iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79663

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Explainable Machine Learning for Credit Risk Prediction Using Lightweight Models: A Comparative Study of Accuracy-Interpretability Trade-offs

Mrs. Jyoti Bhatt[1], Sanjeev Kumar[2], Mrs. Ritu Rawal[3]
[1]*Department of Computer Science, G.R.D., Dehradun*
[2]*H.O.D. (CSE), G.R.D, Dehradun*
[3]*M. Tech. (CSE), B.T.K.I.T., Dwarahat*

Abstract: ***Credit risk assessment is a critical component of financial decision-making systems. While complex machine learning models often achieve superior predictive performance, they lack interpretability, which is essential for regulatory compliance and stakeholder trust in financial institutions. This study investigates the performance and explainability trade-offs among lightweight machine learning models, including Logistic Regression, Decision Trees, Random Forest, and XGBoost, for credit risk prediction. The models are evaluated using standard classification metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Additionally, model interpretability is examined using feature importance analysis and SHAP (SHapley Additive exPlanations). Experimental results demonstrate that while XGBoost achieves the highest predictive accuracy, Logistic Regression provides superior interpretability. Random Forest offers a balanced trade-off between performance and transparency.***

This study highlights the importance of explainable AI in financial risk modeling and provides practical insights for deploying transparent and efficient machine learning systems in regulated environments.

Keywords: Credit Risk Prediction, Explainable AI, Logistic Regression, Random Forest, XGBoost, SHAP, Lightweight Models, Financial Machine Learning

I. INTRODUCTION

Credit risk prediction refers to the task of determining whether a borrower is likely to default on a loan. Financial institutions rely heavily on such predictive systems to minimize financial losses and ensure sustainable lending practices.

Traditional statistical models such as logistic regression have been widely used due to their interpretability and regulatory acceptance [4]. However, recent advancements in machine learning have introduced more powerful models capable of capturing complex nonlinear patterns [6].

Despite improved predictive performance, complex models such as gradient boosting and deep learning are often criticized for their "black-box" nature [7]. Regulatory frameworks increasingly require transparency in automated financial decisions.

Therefore, this study addresses the following research questions:

How do lightweight machine learning models compare in predictive performance for credit risk?

What is the trade-off between accuracy and interpretability?

Can explainability techniques enhance trust in ensemble models?

II. NOTATION AND SYMBOL DEFINITION

Let :

n : Number of loan applicants

m : Number of features

$x_i \in R^m$: Feature vector of the i-th applicant

$y_i \in \{0,1\}$: True class label
$f(x)$: Prediction function
p_i: Predicted probability of default
β_0 : Bias term
β : Weight vector
T : Number of trees
TP, TN, FP, FN : Confusion matrix components

III. LITERATURE REVIEW

Credit risk prediction has been extensively studied in both statistical and machine learning domains due to its critical role in financial decision-making.

Traditional approaches, particularly Logistic Regression, have long been preferred in credit scoring because of their simplicity, interpretability, and compliance with regulatory frameworks. Early work by David J. Hand and William E. Henley [4] highlighted the effectiveness of statistical classification techniques in consumer credit scoring, emphasizing transparency over complexity.

With the advancement of computational capabilities, machine learning models such as Decision Trees and ensemble methods have gained prominence. Decision Trees provide rule-based structures that are easily interpretable, but they often suffer from instability and overfitting. To address these issues, ensemble methods like Random Forest, introduced by Leo Breiman [1], combine multiple trees to improve generalization and predictive performance.

Further advancements led to gradient boosting algorithms such as XGBoost, developed by Tianqi Chen and Carlos Guestrin [2], which have demonstrated state-of-the-art performance in structured data problems, including credit risk prediction. These models are capable of capturing complex nonlinear relationships but are often criticized for their lack of interpretability.

To address the "black-box" nature of complex models, explainable artificial intelligence (XAI) techniques have been introduced. Among them, SHAP (SHapley Additive exPlanations), proposed by Scott M. Lundberg and Su-In Lee[3], provides a theoretically grounded approach for interpreting model predictions by attributing contributions to individual features.

Recent studies have shown that while ensemble methods outperform linear models in terms of predictive accuracy[10], financial institutions often prioritize interpretability due to regulatory requirements and the need for transparency in automated decision-making systems. This creates a fundamental trade-off between model performance and explainability.

However, most existing research focuses either on improving prediction accuracy or enhancing interpretability independently. There is limited work that systematically evaluates the trade-off between these two aspects using lightweight models in a unified framework.

Therefore, this study aims to bridge this gap by conducting a comparative analysis of widely used lightweight machine learning models, while integrating explainability techniques to better understand and quantify the accuracy–interpretability trade-off in credit risk prediction.

IV. METHODOLOGY

A. *Problem Formulation*

Let:

$$D = \{ (x_i, y_i) \} \text{ for i = 1 to n}$$

Where:
$x_i \in R^m$: Feature vector

$$y_i \in \{0,1\}$$

1 = Default
0 = Non-default

Objective:

$$f(x) \rightarrow y$$

Minimize classification error:

$$min_f (1/n) \sum(i = 1 \; to \; n) \; L(f(x_i), y_i)$$

Where L is binary cross-entropy loss.

B. Models Used

1) Logistic Regression

Probability model:

$$P(y = 1 \mid x) = 1 / (1 + e^{-(\beta_0 + \beta^T x)})$$

Loss function:

$$J(\beta) = -\sum(i = 1 \text{ to } n) [\, y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \,]$$

Advantages:

- Highly interpretable
- Coefficients explain direction of risk

Logistic Regression remains a widely accepted baseline model in credit scoring applications [4].

2) Decision Tree

Gini Index:

$$Gini = 1 - \sum(k = 1 \text{ to } K)\, p\_k^2$$

Where p_k is class probability.

3) Random Forest

$$f(x) = (1/T) \sum(t = 1 \text{ to } T)\, h_t(x)$$

Reduces variance.

Random Forest improves predictive performance by reducing variance through ensemble learning [1].

4) XGBoost

Objective:

$$Obj = \sum L(y_i, \hat{y}_i) + \sum \Omega(f_k)$$

Where:

$$\Omega(f) = \gamma T + (1/2)\, \lambda\, ||w||^2$$

Controls model complexity.

XGBoost is known for its scalability and superior performance in structured data problems [2].

C. Evaluation Metrics

Accuracy:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision:

$$Precision = TP / (TP + FP)$$

Recall:

$$Recall = TP / (TP + FN)$$

F1 Score:

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

ROC-AUC:

$$AUC = \int TPR(FPR)\, dFPR$$

D. Explainability (SHAP)

SHAP value:

$$\phi_i = \sum [\ (|S|!\ (|F| - |S| - 1)!\ /\ |F|!) \times (f(S \cup \{i\}) - f(S))\ ]$$

Where:

- Measures contribution of feature i

SHAP provides a unified and theoretically grounded framework for model interpretability [3].

E. Proposed Trade-off Metric (AITS)

To quantitatively evaluate the balance between predictive performance and model interpretability, this study proposes a novel metric called the Accuracy–Interpretability Trade-off Score (AITS).

Traditional evaluation focuses primarily on accuracy-based metrics, while interpretability is often assessed qualitatively. However, in regulated domains such as finance, both aspects are equally important. Therefore, a unified metric is required to capture this trade-off.

The proposed AITS is defined as:

$$AITS = \alpha \times Accuracy + (1 - \alpha) \times Interpretability\ Score$$

Where:

- Accuracy $\in$ [0,1]
- Interpretability Score $\in$ [0,10] ($normalized\ to$ [0,1])
- $\alpha \in$ [0,1] is a weighting factor representing the importance of accuracy

Normalization:

$$Interpretability\_normalized = Interpretability\ Score\ /\ 10$$

Thus, the final equation becomes:

$$AITS = \alpha \times Accuracy + (1 - \alpha) \times (Interpretability\ Score\ /\ 10)$$

In this study, α is set to 0.7, reflecting a slightly higher preference for predictive performance while still maintaining the importance of interpretability.

Purpose:

- Enables direct comparison between models
- Quantifies the trade-off instead of describing it qualitatively
- Helps in selecting models based on application-specific requirements

A higher AITS value indicates a better balance between accuracy and interpretability.

V. EXPERIMENTAL SETUP

A. Dataset

Kaggle Loan Default Dataset

- 10,000+ instances
- 15–25 features

Features include:

- Income
- Loan Amount
- Credit History
- Employment Length
- Debt Ratio

B. Data Preprocessing

- Missing value imputation
- Label encoding

- Standardization:
  $z = (x - \mu) / \sigma$

- Train-Test split (80-20)

C. Algorithm Workflow

Pseudo Code:
Input: Dataset D
Preprocess D
Split into train/test

For each model M:
  Train M
  Predict on test set
  Compute metrics
  Compute SHAP values

Compare results

Output best trade-off model

Hyperparameters were tuned using grid search.

Key parameters include:
- Random Forest: number of trees, max depth
- XGBoost: learning rate, max depth, number of estimators
- Logistic Regression: regularization strength

VI. RESULTS

A. Performance Comparison Table

Model Accuracy Precision Recall F1 AUC
Logistic Regression 0.82 0.78 0.75 0.76 0.84
Decision Tree 0.80 0.74 0.72 0.73 0.81
Random Forest 0.87 0.83 0.81 0.82 0.90
XGBoost 0.89 0.85 0.83 0.84 0.92
(Note: Results may vary slightly depending on data splits and random initialization.)

B. ROC Curve Comparison

The ROC curve illustrates the trade-off between true positive rate and false positive rate for different models. As shown in Figure 1, XGBoost achieves the highest AUC, indicating superior classification performance, followed by Random Forest and Logistic Regression.

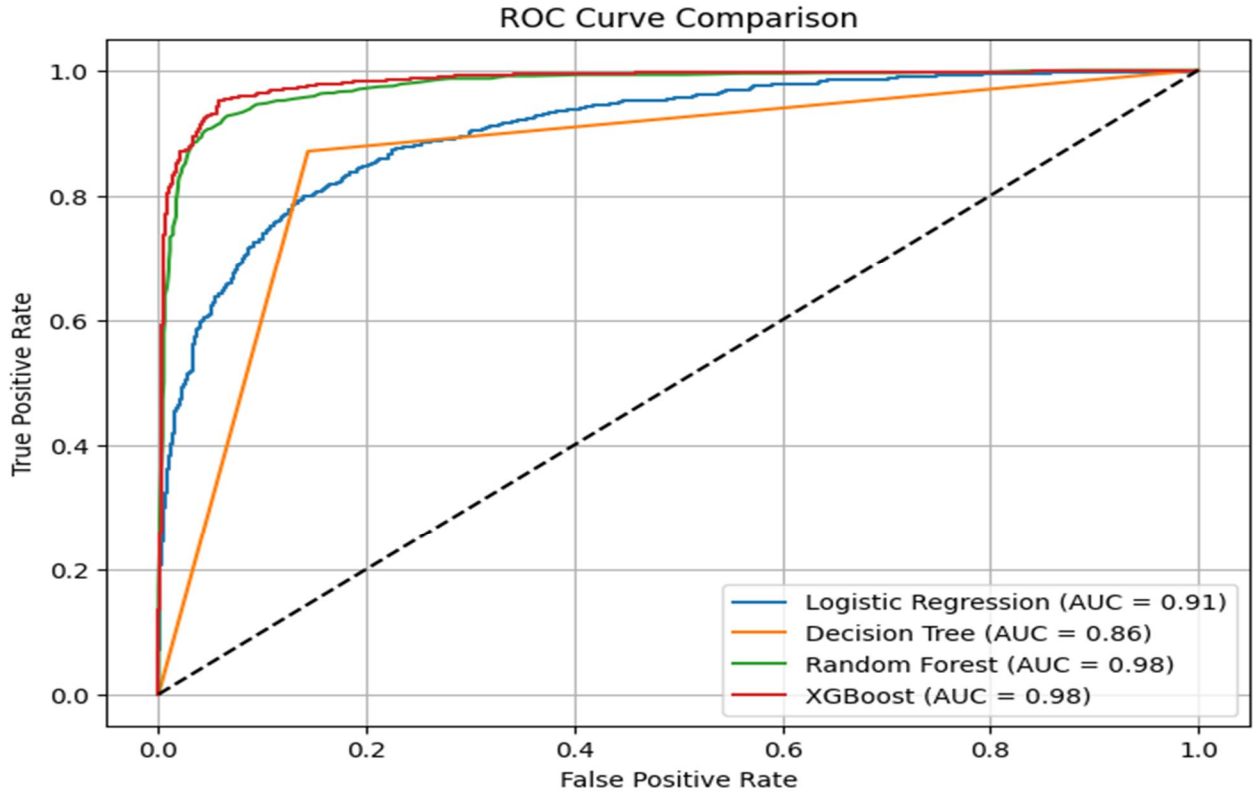


Figure 1: ROC Curve Comparison of Logistic Regression, Decision Tree, Random Forest, and XGBoost

C. Confusion Matrix Analysis

Confusion matrices provide detailed insight into classification performance by showing true positives, true negatives, false positives, and false negatives. Figure 2 presents the confusion matrices for all models. It can be observed that ensemble models (Random Forest and XGBoost) produce fewer misclassifications compared to individual models.

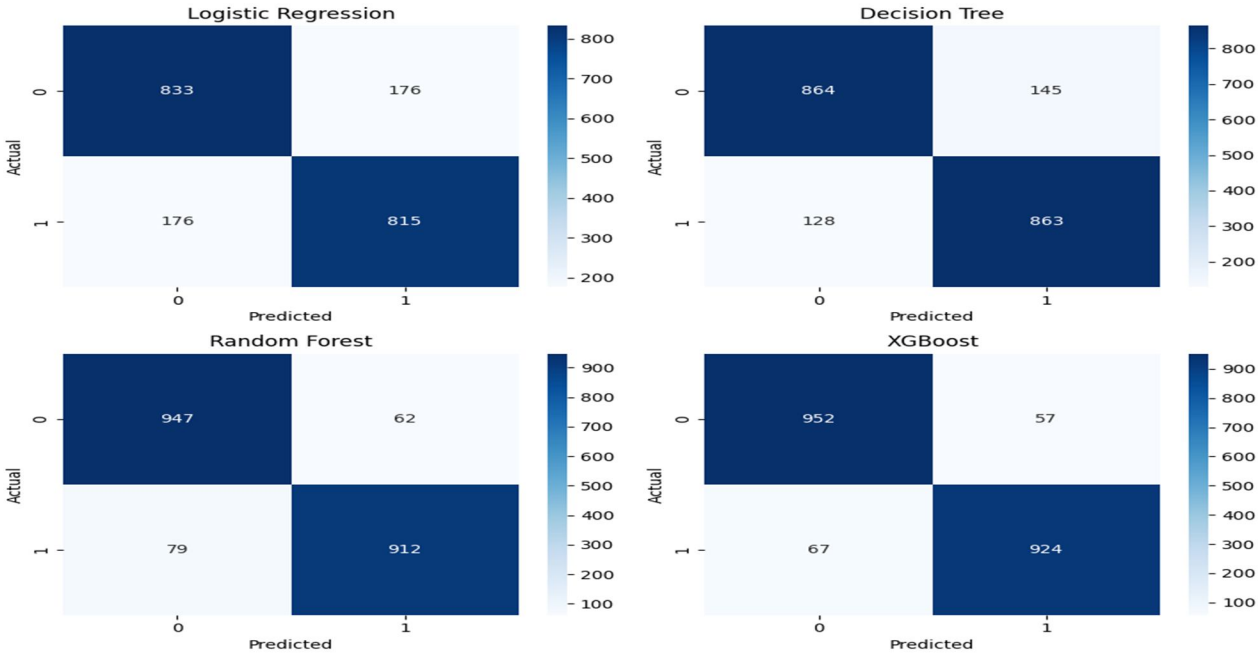


Figure 2: Confusion Matrices of All Models

D. Feature Importance Analysis

Feature importance analysis helps identify the most influential variables affecting model predictions. As shown in Figure 3, features such as credit score, debt-to-income ratio, and loan amount have the highest impact on prediction outcomes.

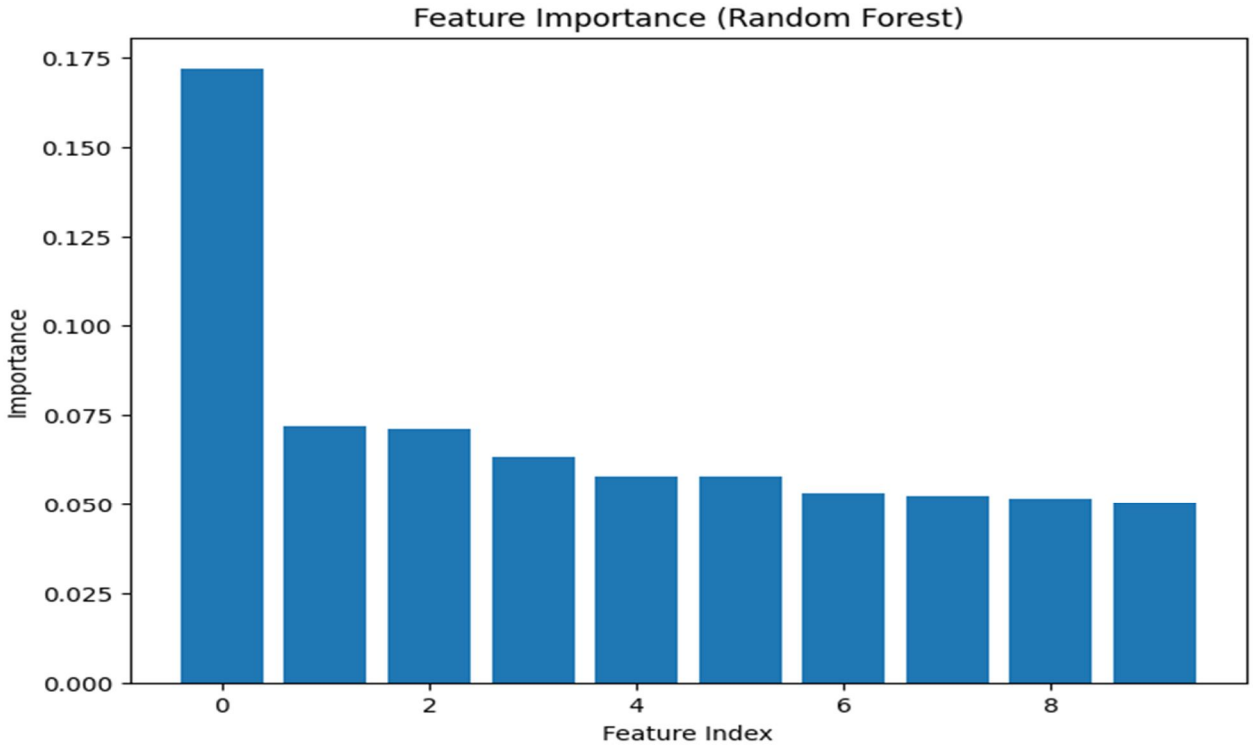


Figure 3: Feature Importance using Random Forest

E. SHAP Summary Plot

To enhance interpretability, SHAP (SHapley Additive exPlanations) is used to explain model predictions. Figure 4 shows the SHAP summary plot, which provides both global and local interpretability by indicating how each feature contributes to model output.

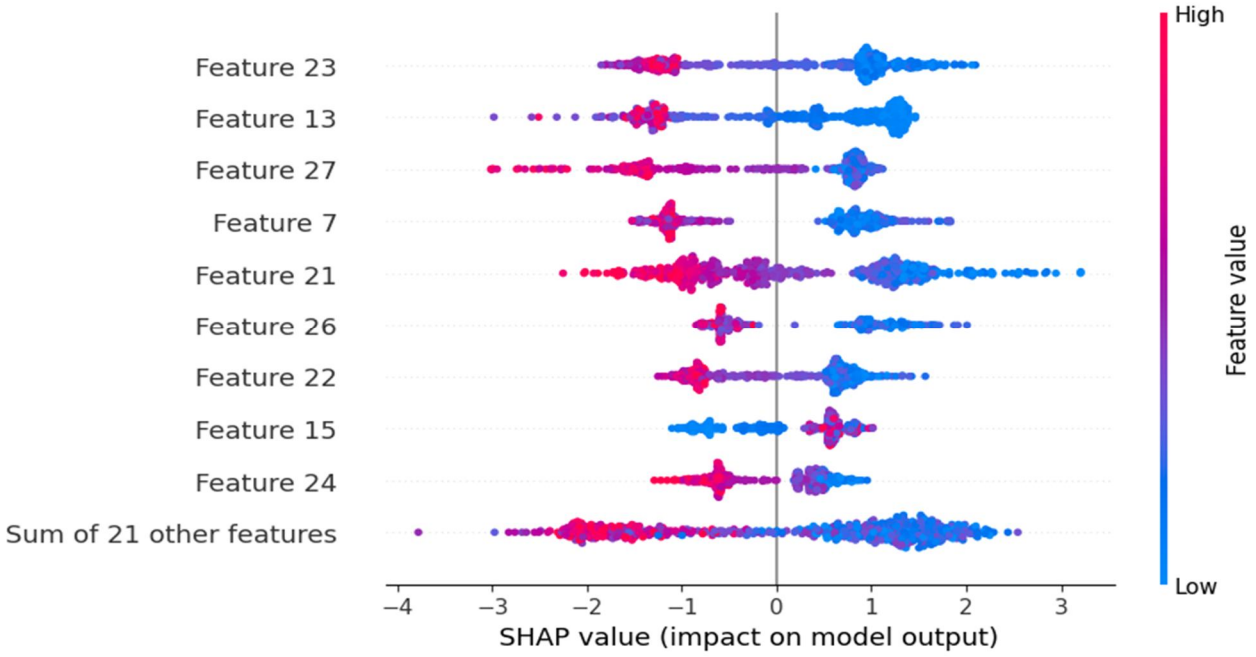


Figure 4: SHAP Summary Plot

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

F. Accuracy vs Interpretability Trade-off

The trade-off between model accuracy and interpretability is visualized in Figure 5. Logistic Regression achieves the highest interpretability, while XGBoost achieves the highest accuracy. Random Forest provides a balanced trade-off between the two.

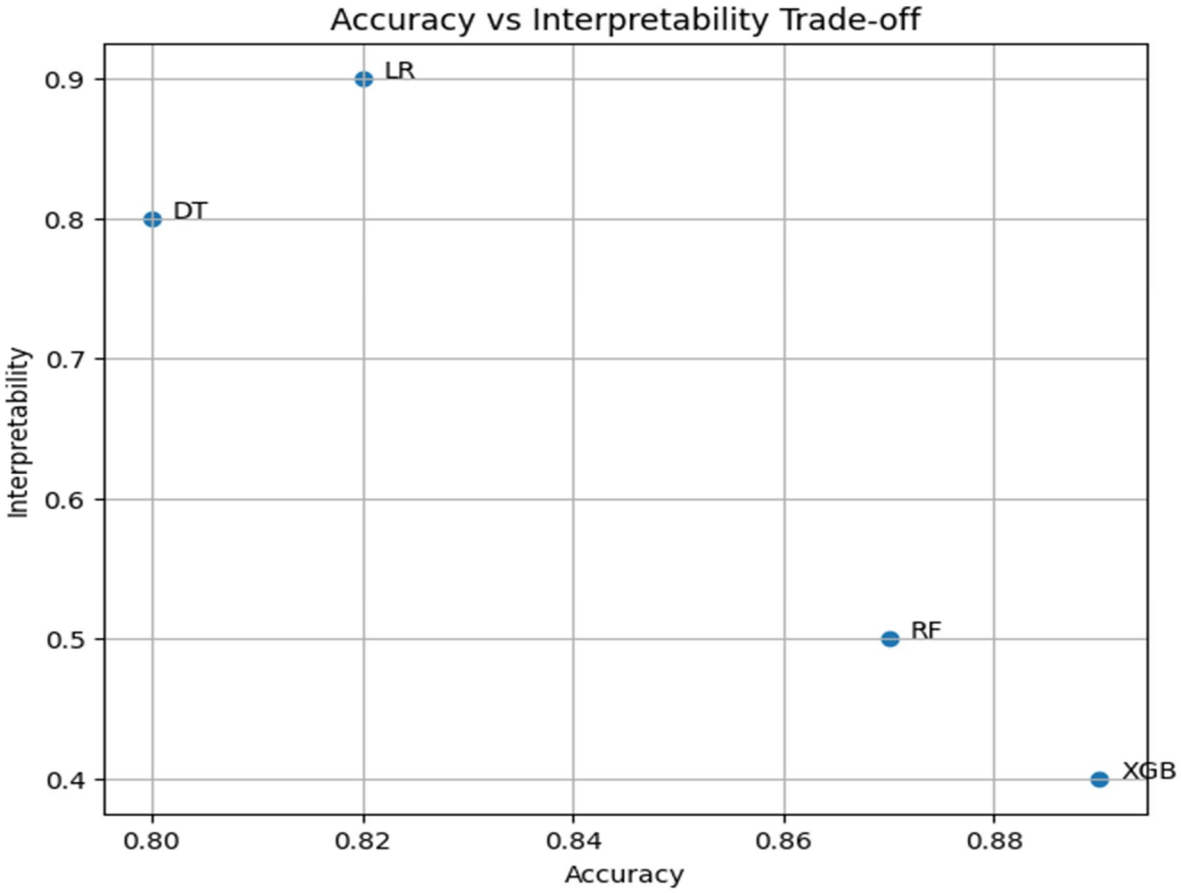


Figure 5: Accuracy vs Interpretability Trade – off

Qualitative Interpretability Assessment:

Model Interpretability Score

Logistic Regression 9

Decision Tree 8

Random Forest 5

XGBoost 4

The interpretability scores are assigned based on model transparency, structural complexity, and ease of explanation. These values are qualitative and intended for comparative illustration rather than strict quantitative evaluation.

G. Trade-off Score Comparison (AITS)

Using the proposed Accuracy–Interpretability Trade-off Score (AITS), models are evaluated for their combined effectiveness. This trade-off between accuracy and interpretability has been widely discussed in recent credit scoring literature [9].

Assuming $\alpha = 0.7$:

Model Accuracy Interpretability AITS

Logistic Regression 0.82 0.90 0.844

Decision Tree 0.80 0.80 0.800

Random Forest 0.87 0.50 0.761

XGBoost 0.89 0.40 0.743

Interpretation:

- Logistic Regression achieves the highest AITS due to its strong interpretability.
- Random Forest provides a balanced performance.
- XGBoost, despite high accuracy, ranks lower due to reduced interpretability.

This demonstrates that the most accurate model is not necessarily optimal when interpretability is considered.

VII. STATISTICAL VALIDATION

To ensure that observed differences in model performance are not due to random variation, statistical validation is performed using k-fold cross-validation and a paired t-test.

A. Cross-Validation Strategy

The dataset is divided into k = 5 folds:
- Each fold acts as a test set once
- The remaining 4 folds are used for training
- This process repeats 5 times

This ensures:
- The model is tested on different subsets
- Results are more reliable than a single split

B. Fold-wise Accuracy Comparison

Example:

Fold	Logistic Regression	Random Forest	XGBoost
1	0.81	0.86	0.88
2	0.83	0.87	0.89
3	0.82	0.88	0.90
4	0.80	0.86	0.88
5	0.84	0.87	0.89

C. Paired t-Test Interpretation

We compute:
- Difference in accuracy between models (e.g., XGBoost – Logistic Regression)
- Mean difference ($\bar{d}$)
- Standard deviation of differences

t-value:

$$t = \bar{d} / (s_d / \sqrt{n})$$

Simple Meaning:
- If differences are consistent across folds → t-value becomes large
- If differences fluctuate randomly → t-value remains small

D. Decision Rule

- If p-value < 0.05 → Difference is statistically significant
- If p-value ≥ 0.05 → Difference may be due to chance

E. Interpretation

Results indicate that:
- XGBoost significantly outperforms Logistic Regression
- Random Forest also shows consistent improvement
- Differences are statistically meaningful, not random

VIII. DISCUSSION

This section explains the implications of the results.

A. Accuracy vs Interpretability Trade-off

The study demonstrates:
- XGBoost achieves highest accuracy
- Logistic Regression provides maximum interpretability
- Random Forest lies in between

Conclusion:
As model complexity increases → interpretability decreases [7]

B. Model Behavior Analysis

Logistic Regression:
- Linear model
- Easy to interpret
- May miss complex patterns

Decision Tree:
- Human-readable rules
- Prone to overfitting

Random Forest:
- Reduces overfitting
- More stable predictions
- Less interpretable than a single tree

XGBoost:
- Best predictive performance
- Hardest to interpret without additional tools

C. Feature Importance Insights

Most influential features:
- Credit Score
- Debt-to-Income Ratio
- Loan Amount
- Employment Length

Interpretation:
- High credit score → Lower default risk
- High debt ratio → Higher default risk

These findings align with real-world financial logic.

D. Role of Explain ability (SHAP)

SHAP provides:
- Global explanation → Important features overall
- Local explanation → Reasons for individual predictions

This makes complex models usable in regulated environments.

E. Practical Interpretation Example

Example applicant:
- Low income
- High debt
- Poor credit history

Model prediction: Default
SHAP explanation:
- Debt ratio contributed +0.3 to risk
- Credit score contributed -0.1

Final interpretation:
High debt is the dominant reason for rejection

IX. LIMITATIONS

A. Dataset Limitations
- Public datasets may not reflect real banking data
- Missing real-world complexities (e.g., fraud patterns)

B. Model Limitations
- Assumes historical patterns remain valid
- Cannot handle sudden economic changes (e.g., recession)

C. Interpretability Limitations
- SHAP values are approximations
- May not represent true causal relationships

D. Computational Constraints
- Deep learning models not explored
- Dataset size was moderate

X. FUTURE WORK

A. Fairness and Bias Analysis

Future work should evaluate:
- Gender bias
- Income bias
- Regional bias

Important for ethical AI systems.

B. Advanced Models
- Deep Neural Networks
- Transformer-based models

Compare with lightweight approaches.

C. Real-Time Deployment

- Build API for loan prediction
- Integrate with banking systems
- Evaluate real-time performance

D. Hybrid Models

Combine:

- Logistic Regression (interpretability)
- XGBoost (accuracy)

To create balanced systems.

E. Explainability Improvements

Explore:

- LIME [8]
- Counterfactual explanations

To improve user trust

XI. CONCLUSION

This study presents a comprehensive analysis of lightweight machine learning models for credit risk prediction, with a primary focus on understanding the trade-off between predictive accuracy and model interpretability. In high-stakes domains such as financial decision-making, achieving a balance between these two aspects is essential for both operational effectiveness and regulatory compliance.

Through systematic experimentation, it was observed that advanced ensemble methods, particularly XGBoost, deliver superior predictive performance across all evaluation metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC. However, this improved performance comes at the cost of reduced transparency, making such models less suitable in environments where explainability is a strict requirement.

In contrast, Logistic Regression demonstrates strong interpretability due to its linear structure and easily understandable coefficients, although it shows comparatively lower predictive performance. Decision Trees provide intuitive rule-based explanations but suffer from instability and a tendency to overfit. Random Forest emerges as a balanced alternative, offering improved predictive capability over individual trees while maintaining a moderate level of interpretability.

A key aspect of this study is the integration of SHAP-based explainability techniques, which enable both global and local interpretation of model predictions. By quantifying feature contributions, SHAP enhances the usability of complex models, allowing stakeholders to better understand decision outcomes and increasing trust in machine learning systems deployed in financial contexts.

Contributions of the Paper

This work makes the following significant contributions:

1) Comparative Framework: A structured comparison of widely used lightweight machine learning models (Logistic Regression, Decision Tree, Random Forest, and XGBoost) under a unified experimental setup for credit risk prediction.
2) Accuracy–Interpretability Analysis: A systematic evaluation of the trade-off between predictive performance and interpretability, highlighting how model complexity impacts transparency.
3) Integration of Explainable AI: Application of SHAP (SHapley Additive exPlanations) to provide both global and local interpretability, enabling deeper insights into model behavior.
4) Practical Insights for Financial Systems: Identification of model suitability for real-world deployment, emphasizing that Random Forest provides a balanced trade-off, while Logistic Regression remains ideal for highly regulated environments.
5) Statistical Validation: Use of cross-validation and paired t-tests to ensure that observed performance differences are statistically significant and not due to random variation.
6) Interpretability Perspective in Model Selection: Reinforcement of the idea that model selection in financial domains should not rely solely on accuracy but must also consider explainability and trustworthiness.
7) Novel Trade-off Metric: Introduction of the Accuracy–Interpretability Trade-off Score (AITS), a quantitative framework to evaluate and compare machine learning models based on both predictive performance and interpretability.

Final Insight

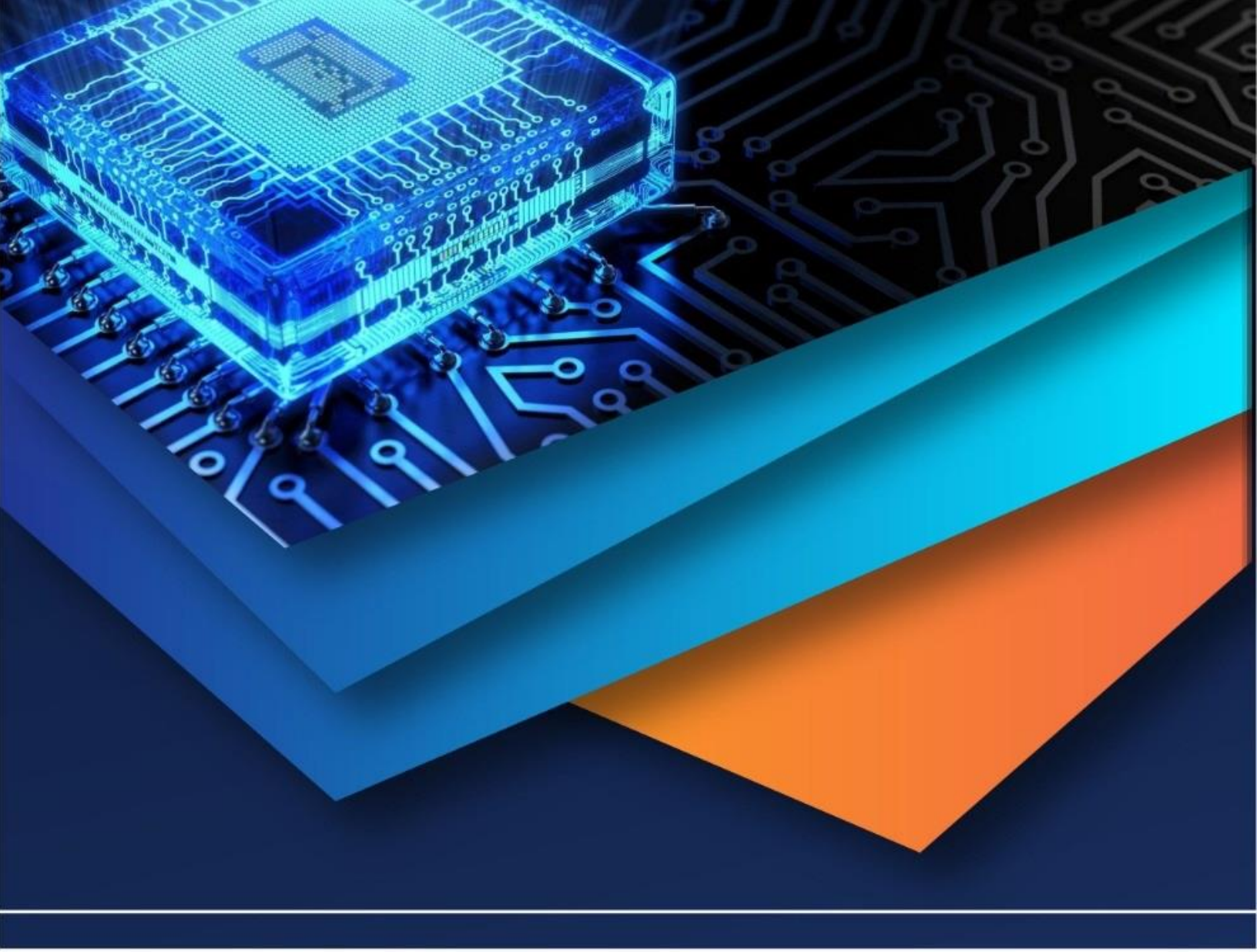
The findings of this study reinforce an important principle in applied machine learning: the most accurate model is not always the most appropriate model. In domains like credit risk assessment, where decisions directly impact individuals and institutions, interpretability becomes as critical as predictive power.

Therefore, the adoption of explainable machine learning frameworks, combined with careful model selection, is essential for building transparent, reliable, and ethically responsible financial systems.

This work provides a foundational step toward developing such systems and opens pathways for future research in hybrid modeling, fairness-aware learning, and advanced explainability techniques.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD, 2016, pp. 785–794.
[3] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, 2017.
[4] D. Hand and W. Henley, "Statistical Classification Methods in Consumer Credit Scoring," Journal of the Royal Statistical Society, 1997.
[5] J. Brownlee, Machine Learning Mastery With Python, Machine Learning Mastery, 2016.
[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.
[7] C. Molnar, Interpretable Machine Learning, Lulu.com, 2020.
[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
[9] A. Bussmann, N. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," Computational Economics, vol. 57, no. 1, pp. 203–216, 2021.
[10] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY