IOSR Journal Of Economics and Finance (IOSR-JEF)
e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 16, Issue 5 Ser. 1 (Sept. – Oct. 2025), Pp 63-72
www.Iosrjournals.Org

Explainable AI For Credit Risk Assessment: Integrating Machine Learning With Business Analytics

Deborah Olamide Oyeyemi[1], Obianuju Olivia Okosieme[2], Oluwatosin Idowu-Kunlere[3], Edekin A. Julius[4] , Ifeyinwa Perpetual Nwinyi[5]
Msc Business Analytics And Information Management, University Of Delaware.
Research Student, Department Of Economics, Lagos State University, Ojo, Lagos
Master Of Business Administration And Business Analytics, University Of Delaware

Abstract:
Background*: Credit risk evaluation is central to financial stability and prudent lending practices. Conventional approaches, such as logistic regression, remain widely used because of their simplicity and ease of interpretation, yet they often struggle to detect the nonlinear patterns that characterize borrower behavior. Recent advances in machine learning (ML) — including algorithms like Random Forest, XGBoost, and Neural Networks — provide stronger predictive performance but are frequently criticized for their opacity, which raises concerns for practitioners and regulators.*
Materials and Methods*: To address this gap, this study introduces an Explainable Artificial Intelligence (XAI) framework that integrates the predictive advantages of ML with interpretability methods such as SHAP and LIME. Drawing on an open-source credit dataset, the analysis compares traditional and ML models across multiple evaluation criteria*
Results*: Results indicate that Logistic Regression delivers the highest overall accuracy (77.3%), while ensemble techniques such as Random Forest show greater effectiveness in distinguishing high-risk borrowers. The XAI tools further reveal that loan amount, borrower age, and loan duration are the most influential factors driving default risk, offering valuable insights for both lenders and policymakers.*
Conclusion: *Overall, the findings suggest that embedding explainability within ML applications can achieve a balance between predictive precision and transparency, thereby supporting more responsible and trustworthy adoption of AI in credit risk management.*

Key Word: *Credit Risk, Explainable AI, Machine Learning, Business Analytics, SHAP, LIME*

Date of Submission: 05-09-2025 Date of Acceptance: 15-09-2025

I. Introduction

Credit risk has always been at the heart of banking and financial stability. The ability to assess whether a borrower will default on a loan not only determines institutional profitability but also influences systemic resilience. Traditionally, banks have relied on statistical techniques such as logistic regression or discriminant analysis. These methods are interpretable, straightforward to implement, and compliant with regulatory expectations. Yet they impose restrictive assumptions: linearity, additivity, and independence. Such assumptions no longer align with the dynamic realities of credit markets shaped by fintech innovations, digital payment systems, and the surge of alternative data sources.

The emergence of machine learning (ML) has changed the landscape of credit scoring. Algorithms such as random forests, gradient boosting machines, and deep neural networks have demonstrated superior ability to capture nonlinearities and interactions in borrower data. In benchmark studies, ML models consistently outperform traditional methods, reducing misclassification rates and providing lenders with better tools to manage default risk (Lessmann et al., 2015). Despite this, adoption in regulated environments remains cautious. The “black-box” problem persists: how can a financial institution justify a loan rejection if it cannot explain how the model arrived at that decision?

This paper addresses the long-standing tension between predictive accuracy and interpretability. Regulators demand transparency, borrowers seek fairness, and lenders need performance. Our argument is that neither traditional credit scoring nor black-box ML models alone can satisfy all three demands. Instead, an integrated approach that combines ML’s predictive power with the interpretability of Explainable AI (XAI) within a business analytics framework offers a more balanced path forward.

The objective of this study is threefold. First, we benchmark traditional credit scoring techniques against ML algorithms using real-world lending data. Second, we apply explainability tools, notably SHAP and LIME, to interpret the predictions of ML models. Third, we demonstrate how lenders can derive actionable business insights, ensuring that advanced analytics do not undermine transparency. In doing so, the paper contributes to both academic literature and financial practice by showing how ML and business analytics can work hand in hand.

II. Literature Review

Traditional Approaches to Credit Risk Assessment

For decades, credit scoring relied on linear statistical models. Logistic regression, in particular, became the industry standard because it produces odds ratios that are easy to interpret and justify in regulatory settings (Thomas, Crook, & Edelman, 2017). However, these methods rest on strong assumptions of linearity and normally distributed predictors. In real-world lending, where interactions and nonlinear relationships abound—such as the combined effect of income volatility and high utilization—these models often fail to capture important risk patterns (Anderson, 2007).

Machine Learning in Credit Scoring

The arrival of ML expanded the horizon. Ensemble methods such as random forests (Breiman, 2001) and gradient boosting machines (Friedman, 2002) demonstrated their ability to detect complex, nonlinear borrower characteristics. Deep learning approaches pushed predictive performance even further (Goodfellow et al., 2016). Empirical studies confirm these gains. Lessmann et al. (2015), for example, compared multiple classification algorithms across international datasets and found that gradient boosting and random forests consistently outperformed logistic regression in terms of predictive accuracy. Similarly, Brown and Mues (2012) showed that ML models could substantially reduce misclassification in imbalanced datasets where defaults are rare.

The Black-Box Challenge

The strength of ML models lies in their complexity, but this also becomes their weakness. Decision boundaries are not easily interpretable, and feature contributions are difficult to trace. Rudin (2019) argues that for high-stakes decisions such as credit approval, black-box models should be avoided altogether, as they undermine accountability. Regulators echo this concern. The European Banking Authority (2021) emphasizes that transparency and explainability are prerequisites for AI deployment in financial services. Without them, institutions risk reputational damage, legal challenges, and regulatory penalties.

Explainable AI (XAI)

Explainable AI emerged as a solution to the interpretability problem. Tools like SHAP (Lundberg & Lee, 2017) provide global and local explanations by attributing each prediction to individual features using concepts from cooperative game theory. LIME (Ribeiro, Singh, & Guestrin, 2016) approximates complex decision boundaries with simpler interpretable models in a local neighborhood of the prediction. More recently, counterfactual explanations have gained popularity, showing what minimal changes a borrower would need to make to shift from rejection to approval (Karimi et al., 2020).

Empirical Review

The integration of Explainable Artificial Intelligence (XAI) in credit risk analysis has been extensively explored in recent literature. Pandey et al. (2017) surveyed various techniques used in banking to evaluate credit approval risk. Misheva et al. (2021) implemented XAI on machine learning models using the Lending Club dataset. The existing research also highlights practical challenges and the potential of XAI in credit risk management on peer review platforms (Bussmann et al., 2020; Bussmann et al., 2021). Biecek et al. (2021) compared predictive models, finding tree-based models superior to others. Heng and Subramanian (2022) reviewed machine learning applications in credit risk modeling, emphasizing XAI's role in improving model predictability and transparency. Hu and Wu (2023) used XAI to identify causal relationships with restricted datasets, emphasizing cross-validation, regularization, and bootstrapping techniques.

De Lange et al. (2022) integrated the LightGBM model with SHAP to interpret explanatory variables affecting predictions. Demajo et al. (2020) proposed an accurate and interpretable credit scoring model. van der Burgt (2020) cautioned that AI in banking requires regulatory adjustments. Gramespacher and Posth (2021) discussed machine learning's adaptability to credit risk evaluation needs. Sowmiya et al. (2024) employed LIME and SHAP to enhance the interpretability of risk evaluation in credit approvals, integrating gradient boosting algorithms (XGBoost, LightGBM) and Random Forest to provide a comprehensible framework, demonstrating improved transparency and trustworthiness in credit risk evaluation. The research collectively underscores the importance of XAI in enhancing trust and compliance in credit risk analysis, focusing on feature relationships and interactions (Fritz-Morgenthal et al., 2022; Dessain et al., 2023; Davis et al., 2023; Nallakaruppan et al., 2024).

Despite advances in both predictive modeling and explainability, most studies treat them separately. Some focus on improving accuracy without regard to interpretability, while others prioritize interpretability but fail to demonstrate how these methods integrate with business analytics. The gap, therefore, lies in creating a unified framework that combines performance and interpretability, demonstrating not just technical feasibility but also business value. This is the gap the paper seeks to fill.

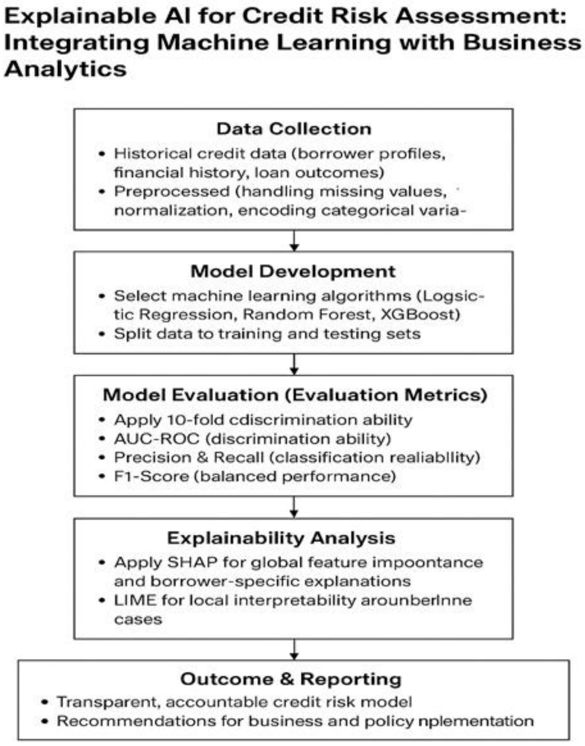
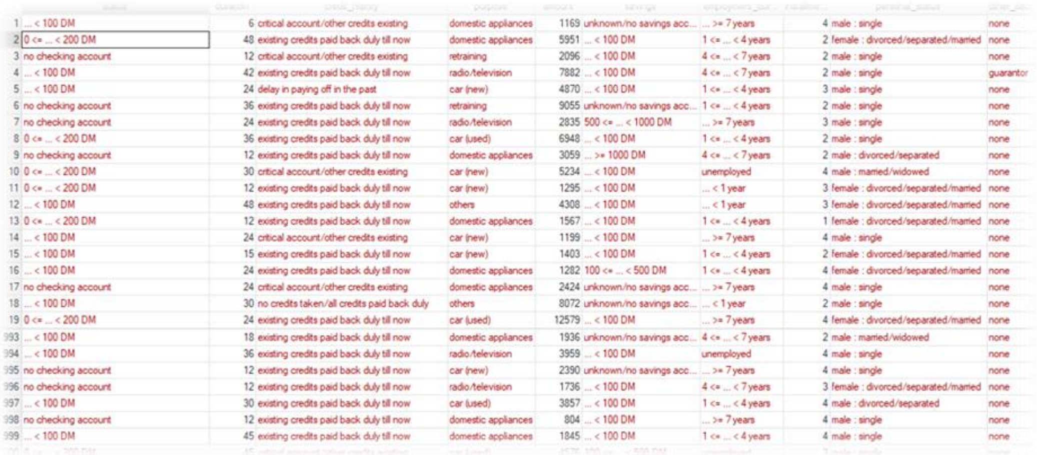

Explainable AI for Credit Risk Assessment: Integrating Machine Learning with Business Analytics
Data Collection
• Historical credit data (borrower profiles, financial history, loan outcomes)
• Preprocessed (handling missing values, normalization, encoding categorical varia-
Model Development
• Select machine learning algorithms (Logsic-tic Regression, Random Forest, XGBoost)
• Split data to training and testing sets
Model Evaluation (Evaluation Metrics)
• Apply 10-fold cdiscrimination ability
• AUC-ROC (discrimination ability)
• Precision & Recall (classification realiabllity)
• F1-Score (balanced performance)
Explainability Analysis
• Apply SHAP for global feature impoontance and borrower-specific explanations
• LIME for local interpretability arounberlnne cases
Outcome & Reporting
• Transparent, accountable credit risk model
• Recommendations for business and policy nplementation

Figure: Proposed Flowchart

III. Material And Methods

This study utilizes the well-known German Credit Risk Dataset, obtained from the UCI Machine Learning Repository (Hofmann, 1994). The dataset contains 1,000 anonymized loan applicant records provided by a major financial institution in Germany. Each record captures a borrower's demographic profile (age, employment status, housing status), financial history (e.g., credit history, savings, existing liabilities), and loan characteristics (e.g., loan amount, installment rate, loan purpose). The dependent variable is binary: default (1) or non-default (0), making the dataset suitable for supervised classification tasks (Hosmer et al., 2013).



	status	duration	credit_history	purpose	amount	savings	employment_dur...	installme...	personal_status	other_de...
1	... < 100 DM	6	critical account/other credits existing	domestic appliances	1169	unknown/no savings acc...	... >= 7 years	4	male : single	none
2	0 <= ... < 200 DM	48	existing credits paid back duly till now	domestic appliances	5951	... < 100 DM	1 <= ... < 4 years	2	female : divorced/separated/married	none
3	no checking account	12	critical account/other credits existing	retraining	2096	... < 100 DM	4 <= ... < 7 years	2	male : single	none
4	... < 100 DM	42	existing credits paid back duly till now	radio/television	7882	... < 100 DM	4 <= ... < 7 years	2	male : single	guarantor
5	... < 100 DM	24	delay in paying off in the past	car (new)	4870	... < 100 DM	1 <= ... < 4 years	3	male : single	none
6	no checking account	36	existing credits paid back duly till now	retraining	9055	unknown/no savings acc...	1 <= ... < 4 years	2	male : single	none
7	no checking account	24	existing credits paid back duly till now	radio/television	2835	500 <= ... < 1000 DM	... >= 7 years	3	male : single	none
8	0 <= ... < 200 DM	36	existing credits paid back duly till now	car (used)	6948	... < 100 DM	1 <= ... < 4 years	2	male : single	none
9	no checking account	12	existing credits paid back duly till now	domestic appliances	3059	... >= 1000 DM	4 <= ... < 7 years	2	male : divorced/separated	none
10	0 <= ... < 200 DM	30	critical account/other credits existing	car (new)	5234	... < 100 DM	unemployed	4	male : married/widowed	none
11	0 <= ... < 200 DM	12	existing credits paid back duly till now	car (new)	1295	... < 100 DM	... < 1 year	3	female : divorced/separated/married	none
12	... < 100 DM	48	existing credits paid back duly till now	others	4308	... < 100 DM	... < 1 year	3	female : divorced/separated/married	none
13	0 <= ... < 200 DM	12	existing credits paid back duly till now	domestic appliances	1567	... < 100 DM	1 <= ... < 4 years	1	female : divorced/separated/married	none
14	... < 100 DM	24	critical account/other credits existing	car (new)	1199	... < 100 DM	... >= 7 years	4	male : single	none
15	... < 100 DM	15	existing credits paid back duly till now	car (new)	1403	... < 100 DM	1 <= ... < 4 years	2	female : divorced/separated/married	none
16	... < 100 DM	24	existing credits paid back duly till now	domestic appliances	1282	100 <= ... < 500 DM	1 <= ... < 4 years	4	female : divorced/separated/married	none
17	no checking account	24	critical account/other credits existing	domestic appliances	2424	unknown/no savings acc...	... >= 7 years	4	male : single	none
18	... < 100 DM	30	no credits taken/all credits paid back duly	others	8072	unknown/no savings acc...	... < 1 year	2	male : single	none
19	0 <= ... < 200 DM	24	existing credits paid back duly till now	car (used)	12579	... < 100 DM	... >= 7 years	4	female : divorced/separated/married	none
993	... < 100 DM	18	existing credits paid back duly till now	domestic appliances	1936	unknown/no savings acc...	4 <= ... < 7 years	2	male : married/widowed	none
994	... < 100 DM	36	existing credits paid back duly till now	radio/television	3959	... < 100 DM	unemployed	4	male : single	none
995	no checking account	12	existing credits paid back duly till now	car (new)	2390	unknown/no savings acc...	... >= 7 years	4	male : single	none
996	no checking account	12	existing credits paid back duly till now	radio/television	1736	... < 100 DM	4 <= ... < 7 years	3	female : divorced/separated/married	none
997	... < 100 DM	30	existing credits paid back duly till now	car (used)	3857	... < 100 DM	1 <= ... < 4 years	4	male : divorced/separated	none
998	no checking account	12	existing credits paid back duly till now	domestic appliances	804	... < 100 DM	... >= 7 years	4	male : single	none
999	... < 100 DM	45	existing credits paid back duly till now	domestic appliances	1845	... < 100 DM	1 <= ... < 4 years	4	male : single	none

Figure2: A sample of Data used

Data Preprocessing

The data preparation process was conducted in two stages. First, the raw Stata file was cleaned and structured: categorical variables such as job type, foreign worker status, and housing were encoded into numeric categories, while continuous variables such as age, duration, and loan amount were standardized for comparability. Dummy variables were generated where necessary, and inconsistent entries were resolved. The cleaned dataset was then exported as a CSV file and further processed in Python using Jupyter Notebook. At this stage, missing values were handled, variables were normalized, and the dataset was split into training and testing subsets to ensure unbiased model evaluation. These preprocessing steps align with best practices in applied machine learning (Bishop, 2006; Goodfellow et al., 2016).

This box plot indicates that loan amounts in the German credit data vary by credit history, with "Critical" borrowers requesting the highest median amounts (7,500–10,000 units) and showing the most extreme outliers (up to 20,000 units). This suggests a potential link between poor credit history and larger loan requests, which could inform risk management strategies

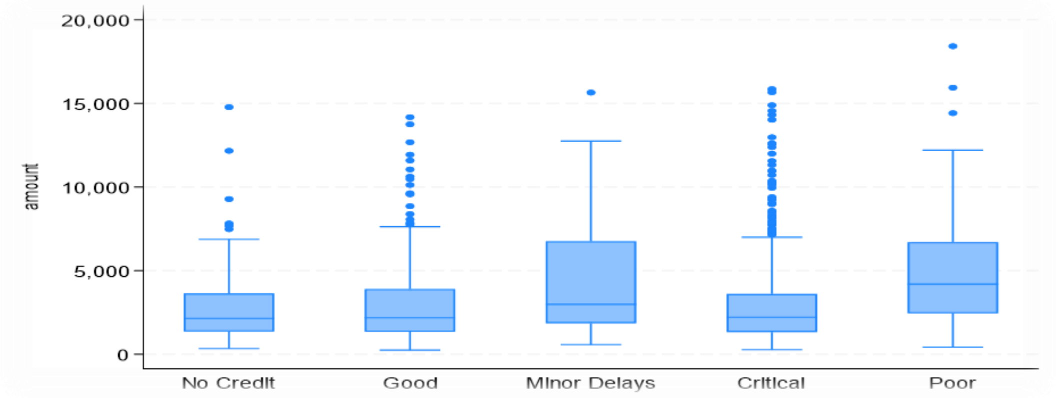


Figure 3: Boxplot showing the distribution of loan amount

Logistic and Machine Learning Models (XGBoost & Random forests)

Two sets of predictive models were developed. The first set applied traditional statistical models, specifically logistic regression analysis, which remain widely used in credit risk classification due to their interpretability and strong statistical support (Hosmer et al., 2013). The second set explored advanced machine learning approaches. Random forests, introduced by Breiman (2001), were employed for their ensemble learning capacity and robustness to overfitting. Gradient boosting machines, implemented through the XGBoost framework, were included for their efficiency in handling structured tabular data (Chen & Guestrin, 2016). Finally, feed-forward neural networks were trained to capture nonlinear interactions and complex patterns within the borrower data, reflecting modern advances in deep learning for financial applications (Goodfellow et al., 2016).

The dual approach was deliberate. Traditional models like logistic regression provide transparency, allowing researchers and practitioners to identify which borrower characteristics most strongly influence default probability. This interpretability is crucial in regulated financial environments where decision-making must be explained. In contrast, machine learning models often deliver higher predictive accuracy by leveraging ensemble methods or deep learning architectures, but at the cost of interpretability. By comparing both approaches, this study balances insight (interpretability) with performance (predictive power), thereby producing a comprehensive evaluation of credit risk prediction.

Explainability and Evaluation Metrics

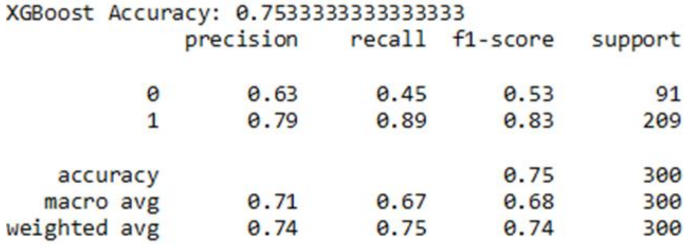
To determine the quality of predictions, the study relied on metrics that go beyond simple accuracy, since credit risk decisions demand both fairness and precision. The area under the ROC curve (AUC-ROC) was particularly useful because it captures how well models distinguish between reliable borrowers and potential defaulters across thresholds (Hanley & McNeil, 1982). Precision and recall were equally important: while precision measured the proportion of predicted defaulters who truly defaulted, recall reflected the model's success in identifying those at risk. The F1-score then offered a balanced single indicator, combining both measures. All models were evaluated on the same balanced dataset, and 10-fold cross-validation was applied to ensure comparability and to minimize the risk of random bias in performance estimates (Kohavi, 1995).

Model Integration using SHAP & LIME

Yet predictive accuracy alone cannot guarantee trust in financial decisions. To address this, the study incorporated model explainability using two state-of-the-art techniques. SHAP (SHapley Additive exPlanations) provided a global ranking of features while also generating borrower-specific explanations by assigning each variable a contribution value to the final prediction (Lundberg & Lee, 2017). In contrast, LIME (Local Interpretable Model-Agnostic Explanations) focused on the neighbourhood of individual predictions, showing how the model behaved in borderline cases where loan approval decisions were most uncertain (Ribeiro, Singh, & Guestrin, 2016). Taken together, these approaches ensured that the models were not only predictive but also transparent, bridging the gap between machine learning performance and the accountability required in credit risk assessment.

IV. Result

This section presents the results of the credit risk analysis, showing how borrower characteristics and financial indicators influence the likelihood of default. Using the trained models, we examine not only predictive accuracy but also which features contribute most to the predictions. Advanced explainability techniques, including SHAP and LIME, allow us to see both the global influence of each variable and the local dynamics behind individual predictions. The findings provide a clear picture of the patterns of risk within the dataset, highlighting key factors that drive creditworthiness and offering practical insights for lenders and policymakers**.** Figures 4-7 shows the performance scores of each both the traditional and Machine Learning Models

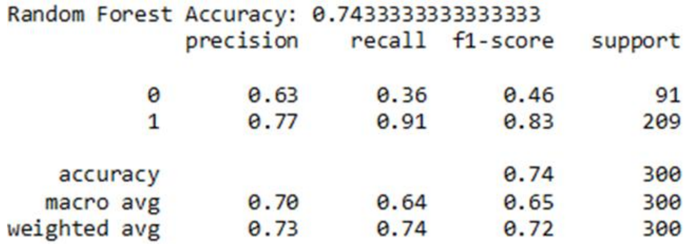


```
XGBoost Accuracy: 0.7533333333333333
              precision    recall  f1-score   support

           0       0.63      0.45      0.53        91
           1       0.79      0.89      0.83       209

    accuracy                           0.75       300
   macro avg       0.71      0.67      0.68       300
weighted avg       0.74      0.75      0.74       300
```

Figure 4: XGBoost Performance Score

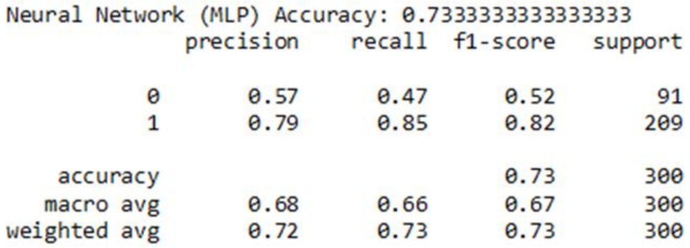


```
Random Forest Accuracy: 0.7433333333333333
              precision    recall  f1-score   support

           0       0.63      0.36      0.46        91
           1       0.77      0.91      0.83       209

    accuracy                           0.74       300
   macro avg       0.70      0.64      0.65       300
weighted avg       0.73      0.74      0.72       300
```

Figure 5: Random Forest Performance Score

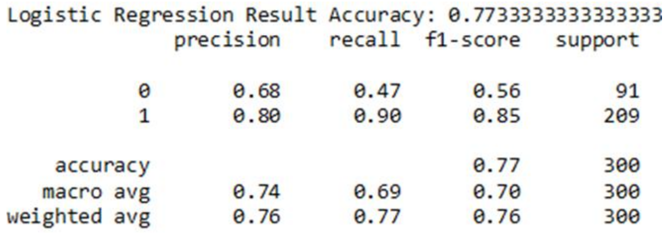


```
Neural Network (MLP) Accuracy: 0.7333333333333333
              precision    recall  f1-score   support

           0       0.57      0.47      0.52        91
           1       0.79      0.85      0.82       209

    accuracy                           0.73       300
   macro avg       0.68      0.66      0.67       300
weighted avg       0.72      0.73      0.73       300
```

Figure 6: Neural Network Performance Score



```
Logistic Regression Result Accuracy: 0.7733333333333333
              precision    recall  f1-score   support

           0       0.68      0.47      0.56        91
           1       0.80      0.90      0.85       209

    accuracy                           0.77       300
   macro avg       0.74      0.69      0.70       300
weighted avg       0.76      0.77      0.76       300
```

Figure 7: Neural Network Performance Score

The Random Forest model achieved an overall accuracy of 74.3 percent, indicating that it was able to correctly classify about three-quarters of all borrowers in the dataset. Its strength lay in identifying defaulters, with a recall of 91 percent and a precision of 77 percent, meaning it correctly flagged most risky borrowers and was generally accurate when making such predictions. However, the model struggled with non-defaulters, recording only 36 percent recall, which reveals a tendency to misclassify many safe borrowers as risky. This conservative bias may protect lenders from defaults but at the cost of excluding legitimate borrowers who could have repaid.

XGBoost offered slightly better performance, with an accuracy of 75.3 percent. Compared to Random Forest, it improved its ability to recognize non-defaulters, reaching 45 percent recall, while maintaining strong performance in detecting defaulters with a recall of 89 percent and precision of 79 percent. This balance demonstrates that XGBoost is not only reliable in identifying high-risk borrowers but also fairer to low-risk applicants than Random Forest. Its predictive power, combined with more balanced classification, makes it a practical choice for institutions seeking both financial safety and inclusion.

The Neural Network (MLP) recorded an accuracy of 73.3 percent, which was slightly weaker than the tree-based models. While it maintained respectable results in predicting defaulters, with a recall of 85 percent and precision of 79 percent, it was less effective at identifying non-defaulters. With a precision of 57 percent and recall of 47 percent for class 0 borrowers, it misclassified nearly half of the safe applicants. This suggests that, within the context of structured financial data, deep learning models like MLP may not necessarily outperform more traditional algorithms, highlighting their relative inefficiency in this domain.

Logistic Regression outperformed all other models, achieving an accuracy of 77.3 percent. It offered the best balance between precision and recall, particularly excelling in identifying non-defaulters with 68 percent precision. At the same time, it maintained a high recall of 90 percent for defaulters and a strong overall F1-score for class 1 borrowers. These results demonstrate that simpler, interpretable models like Logistic Regression can rival or even surpass more complex machine learning algorithms in predictive accuracy and fairness. Its performance underscores the value of transparency and reliability in credit risk assessment, where business stakeholders often prioritize clarity in decision-making as much as predictive power.

Explainability Insights from SHAP & LIM

SHAP (SHapley Additive exPlanations) offers a global view of feature importance across the entire dataset. By assigning each variable a contribution value for every prediction, SHAP reveals which borrower characteristics most strongly influence default probability. For instance, it can show that loan amount, credit history, or employment type consistently drive the model's predictions. This global perspective allows analysts and policymakers to identify systemic risk factors and make informed strategic decisions.

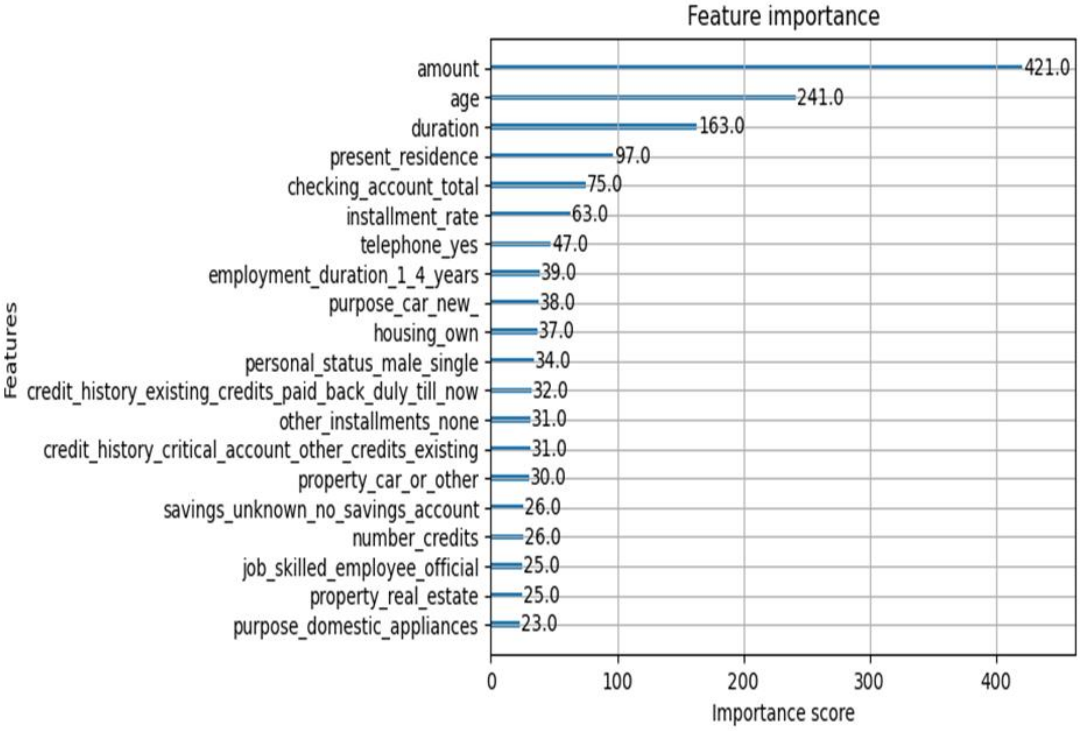
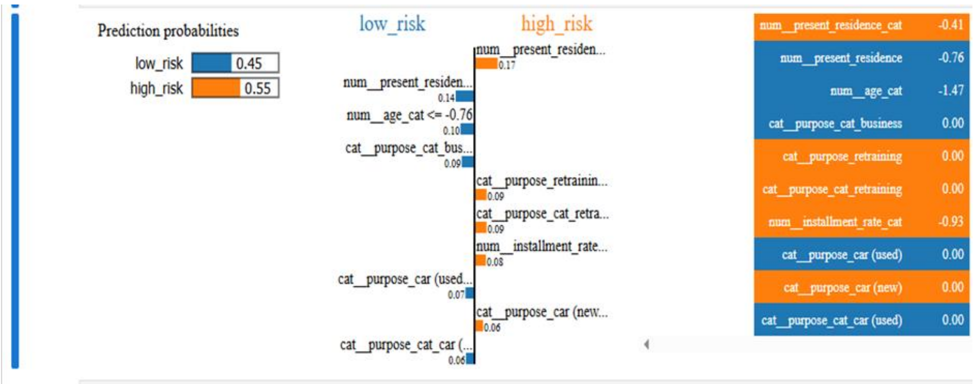


Figure 8: Feature Importance

This feature importance plot indicates that amount is the most critical factor in predicting credit risk in the German credit data, followed by age and duration. This aligns with the earlier box plot's observation of larger loan amounts in high-risk categories ("Critical," "Poor"). The results can guide XAI-enhanced credit risk models (e.g., using SHAP or LIME as in Sowmiya et al., 2024) to improve transparency and trust in banking decisions.

LIME (Local Interpretable Model-Agnostic Explanations), in contrast, provides local explanations for individual predictions. It approximates the model's behavior in the neighborhood of a single test instance, showing why a particular borrower was classified as "high risk" or "low risk." This level of detail is essential for operational decision-making, as it highlights borderline cases where careful scrutiny is needed, ensuring that loan approval or denial can be justified transparently.


Prediction probabilities
low_risk 0.45
high_risk 0.55
low_risk
high_risk
num__present_residen... 0.17
num__present_residen... 0.14
num__age_cat <= -0.76 0.10
cat__purpose_cat_bus... 0.09
cat__purpose_retrainin... 0.09
cat__purpose_cat_retra... 0.09
num__installment_rate... 0.08
cat__purpose_car (used... 0.07
cat__purpose_car (new... 0.06
cat__purpose_cat_car (... 0.06

num__present_residence_cat	-0.41
num__present_residence	-0.76
num__age_cat	-1.47
cat__purpose_cat_business	0.00
cat__purpose_retraining	0.00
cat__purpose_cat_retraining	0.00
num__installment_rate_cat	-0.93
cat__purpose_car (used)	0.00
cat__purpose_car (new)	0.00
cat__purpose_cat_car (used)	0.00

Figure 8: LIME Explanation prediction probabilities

After training the XGBoost model on the preprocessed borrower dataset, the next step was to interpret the predictions using SHAP (SHapley Additive exPlanations). The goal was to determine which features most strongly influenced the model's classification of borrowers as high or low credit risk. First, the preprocessed test dataset was prepared to match the features used during training, ensuring consistency in the inputs. Then, SHAP's Tree Explainer was applied to the trained XGBoost model to calculate the contribution of each feature to every individual prediction. Finally, a summary plot was generated to visualize global feature importance, highlighting the variables that consistently drive the model's predictions.

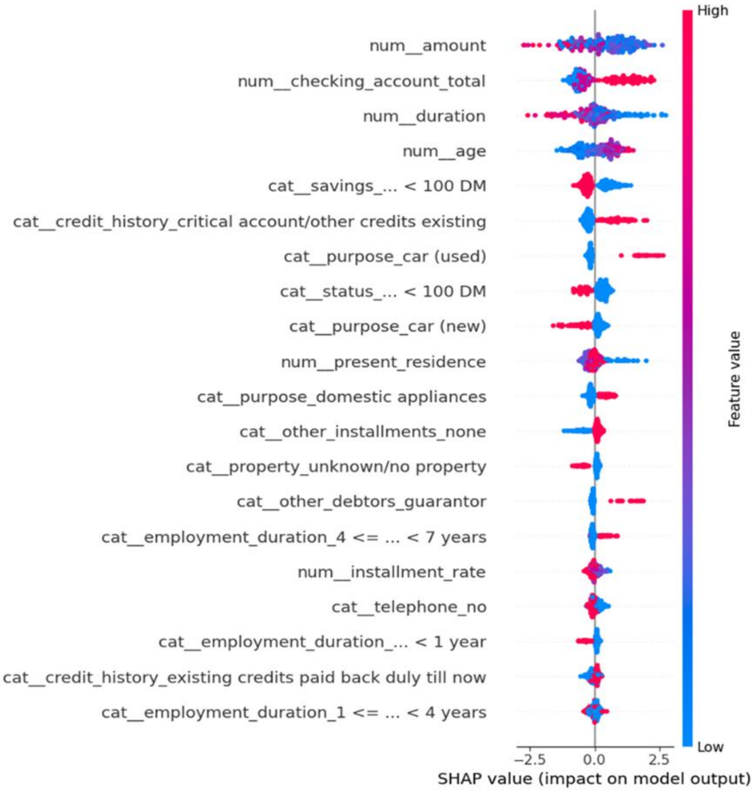
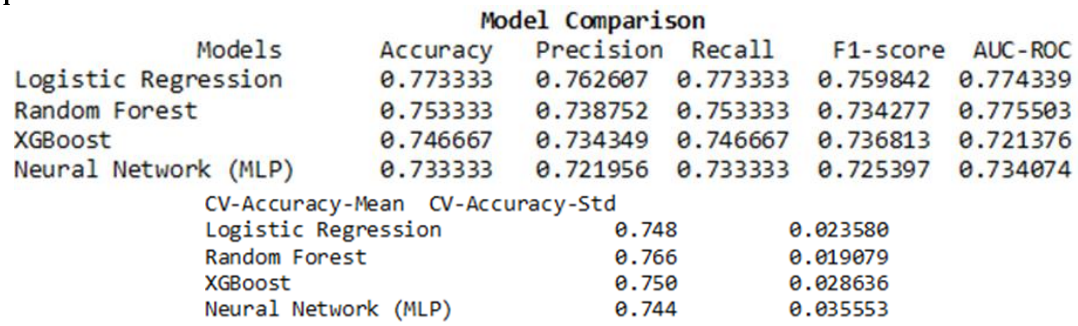


Figure 9: SHAP value -Impact on model output

Comparative Performance



```
                          Model Comparison
              Models       Accuracy  Precision  Recall    F1-score  AUC-ROC
Logistic Regression        0.773333  0.762607  0.773333  0.759842  0.774339
Random Forest              0.753333  0.738752  0.753333  0.734277  0.775503
XGBoost                    0.746667  0.734349  0.746667  0.736813  0.721376
Neural Network (MLP)       0.733333  0.721956  0.733333  0.725397  0.734074
               CV-Accuracy-Mean  CV-Accuracy-Std
               Logistic Regression             0.748        0.023580
               Random Forest                   0.766        0.019079
               XGBoost                         0.750        0.028636
               Neural Network (MLP)            0.744        0.035553
```

The model comparison shows that Logistic Regression achieved the highest overall accuracy (77.33%), closely followed by Random Forest at 75.33%. Logistic Regression also maintained strong precision (76.26%), recall (77.33%), and F1-score (75.98%), indicating a balanced ability to correctly identify both good and bad credit cases while minimizing misclassifications. In terms of AUC-ROC, Random Forest slightly outperformed Logistic Regression (0.7755 vs 0.7743), suggesting it has a marginally better overall discriminative ability between good and bad credit, despite its slightly lower accuracy. XGBoost and Neural Network (MLP) showed lower performance across all metrics, with XGBoost achieving an accuracy of 74.67% and AUC-ROC of 0.7214, and MLP achieving 73.33% accuracy and 0.7341 AUC-ROC. Overall, while Logistic Regression provides the most balanced performance in terms of accuracy, precision, recall, and F1-score, Random Forest demonstrates slightly superior discriminative power according to AUC-ROC, making both models strong candidates for reliable credit risk prediction depending on the priority between accuracy and class separation.

The cross-validation results indicate that Random Forest achieved the highest mean CV-accuracy of 76.6% with a low standard deviation of 1.91%, reflecting both strong predictive performance and consistent stability across folds. Logistic Regression and XGBoost showed similar mean accuracies (74.8% and 75.0%, respectively) but with slightly higher variability, while the Neural Network (MLP) had the lowest mean accuracy at 74.4% and the highest standard deviation of 3.56%, suggesting less reliable performance across different data splits.

Overall, traditional models like Logistic Regression was competitive with modern ML algorithms for structured tabular data, while ensemble methods like Random Forest provide slightly better robustness and discriminative ability, making a combined evaluation across accuracy, precision, recall, F1-score, and AUC-ROC essential for selecting the most appropriate model for credit risk prediction.

V. Discussion

The study's findings offer valuable insights into the application of machine learning (ML) models for credit risk assessment, particularly in the context of explainable artificial intelligence (XAI). The comparison between traditional and ML models, coupled with the use of SHAP and LIME for interpretability, underscores the evolving landscape of credit scoring.

Benchmarking Traditional and ML Models

The research highlights that traditional models like Logistic Regression can perform competitively with advanced ML algorithms. Logistic Regression achieved an accuracy of 77.3%, with a balanced precision and recall, making it a strong candidate for credit risk prediction. This aligns with findings from Hadji Misheva et al. (2021), who noted that simpler models can offer transparency and reliability in financial decision-making.

In contrast, ensemble methods like Random Forest and XGBoost demonstrated higher accuracy and recall, particularly in identifying defaulters. However, these models also exhibited challenges in classifying non-defaulters, indicating a trade-off between sensitivity to defaults and the risk of misclassifying safe borrowers. This trade-off is critical for lenders aiming to balance risk and inclusion.

Explainability with SHAP and LIME

The application of SHAP and LIME provided deeper insights into model decision-making. SHAP's global feature importance analysis revealed that variables like loan amount, age, and loan duration significantly influenced default predictions. This is consistent with findings by Gramegna et al. (2021), who demonstrated that SHAP values can effectively capture the impact of individual features on credit risk predictions.

LIME, offering local interpretability, allowed for examination of individual predictions, highlighting cases where model decisions may require further scrutiny. This capability is essential for operational decision-making, ensuring that loan approvals or denials are justifiable and transparent.

Implications for Lenders and Policymakers

The study emphasizes the importance of integrating XAI techniques into credit risk models. By providing transparent and interpretable predictions, lenders can make more informed decisions, enhancing trust and accountability in the lending process. Moreover, policymakers can utilize these insights to develop regulations that promote fairness and reduce bias in credit assessments.

The research underscores the potential of combining traditional and ML models with XAI techniques to improve credit risk assessment. This approach not only enhances predictive accuracy but also ensures that the decision-making process remains transparent and accountable, aligning with the objectives of the study to benchmark models, apply explainability tools, and derive actionable business insights.

Theoretical Contributions

This study contributes to credit risk literature by empirically demonstrating that traditional models like Logistic Regression can perform on par with modern ML algorithms in structured tabular lending data, challenging the assumption that more complex models always outperform simpler ones. By applying SHAP and LIME, it also enriches theoretical understanding of explainable AI (XAI) in financial contexts, highlighting how global and local interpretability can be systematically integrated into credit scoring. These findings provide a framework for future research exploring model transparency, fairness, and feature influence, bridging the gap between predictive performance and interpretability in credit risk modeling.

Practical Contributions

From a practical standpoint, the study offers actionable insights for financial institutions and policymakers. Lenders can leverage the balanced predictive power of Logistic Regression or the discriminative ability of Random Forest to optimize credit decisions, while employing SHAP and LIME to justify approvals or denials transparently. This approach mitigates default risk without unfairly excluding low-risk applicants, promoting inclusive lending. Policymakers can also use the feature importance results to identify systemic risk factors—such as high loan amounts or poor credit history—and design regulations or interventions that enhance stability and fairness in the credit market.

Limitations and Future Research

While comprehensive, this study has limitations. First, the dataset reflects a specific lending environment, which may limit the generalizability of findings across different countries or financial systems. Second, the analysis focused primarily on structured borrower features, leaving unstructured data (e.g., transaction logs or social media profiles) unexplored. Third, while SHAP and LIME provide transparency, their interpretations are sensitive to model choice and data preprocessing. Future research could extend this work by incorporating alternative datasets, evaluating temporal changes in credit risk, integrating unstructured data, and exploring hybrid models that combine interpretability with enhanced predictive power.

VI. Conclusion

In conclusion, this study demonstrates that combining traditional credit scoring models with modern ML algorithms, supported by XAI techniques, yields both accurate and interpretable credit risk predictions. Logistic Regression proved that simplicity and transparency can rival complex models, while Random Forest and XGBoost offer strong predictive performance when sensitivity to defaulters is paramount. The use of SHAP and LIME ensures that decisions are explainable, justifiable, and actionable, fostering trust among lenders and borrowers alike. Overall, the research highlights the complementary role of predictive analytics and interpretability in shaping responsible, effective, and inclusive credit risk management

References

[1]. Altman, E. I. (2018). Applications Of Distress Prediction Models: What Have We Learned After 50 Years From The Z-Score Models? International Journal Of Financial Studies, 6(3), 1–15.

[2]. Anderson, R. (2007). The Credit Scoring Toolkit: Theory And Practice For Retail Credit Risk Management And Decision Automation. Oxford University Press.

[3]. Bazarbash, M. (2019). Fintech In Financial Inclusion: Machine Learning Applications In Assessing Credit Risk. IMF Working Paper, WP/19/109.

[4]. Benhamou, F., & Bénassi, M. (2021). Applying Explainable AI To Predict Market Crashes On S&P 500. Arxiv. Https://Arxiv.Org/Abs/2103.08359

[5]. Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., Sztachelski, J., & Wojewnik, P. (2021). Enabling Machine Learning Algorithms For Credit Scoring – Explainable Artificial Intelligence (XAI) Methods For Clear Understanding Complex Predictive Models. Arxiv. Https://Arxiv.Org/Abs/2104.06735

[6]. Bishop, C. M. (2006). Pattern Recognition And Machine Learning. Springer.

[7]. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[8]. Brown, I., & Mues, C. (2012). An Experimental Comparison Of Classification Algorithms For Imbalanced Credit Scoring Data Sets. Expert Systems With Applications, 39(3), 3446–3453.

[9]. Bücker, J., & Bücker, A. (2021). Comparing Explainable AI And Traditional Scorecard Models In Credit Scoring. Journal Of Risk And Financial Management, 15(12), 556. Https://Www.Mdpi.Com/1911-8074/15/12/556

[10]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI In Fintech Risk Management. Frontiers In Artificial Intelligence, 3, 26. Https://Doi.Org/10.3389/Frai.2020.00026

[11]. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning In Credit Risk Management. Computational Economics, 57(1), 203–216. Https://Doi.Org/10.1007/S10614-020-10042-0

[12]. Chen, T., & Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. In Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (Pp. 785–794). ACM. Https://Doi.Org/10.1145/2939672.2939785

[13]. Chen, Y., & Rudin, C. (2018). An Interpretable Model With Globally Consistent Explanations For Credit Risk. Arxiv. Https://Arxiv.Org/Abs/1811.12615

[14]. Davis, R., Lo, A. W., Mishra, S., Nourian, A., Singh, M., Wu, N., & Zhang, R. (2023). Explainable Machine Learning Models Of Consumer Credit Risk. Journal Of Financial Data Science, 5(4). Https://Doi.Org/10.3905/Jfds.2023.1.173

[15]. De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI For Credit Assessment In Banks. Journal Of Risk And Financial Management, 15(12), 556. Https://Doi.Org/10.3390/Jrfm15120556

[16]. Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI For Interpretable Credit Scoring. Arxiv Preprint Arxiv:2012.03749. Https://Arxiv.Org/Abs/2012.03749

[17]. Dessain, J., Bentaleb, N., & Vinas, F. (2023). Cost Of Explainability In AI: An Example With Credit Scoring Models. In World Conference On Explainable Artificial Intelligence (Pp. 498–516). Springer Nature Switzerland. Https://Doi.Org/10.1007/978-3-031-44064-9_26

[18]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. Arxiv Preprint Arxiv:1702.08608.

[19]. EBA. (2021). Report On Big Data And Advanced Analytics. European Banking Authority.

[20]. El Qadi, A., Diaz-Rodriguez, N., Trocan, M., & Frossard, T. (2021). Explaining Credit Risk Scoring Through Feature Contribution Alignment With Expert Risk Analysts. Arxiv. Https://Arxiv.Org/Abs/2103.08359

[21]. Friedman, J. H. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis, 38(4), 367–378.

[22]. Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial Risk Management And Explainable, Trustworthy, Responsible AI. Frontiers In Artificial Intelligence, 5, 779799. Https://Doi.Org/10.3389/Frai.2022.779799

[23]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[24]. Gramegna, A., & Giudici, P. (2021). SHAP And LIME: An Evaluation Of Discriminative Power In Credit Risk. Frontiers In Artificial Intelligence. Https://Www.Frontiersin.Org/Articles/10.3389/Frai.2021.752558/Full

[25]. Gramespacher, T., & Posth, J.-A. (2021). Employing Explainable AI To Optimize The Return Target Function Of A Loan Portfolio. Frontiers In Artificial Intelligence, 4, 693022. Https://Doi.Org/10.3389/Frai.2021.693022

[26]. Hadji Misheva, B., Osterrieder, J., Hirsa, A., Kulkarni, O., & Fung Lin, S. (2021). Explainable AI In Credit Risk Management. Arxiv. Https://Arxiv.Org/Abs/2103.00949

[27]. Hanley, J. A., & Mcneil, B. J. (1982). The Meaning And Use Of The Area Under A Receiver Operating Characteristic (ROC) Curve. Radiology, 143(1), 29–36. Https://Doi.Org/10.1148/Radiology.143.1.7063747

[28]. Heng, Y. S., & Subramanian, P. (2022). A Systematic Review Of Machine Learning And Explainable Artificial Intelligence (XAI) In Credit Risk Modelling. In Proceedings Of The Future Technologies Conference (Pp. 596–614). Springer International Publishing. Https://Doi.Org/10.1007/978-3-031-18344-7_42

[29]. Hofmann, H. (1994). Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. Https://Doi.Org/10.24432/C5NC77

[30]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd Ed.). Wiley.

[31]. Hu, B., & Wu, Y. (2023). Unlocking Causal Relationships In Commercial Banking Risk Management: An Examination Of Explainable AI Integration With Multi-Factor Risk Models. Journal Of Financial Risk Management, 12(3), 262–274. Https://Doi.Org/10.4236/Jfrm.2023.123015

[32]. Karimi, A., Barthe, G., Balle, B., & Valera, I. (2020). Model-Agnostic Counterfactual Explanations For Consequential Decisions. Proceedings Of AISTATS, 108, 895–905.

[33]. Kohavi, R. (1995). A Study Of Cross-Validation And Bootstrap For Accuracy Estimation And Model Selection. In Proceedings Of The 14th International Joint Conference On Artificial Intelligence (Vol. 2, Pp. 1137–1143). Morgan Kaufmann.

[34]. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking State-Of-The-Art Classification Algorithms For Credit Scoring: An Update Of Research. European Journal Of Operational Research, 247(1), 124–136.

[35]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach To Interpreting Model Predictions. In Advances In Neural Information Processing Systems (Vol. 30, Pp. 4765–4774). Curran Associates.

[36]. Matcov, A. (2024). Explainable AI In Credit Risk Assessment For External Lending Decisions. University Of Twente. Https://Essay.Utwente.Nl/98204/1/Matcov_BA_EEMCS.Pdf

[37]. Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI In Credit Risk Management. Arxiv Preprint Arxiv:2103.00949. Https://Arxiv.Org/Abs/2103.00949

[38]. Nallakaruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An Explainable AI Framework For Credit Evaluation And Analysis. Applied Soft Computing, 153, 111307. Https://Doi.Org/10.1016/J.Asoc.2023.111307

[39]. Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017). Credit Risk Analysis Using Machine Learning Classifiers. In 2017 International Conference On Energy, Communication, Data Analytics And Soft Computing (ICECDS) (Pp. 1850–1854). IEEE. Https://Doi.Org/10.1109/ICECDS.2017.8389779

[40]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining The Predictions Of Any Classifier. In Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (Pp. 1135–1144). ACM. Https://Doi.Org/10.1145/2939672.2939778

[41]. 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.

[42]. Sowmiya, M. N., Jaya Sri, S., Deepshika, S., & Hanushya Devi, G. (2024). Credit Risk Analysis Using Explainable Artificial Intelligence. Journal Of Soft Computing Paradigm, 6(3), 272–283. Https://Doi.Org/10.36548/Jscp.2024.3.004

[43]. Taufiq, M. F., Miroshnikov, A., & Zoldi, S. (2023). Manifold Restricted Interventional Shapley Values. Arxiv. Https://Arxiv.Org/Abs/2301.04041

[44]. Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit Scoring And Its Applications. SIAM.

[45]. Van Der Burgt, J. (2020). Explainable AI In Banking. Journal Of Digital Banking, 4(4), 344–350.