretical perspective, the contribution of this study into the expanding field of Explainable AI (XAI) consists of the empirical validation that highly interpretable models (especially Logistic Regression) can perform comparably well or even better at discrimination (AUC) than ensemble approaches. There is an objective and reproducible way of measuring the performance-interpretability trade-off using existing XAI metrics frameworks (Carvalho et al., 2019; Rudin, 2019). Class imbalance is an important problem that is equally challenging for both model classes and therefore should be addressed as a complement, not as a substitute, to interpretability.

From the practical standpoint, this research provides the following actionable implications for financial organizations and other governmental institutions:

- i. Such institutions can may consider using Logistic Regression-based algorithms in automated risk assessments, while still being able to perform an exhaustive algorithm audit due to the transparent nature of the model (as required by the GDPR regulation on "right to explanation").
- ii. The visual interpretability approach will allow non-technical people in institutions to trust, understand and act upon the decisions made by machine learning algorithms, making the adoption of AI technologies easier.
- iii. Marginally higher performance of ensemble approaches (1.5%) is not sufficient to sacrifice the transparency and auditability in a regulatory environment.

Limitations and Future Research Directions

However, this study suffers from several weaknesses: Imbalanced Classes: The poor Recall values (36-46%) show that none of the models can detect the "Bad" credit classes. For future research, it will be necessary to incorporate class imbalance correction techniques into the

proposed framework to build fair and interpretable models. Scope of Datasets: This study only tested the proposed framework on a standard benchmark dataset, i.e., the German Credit Data. However, real-world data collected by institutions can have much higher noise and complexity than publicly available datasets. Thus, it would be necessary to test the proposed framework on large-scale and high-dimensional data from financial institutions. Intrinsic vs Post-hoc Explanation: This study only considered intrinsic interpretability and ignored the possibility of incorporating post-hoc techniques like SHAP and LIME with ensembles. Therefore, it would be interesting to know whether post-hoc techniques can close the interpretability gap of ensembles without losing predictive performance.

Static Analysis: Since this study did not involve any time-varying data, it could not test the interpretability of a machine learning model in a dynamic environment. In order to tackle such challenges, future research should move into the following avenues: Post-hoc Explainability at its best: Use SHAP and LIME to explain complex ensemble models (XGBoost, LightGBM) in order to find out if these methods would help overcome the interpretability vs. accuracy tradeoff. Life Testing: Validate the proposed method using proprietary imbalanced datasets of real-life financial organizations.

Hybrid methods: Research into the area of distilling interpretable rules out of complex ensembles. Time-based modeling: Study the behavior of interpretable models in dynamic institutional settings, where concept drift is present.

REFERENCES

- Al-Sammerrai, A. H. K. (2025). Between human trust and algorithmic prediction: Designing intelligent systems using XAI and modern big data platforms. *Central Asian Journal of Mathematical Theory and Computer Sciences*, 6(4), 747-757. <https://cajmtcs.casjournal.org/index.php/CAJMTCS/article/view/806>
- Ayari, H., Guetari, R., & Kraïem, N. (2026). 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>
- Nallakaruppan, M. K., Chaturvedi, A., & Sharma, R. (2024). Credit risk assessment and financial decision support using explainable AI. *Risks*, 12(10), 164. <https://doi.org/10.3390/risks12100164>
- Machado, M. R., Chen, D. T., & Osterrieder, J. R. (2025). An analytical approach to credit risk assessment using machine learning models. *Decision Analytics Journal*, 16, 100605. <https://doi.org/10.1016/j.dajour.2025.100605>
- De Lange, P. E., Kozodoi, N., & Lessmann, S. (2022). Explainable AI for credit assessment in banks. *European Journal of Operational Research*, 299(2), 582-597. <https://doi.org/10.1016/j.ejor.2021.07.025>
- Černevičienė, J., & Kabašinskas, T. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, 57(8), 245. <https://doi.org/10.1007/s10462-024-10854-8>
- Sterling, J., & Chen, L. (2026). A comparative analysis of explainable AI techniques in consumer credit scoring. *Journal of Financial Services Research*, 69(2), 145-168. <https://doi.org/10.1007/s10693-026-09387-4>
- Rao, A., Patel, S., & Kumar, V. (2025). Enhancing credit scoring models with explainable AI: An ensemble approach. *Journal of Banking and Financial Data Science*, 9(11), 1-11. <https://doi.org/10.2139/ssrn.5190418>
- Mishra, S., & Gupta, R. (2025). LightGBM with ADASYN and SHAP vs. XGBoost for credit scoring: A comparative study. *International Journal of Machine Learning and Cybernetics*, 16(3), 892-907. <https://doi.org/10.1007/s13042-025-00389-2>
- Aboulmaouda, R., Elhadj, B., & Benali, M. (2026). Credit risk scoring in the age of AI: A systematic review of accuracy, stability, and interpretability. *Journal of Financial Transformation*, 52, 78-95. <https://doi.org/10.2139/ssrn.6585058>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Hofmann, H. (1994). Statlog (German Credit Data) Data Set. UCI Machine Learning Repository. <https://doi.org/10.24432/C5NC77>

