

University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria

Volume 15 No. 1, September 2025

journals.ui.edu.ng/uijslictr

<http://uijslictr.org.ng/>

uijslictr@gmail.com



Prediction of Loan Defaulters Using Machine Learning

¹Eze L. Oluchi, ²Mutiat A. Ogunrinde and ³Solomon O. Akinola

^{1,3}Department of Computer Science, University of Ibadan, Ibadan, Nigeria.

Email: ¹Solom202@yahoo.co.uk

²Department of Mathematical and Computer Sciences, Fountain University, Osogbo, Nigeria.

Email: ²Ogunrinde.mutiat@fuo.edu.ng, ²bogunrinde@gmail.com

Abstract

Financial institutions face significant challenges in accurately assessing the risk of loan defaults, which can lead to substantial financial losses and impact overall stability. The primary objective of this study is to develop predictive models that accurately identify potential loan defaulters, enabling lenders to make more informed lending decisions. The study addresses the critical need for more reliable and data-driven credit risk assessment tools by employing logistic regression, random forest, and decision tree algorithms. The research design involves a systematic approach to data collection, preprocessing, feature selection, model development, and evaluation. The dataset, sourced from Coursera's Loan Default Prediction Challenge, includes 255,347 instances and 18 features relevant to loan default prediction. The study employed an under sampling technique to address class imbalance and used train-test split to evaluate model performance. Logistic regression, random forest, and decision tree models were trained and assessed for their predictive capabilities. The results indicate that Logistic regression and random forest models demonstrated superior performance, with accuracy rates of approximately 69% and 68%, respectively. The feature importance analysis revealed key factors influencing loan defaults, such as credit score, loan amount, and employment history.

Keywords: Defaulters, Linear Model, Performance, Financial Institutions, Performance

1. Introduction

A loan is a financial arrangement in which one party (the lender) provides money or assets to another party (the borrower) under the condition that it will be repaid in the future, often with interest. According to Frederic and Apostolos [1], "A loan represents an agreement where the lender gives a borrower resources in return for a promise of future repayment with added interest, which compensates for the time and risk involved in the transaction." When loans are given, the lender expects the borrower to pay back within the agreed-upon time. Most of the time, the borrowers default on the agreement. Many factors usually lead to default on loan repayment, including but not limited to a lousy economy and natural disasters. A defaulter is an individual or entity that fails to meet the legal obligation of repaying a debt or loan according to the agreed-upon terms. Hull

[2] states, "a defaulter refers to a borrower who does not fulfil their repayment commitments as specified in the loan contract, leading to potential legal and financial consequences."

Although managing loans can be challenging, they are essential to financial institutions and the global economy. The danger of default in which borrowers do not return their loans in accordance with the terms of the agreement, is one of the primary obstacles and causes financial losses for lenders. For financial institutions to reduce risks and make wise lending decisions, it is becoming more and more crucial to forecast loan defaulters. Lenders can reduce credit risk and identify possible defaulters with the aid of advanced analytics and machine learning approaches. More precise and predictive models are required to detect possible loan defaulters and enhance credit risk management procedures because traditional credit scoring algorithms frequently rely on incomplete data. The purpose of this study is to use machine learning techniques to create a predictive model for identifying possible loan defaulters.

Eze L. Oluchi, Mutiat A. Ogunrinde and Solomon O. Akinola (2025). Prediction of Loan Defaulters Using Machine Learning, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 15 No. 1, pp. 183 – 197.

©U IJSLICTR Vol. 15, No. 1, September 2025

The application of machine learning (ML) in loan prediction is a rapidly evolving field, leveraging advanced algorithms to enhance the accuracy and reliability of credit risk assessments. However, despite significant progress, several research gaps still remain unaddressed. This gap includes areas like encompassing data quality, model interpretability, algorithm performance, fairness and bias, real-time prediction capabilities, and the integration of alternative data sources. This study is limited to algorithm performance and prediction capabilities.

2. Related Works

Zhang *et. al.* [3] asserted that reliable and effective loan default risk prediction could help regulators and lenders effectively identify risky loan applicants and develop proactive and timely response measures to enhance the stability of the financial system. Traditional prediction models concentrated more on improving loan default prediction accuracy while neglecting to take profit maximization as the goal and evaluation measure of model construction. The study proposed a novel profit-driven prediction model, taking a profit indicator as the optimization objective of the Bayesian optimization to optimize the hyper-parameters of the predictor - categorical boosting.

The Shapley additive explanations (SHAP) value was then calculated to further interpret the relationship between the input variables and the predicted values. Based on two datasets from Renrendai and Lending Club, the results indicated that the proposed model achieved the highest profit-related evaluation metrics values, with mean average extra profit rate values of 3.0872% and 2.1858%, respectively, and mean Profit values of 5168.8762 and 352.9787 in the two datasets, respectively. The SHAP value further revealed the key factors that impacted predictive output for lenders to identify possible defaulters. The argument underscored that incorporation of profit-driven objectives in predictive modeling enhances the financial performance and stability.

Stevenson *et. al.* [4] compared two consumer lending model and found out that mSME credit risk modeling is challenging due to limited data availability. They used Deep Learning and NLP,

including the BERT model, to extract information from textual loan assessments. The study found that text alone was effective in predicting defaulter, but when combined with traditional data, it did not yield additional predictive capability. Their Deep Learning model appeared robust to text quality, suggesting partial automation of the mSME lending process.

Xia *et. al.* [5] proposed a credit scoring model for P2P lending using CatBoost and narrative data extraction techniques. They found that variables extracted from narrative data were powerful features, significantly improving predictability compared to using only hard information. The study suggested that a small number of clusters for soft information extraction were preferred for model performance, computational cost, and comprehensibility.

Bhatore *et. al.* [6] reviewed existing research methods and ML techniques for credit risk evaluation, focusing on credit scoring, NPA prediction, and fraud detection. They found that Ensemble and Hybrid models with neural networks and SVM were more adopted for credit scoring. Lack of comprehensive public datasets was identified as a concern for researchers.

Alam *et. al.* [7] developed a model for credit default prediction using various credit-related datasets. They used Min-Max normalization for feature scaling and data level resampling techniques to address data imbalance. The study demonstrated the effectiveness of different machine learning models and resampling techniques in improving prediction accuracy.

Sheikh *et. al.* [8] focused on the importance of predicting loan defaulters for banks to reduce Non-Performing Assets. The study applied logistic regression on Kaggle data to predict defaulters, finding that including personal attributes improved model performance. The study highlighted the practical relevance for banks, demonstrating the need for predictive models beyond traditional credit scoring methods.

Moscattelli *et. al.* [9] compared machine learning models with statistical models like

logistic regression in predicting default risk. They found that machine learning models performed better when limited information was available, but this advantage diminished with access to confidential data or small datasets. The study argued that while machine learning offered benefits, its superiority depended on the context and availability of data, highlighting the need for nuanced analysis in credit risk management.

Wang *et. al.* [10] compared five machine learning classifiers for credit scoring, finding that Random Forest performed best in terms of precision, recall, AUC, and accuracy. The study centered on the comparative analysis of classifiers, highlighting the strengths of Random Forest in credit scoring applications.

Maheswari and Narayana [11] noted that with the progress of technology and the implementation of Data Science in banking, the face of the banking industry changed significantly. Most of the banking, financial sectors, and social lending platforms actively invested in lending. However, financial institutions might have faced huge capital losses if they approved loans without having any prior assessment of default risk. Financial institutions always needed a more accurate predictive system for various purposes. Predicting loan defaulters was a crucial task for the banking industry. Banks had immensely large amounts of data, including customer data and transaction behavior. Data Science

emerged as a promising area to process the data and extract hidden patterns using machine learning techniques. This study used statistical measures to preprocess the data and build an effective model that predicted loan defaulters accurately. The argument built in this study underscored the necessity of integrating advanced data science techniques to enhance the accuracy of predictive models, thus safeguarding financial institutions against potential losses.

3. Methodology

The methodology employed in this study includes data collection, pre-processing, feature selection, model development and evaluation as depicted in Figure 1.

3.1 Data Collection

Data collection is the first critical step in developing a predictive model for loan defaulters. Data was sourced from Coursera's Loan Default Prediction Challenge dataset, available on Kaggle, which comprises of 255347 rows and 18 columns, offering a comprehensive view of loan applications and their associated attributes. The dataset comprises various features that are crucial for understanding the profile of each loan applicant. These features collectively provide a comprehensive profile of each loan applicant, enabling the predictive models to assess the risk of loan default accurately. Each feature contributes uniquely to the overall prediction:

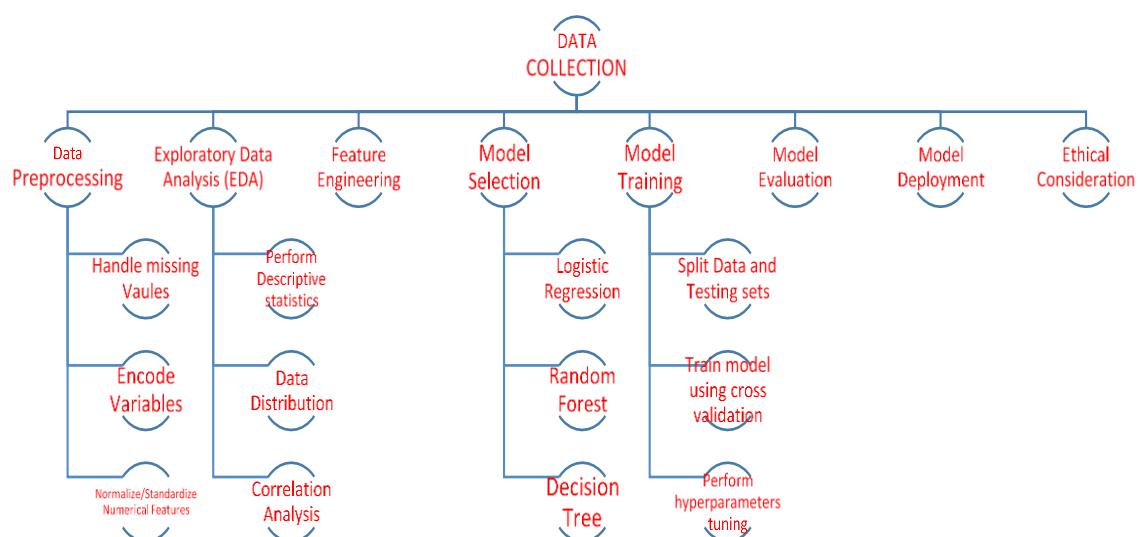


Figure 1: Flowchart Process

1. Age and Income help gauge the financial maturity and capacity of the borrower.
2. LoanAmount and InterestRate directly affect the repayment burden.
3. CreditScore and MonthsEmployed provide insights into the borrower's credit worthiness and job stability.
4. NumCreditLines and DTIRatio highlight the borrower's existing financial obligations.
5. LoanTerm determines the repayment duration, influencing the borrower's ability to manage monthly payments.

Each feature was carefully selected to enhance the predictive accuracy of the models, considering their relevance and impact on loan default risk. Part of the dataset table is shown in Figure 2.

3.2 Data Preprocessing

Data preprocessing is essential in preparing the collected data for analysis and modelling. The preprocessing phase involves several critical tasks to transform raw data into a clean, structured format suitable for machine learning, including handling missing values common in large datasets, using the Standard Scale to standardize their values and preventing features with larger magnitudes from dominating the model training process as shown in Figure 3. Meanwhile, the categorical features were encoded using a categorical encoding method to convert them into numerical format, and

balancing the data set which is required by machine learning algorithms

3.3 Feature Selection

A correlation matrix heat-map was created to visualize the correlation between different features as shown in Figure 4. This helps in identifying highly correlated features that might be redundant for the model, aiding in feature selection and improving the model's efficiency. All the features were negatively correlated, meaning there is homeostaticity and identity in the data.

3.4 Handling Class Imbalance

The dataset exhibited class imbalance, with a larger number of non-default cases compared to default cases. To address this, the majority class (non-default) was down-sampled to match the minority class (default) using the resample function. This step helped balance the class distribution and prevent the model from being biased towards the majority class as in Figure 5.

3.5 Train-Test Split

The final step involved splitting the balanced dataset into training and testing sets using the train_test_split function. This step ensures that the model's performance is evaluated on unseen data, helping to assess its generalization capabilities. The training set comprised 44,479 samples, while the testing set comprised 14,827. This is shown in Figure 6.

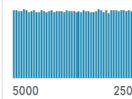
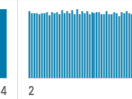
# Age	# Income	# LoanAmount	# CreditScore	# MonthsEmployed	# NumCreditLines	# InterestRate	# LoanTerm
The age of the borrower	The annual income of the borrower	The amount of money being borrowed	The credit score of the borrower	The number of months the borrower has been employed	The number of credit lines the borrower has open	The interest rate for the loan	The term length of the loan in months
							
18 69	15.0k 150k	5000 250k	300 849	0 119	1 4	2 25	12 60
56	85994	58587	528	88	4	15.23	36
69	58432	124448	458	15	1	4.81	68
46	84288	129188	451	26	3	21.17	24
32	31713	44799	743	8	3	7.87	24
68	28437	9139	633	8	4	6.51	48
25	98298	98448	728	18	2	22.72	24
38	111188	177825	429	88	1	19.11	12
56	126882	155511	531	67	4	8.15	68
36	42853	92357	827	83	1	23.94	48
48	132784	228518	488	114	4	9.89	48

Figure 2 Pictorial diagram of the Dataset Table

```

# Drop the LoanID column
loan_data.drop('LoanID', axis=1, inplace=True)

# Select X and Y
X = loan_data.drop(['Default'], axis=1)
Y = loan_data['Default']

# Separate numerical and categorical columns
numerical_cols = X.select_dtypes(include=['int64', 'float64']).columns.tolist()
categorical_cols = X.select_dtypes(include=['object']).columns.tolist()

# Scale the numerical columns
scaler = StandardScaler()
X_scaled = pd.DataFrame(scaler.fit_transform(X[numerical_cols]), columns=numerical_cols)

for col in categorical_cols:
    X[col] = X[col].astype('category').cat.codes

# Combine the scaled numerical columns with the categorical columns
X = pd.concat([X_scaled, X[categorical_cols]], axis=1)

X.head()

```

	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Education	EmploymentType	MaritalStatus	HasMo
0	0.833990	0.089693	-1.086833	-0.341492	0.590533	1.341937	0.261771	-0.001526	-0.260753	0	0	0	
1	1.701221	-0.823021	-0.044309	-0.731666	-1.285731	-1.343791	-1.308350	1.412793	0.778585	2	0	1	
2	0.166888	0.043854	0.022715	-0.775718	-0.968209	0.446694	1.156831	-0.708685	-0.823728	2	3	0	
3	-0.767053	-1.303452	-1.168538	1.061875	-1.718715	0.446694	-0.967805	-0.708685	-1.170174	1	0	1	
4	1.100830	-1.592855	-1.671921	0.369631	-1.487790	1.341937	-1.052188	0.705634	0.995114	0	3	0	

Figure 3: Scaling Data and Encoding

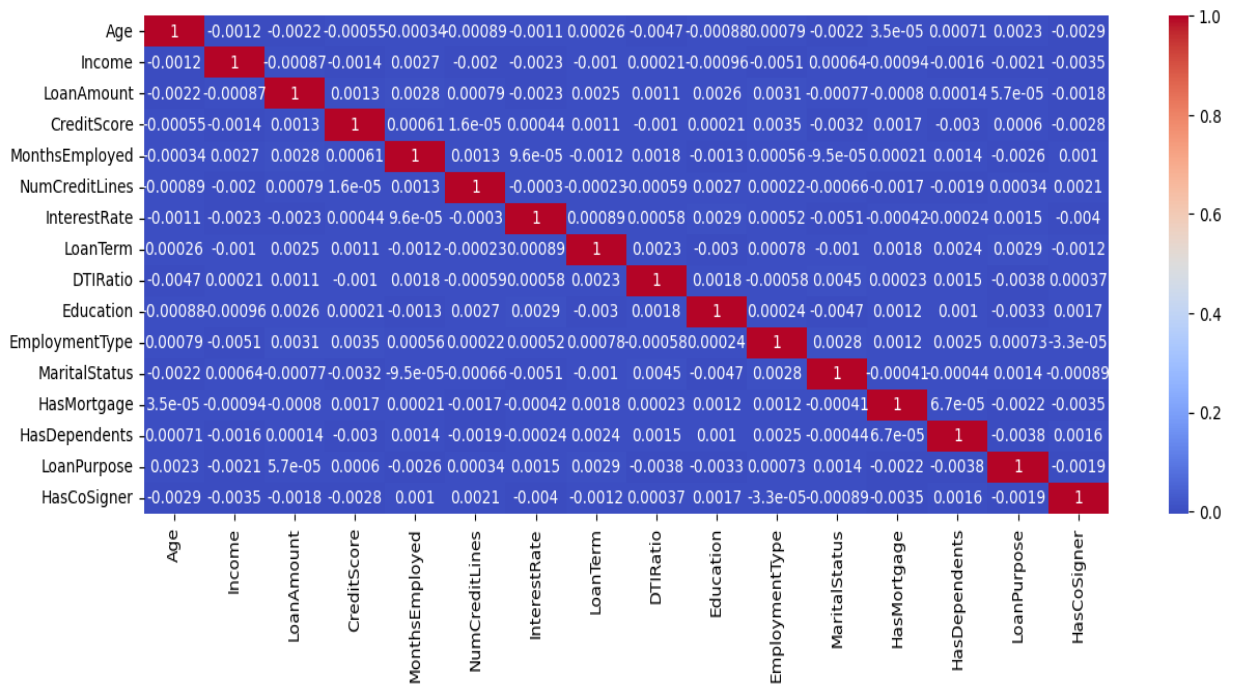


Figure 4: Correlation Matrix

```

from sklearn.utils import resample

# Separate majority and minority classes
majority = loan_data[loan_data['Default'] == 0]
minority = loan_data[loan_data['Default'] == 1]

# Downsample majority class
majority_downsampled = resample(majority, replace=False, n_samples=len(minority), random_state=42)

# Combine minority and downsampled majority
loan_data_balanced = pd.concat([majority_downsampled, minority])

# Shuffle the rows
loan_data_balanced = loan_data_balanced.sample(frac=1, random_state=42)

# Check the class distribution
loan_data_balanced['Default'].value_counts()

```

```

Default
0      29653
1      29653
Name: count, dtype: int64

```

Figure 5: Handling Imbalance

```

X_train, X_test, Y_train, Y_test = train_test_split(loan_data_balanced[numerical_cols + categorical_cols], loan_data_balanced['Default'], test_size=0.25, r

```

```

print('Training data shape:', X_train.shape)
print('Training labels shape:', Y_train.shape)
print('Testing data shape:', X_test.shape)
print('Testing labels shape:', Y_test.shape)

```

```

Training data shape: (44479, 16)
Training labels shape: (44479,)
Testing data shape: (14827, 16)
Testing labels shape: (14827,)

```

Figure 6: Data Splitting

3.6 Feature Selection

Feature selection is identifying the most relevant variables that significantly impact the prediction of loan defaults. In this study, feature selection was performed using advanced statistical and machine learning techniques to ensure that the model focuses on the most informative attributes. Methods such as correlation analysis was used to identify relationships between features and the target variable (loan default status). Additionally, techniques like Principal Component Analysis (PCA) can help reduce the dataset's dimensionality. Recursive Feature Elimination (RFE) is another technique that was employed to iteratively remove less important features, thus refining the model to include only the most significant predictors.

3.7 Model Development

Model development is the phase where the actual predictive models are created using the preprocessed and selected features, various machine learning algorithms were explored to identify the most effective model for predicting

loan defaulters. Algorithms such as logistic regression, decision trees, random forests, and neural networks were evaluated. The model development phase involves training these algorithms on the preprocessed dataset, allowing them to learn the underlying patterns that distinguish between defaulters and non-defaulters. The training process involves splitting the data into training and validation sets to fine-tune the model parameters and avoid overfitting. Techniques such as cross-validation was used to ensure that the model generalizes well to unseen data. The ultimate goal of model development in this study is to create a robust and accurate predictive model that can effectively identify potential loan defaulters, thereby enhancing the credit risk management practices of banks.

3.8 Model Evaluation

Model evaluation is a crucial step to assess the performance and effectiveness of the developed predictive models. The evaluation was conducted using several four (4) key metrics, including accuracy, precision, recall, and F1-

score. Accuracy measures the overall correctness of the model, while precision indicates the proportion of true positive predictions among all positive predictions made by the model. Recall, on the other hand, measures the model's ability to identify all actual positive cases (i.e., actual defaulters). The F1-score provides a balance between precision and recall, offering a single metric that considers both false positives and false negatives.

4.0 Results And Findings

4.1 Dataset Information

The dataset used in this study was sourced from Coursera's Loan Default Prediction Challenge, providing a unique opportunity to tackle a real-world machine learning problem in the financial sector. Figure 7 shows the dataset comprises 255,347 rows and 18 columns, offering a comprehensive view of loan applications and their associated attributes. Each row represents a loan application, while the columns include features such as applicant demographics, loan details, and historical payment information.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain insights into the dataset and understand the

relationships between variables. This involved visualizing the distribution of key features, identifying outliers, and examining correlations between variables. EDA insights helped understand the data better and guide the feature selection process for building predictive models.

The histogram in Figure 8 shows the distribution of ages among loan applicants. The majority of applicants fall between the ages of 30 and 60, with a peak around the age of 40. This distribution suggests that the dataset contains a relatively balanced representation of different age groups, which is important for ensuring that the predictive model is not biased towards any particular age group.

The histogram in Figure 9 displays the distribution of income levels among loan applicants. The distribution is right-skewed, with the majority of applicants having lower to moderate income levels. This skewness indicates that there may be a larger number of lower-income applicants in the dataset compared to higher-income applicants, which could impact the model's predictions regarding loan defaults.

	Column_name	Column_type	Data_type	Description
0	LoanID	Identifier	string	A unique identifier for each loan.
1	Age	Feature	integer	The age of the borrower.
2	Income	Feature	integer	The annual income of the borrower.
3	LoanAmount	Feature	integer	The amount of money being borrowed.
4	CreditScore	Feature	integer	The credit score of the borrower, indicating their creditworthiness.
5	MonthsEmployed	Feature	integer	The number of months the borrower has been employed.
6	NumCreditLines	Feature	integer	The number of credit lines the borrower has open.
7	InterestRate	Feature	float	The interest rate for the loan.
8	LoanTerm	Feature	integer	The term length of the loan in months.
9	DTIRatio	Feature	float	The Debt-to-Income ratio, indicating the borrower's debt compared to their income.
10	Education	Feature	string	The highest level of education attained by the borrower (PhD, Master's, Bachelor's, High School).
11	EmploymentType	Feature	string	The type of employment status of the borrower (Full-time, Part-time, Self-employed, Unemployed).
12	MaritalStatus	Feature	string	The marital status of the borrower (Single, Married, Divorced).
13	HasMortgage	Feature	string	Whether the borrower has a mortgage (Yes or No).
14	HasDependents	Feature	string	Whether the borrower has dependents (Yes or No).
15	LoanPurpose	Feature	string	The purpose of the loan (Home, Auto, Education, Business, Other).
16	HasCoSigner	Feature	string	Whether the loan has a co-signer (Yes or No).
17	Default	Target	integer	The binary target variable indicating whether the loan defaulted (1) or not (0).

Figure 7 Dataset Information

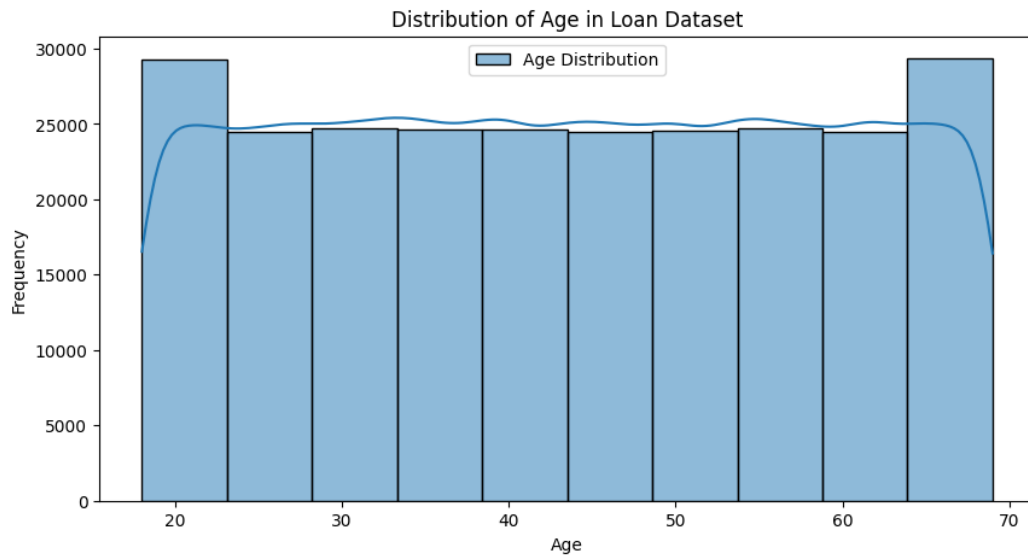


Figure 8: Distribution of Age in Loan Dataset

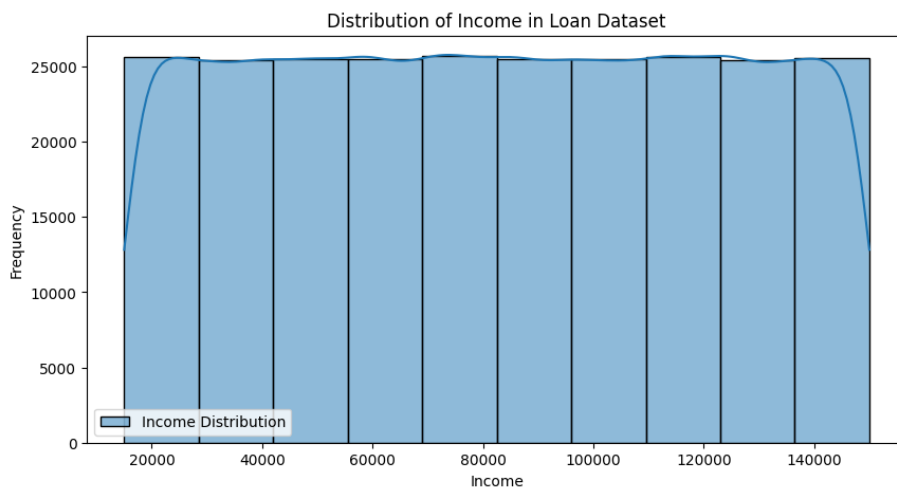


Figure 9: Distribution of Income in Loan Dataset

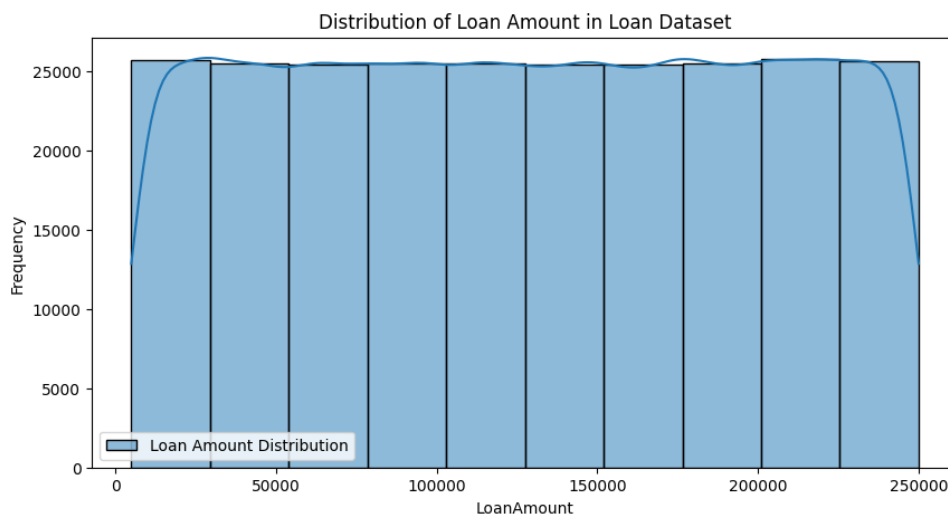


Figure 10: Distribution of Loan Amount in Loan Dataset

The histogram in Figure 10 illustrates the distribution of loan amounts among applicants. The distribution is relatively uniform, indicating that loan amounts are spread evenly across applicants. This even distribution suggests that the dataset contains a diverse range of loan amounts, which is important for capturing the full range of loan default behaviors in the model.

The bar chart in Figure 11 represents the distribution of education levels among loan

applicants. It shows the frequency of different education categories, which include categories like high school, college, and graduate education. This distribution provides insights into the educational background of loan applicants, which can be an important factor in predicting loan defaults. Understanding the education level of applicants helps in assessing their financial literacy and stability, which are crucial factors in determining their likelihood of defaulting on loans.

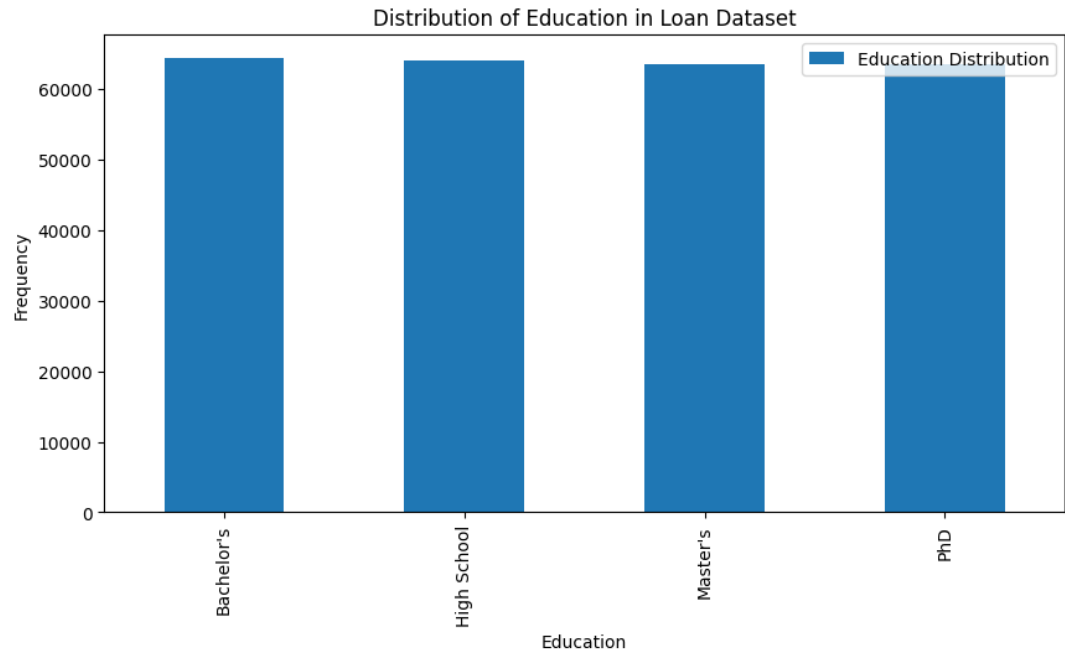


Figure 11: Distribution of Education in Loan Dataset

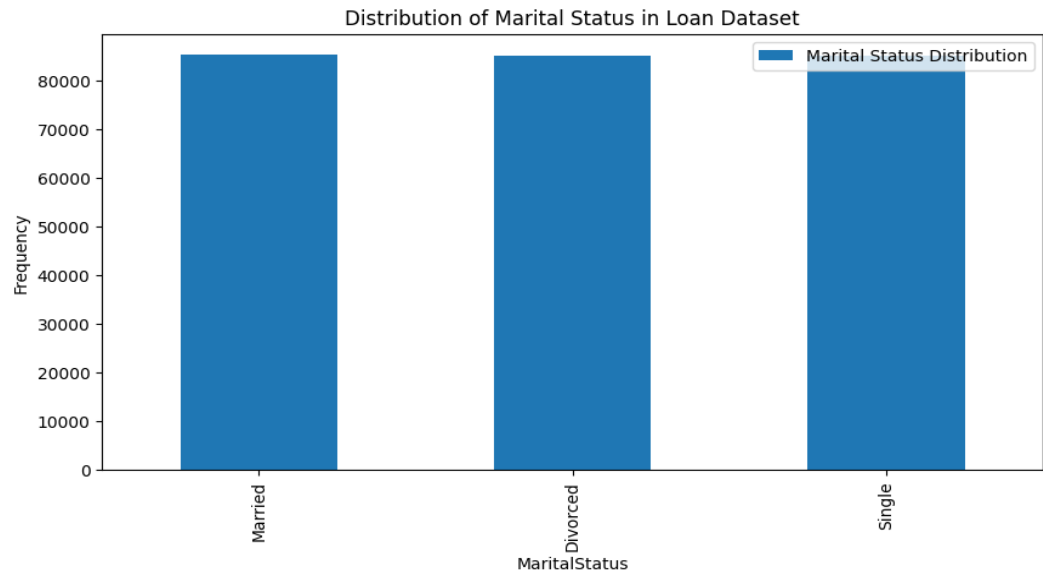


Figure 12: Distribution of Marital Status in Loan Dataset

Figure 12 illustrates the distribution of marital status among loan applicants. The bar chart shows the frequency of different marital status categories, such as single, married, divorced, or widowed. Marital status can be a significant factor in predicting loan defaults, as it can reflect stability and support systems available to the applicant. For instance, married individuals may have dual incomes and more financial stability, potentially reducing their risk of default compared to single individuals.

The bar chart in Figure 13 displays the distribution of default status among loan applicants. It shows the frequency of applicants who defaulted on their loans compared to those who did not. This distribution is crucial for understanding the prevalence of loan defaults in

the dataset and provides a baseline for evaluating the performance of predictive models. A balanced distribution of default and non-default cases is essential for training the model effectively and ensuring that it can accurately predict defaults.

Figure 14 is a scatter plot showing the relationship between loan amount and income among loan applicants. The plot helps visualize the distribution of loan amounts relative to income levels and provides insights into any potential patterns or trends. Understanding this relationship is crucial for predicting loan defaults, as individuals with higher loan amounts relative to their income may be at a higher risk of defaulting.

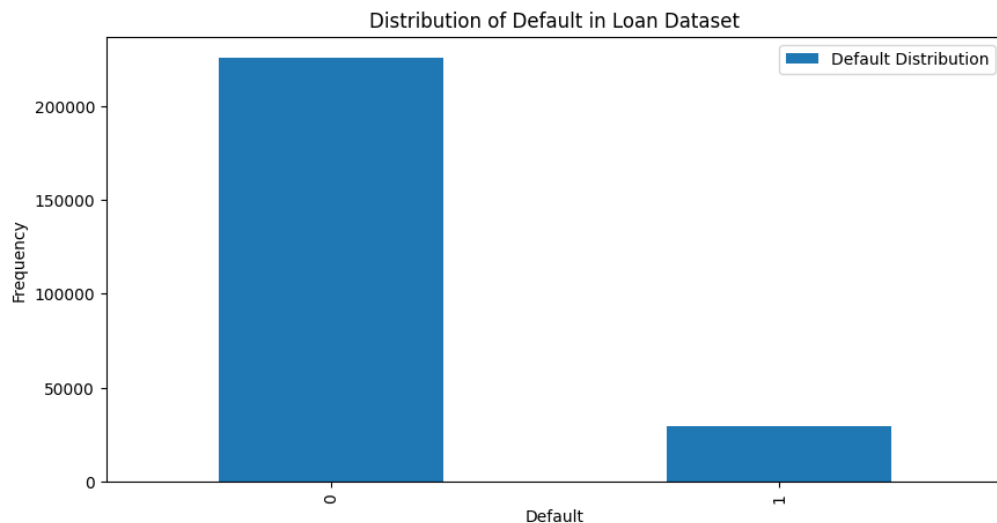


Figure 13: Distribution of Default in Loan Dataset

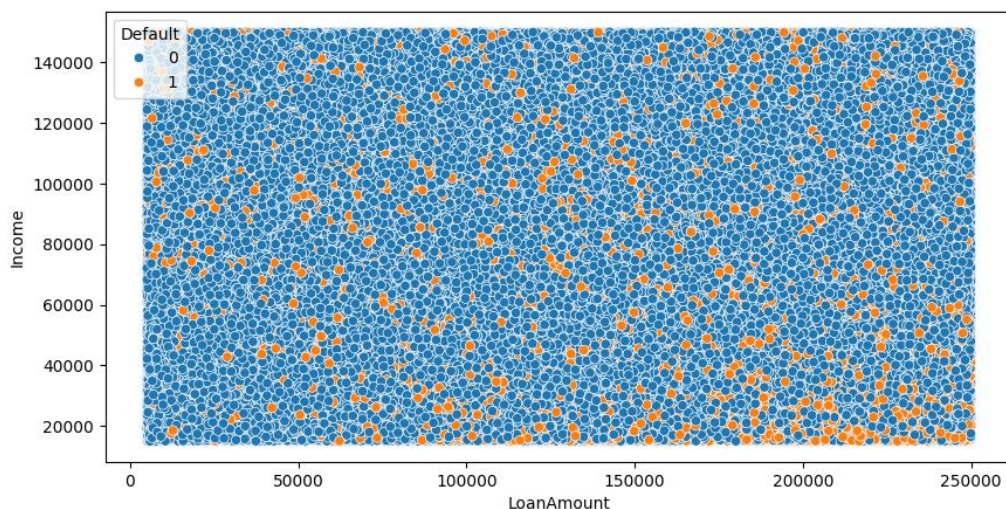


Figure 14: Relationship between Loan Amount and Income

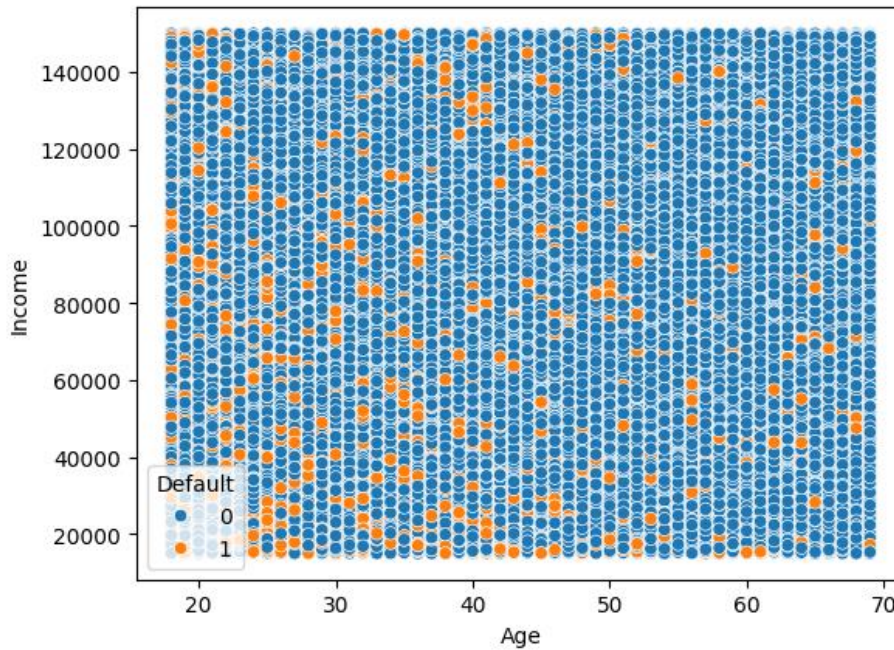


Figure 15: Relationship between Age and Income

The scatter plot in Figure 15 illustrates the relationship between age and income among loan applicants. It helps visualize how income levels vary across different age groups and provides insights into income trends among applicants. Age can be a significant factor in predicting loan defaults, as older individuals may have more stable incomes and financial behaviors compared to younger individuals.

4.3 Model Creation And Training

Three machine learning models were created and trained using the preprocessed dataset to predict loan defaults: Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. Logistic Regression is a linear model used for binary classification tasks. It models the probability of the default class using a logistic function. The logistic regression model was trained using the Logistic Regression class from scikit-learn, with default parameters and a random state of 42.

Random Forest Classifier is an ensemble learning method that constructs a multitude of decision trees during training. It outputs the class that is the mode of the classes predicted by individual trees. The random forest model was trained using the RandomForestClassifier

class from scikit-learn, with default parameters and a random state of 42. Decision Tree Classifier is a tree-like model where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. The decision tree model was trained using the DecisionTreeClassifier class from scikit-learn, with default parameters and a random state of 42.

The next step after training the models is to evaluate their performance using the testing set to determine which model performs best in predicting loan defaults. This evaluation will be based on metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the models' effectiveness and help in selecting the best model for predicting loan defaults in practical applications.

4.4. Model Evaluation

The evaluation of the predictive models involved assessing their performance on the testing set using various metrics, including accuracy, precision, recall, and F1-score. These metrics provided a comprehensive understanding of each model's ability to predict loan defaults accurately.

```
logistic_model = LogisticRegression(random_state=42)
logistic_model.fit(X_train, Y_train)
```

▼ **LogisticRegression**
LogisticRegression(random_state=42)

```
random_model = RandomForestClassifier(random_state=42)
random_model.fit(X_train, Y_train)
```

▼ **RandomForestClassifier**
RandomForestClassifier(random_state=42)

```
decision_model = DecisionTreeClassifier(random_state=42)
decision_model.fit(X_train, Y_train)
```

▼ **DecisionTreeClassifier**
DecisionTreeClassifier(random_state=42)

Figure 16: Model Training

Table 1 summarizes the accuracy, precision, recall, and F1-score for each model:

Table 1 Performance Metrics

Metrics	Logistic Regression	Random Forest	Decision Tree
Accuracy	0.69	0.68	0.58
Precision (0)	0.69	0.68	0.57
Recall (0)	0.67	0.69	0.58
F1-score (0)	0.68	0.68	0.58
Precision (1)	0.68	0.69	0.58
Recall (1)	0.70	0.68	0.57
F1-score (1)	0.69	0.68	0.58

Comparison Chart

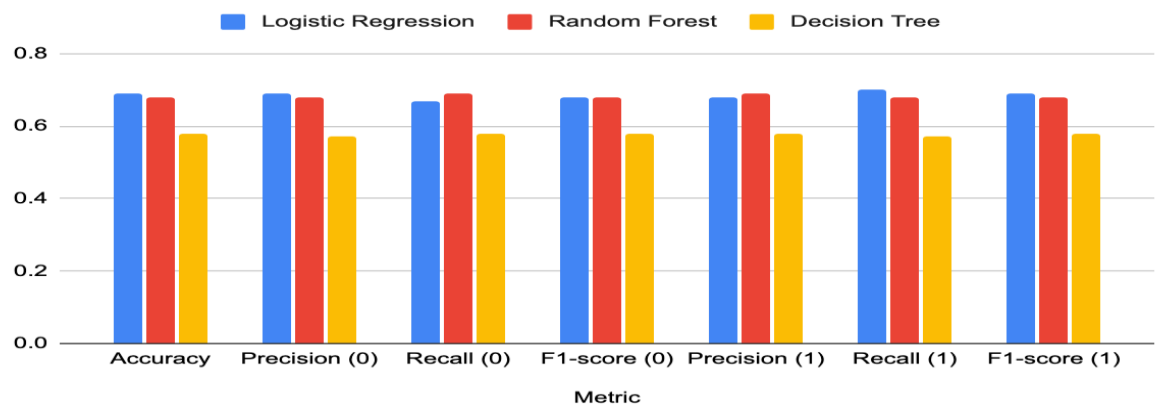


Figure 17 Comparison Chart

These results indicated that logistic regression and random forest models outperformed the decision tree model.

4.5 Confusion Matrix Visualization

To further understand the model's performance, confusion matrices were plotted for each model, illustrating the number of true positive, true negative, false positive, and false negative predictions. Figure 18 shows the confusion matrices for logistic regression, random forest, and decision tree models

4.6 Discussion of Results

The predictive models were evaluated to determine their ability to accurately identify potential loan defaulters. The three models assessed were Logistic Regression, Random Forest, and Decision Tree, using accuracy, precision, recall, and F1-score metrics. These metrics provided a comprehensive understanding of each model's performance.

The Logistic Regression model demonstrated balanced performance, achieving an accuracy of 69%. The precision and recall values were close for defaulters and non-defaulters, with an F1-score of 0.68 for non-defaulters and 0.69 for defaulters. This indicates that the model was reasonably effective in distinguishing between the two classes. This finding aligns with the work of Patel *et. al.* [14].

The Random Forest model also showed balanced performance, with an accuracy of

68%. The precision and recall values were nearly equal for both classes, and the F1-score was 0.68 for both non-defaulters and defaulters. This model's ability to handle diverse data points and provide consistent performance is supported by Maheswari and Narayana [11].

The Decision Tree model exhibited lower performance, with an accuracy of 58%. The precision and recall were similar for both classes but significantly lower than those of the Logistic Regression and Random Forest models. The F1-score was 0.58 for both non-defaulters and defaulters, reflecting the model's limited capability in accurately predicting loan defaults. This lower performance is consistent with findings from Aslam *et. al.* [13] who noted that Decision Trees are useful for understanding data and making quick decisions, they often fall short in accuracy compared to more complex models like Random Forest.

Comparing these models highlights the importance of using advanced machine learning techniques for predicting loan defaults. The results showed that Logistic Regression and Random Forest outperformed the Decision Tree model, providing more accurate and reliable predictions. This aligns with the conclusions of Moscatelli *et. al.* [9], who compared machine learning models with traditional statistical models and found that machine learning models, such as Random Forest, generally performed better when handling large and complex datasets.

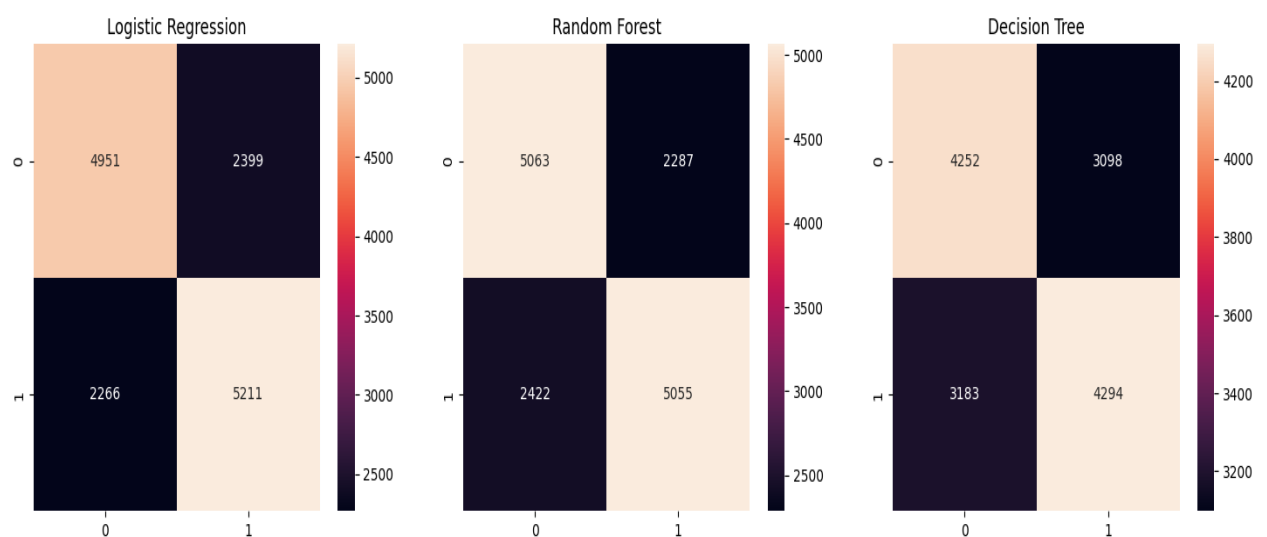


Figure 18: Confusion Matrix

This study's approach provided valuable insights for lenders to identify potential defaulters, emphasizing the importance of considering financial gains alongside prediction accuracy. The results of this study, coupled with insights from previous researches like Zhang *et. al.* [3], underscore the critical role of advanced predictive models in the banking sector. Accurate prediction of loan defaulters enables banks and financial institutions to make informed decisions, reduce non-performing assets, and enhance overall financial stability. The balanced performance of Logistic Regression and Random Forest models in this study highlights their effectiveness in predicting loan defaults. It supports the argument for adopting advanced machine learning techniques in credit risk management.

5. Conclusion

This study successfully developed machine learning models for predicting loan defaults, demonstrating their potential to enhance credit risk management. The balanced performance of Logistic Regression and Random Forest models highlights their suitability for practical applications in the financial sector. These findings underscore the transformative role of machine learning in mitigating credit risks and promoting financial stability.

Furthermore, Handling missing values, encoding categorical variables, and scaling numerical features were critical steps in preparing the dataset for model training. The use of under-sampling to address class imbalance ensured that the models were trained on a balanced dataset, improving their ability to identify defaulters accurately.

As part of recommendations for further study in this research domain, the following recommendations are suggested

1. Incorporating additional data sources into predictive models that covers external data such as social media activity, transaction history, and macroeconomic indicators could further enhance the accuracy of default predictions.
2. Future research should investigate the use of advanced machine learning techniques such as deep learning and ensemble learning.
3. Performing Longitudinal studies evaluating the long-term effectiveness of

predictive models would provide valuable insights.

References

- [1] Frederic S. Mishkin and Apostolos Serletis (2019). *The Economics of Money, Banking and Financial Markets*, Pearson Education Canada, 28 Jan 2019 - Business & Economics - 736 pages
- [2] Hull John (2018). *Risk Management and Financial Institutions*, Wiley.
- [3] Zhang, L., Wang, J., and Liu, Z. (2023). What should lenders be more concerned about? *Developing a profit-driven loan default prediction model. Expert Systems with Applications*, 213, 118938.
- [4] Stevenson, M., Mues, C. and Bravo, C., 2021. *The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research*, 295(2), pp.758-771.
- [5] Xia, Y., He, L., Li, Y., Liu, N. and Ding, Y., 2020. *Predicting loan default in peer-to-peer lending using narrative data. Journal of Forecasting*, 39(2), pp.260-280.
- [6] Bhatore, S., Mohan, L. and Reddy, Y.R., 2020. *Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology*, 4(1), pp.111-138.
- [7] Alam, T.M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U., Shabbir, S., Li, J. and Khushi, M., 2020. *An investigation of credit card default prediction in the imbalanced datasets. Ieee Access*, 8, pp.201173-201198.
- [8] Sheikh, M.A., Goel, A.K. and Kumar, T., 2020, July. *An approach for prediction of loan approval using machine learning algorithm. In 2020 international conference on electronics and sustainable communication systems (ICESC) (pp. 490-494). IEEE*.
- [9] Moscatelli, M., Parlapiano, F., Narizzano, S. and Viggiano, G., 2020. *Corporate default forecasting with machine learning. Expert Systems with Applications*, 161, p.113567.
- [10] Wang, Y., Zhang, Y., Lu, Y. and Yu, X., 2020. *A Comparative Assessment of Credit Risk Model Based on Machine Learning a case study of bank loan data. Procedia Computer Science*, 174, pp.141-149.
- [11] Maheswari, P., and Narayana, C. V. (2020, October). *Predictions of loan defaulter-A data science perspective. In 2020 5th International*

Conference on Computing, Communication and Security (ICCCS) (pp. 1-4). IEEE.

- [12] Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K., 2019. *A study on predicting loan default based on the random forest algorithm*. *Procedia Computer Science*, 162, pp.503-513.
- [13] Aslam, U., Tariq Aziz, H.I., Sohail, A. and Batcha, N.K., 2019. *An empirical study on loan default prediction models*. *Journal of Computational and Theoretical Nanoscience*, 16(8), pp.3483-3488.
- [14] Patel, B., Patil, H., Hembram, J., and Jaswal, S. (2020, June). *Loan default forecasting using data mining*. In 2020 international conference for emerging technology (INCET) (pp. 1-4). IEEE.