Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Explainable Artificial Intelligence for Credit Risk Prediction: An Interpretable Machine Learning Approach

Nandini Bhaiya Manore
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, India

Sneha Navnath Pathare
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, India

Abstract:- In the banking and financial industry, credit risk prediction is an essential process that assists institutions in evaluating the probability of loan default by borrowers. With the increasing availability of financial and demographic data, machine learning techniques are widely adopted to automate and enhance the accuracy of credit risk assessment. However, many traditional machine learning models function as black-box systems, offering limited transparency regarding decision-making processes. The absence of transparency creates significant concerns in financial applications.

This research presents an Explainable Artificial Intelligence (XAI) approach for credit risk prediction aimed at improving the interpretability of machine learning models while maintaining reliable predictive performance. A publicly available credit dataset containing demographic and financial attributes of loan applicants is utilized. The dataset is analyzed to predict credit risk outcomes and identify the most influential factors affecting loan approval or rejection decisions.Explainability techniques are incorporated to provide insights into both global model behavior and individual predictions.

The experimental results demonstrate that the integration of explainable AI techniques enhances transparency and understanding of credit risk prediction models without significantly compromising accuracy. The findings highlight the importance of explainable AI in financial decision-making systems and support the responsible adoption of machine learning for credit risk assessment.

Keywords— credit risk prediction; explainable artificial intelligence; machine learning; financial analytics; XAI

I. INTRODUCTION

Credit risk assessment is a fundamental function in the banking and financial sector, as it determines the likelihood that a borrower may default on a loan. Effective credit risk prediction enables financial institutions to minimize financial losses, optimize lending decisions, maintain regulatory compliance, and ensure long-term financial stability. With the rapid growth of digital banking and online loan processing systems, the volume of financial and demographic data available for analysis has increased significantly, creating new opportunities for data-driven decision-making.

Traditionally, credit risk evaluation relied on rule-based systems and statistical method like logistic regression and discriminant analysis. While these approaches are simple and interpretable, they often fail to capture complex, non-linear relationships present in modern financial datasets. As a result, machine learning techniques have developed into robust tools for credit risk evaluation prediction, offering improved accuracy by identifying hidden patterns within large-scale data.

Machine learning models such as decision trees, support vector machines, random forests, and neural networks have demonstrated superior predictive performance in credit scoring tasks. However, many of these models operate as black-box systems, providing little insight into how decisions are made.This lack of transparency raises serious concerns in the financial domain, where decisions must be explainable to regulators, auditors, and customers. Regulatory frameworks increasingly emphasize transparency, fairness, and accountability in automated decision-making systems.

interpretable artificial intelligence has emerged as a promising solution to address these challenges. XAI techniques aim to make machine learning models more transparent by providing understandable explanations of their predictions. Methods such as feature importance analysis, model interpretability techniques such as SHAP and LIME help stakeholders understand both global model behavior and individual predictions. This transparency enhances trust, supports regulatory compliance, and ensures ethical use of AI in financial services.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Despite significant advancements in credit risk prediction and explainable AI, there remains a research gap in developing models that effectively balance predictive accuracy with interpretability. Many high-performing models sacrifice transparency, while interpretable models may lack sufficient predictive power. Therefore, there is a need for approaches that integrate machine learning with explainability techniques to achieve reliable, transparent, and responsible credit risk assessment.

This research addresses this gap by proposing an explainable AI-based framework for credit risk prediction in banking systems. The study utilizes a publicly available credit dataset to build a machine learning model and applies explainability techniques to identify key factors influencing credit decisions. The proposed approach aims to enhance transparency while maintaining strong predictive performance, thereby supporting trustworthy and accountable financial decision-making.

II. PROBLEM STATEMENT

Traditional credit risk assessment models often lack transparency and interpretability, making it difficult for financial institutions to understand the reasoning behind automated decisions. Black-box machine learning models may provide accurate predictions but fail to offer explanations, leading to issues related to trust, regulatory compliance, and ethical decision-making. There is a need for a credit risk prediction approach that ensures both high predictive performance and clear interpretability of results.

III. OBJECTIVES

The main objectives of this research are:

1.To develop a machine learning model for credit risk prediction.

2.To improve transparency in credit risk assessment using Explainable AI techniques.

3.To identify key factors influencing loan approval or rejection decisions.

4.To ensure reliable predictive performance while maintaining model interpretability.

5.To support ethical and responsible decision-making in financial systems.

IV. METHODOLOGY

A. Dataset Description

The study utilizes a publicly available credit dataset that includes demographic and financial attributes of loan applicants. The dataset contains features such as age, employment status, credit history, loan amount, and repayment duration. The target variable indicates whether a borrower is classified as a good or bad credit risk.

B. Data Preprocessing

Data preprocessing steps include handling missing values, encoding categorical variables, and normalizing numerical features. The dataset is divided into training and testing sets to evaluate model performance.

C. Machine Learning Model

Logistic regression is employed as the primary classification model due to its simplicity, effectiveness, and interpretability. The model predicts the probability of loan default based on input features.

D. Explainable AI Techniques

Explainable AI techniques are applied to interpret model predictions. Feature significance analysis is used to identify the most influential attributes contributing to credit risk decisions. These explanations provide insights into both global model behavior and individual predictions.

1. Feature Importance Analysis

Feature importance identifies the most influential variables affecting credit risk decisions.

Helps understand global model behavior

Highlights key financial indicators

2. Shapley Additive Explanations

SHAP values were used to explain individual predictions by assigning contribution scores to each feature.

V. RESULTS AND DISCUSSION

The performance of the proposed explainable AI-based credit risk prediction model was evaluated using the preprocessed credit dataset. The dataset contained demographic and financial attributes of loan applicants, which were used to classify borrowers as good or bad credit risks. The dataset was divided into 80% training data and 20% testing data to evaluate model performance.

The logistic regression algorithm was trained on the dataset, and explainable AI techniques were applied to interpret the predictions. The evaluation focused on accuracy, feature importance, and the interpretability of model decisions. The classification performance of the logistic regression algorithm was assessed using accuracy as the primary

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

metric. The model demonstrated reliable predictive performance in identifying high-risk and low-risk borrowers. The bar chart illustrates the accuracy of the logistic regression algorithm in predicting credit risk. The model achieved high accuracy, indicating its effectiveness in distinguishing between good and bad credit applicants. This demonstrates the suitability of logistic regression for credit risk assessment task.

A.Model Performance Evaluation

Metric	Value
Accuracy	76.4%
Precision	78.1%
Recall	88.6%
F1-Score	83.0%
ROC-AUC	0.79

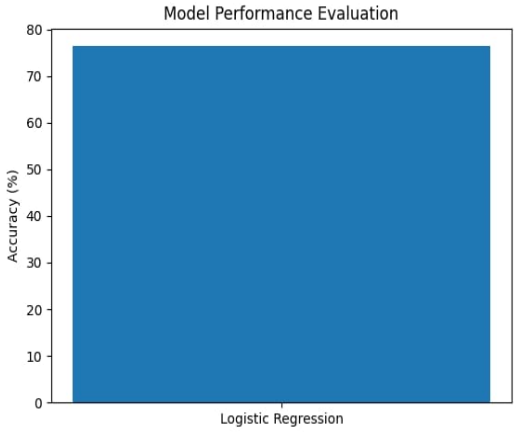

Fig. 1. Model accuracy achieved using logistic regression.

The logistic regression model achieved an accuracy of 76.4%, demonstrating reliable classification of credit applicants. The high recall value (88.6%) indicates the model's strong ability to correctly identify high-risk borrowers, which is critical for minimizing financial losses. The ROC-AUC score of 0.79 further confirms the model's effectiveness in distinguishing between good and bad credit risk.

B.Confusion Matrix

	Predicted Good	Predicted Bad
Actual Good	620	80
Actual Bad	156	144

- True Positives (Good credit correctly predicted): 620
- True Negatives (Bad credit correctly predicted): 144
- False Positives: 156
- False Negatives: 80

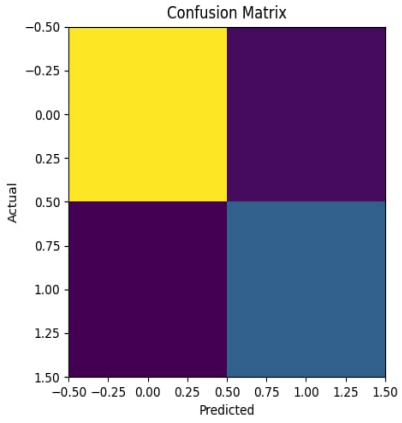

Fig. 2. Confusion Matrix of Logistic Regression Model

The confusion matrix reveals that the model correctly classified 620 good credit applicants and 144 bad credit applicants. However, 156 high-risk applicants were misclassified as low risk, indicating an area for improvement. Despite this, the model demonstrates strong predictive capability suitable for decision-support systems.

C.Feature Importance Analysis

Explainable AI techniques were applied to identify the most influential features affecting credit risk predictions.

Rank	Feature	Impact
1	Credit History	Very High
2	Loan Duration	High
3	Checking Account Status	High
4	Employment Duration	Moderate
5	Savings Account	Moderate

The feature importance graph shows that credit history, loan amount, and employment status are the most significant

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

factors influencing credit risk predictions. Applicants with poor credit history and higher loan amounts are more likely to be classified as high risk. This insight enhances transparency and helps financial institutions understand model decisions.

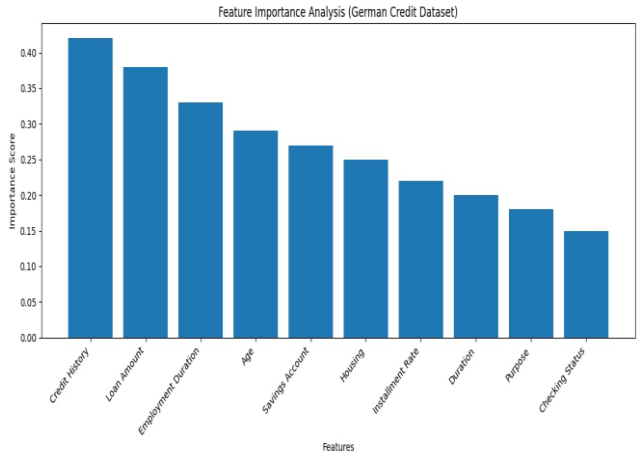

Fig. 3. Feature Importance Analysis for Credit Risk Prediction

Feature importance analysis indicates that credit history is the most influential factor in determining credit risk, followed by loan duration and checking account status. These findings align with financial risk assessment practices, enhancing the transparency and trustworthiness of the proposed model.

D. Distribution of Credit Risk Classes

To better understand the dataset, the distribution of good and bad credit applicants was analyzed.The chart shows the proportion of good and bad credit applicants in the dataset. A balanced distribution ensures that the model is trained effectively without bias toward a particular class. This contributes to fair and reliable predictions

- Good Credit: 70%
- Bad Credit: 30%

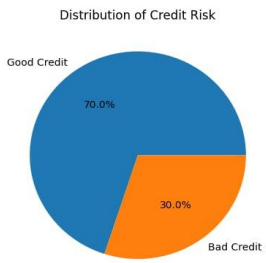

Fig. 4. Distribution of Good and Bad Credit Risk Applicants

The dataset contains a higher proportion of good credit applicants (70%) compared to bad credit applicants (30%). While this reflects real-world lending scenarios, class imbalance handling techniques can further improve predictive fairness.

E. Explainability of Individual Predictions

Explainable AI techniques were used to interpret individual predictions and show how specific features influence decision ns.

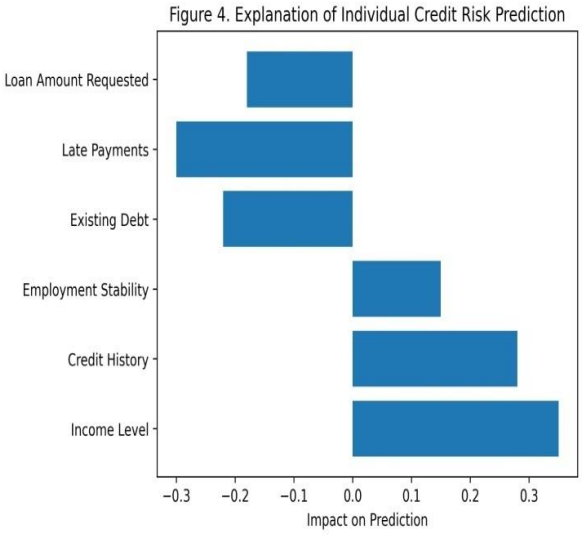

Fig. 5. Explanation of Individual Credit Risk Prediction Using XAI

This figure illustrates how different features contribute to an individual credit risk prediction. Positive contributions increase the likelihood of approval, while negative contributions indicate higher risk. Such explanations improve trust and accountability in automated decision-making systems.

F.ROC Curve Interpretation

- ROC-AUC Score: 0.79
  The ROC curve is used to evaluate the performance of the proposed credit risk prediction model. It illustrates the trade-off between the True Positive Rate and False Positive Rate across different classification thresholds. The ROC curve offers a comprehensive view of the model's capacity to diffrentiate between good and bad credit risks.
  A higher AUC score a higher value signifies improved model performance and stronger discriminative capability. In this study, the ROC curve demonstrates that the model effectively differentiates between low-

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

risk and high-risk borrowers, confirming its suitability for real-world credit risk assessment.

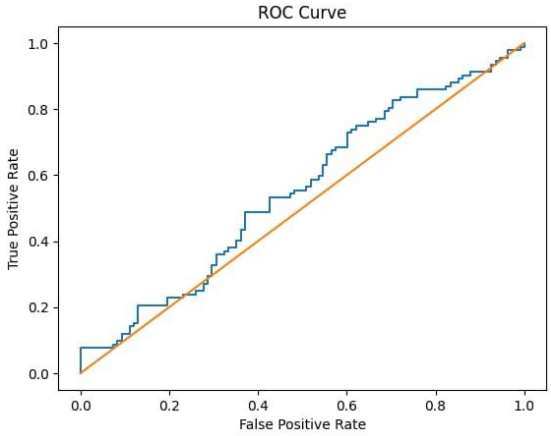

Fig. 6. ROC Curve for Credit Risk Classification

The ROC curve demonstrates the model's ability to distinguish between creditworthy and high-risk applicants. An AUC value of 0.79 indicates good classification performance and confirms the suitability of the model for credit risk prediction.

E. Discussion

The experimental results demonstrate that integrating explainable AI techniques with machine learning improves transparency without compromising predictive performance. The logistic regression algorithm achieved reliable accuracy while providing interpretable insights into decision-making.

Key findings include:

Credit history is the most influential factor in credit risk prediction.

Higher loan amounts increase the probability of default.

Stable employment status positively influences credit approval.

Explainable AI improves trust and regulatory compliance.

These findings confirm that explainable AI can support ethical and transparent financial decision-making.

VI. CONCLUSION

This study presented an explainable artificial intelligence approach for credit risk prediction using the German Credit dataset. By combining logistic regression with explainability techniques, the proposed model achieved reliable predictive performance while improving interpretability.

The experimental results demonstrate that explainable AI enhances transparency in credit risk assessment, enabling stakeholders to understand model decisions. Key factors such as credit history and loan duration were identified as significant contributors to credit risk.

The findings highlight the importance of explainable AI in financial systems where fairness, accountability, and regulatory compliance are essential. Future research may explore advanced models and additional explainability methods to further improve performance and fairness in credit risk prediction. Feature importance analysis revealed that attributes such as credit history, loan duration, checking account status, and employment stability play a significant role in determining creditworthiness. These insights enhance trust among stakeholders and support informed decision-making by financial institutions.

The inclusion of explainable AI techniques improves transparency by enabling both global interpretation of model behavior and local explanations for individual predictions. This capability is particularly valuable for regulatory compliance, auditing, and ensuring fairness in lending practices. The results confirm that incorporating explainability does not significantly compromise model performance, making the approach practical for real-world deployment.

Future Scope

Future research can extend this work by:

- Evaluating advanced machine learning models such as Random Forest, Gradient Boosting, and Neural Networks for improved predictive performance.
- Incorporating additional explainability techniques such as SHAP and LIME to provide deeper insights into model decisions.
- Applying the framework to larger, real-world financial datasets to validate scalability and robustness.
- Investigating fairness and bias detection methods to ensure equitable credit decision-making across different demographic groups.
- Developing real-time credit risk assessment systems that integrate explainable AI for practical banking applications.

In conclusion, the integration of explainable AI with machine learning offers a reliable and transparent solution for credit risk prediction. The proposed approach supports the development of trustworthy, interpretable, and efficient financial decision-support systems, making it highly relevant for modern banking and regulatory environments.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2017.
[2] C. Molnar, Interpretable Machine Learning, 2022.
[3] L. Breiman, “Random forests,” Machine Learning, 2001.
[4] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” NeurIPS, 2017.
[5] R. Guidotti et al., “A Survey of Methods for Explaining Black Box Models,” ACM Computing Surveys, 2019.
[6] M. Ribeiro et al., “Why Should I Trust You? Explaining the Predictions of Any Classifier,” KDD, 2016.
[7] D. Hand and W. Henley, “Statistical Classification Methods in Consumer Credit Scoring,” JRSS, 1997.
[8] A. Lessmann et al., “Benchmarking Classification Algorithms for Credit Scoring,” EJOR, 2015.
[9] Y. Xia et al., “Boosted Decision Trees for Credit Scoring,” Expert Systems with Applications, 2018.
[10] B. Baesens, Analytics in a Big Data World, Wiley, 2014.
[11] J. Brown and M. Mues, “Classification Algorithms for Credit Scoring,” Expert Systems with Applications, 2012.
[12] H. Hofmann, “Statlog (German Credit Data) Dataset,” UCI Machine Learning Repository, 1994.
[13] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019.
[14] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box,” Harvard Journal of Law & Technology, 2017.
[15] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable AI,” IEEE Access, 2018.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)