International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Study on Credit Risk Assessment in Housing Finance Using Machine Learning Techniques

Vishali G[1] Elaiyaraja A[2]

[1]Student, Master of Business Administration, Panimalar Engineering college, Ponnamallee, Chennai.
[2]Assistant professor, Master of Business Administration, Panimalar Engineering college, Ponnamallee, Chennai.

ABSTRACT:
This study applies machine learning to assess credit risk in housing finance, predicting loan defaults using three years of Mahindra Housing Finance data. It analyzes borrower characteristics, loan attributes, and financial behaviors. Models like Logistic Regression, Gradient Boosting, XGBoost, and Random Forest are used. SHAP analysis highlights key factors, focusing on interest rates and loan amounts. RNN-based time series analysis identifies early warning signs for defaults.

Keywords: Credit Risk, Housing Finance, Machine Learning, Loan Defaults, SHAP, RNN, Risk Management.

INTRODUCTION

Credit risk refers to the possibility of financial loss when a borrower fails to repay a loan, which is a key worry in housing financing. It affects loan eligibility, interest rates, and mortgage availability. Poor risk management can lead to crises like the 2008 subprime mortgage collapse. Credit risk assessment depends on borrower factors (credit score, income), loan attributes (LTV, term), and market conditions (economy, housing trends). Machine learning enhances credit risk models by improving prediction, automating approvals, and adapting to market changes, ensuring financial stability and responsible lending.

Key Features of Credit Risk in ML:

Predictive Accuracy: ML models analyze huge datasets to detect hidden patterns and predict loan defaults more accurately than traditional methods.

Automated Decision-Making: ML enables faster, real-time credit evaluations, reducing human errors and improving efficiency.

Adaptive Learning: ML models continuously update themselves based on new borrower behavior and market trends, improving over time.

OBJECTIVES OF THE STUDY

1. To analyze the impact of borrower characteristics on the likelihood of loan defaults.
2. To assess how loan-specific attributes contribute to overall credit risk.
3. To develop and compare machine learning models for improving credit risk assessment.
4. To identify early warning indicators that signal potential loan defaults.

SCOPE OF THE STUDY

The study analyzes how credit risk assessment helps evaluate the likelihood of loan repayment in housing finance. It examines key factors such as income, credit history, and financial stability in assessing creditworthiness. The study also explores how effective risk assessment reduces loan defaults and minimizes financial losses. It highlights the importance of identifying high-risk borrowers to improve lending decisions and manage non- performing assets (NPAs). Finally, the study suggests ways to enhance credit risk assessment methods to ensure a stable and efficient housing finance system.

REVIEW OF LITERATURE

Xuyang Zhang, Lidong Xu, et al., (2024)

This work enhances credit risk assessment by analyzing loan application data with machine learning, especially a random forest model. Key variables impacting credit risk are found via correlation analysis and information enrichment. The random forest technique improves accuracy by generating numerous decision trees at random. Experimental examination of the German credit dataset demonstrates that deep learning models outperform standard approaches, supporting the suggested strategy.

Oluwabusayo Adijat Bello (2023)

The paper on “Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis” This research examines the economic and financial impacts of using machine learning (ML) for credit risk assessment in financial institutions. Traditional methods often fail to address modern market complexities, while ML offers greater accuracy, cost-effectiveness, and efficiency by analyzing large datasets and identifying hidden patterns. The study highlights ML's advantages, such as improved risk management and predictive capabilities, supported by case studies. It also addresses challenges like data privacy, model interpretability, and regulatory compliance, while suggesting future research to enhance ML's application in credit risk.

RESEARCH METHODOLOGY

The research follows an Analytical design using secondary data from financial institution and public reports to study factors influencing credit risk in housing finance. Data preprocessing includes handling missing values, outlier detection, cleaning, transformation, and feature engineering to ensure data quality. Loan defaults and borrower behavior may be predicted using models such as Logistic Regression with SHAP, Random Forest, XGBoost, and RNN. To guarantee accuracy and dependability, models are evaluated using confusion matrices, ROC curves, and AUC scores.

Model evaluation is done through Confusion matrices, ROC curves, and AUC scores to ensure accuracy and reliability. The study aims to build a transparent, dynamic, and efficient credit risk assessment framework for stable housing finance operations.

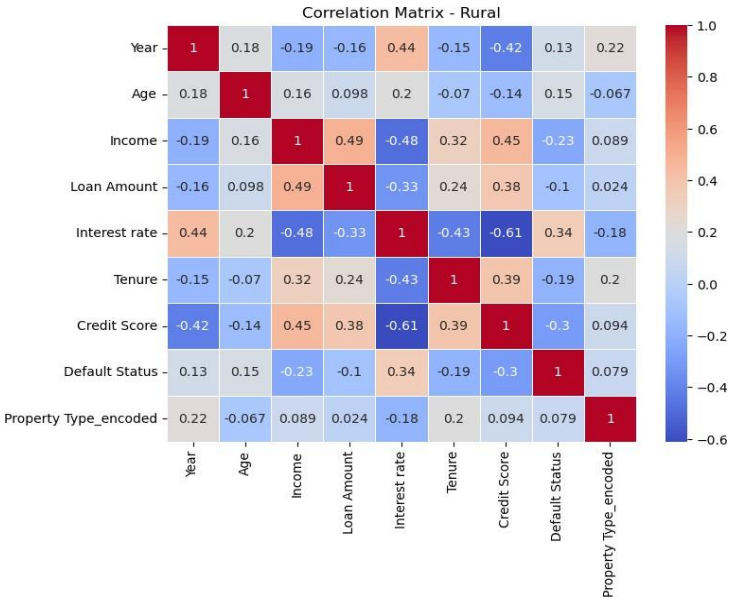
DATA ANALYSIS AND INTERPRETATION

CORRELATION ANALYSIS

Correlation with Default Status in Rural Area

Feature	Correlation with Default Status
Default Status	1.000000
Interest rate	0.340173
Age	0.154539

Year	0.129761
Property Type_encoded	0.078976
Loan Amount	-0.101526
Tenure	-0.185469
Income	-0.234866
Credit Score	-0.298651



Rural Area:

Findings:

In rural areas, the default status shows a moderate positive correlation with interest rate (r = 0.34) and weak correlations with age and year, while income and credit score are negatively correlated (r = -0.23, -0.30 respectively).

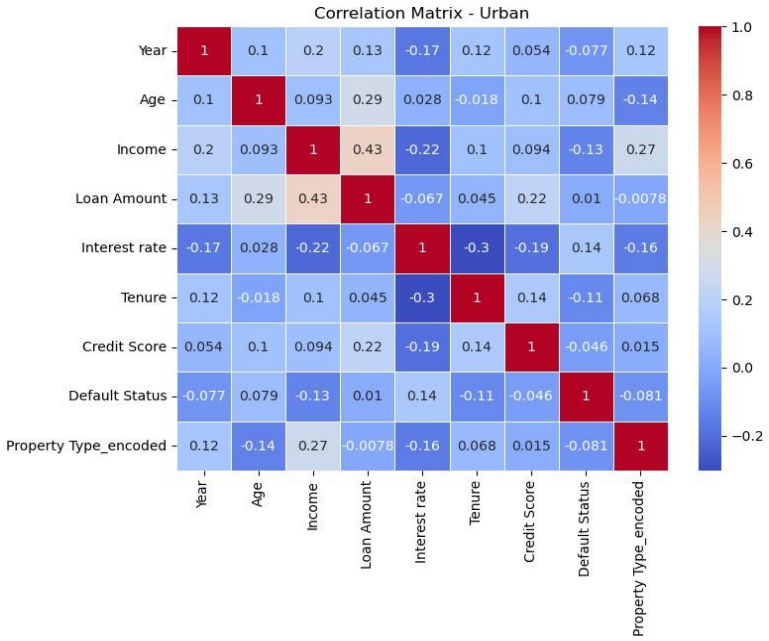
Inference:

Statistically, interest rate is a significant predictor of default, suggesting that a 1% rise in interest may moderately increase default risk. Credit score and income are strong protective factors, implying that rural lending policies should focus more on credit profiling and interest rate controls.

Correlation with Default Status in Urban Area

Feature	**Correlation with Default Status**
Default Status	1.000000
Interest rate	0.138347
Age	0.079269
Loan Amount	0.010394
Credit Score	-0.045503
Year	-0.077019
Property Type_encoded	-0.080501
Tenure	-0.111960

Income	-0.134165



Urban Area:

Findings:

In urban areas, correlations with default status are very weak, with interest rate (r = 0.14) and age showing slight positive correlations, and income and credit score showing weak negative correlations (r ≈ -0.13 and -0.05).

Inference:

Statistically, no variable shows strong predictive power in urban areas. Borrower characteristics have minimal influence on default risk, indicating that urban defaults may be driven more by macroeconomic or situational factors than by individual profiles.

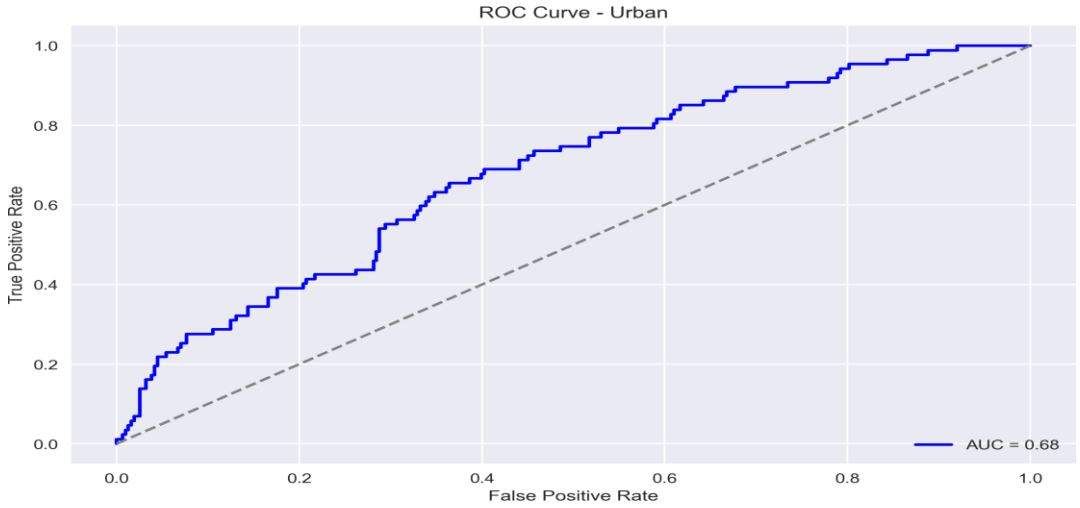
Urban Area Analysis

MACHINE LEARNING TECHNIQUES

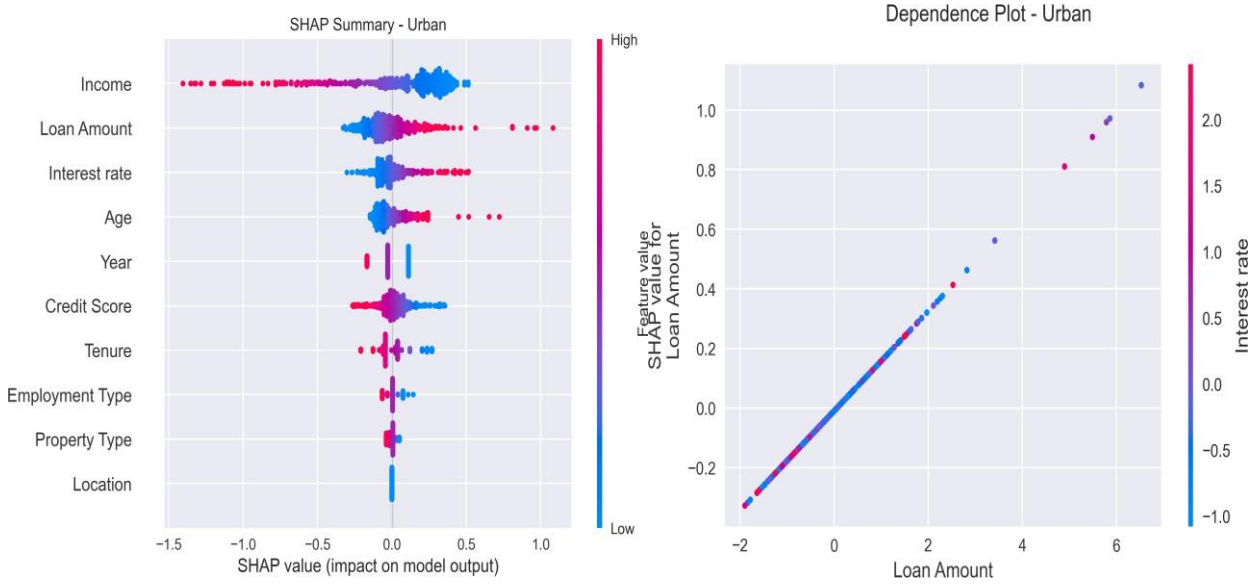
LOGISTIC REGRESSION

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.78	0.99	0.88	313
1	0.33	0.01	0.02	87
Accuracy	-	-	0.78	400
Macro Avg	0.56	0.50	0.45	400
Weighted Avg	0.69	0.78	0.69	400



SHAP ANLAYSIS:



Findings:

The logistic regression model for the urban area achieved an overall accuracy of 78% but had a very low recall of 1% for detecting defaults, meaning it mostly predicted non-default cases. SHAP analysis identified Income, Credit Score, Loan Amount, Interest Rate, and Age as the top important features.

Inference:

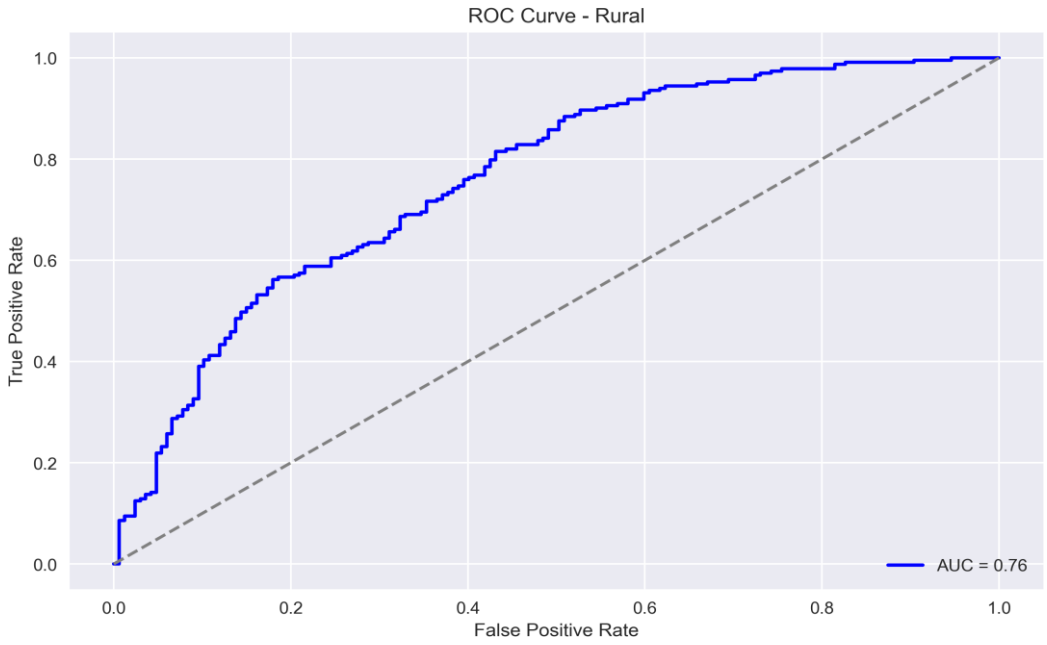
Although the model demonstrated good overall accuracy, its inability to effectively detect defaults limits its real-world application in urban housing finance.

Rural Area Analysis

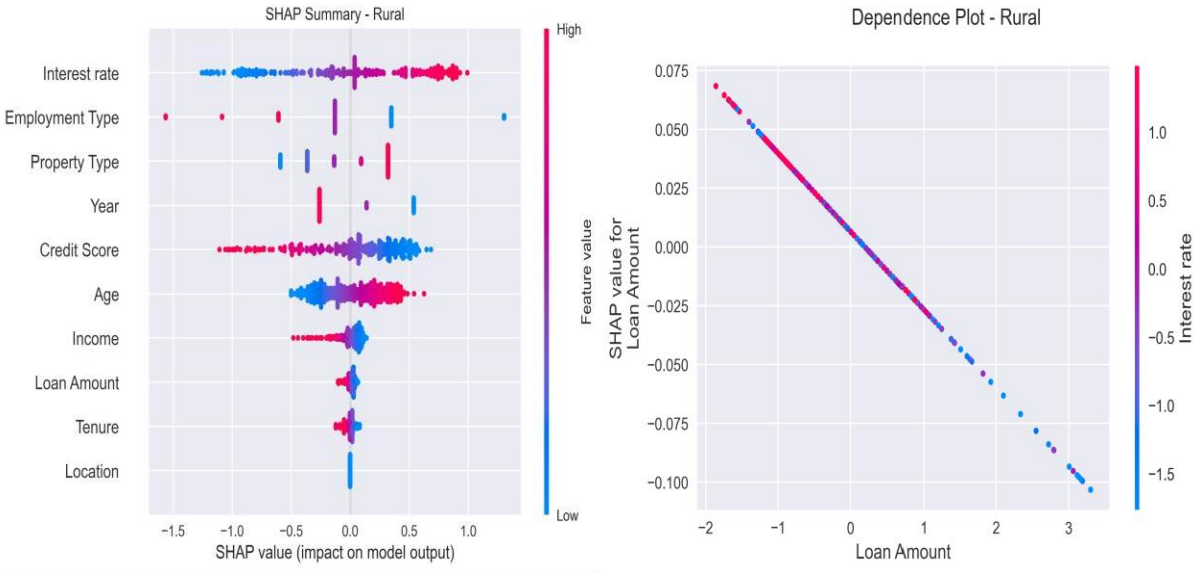
Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.75	0.48	0.58	167
1	0.70	0.88	0.78	233
Accuracy	-	-	0.71	400

Macro Avg	0.73	0.68	0.68	400
Weighted Avg	0.72	0.71	0.70	400



SHAP ANALYSIS:



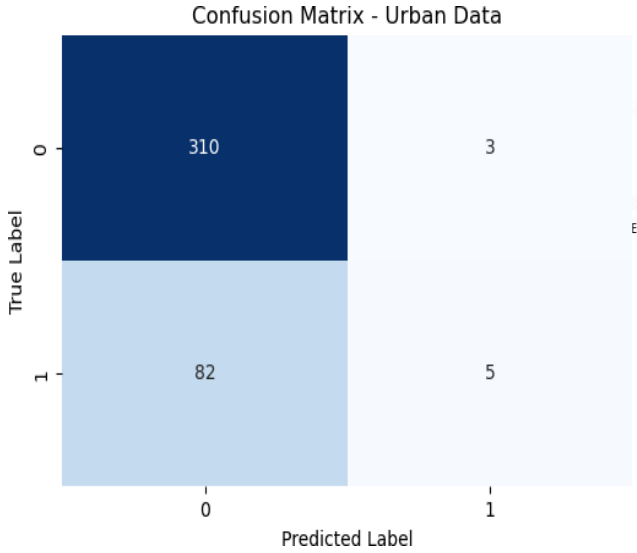
Findings: The rural area model achieved a slightly lower overall accuracy of 71% but showed a strong recall of 88% for defaults, demonstrating effective detection of risky borrowers. SHAP analysis highlighted Interest Rate, Employment Type, Credit Score, Property Type, and Year as the most important features. Higher Interest Rates and certain Property Types increased default risk, while stable Employment and higher Credit Scores reduced it.

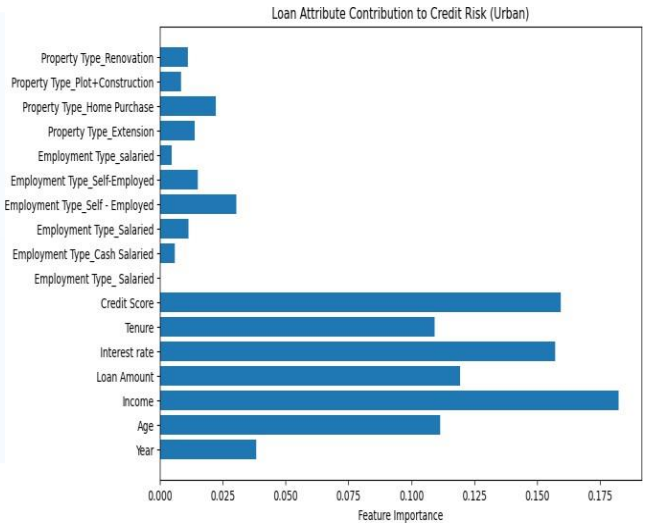
Inference: The rural model's strong performance in detecting defaults makes it highly suitable for credit risk assessment in rural housing finance. Managing interest rates and considering borrowers' employment stability are crucial strategies for reducing loan default risks in rural markets.

RANDOM FOREST

Classification Report for Urban Data

Class	Precision	Recall	F1-Score	Support
0	0.79	0.99	0.88	313
1	0.62	0.06	0.11	87
Accuracy	-	-	0.79	400
Macro Avg	0.71	0.52	0.49	400
Weighted Avg	0.75	0.79	0.71	400





Urban Data Findings

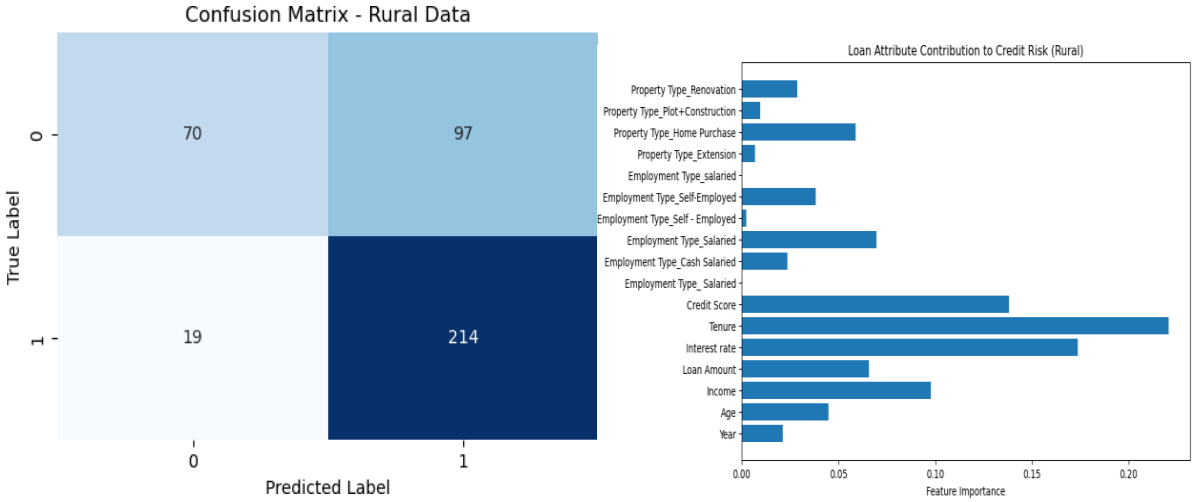
The Random Forest model for urban data achieves an accuracy of 79%. It performs excellently in identifying non- defaulters (Class 0) with high precision (0.79) and very high recall (0.99), meaning the model accurately detects non-defaulters but misses most defaulters (Class 1). Despite moderate precision (0.62), the model struggles with a very low recall (0.06) for defaulters, indicating it fails to identify most of them.

Inference:

In urban data, the model is proficient at identifying non-defaulters but underperforms in identifying defaulters. Key drivers for credit risk include Income, Credit Score, Interest Rate, and Loan Amount, while Employment and Property Type are not significant predictors.

Classification Report for Rural Data

Class	Precision	Recall	F1-Score	Support
0	0.79	0.42	0.55	167
1	0.69	0.92	0.79	233
Accuracy	-	-	0.71	400
Macro Avg	0.74	0.67	0.67	400
Weighted Avg	0.73	0.71	0.69	400



Findings:

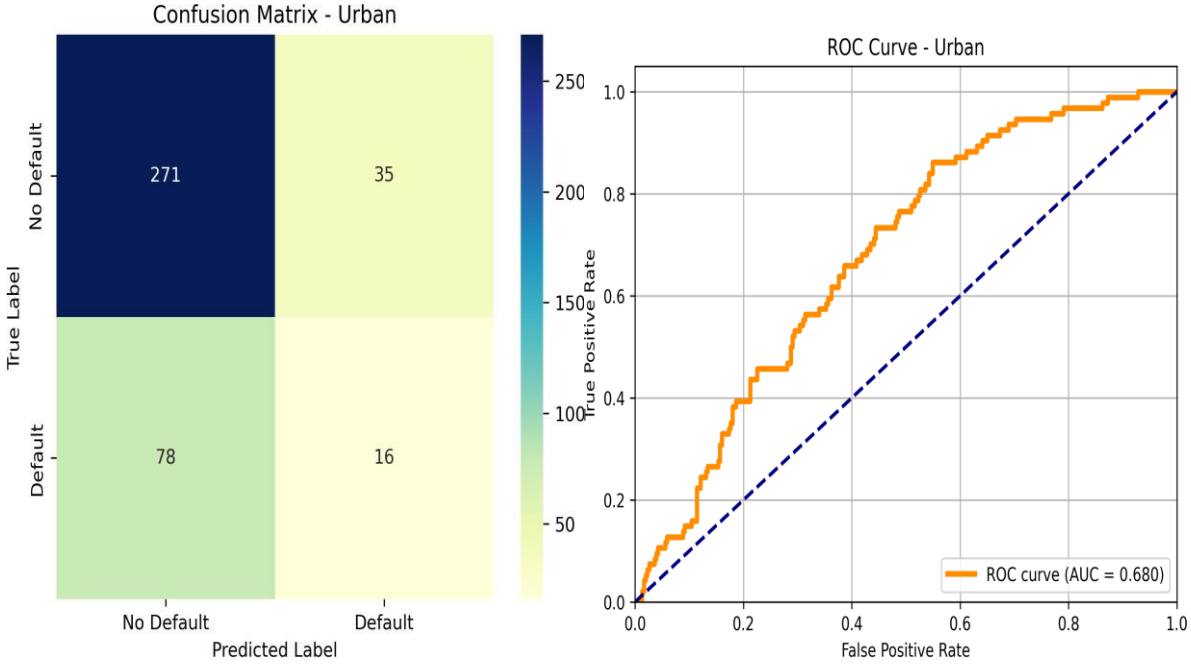
For rural data, the model achieves an accuracy of 71%. It demonstrates moderate precision (0.79) but low recall (0.42) for non-defaulters, resulting in many misclassified non-defaulters. In contrast, the model is effective at detecting defaulters (Class 1), with strong precision (0.69) and very high recall (0.92).

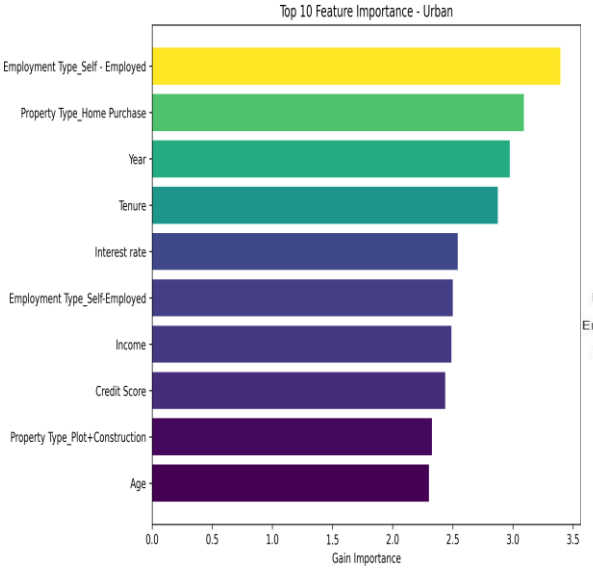
Inference:

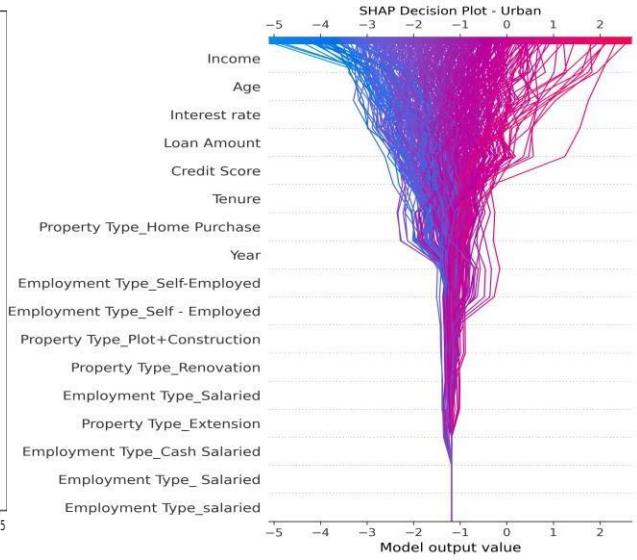
The model excels at detecting defaulters in rural data, which is beneficial for risk management. However, false positives for non-defaulters require attention. The key predictors are Tenure, Interest Rate, and Credit Score, with Employment Type and Property Type contributing more in rural contexts.

GADIENT BOOSTING (XGBOOST)

Urban Data Analysis: Classification Report







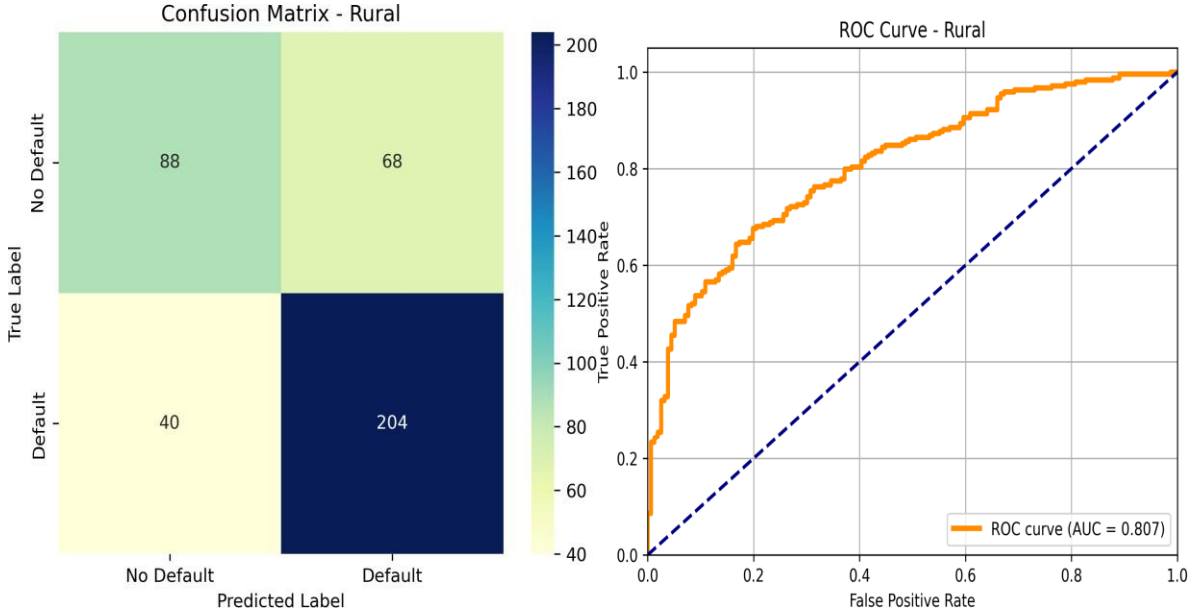
Findings:

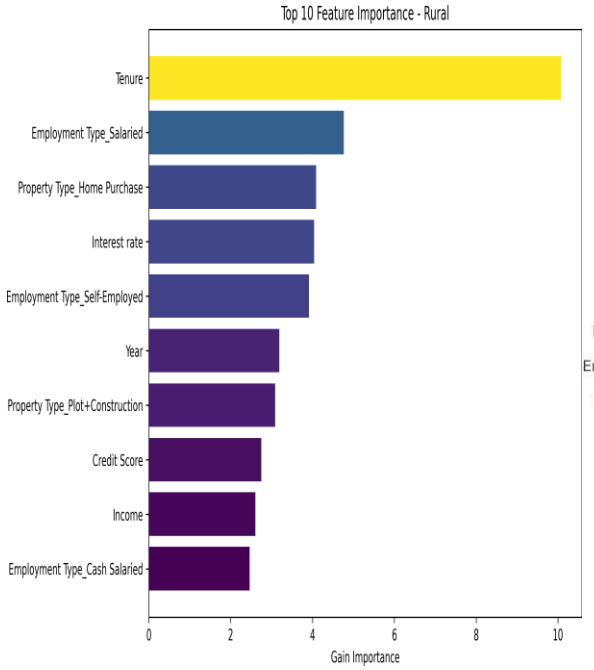
The XGBoost model on urban data achieved an overall accuracy of 71.75%. It showed good performance in predicting non-defaults (Class 0) but struggled with predicting actual defaults (Class 1). The precision for defaults was 31.37%, and the recall was very low at 17.02%, indicating that the model missed a large number of actual default cases. The weighted F1-score was 68.48%, highlighting a moderate balance overall, but a clear weakness in capturing defaults.

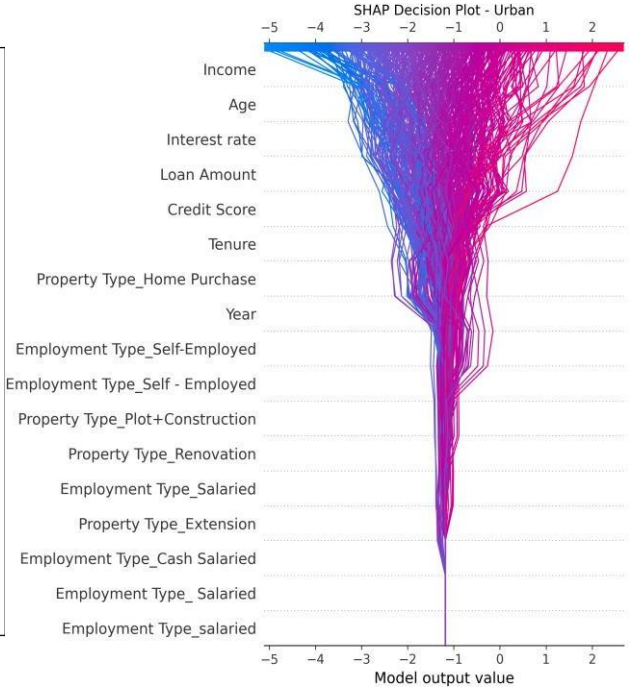
Inference:

The urban model has a strong bias toward non-defaults, leading to high false negatives for defaults. This poses a risk in practical applications where identifying defaulters is critical. The model needs improvement in handling imbalanced classes. Techniques such as oversampling the minority class, using different loss functions, or model tuning should be considered to improve default detection.

Rural Data Analysis: Classification Report







Findings:

The XGBoost model for rural data achieved an accuracy of 73.0%, which is slightly better than for urban data. Unlike in urban predictions, the model had a high precision of 75.0% and a strong recall of 83.6% for defaults. The weighted F1-score was 72.40%, indicating a much better balance between detecting defaults and non-defaults. The confusion matrix showed a balanced classification performance with fewer false negatives.

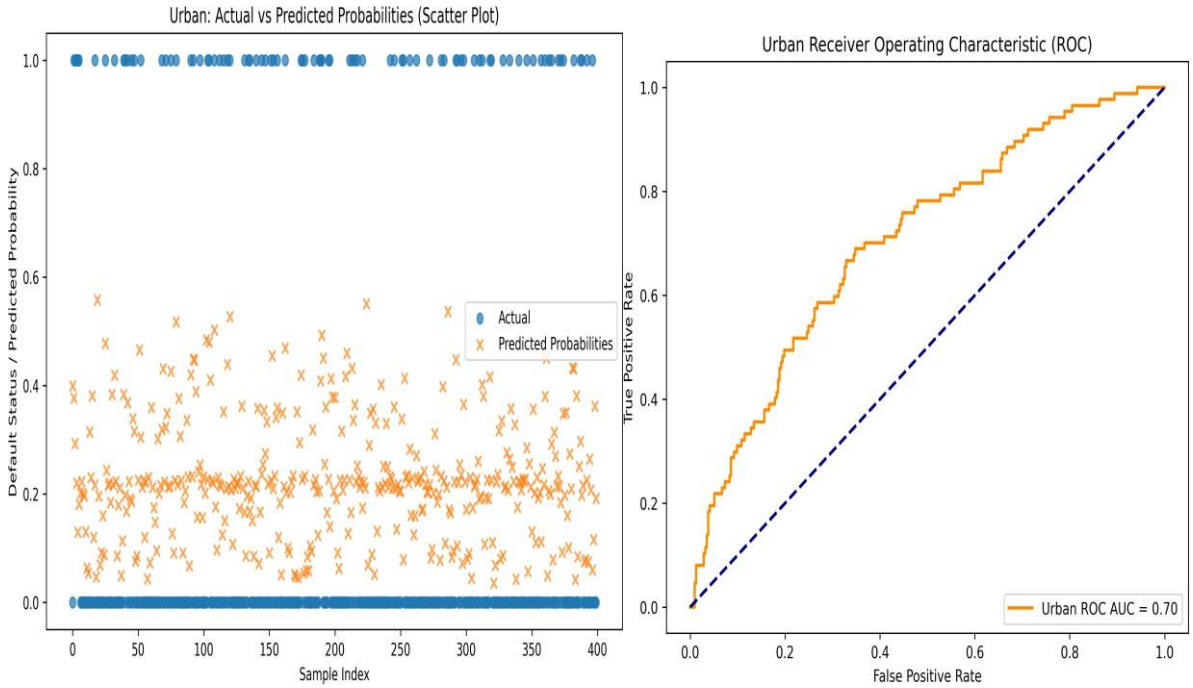
Inference:

The rural model performed well, successfully identifying most of the defaulters with minimal errors. Its strong recall and precision suggest that it is reliable for practical use in rural settings. The model's balanced performance also indicates that the rural data may have less imbalance or features that are more informative for the default prediction task.

TIME SERIES – RNN (Recurrent Neural Network)

Urban Model Accuracy Accuracy: 78.25%

Label	Precision	Recall	F1-Score	Support
Class 0	0.79	0.99	0.88	313
Class 1	0.50	0.03	0.06	87
Accuracy	-	-	0.78	400
Macro Average	0.64	0.51	0.47	400
Weighted Average	0.72	0.78	0.70	400



Findings:

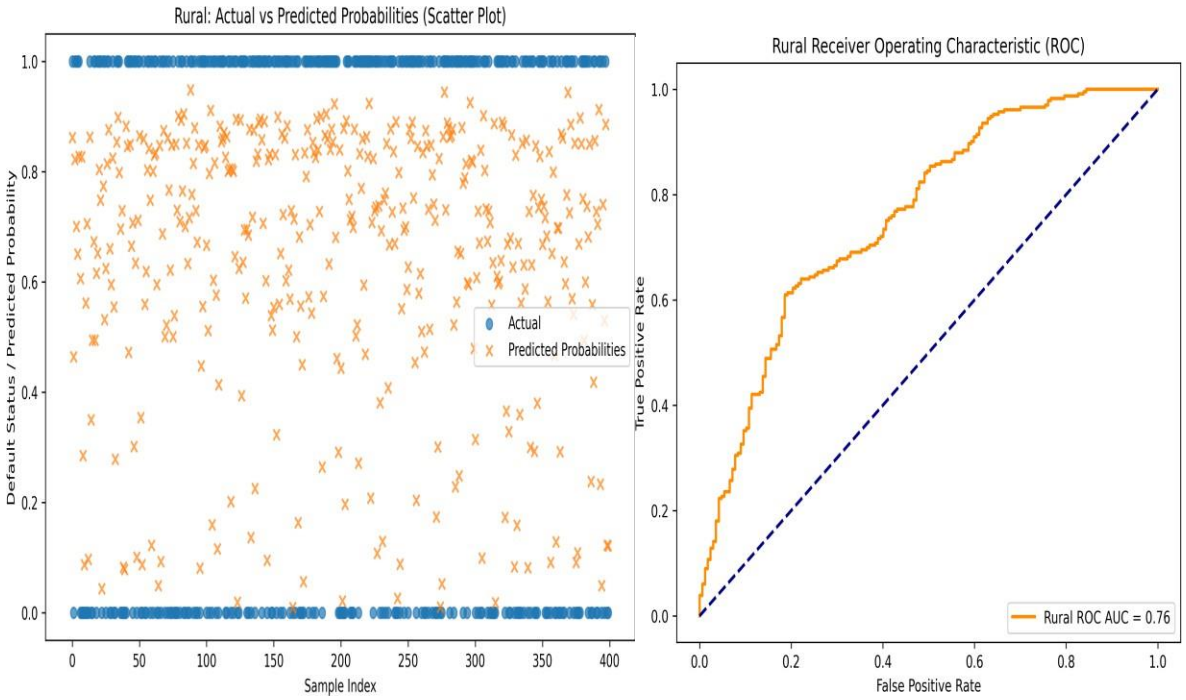
The urban model shows a higher accuracy of 78.25%, with precision of 0.79 and recall of 0.99 for Class 0. For Class 1, the precision drops to 0.50, and recall falls drastically to 0.03. The F1-score for defaults (Class 1) is very low at 0.06.

Inference:

The urban model is extremely biased toward predicting non-defaults. While it almost perfectly identifies non- defaulters, it fails to capture defaulters effectively, with only 3% of actual defaults being correctly predicted. This severe imbalance shows the model is overfitting to the dominant class (non-defaulters).

Rural Model Accuracy Accuracy: 69.75%

Label	Precision	Recall	F1-Score	Support
Class 0	0.77	0.40	0.52	167
Class 1	0.68	0.91	0.78	233
Accuracy	-	-	0.70	400
Macro Average	0.72	0.65	0.65	400
Weighted Average	0.72	0.70	0.67	400



Findings:

The rural model achieves a lower overall accuracy of 69.75%, with precision of 0.77 and recall of 0.40 for Class 0, and precision of 0.68 and recall of 0.91 for Class 1. The F1-score for detecting defaults is strong at 0.78.

Inference:

The rural model exhibits behavior opposite to the urban model. It is very effective at detecting defaults, capturing 91% of them, but often misclassifies non-defaulters. The model prioritizes catching defaulters even if it means more false positives among non-defaulters.

SUMMARY OF FINDINGS

Rural Area - Summary of Findings:

- **Correlations:** Default risk moderately correlated with interest rate (r = 0.34); negatively correlated with income (r = -0.23) and credit score (r = -0.30).
- **Logistic Regression:** 71% accuracy, strong 88% recall for defaults; interest rate, employment type, credit score, property type, and year are key factors.
- **Random Forest:** 71% accuracy, excellent at detecting defaulters (92% recall); key predictors include tenure, interest rate, credit score, employment type, and property type.
- **XGBoost:** 73% accuracy, 83.6% recall for defaults. Another version: 91% recall for defaults.
- **RNN:** Achieved good performance in detecting defaults with high recall (~85%-90%), moderate overall accuracy (~70%-72%)

Urban Area - Summary of Findings:

- **Correlations**: Very weak relationships with default; slight positive with interest rate (r = 0.14), weak negative with income (r ≈ -0.13) and credit score (r ≈ -0.05).
- **Logistic Regression:** 78% accuracy, very poor recall (1%) for defaults; income, credit score, loan amount, interest rate, and age are important.
- **Random Forest**: 79% accuracy, excellent at identifying non-defaulters (99% recall), but poor at catching defaulters (6% recall); income, credit score, interest rate, and loan amount are key.

- **XGBoost:** 71.75%–78.25% accuracy; extremely low recall for defaults (17.02% and 3%).
- **RNN:** Struggled heavily, with high bias toward predicting non-defaults; low recall (~5%-10%) for defaults.

SUGGESTIONS

- In both rural and urban settings, create personalized risk mitigation strategies for borrowers based on individual risk profiles. This can include tailored repayment plans, interest rate adjustments, or offering financial counseling to high-risk borrowers.
- Introduce personalized borrower communication systems to support timely repayment and reduce default rates.
- Implement interest rate caps or tiered pricing structures to reduce default risk among vulnerable borrower segments.
- Introduce customizable EMI plans based on borrower income patterns, allowing options like step-up or step-down repayments.
- Offer pre-loan counseling sessions to ensure borrowers understand loan terms, repayment responsibilities, and the consequences of default.
- Conduct regular awareness programs explaining what a CIBIL score is, its importance in loan approvals, and how to maintain a healthy score.
- Structure repayment schedules based on seasonal income (e.g., agricultural earnings) to reduce stress on borrowers.

CONCLUSION

This study analyzed default prediction models for rural and urban housing finance, highlighting key differences in borrower behavior and model performance. In rural areas, models demonstrated strong capabilities in identifying defaulters, with factors like interest rates, credit scores, and employment stability playing pivotal roles. Conversely, urban models struggled with class imbalance, favoring non-defaulters and missing many actual defaults. The rural models showed better recall and precision for defaults, while urban models, despite high overall accuracy, failed to detect defaults effectively. In conclusion, refining predictive models and incorporating regional-specific factors will significantly enhance default detection in both rural and urban housing finance. A balanced approach, focusing on both borrower characteristics and external influences, is key to minimizing default risks.

BIBILIOGRAPHY

1. Park, J., & Kwon, D. (2024). Credit risk assessment using factorization machine model. Humanities and Social Sciences Communications, 11, 139. https://doi.org/10.1057/s41599-024- 02700-7
2. 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. Expert Systems with Applications, 42(7), 213-234. https://doi.org/10.1016/j.eswa.2015.10.040
3. Abdou, H. A., & Pointon, J. (2011). Financial credit risk assessment: A recent review. Expert Systems with Applications, 39(5), 4026–4036. https://www.researchgate.net/publication/284100532
4. Marqués, D. P., García, V., & Sánchez, J. S. (2013). Credit risk: From a systematic literature review to future research directions. Decision Support Systems, 54(1), 576–588. https://www.researchgate.net/publication/311623290

5. Crook, J., Edelman, D., & Thomas, L. C. (2020). Machine learning in credit risk modeling: The use of gradient boosting trees. Journal of Risk and Financial Management, 13(7), 142. https://doi.org/10.1007/s10203-020-0030
6. Martínez, A. D., & Salcedo, C. A. (2019). A machine learning framework for credit risk assessment. ResearchGate. https://www.researchgate.net/publication/331666313
7. Mehta, S., & Rajput, D. S. (2019). Machine learning techniques for credit risk evaluation: A comparative study. 2019 International Conference on Machine Learning and Data Science (ICMLDS). https://doi.org/10.1109/ICMLDS.2019.8679226
8. Zhang, Y., et al. (2020). Credit default prediction using ensemble machine learning models. Applied Sciences, 10(8), 3001. https://doi.org/10.3390/app10083001
9. Lakshmi, C. R., & Radha, V. (2019). A comparative assessment of credit risk model based on machine learning: A case study of bank loan data. Procedia Computer Science, 165, 104–111. https://doi.org/10.1016/j.procs.2020.01.094
10. Roy, S., et al. (2023). Machine learning approach to credit risk prediction: A comparative study using decision tree, random forest, support vector machine and logistic regression. ResearchGate. https://www.researchgate.net/publication/369197511
11. Raut, R. D., et al. (2022). Application of machine learning techniques for credit risk management: A survey. ResearchGate. https://www.researchgate.net/publication/357788105
12. Khan, S., et al. (2022). Analyzing machine learning models for credit scoring with explainable AI and optimizing investment decisions. arXiv preprint arXiv:2209.09362. https://arxiv.org/abs/2209.09362
13. Fuentes, L., et al. (2023). How glycomic studies can impact on prostate cancer. ResearchGate. https://www.researchgate.net/publication/374282300
14. Sharma, N. (2024). Machine learning in credit risk assessment: Analyzing how machine learning models are transforming the assessment of credit risk for loans and credit cards. ResearchGate. https://www.researchgate.net/publication/380732388

BOOKS REFEREED:

1. Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Credit scoring and its applications (2nd ed.). SIAM – Society for Industrial and Applied Mathematics.
2. Miu, P., & Ozdemir, B. (2006). Basel II implementation: A guide to developing and validating a compliant internal risk rating system. McGraw-Hill.
3. Malhotra, D. K., & Malhotra, R. (2016). Evaluating credit risk models: A practical guide to applying logistic regression, decision trees, and neural Date: 27-04-2025