Predicting Loan Default Using Machine Learning: A Comprehensive Analysis* Nile University

Ahmed Eltohamy
A.Magdyabdelsatar@nu.edu.eg

Ahmed Sarg
A.Sarg@nu.edu.eg

Yousef Farouk
Y.Farouk@nu.edu.eg

Youssef Haitham
Y.Haitham@nu.edu.eg

Yousef Ayman
Y.Abozeid@nu.edy.eg

Abstract—**This paper presents a detailed study on predicting loan default using machine learning techniques. Leveraging a dataset containing demographic, financial, and credit history information of loan applicants, we develop predictive models to identify the likelihood of loan default. We explore various machine learning algorithms, evaluate their performance, and discuss the implications of our findings. The results demonstrate that our proposed solution effectively predicts loan defaults, providing valuable insights for financial institutions to mitigate risk.**

I. Introduction

Loan default prediction is critical for financial institutions to manage risk and maintain financial stability. Accurately predicting whether a borrower will default on a loan can help lenders make informed decisions, reduce losses, and improve overall loan portfolio quality. This study uses a dataset with detailed borrower information, including age, income, home ownership status, employment length, loan intent, loan grade, loan amount, interest rate, and credit history. We aim to develop a machine learning model that can predict loan defaults with high accuracy.

II. Related Work

Several studies have explored the use of machine learning for loan default prediction. Traditional methods, such as logistic regression and decision trees, have been widely used due to their interpretability. Recent advancements in machine learning have introduced more complex models, including random forests, gradient boosting machines, and neural networks, which often provide higher predictive accuracy. However, these models come with increased complexity and reduced interpretability. Our study aims to balance accuracy and interpretability by evaluating multiple models and selecting the best-performing one for practical application.

III. Methodology

To predict loan defaults, we followed a comprehensive methodology, which included data preprocessing, feature engineering, model training, and evaluation.

A. Data Preprocessing

- Removing Duplicates and Handling Missing Values: We first removed any duplicate entries from the dataset to ensure that each loan application was unique. We then selected key columns relevant to our analysis: loan interest rate, employment length, and loan status. We computed the correlation matrix for these variables to understand the relationships between them and printed the correlation values for review.
- Handling Missing Data: To handle missing data in the employment length column, we used a strategy where missing values were replaced with the median value of the column. This approach helps to maintain the central tendency of the data without skewing the results.
- Outlier Removal: We identified and removed outliers in the age and income columns using the Interquartile Range (IQR) method. This involved calculating the first quartile (Q1) and the third quartile (Q3) for both age and income, and then computing the IQR. Outliers were defined as values outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR. By filtering out these outliers, we ensured that extreme values did not adversely affect our model.
- Balancing the Classes: Given that loan defaults might be a minority class, we used the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. SMOTE generates synthetic examples of the minority class to ensure that the machine learning model has enough examples to learn from. This technique helps in mitigating the bias towards the majority class.
- Splitting the Data: The balanced data was then split into training and testing sets. This step involved dividing the dataset such that 80 % of the data was used for training the models and 20 % was reserved for testing the models' performance. This split helps in evaluating how well the model generalizes to unseen data.

B. Model Training

We trained three different machine learning models: Logistic Regression, Decision Tree, and Gradient Boosting. These models were chosen for their varying levels of complexity and interpretability.

Identify applicable funding agency here. If none, delete this.

- Logistic Regression: This model is widely used for binary classification problems and provides a baseline for comparison.
- Decision Tree: A model that splits the data into branches to make predictions, known for its simplicity and interpretability.
- Gradient Boosting: An advanced ensemble technique that builds multiple weak learners (typically decision trees) and combines their predictions to improve accuracy.

C. Model Evaluation

We evaluated the trained models using several metrics, including accuracy, precision, and recall. These metrics provide a comprehensive view of the model's performance:

- Accuracy: The ratio of correctly predicted instances to the total instances.
- Precision: The ratio of correctly predicted positive observations to the total predicted positives, indicating how many of the predicted defaults were actual defaults.
- Recall: The ratio of correctly predicted positive observations to all actual positives, indicating how many of the actual defaults were correctly identified by the model.

D. Confusion Matrix and ROC Curve

a) -: To further assess the models, we plotted the Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve (AUC) for each model. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, while the AUC provides a single scalar value to compare model performance. A higher AUC indicates better model performance.

IV. Results

TABLE I
Logistic Regression Classification Report

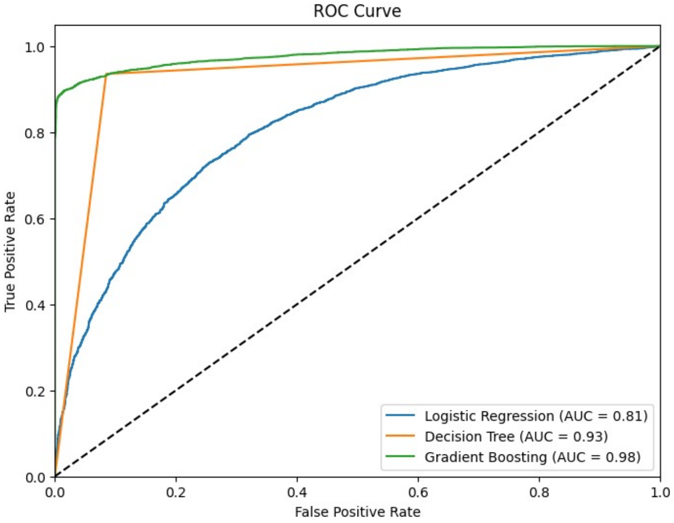
	Precision	Recall	F1-score	Support
0	0.76	0.68	0.72	4536
1	0.72	0.79	0.75	4614
Accuracy			0.74	9150
Macro Avg	0.74	0.74	0.73	9150
Weighted Avg	0.74	0.74	0.73	9150

TABLE II
Decision Tree Classification Report

	Precision	Recall	F1-score	Support
0	0.93	0.92	0.92	4536
1	0.92	0.94	0.93	4614
Accuracy			0.93	9150
Macro Avg	0.93	0.93	0.93	9150
Weighted Avg	0.93	0.93	0.93	9150

TABLE III
Gradient Boosting Classification Report

	Precision	Recall	F1-score	Support
0	0.91	0.97	0.94	4536
1	0.97	0.90	0.93	4614
Accuracy			0.94	9150
Macro Avg	0.94	0.94	0.94	9150
Weighted Avg	0.94	0.94	0.94	9150



V. Discussions

Our findings suggest that advanced machine learning models, such as Gradient Boosting, provide significant improvements in predicting loan defaults compared to traditional methods. Feature importance analysis revealed that loan grade, interest rate, and historical default status were the most influential factors in predicting default. However, the complexity of these models may pose challenges for practical implementation. Future work should focus on improving model interpretability and exploring the use of explainable AI techniques.

VI. Conclusion

This study demonstrates the effectiveness of machine learning in predicting loan defaults. By leveraging a comprehensive dataset and advanced algorithms, we developed a robust predictive model that can assist financial institutions in making informed lending decisions. Our results highlight the importance of incorporating various borrower attributes and credit history information in predictive modeling. Further research should aim to refine these models and address the interpretability challenges to enhance their practical applicability.

VII. References

- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.