# Loan Default Prediction Using Machine Learning

**Submitted by Group 1:**
Pragya Gupta RBA55
Sahil Jain RBA39
Aman Dwivedi RBA24
Krishna Agrawal RBA23

**Abstract**

This study develops and evaluates multiple machine learning models to predict loan default outcomes using borrower demographic, financial, and credit-related attributes. Using a dataset of 45,000 loan records, the analysis follows a structured pipeline comprising exploratory data analysis (EDA), statistically driven outlier treatment, feature engineering, multicollinearity management, and comparative model evaluation. All insights are supported strictly by visualizations generated within the accompanying notebook, ensuring analytical consistency and reproducibility.

From a business perspective, accurate loan default prediction is critical for minimizing credit losses, optimizing risk-based pricing, and improving portfolio quality. Logistic regression and tree-based ensemble models are assessed using accuracy, precision, recall, F1-score, and ROC–AUC. The results highlight a trade-off between interpretability and predictive performance, with tree-based ensembles delivering superior risk discrimination, while regularized logistic regression remains valuable for transparent, regulator-friendly credit decisions.

## 1. Introduction

Loan default prediction is a core problem in credit risk management, directly impacting profitability, regulatory capital, and financial stability. Traditional credit scoring models rely on linear assumptions, which may fail to capture complex, non-linear borrower behavior. This project investigates whether modern machine learning models can improve predictive performance while maintaining interpretability suitable for academic and regulatory contexts.

**Objective:**

- Predict loan default status (binary classification)
- Compare linear, regularized, and tree-based classification models
- Identify key risk drivers influencing loan rejection

## 2. Dataset Description

The dataset consists of **45,000 observations and 14 original variables**, capturing borrower demographics, financial capacity, credit history, and loan characteristics. These variables closely mirror information typically collected during retail loan underwriting.

**Table 1: Original Variables and Their Definitions**

| Variable Name | Type | Definition | Business Interpretation |
|---|---|---|---|
| person_age | Numerical | Age of the loan applicant (in years) | Proxy for life-stage risk and income stability |
| person_gender | Categorical | Gender of the applicant (male/female) | Used for demographic analysis (not a primary risk driver) |
| person_education | Categorical | Highest education level attained | Indicator of earning potential and job stability |
| person_income | Numerical | Annual income of the applicant | Core measure of repayment capacity |
| person_emp_exp | Numerical | Employment experience in years | Reflects income stability and career maturity |
| person_home_ownership | Categorical | Home ownership status (Rent, Own, Mortgage, Other) | Proxy for financial stability and asset backing |
| loan_amnt | Numerical | Loan amount requested | Higher values imply greater exposure and potential loss |
| loan_intent | Categorical | Purpose of the loan | Risk varies by usage (e.g., medical vs personal) |
| loan_int_rate | Numerical | Interest rate charged on the loan | Higher rates signal higher perceived credit risk |
| loan_percent_income | Numerical | Loan amount as a proportion of income | Measures repayment burden on the borrower |
| cb_person_cred_hist_length | Numerical | Length of credit history (years) | Longer histories provide better risk visibility |
| credit_score | Numerical | Credit score of the applicant | Primary indicator of past credit behavior |
| previous_loan_defaults_on_file | Categorical | Whether the applicant defaulted previously | Strong predictor of future default risk |
| loan_status | Binary (Target) | Loan outcome (0 = approved, 1 = rejected) | Business decision variable |

**3. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was conducted to understand borrower behavior, assess data quality, and identify patterns relevant to credit risk assessment. From a lending institution's perspective, this step is essential to ensure that models are built on economically meaningful and statistically stable inputs.

### 3.1 Data Quality Checks

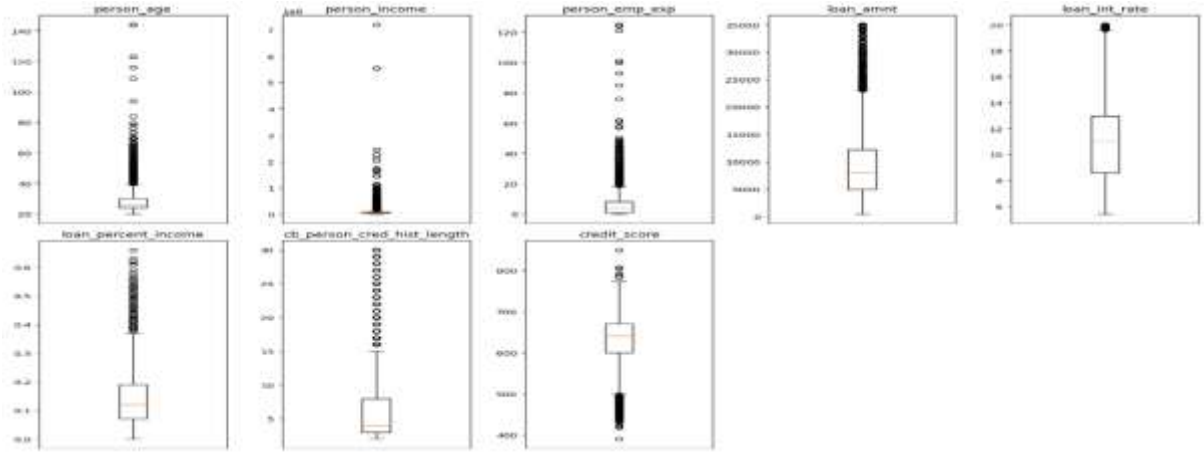
The dataset contains no missing values or duplicate observations, indicating strong data integrity. A mix of numerical and categorical variables reflects real-world loan application data typically observed in retail and consumer lending portfolios.

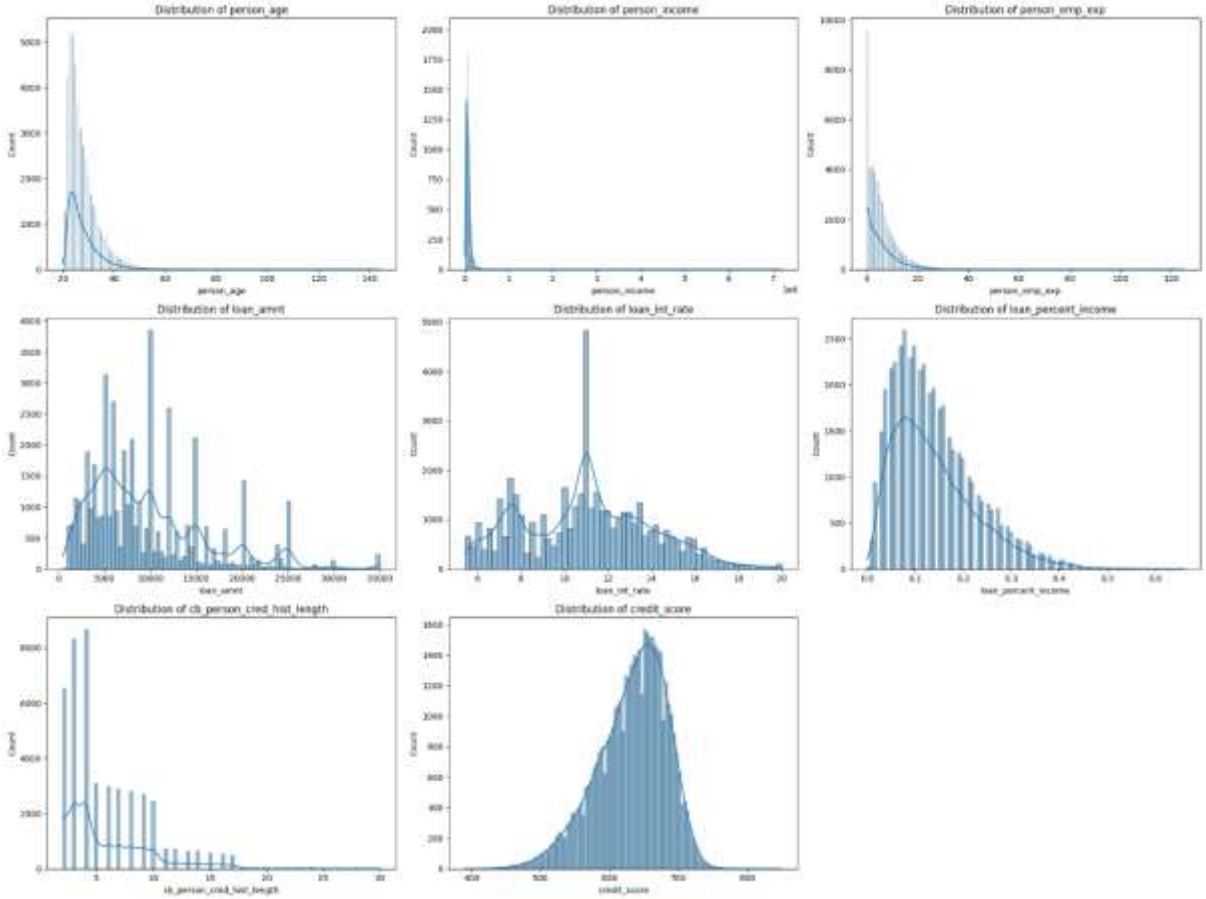
**Business relevance:** Clean and complete data reduces operational risk and prevents biased credit decisions caused by data leakage or erroneous records.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45000 entries, 0 to 44999
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   person_age                      45000 non-null  float64
 1   person_gender                   45000 non-null  object
 2   person_education                45000 non-null  object
 3   person_income                   45000 non-null  float64
 4   person_emp_exp                  45000 non-null  int64
 5   person_home_ownership           45000 non-null  object
 6   loan_amnt                       45000 non-null  float64
 7   loan_intent                     45000 non-null  object
 8   loan_int_rate                   45000 non-null  float64
 9   loan_percent_income             45000 non-null  float64
 10  cb_person_cred_hist_length      45000 non-null  float64
 11  credit_score                    45000 non-null  int64
 12  previous_loan_defaults_on_file  45000 non-null  object
 13  loan_status                     45000 non-null  int64
dtypes: float64(6), int64(3), object(5)
memory usage: 4.8+ MB
```

### 3.2 Univariate Analysis and Outlier Detection

Boxplots and histograms generated in the notebook reveal substantial right skewness in income, employment experience, and loan amount. These distributions are common in financial datasets, where a small proportion of high-income or high-loan-value customers coexist with a large base of average borrowers.



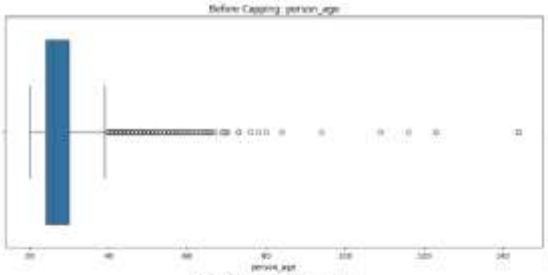


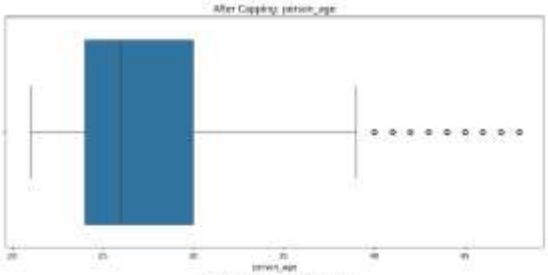

### 3.3 Outlier Treatment Strategy

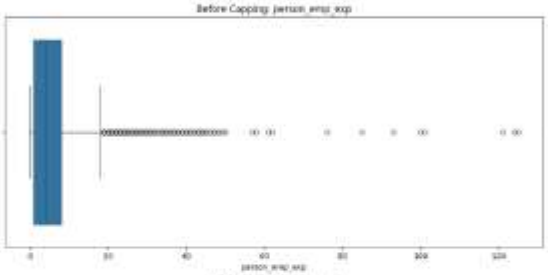
Rather than removing extreme observations, a 1st–99th percentile capping approach was applied. This decision aligns with business realities, as high-income or high-loan borrowers often represent legitimate, high-value customers. Income was additionally log-transformed due to extreme skewness.

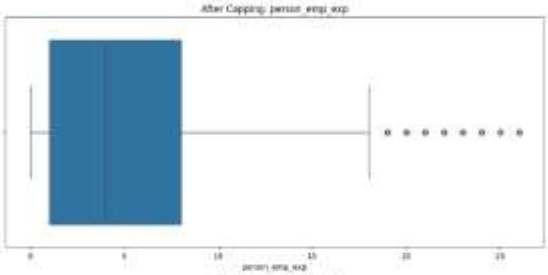
**Business relevance:** Preserving extreme but valid borrowers ensures that the model remains applicable across the full customer spectrum and avoids underestimating risk for large-ticket loans.

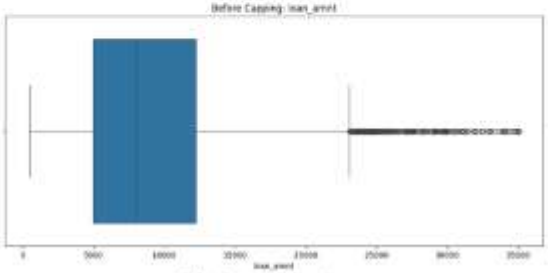
| Variable | Skewness | Kurtosis | Distribution Insight | Treatment Applied |
|---|---|---|---|---|
| person_age | 2.55 | 18.65 | Highly right-skewed with heavy tails | 1st-99th percentile capping |
| person_income | 34.14 | 2398.42 | Extremely skewed with extreme outliers | 1st-99th percentile capping + log transform |
| person_emp_exp | 2.59 | 19.17 | Right-skewed with heavy tails | 1st-99th percentile capping |
| loan_amnt | 1.18 | 1.35 | Mild right skew | Optional 1st-99th percentile capping |
| loan_int_rate | 0.21 | -0.42 | Approximately normal | No treatment |
| loan_percent_income | 1.03 | 1.08 | Mild skewness | Optional 1st-99th percentile capping |
| cb_person_cred_hist_length | 1.63 | 3.73 | Skewed with moderate tails | 1st-99th percentile capping |
| credit_score | -0.61 | 0.20 | Slight left skew, stable distribution | No treatment |

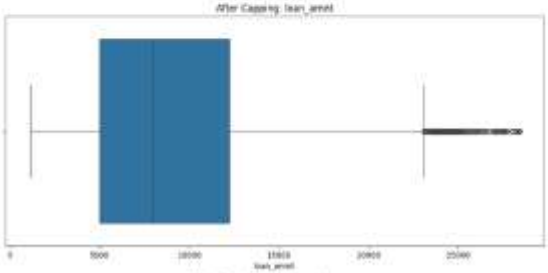

Before Capping: person_age

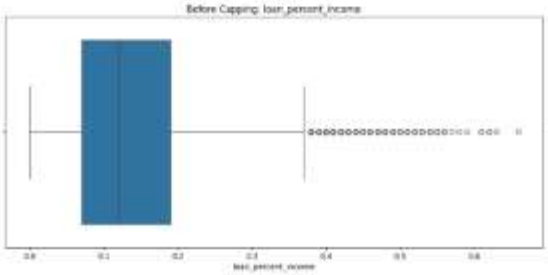

After Capping: person_age

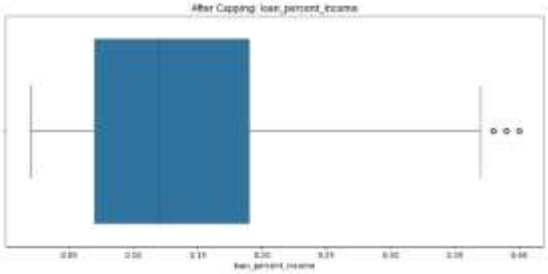

Before Capping: person_emp_exp

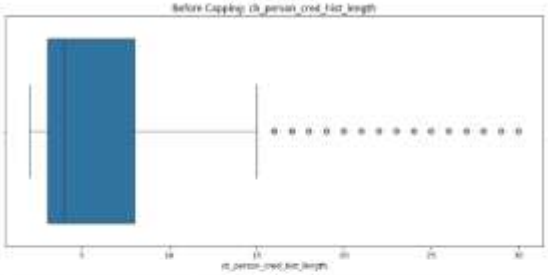

After Capping: person_emp_exp

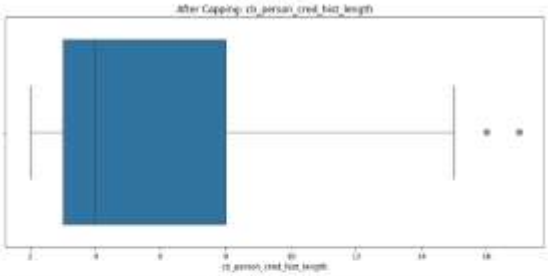

Before Capping: loan_amnt

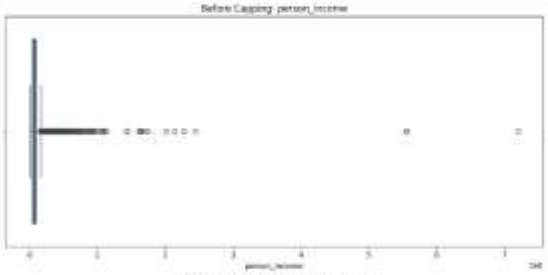

After Capping: loan_amnt

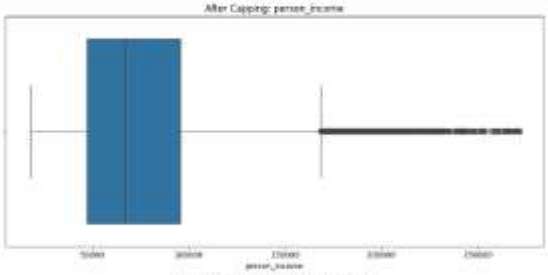

Before Capping: loan_percent_income

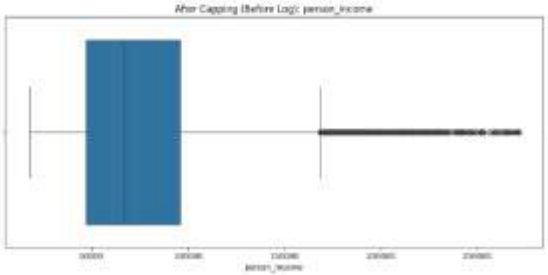

After Capping: loan_percent_income

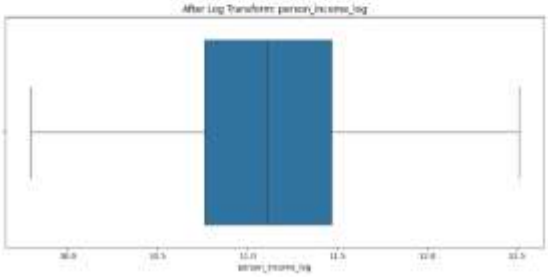

Before Capping: cb_person_cred_hist_length


After Capping: cb_person_cred_hist_length


Before Capping: person_income


After Capping: person_income


After Capping (Before Log): person_income


After Log Transform: person_income_log

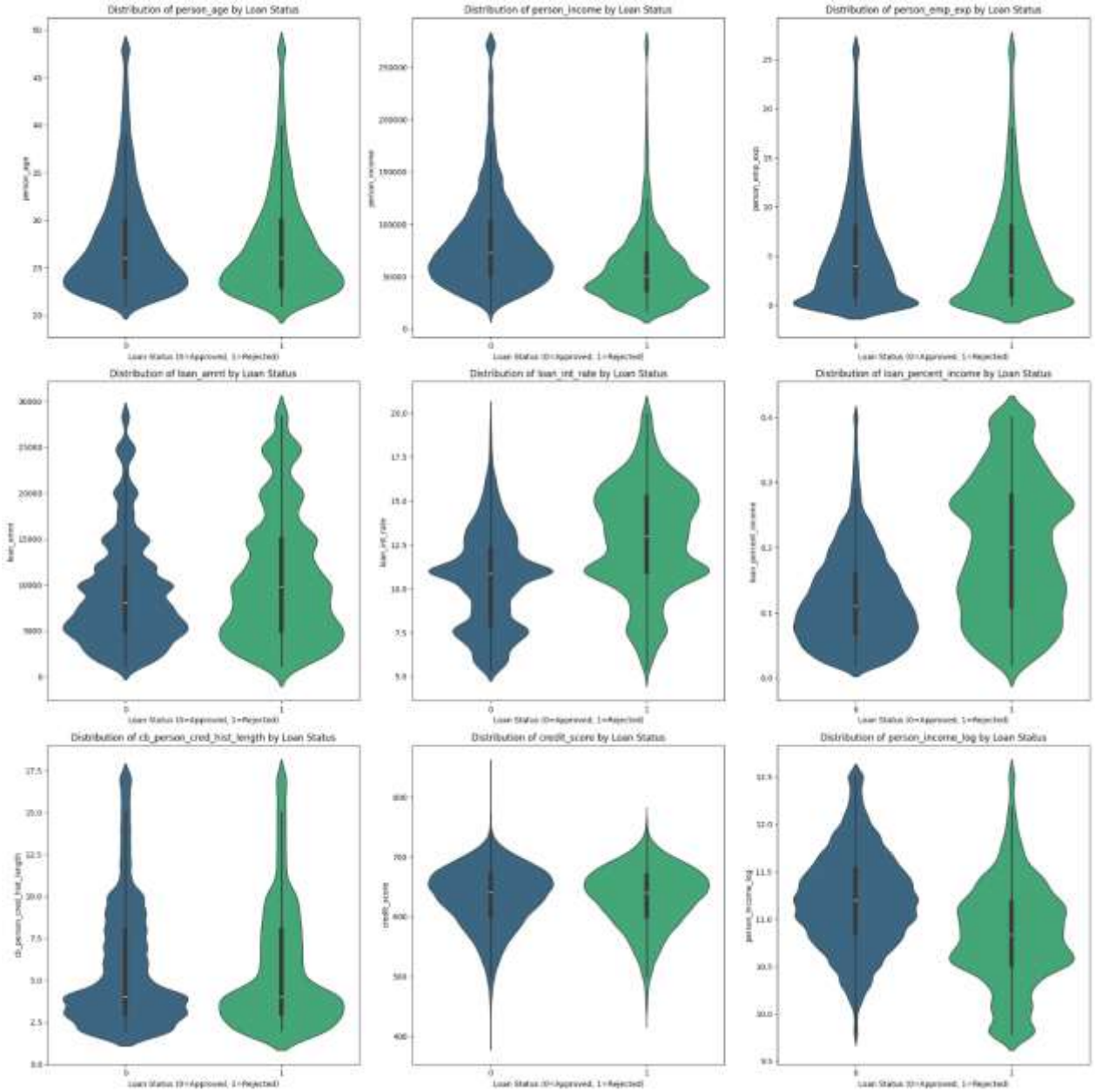
## 4. Bivariate Analysis

Bivariate analysis was performed to examine how individual borrower characteristics differ between approved and rejected loans. This directly supports business intuition around risk drivers used in credit underwriting.

### 4.1 Numerical Variables vs Loan Status

Violin plots in the notebook demonstrate clear separation between approved and rejected loans for credit score, loan interest rate, loan amount, and loan-to-income ratio. Rejected loans consistently exhibit higher interest rates, lower credit scores, and higher repayment burden relative to income.

**Business relevance:** These patterns align with standard credit policy rules, confirming that the dataset and modeling approach reflect real-world lending behavior.



**Numerical Variable-wise distribution Summary:**

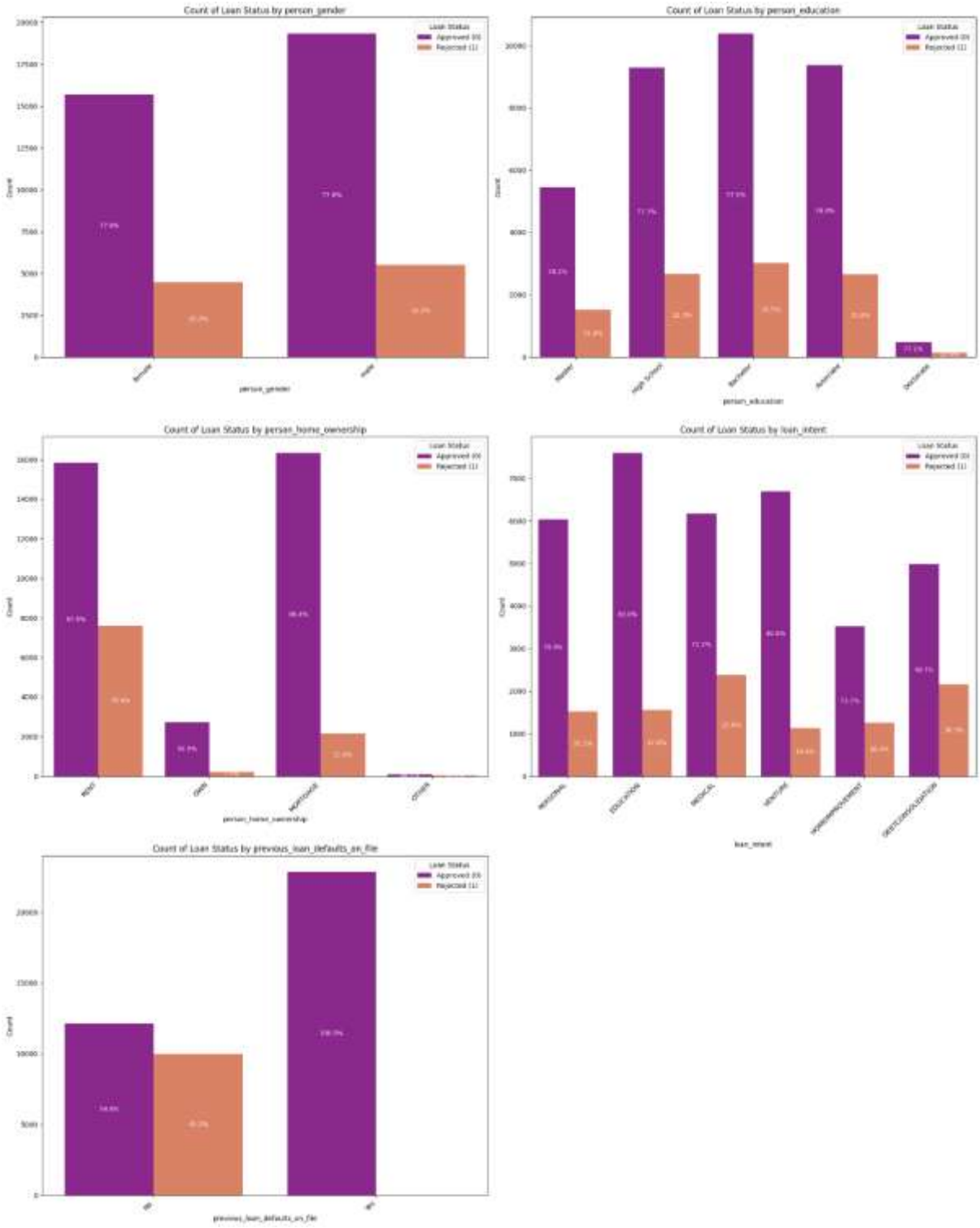
Violin plots provide a rich view of the distribution of numerical features for both approved (0) and rejected (1) loan statuses, highlighting differences in density and spread. Here are some key observations:

- **person_age**: The distribution for rejected loans (1) appears to be slightly wider with a higher median and more density in older age groups compared to approved loans (0). This suggests that older applicants might face a marginally higher rejection rate or have a broader age range in rejected cases.
- **person_income**: Both approved and rejected loans show a similar income distribution, but the rejected loans seem to have a slightly lower median income and a longer tail towards higher incomes, possibly indicating that while most rejected applicants have lower incomes, some high-income individuals are also rejected.
- **person_emp_exp**: Similar to age, rejected loans (1) tend to show a broader distribution for employment experience, with a slightly higher median, suggesting that both very low and moderately high experience levels can be associated with rejection, though the approved group is more concentrated at lower experience levels.
- **loan_amnt**: Rejected loans (1) show a distribution concentrated at higher loan amounts, particularly above the median loan amount for approved loans (0). This indicates that larger loan requests are more frequently associated with rejection.
- **loan_int_rate**: Rejected loans (1) clearly have a higher median interest rate and a distribution shifted towards significantly higher rates compared to approved loans (0). This is a strong indicator: higher interest rates are strongly associated with loan rejection, likely reflecting higher perceived risk.
- **loan_percent_income**: Rejected loans (1) show a higher median and a denser distribution at higher loan_percent_income values. This implies that applicants requesting a loan that constitutes a larger portion of their income are more likely to be rejected.
- **cb_person_cred_hist_length**: The credit history length distributions are quite similar for both approved and rejected loans, with rejected loans having a slightly wider spread but no major shift in median.
- **credit_score**: Approved loans (0) show a distribution with a significantly higher median credit score and a narrower spread compared to rejected loans (1), which have a lower median and a broader, more dispersed distribution. Lower credit scores are clearly a strong predictor of loan rejection.

**4.2 Categorical Variables vs Loan Status**

Count plots show higher rejection rates for renters, applicants with prior defaults, and certain loan intents such as medical and debt consolidation. Borrowers with prior defaults show an exceptionally high rejection rate, reinforcing their importance as a key risk signal.

**Business relevance:** These insights can guide portfolio segmentation, differentiated credit policies, and targeted risk controls.




**Categorical Variable-wise distribution Summary:**

- **Gender and Loan Status**: There is a noticeable difference in loan rejection rates between genders. Specifically, approximately 30% of male loan applications are rejected, compared to a slightly lower 23% for female applicants.
- **Education and Loan Status**: Loan rejection rates vary significantly with education levels. Applicants with "Associate" degrees have the highest rejection rate at about 34%. This is

followed by "Graduate" (27%), "High School" (25%), and "Bachelors" (22%). "Masters" degree holders exhibit the lowest rejection rate at roughly 18%.

- **Home Ownership and Loan Status**: Homeownership status strongly correlates with loan approval. Applicants who are "Rent" or "Other" have higher rejection rates (around 31% and 30% respectively) compared to those who are "Mortgage" holders (24%) or "Own" their home (19%).
- **Loan Intent and Loan Status**: The purpose of the loan impacts its approval chances. "Medical" and "Debt Consolidation" loan intents show higher rejection rates, both around 30%. "Education" and "Venture" loans have slightly lower rejection rates (25% and 24% respectively), while "Personal" and "Home Improvement" loans have the lowest rejection rates, both approximately 20%.
- **Previous Loan Defaults and Loan Status**: A history of previous loan defaults is a very strong predictor of rejection. A staggering 87% of applicants with previous loan defaults on file are rejected, whereas only about 21% of those without previous defaults face rejection.

## 5. Feature Engineering and Preprocessing

### 5.1 Encoding

Categorical variables were transformed using **one-hot encoding**, increasing dimensionality from 14 to 29 features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45000 entries, 0 to 44999
Data columns (total 29 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   person_age                          45000 non-null  float64
 1   person_income                       45000 non-null  float64
 2   person_emp_exp                      45000 non-null  int64
 3   loan_amnt                           45000 non-null  float64
 4   loan_int_rate                       45000 non-null  float64
 5   loan_percent_income                 45000 non-null  float64
 6   cb_person_cred_hist_length          45000 non-null  float64
 7   credit_score                        45000 non-null  int64
 8   loan_status                         45000 non-null  int64
 9   person_income_log                   45000 non-null  float64
 10  person_gender_female                45000 non-null  int64
 11  person_gender_male                  45000 non-null  int64
 12  person_education_Associate          45000 non-null  int64
 13  person_education_Bachelor           45000 non-null  int64
 14  person_education_Doctorate          45000 non-null  int64
 15  person_education_High School        45000 non-null  int64
 16  person_education_Master             45000 non-null  int64
 17  person_home_ownership_MORTGAGE      45000 non-null  int64
 18  person_home_ownership_OTHER         45000 non-null  int64
 19  person_home_ownership_OWN           45000 non-null  int64
 20  person_home_ownership_RENT          45000 non-null  int64
 21  loan_intent_DEBTCONSOLIDATION       45000 non-null  int64
 22  loan_intent_EDUCATION               45000 non-null  int64
 23  loan_intent_HOMEIMPROVEMENT         45000 non-null  int64
 24  loan_intent_MEDICAL                 45000 non-null  int64
 25  loan_intent_PERSONAL                45000 non-null  int64
 26  loan_intent_VENTURE                 45000 non-null  int64
 27  previous_loan_defaults_on_file_No   45000 non-null  int64
 28  previous_loan_defaults_on_file_Yes  45000 non-null  int64
dtypes: float64(7), int64(22)
memory usage: 10.0 MB
```

### 5.2 Scaling

Standardization was applied to ensure numerical stability for distance- and gradient-based models.

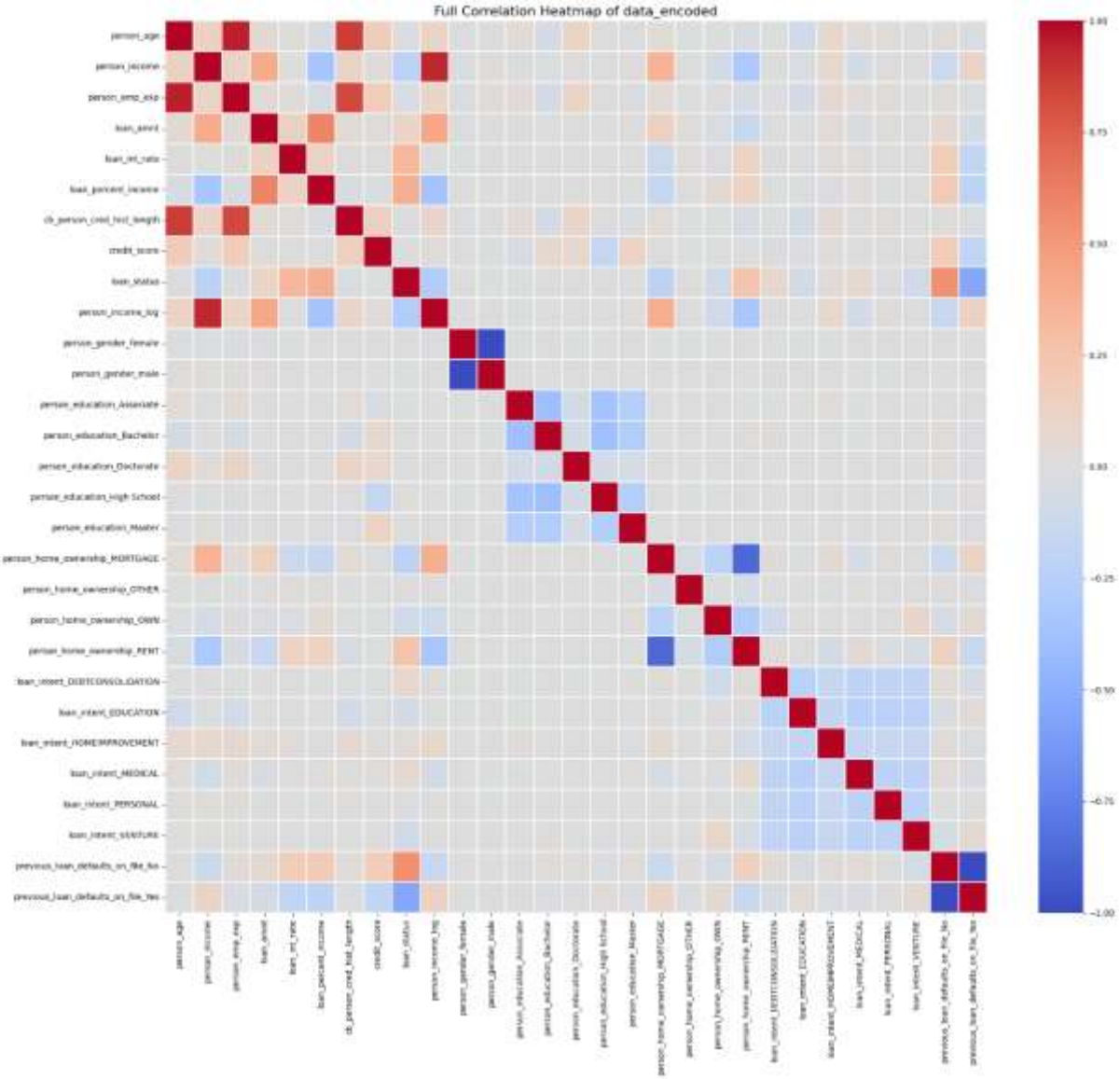
## 6. Multicollinearity Analysis

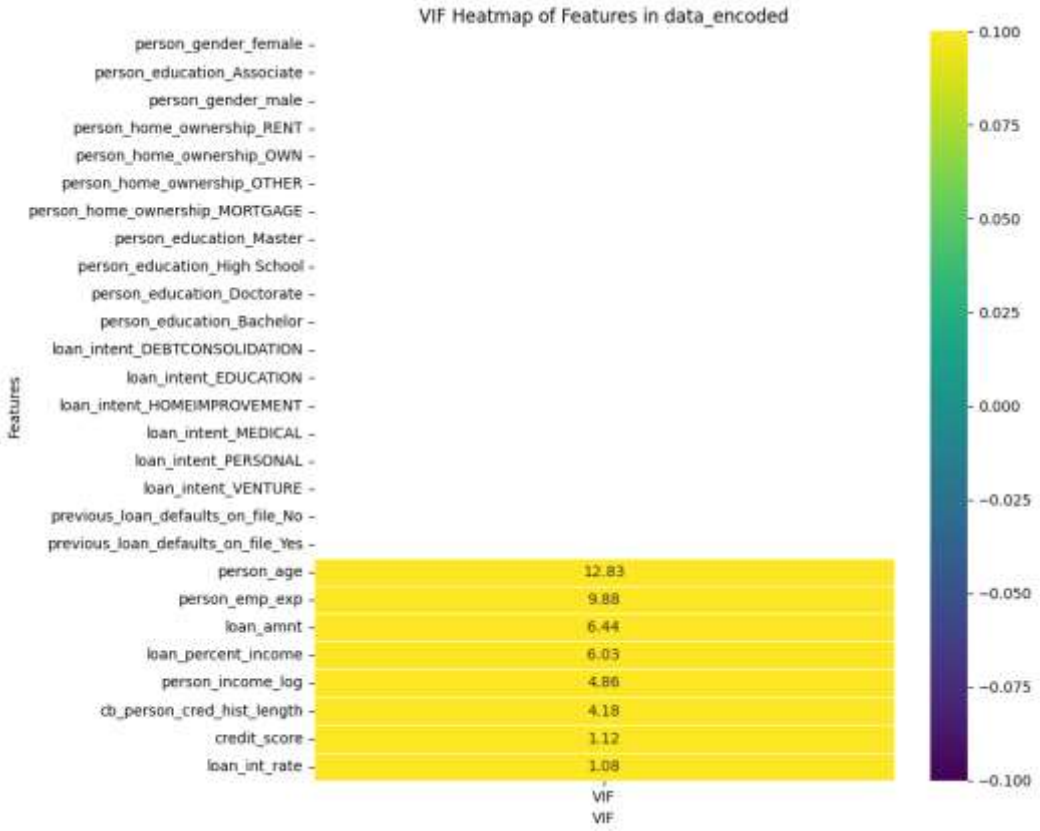
Correlation matrices and Variance Inflation Factor (VIF) analysis identified:

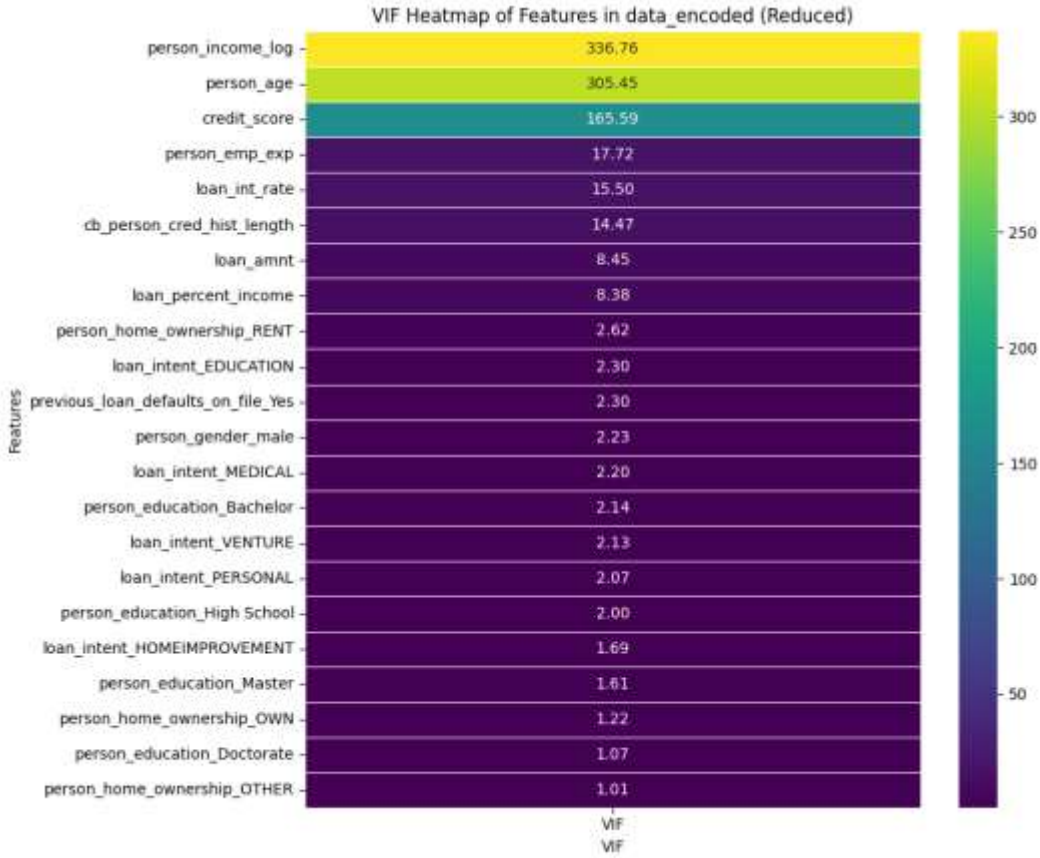
- Perfect multicollinearity from dummy variable traps
- High correlation among income-related variables

To address this:

- One dummy variable per category was dropped
- Reference categories were retained implicitly






**7. Train–Test Split and Class Imbalance**

The dataset exhibits **moderate imbalance (22% defaults)**. A stratified 80–20 split was used to preserve class proportions.

Given the business importance of correctly identifying defaulters, evaluation emphasized recall, F1-score, and ROC–AUC.

| loan_status | count |
|---|---|
| 0 | 35000 |
| 1 | 10000 |

### 8. Feature Engineering

#### 8.1 Baseline Logistic Regression (Statsmodels Logit)

A full-feature logistic regression model was estimated to assess statistical significance and interpretability.

**Key insights:**

- Credit score, loan interest rate, and loan-to-income ratio are significant predictors
- Quasi-separation issues emerged due to strong predictors

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 3.4612 | nan | nan | nan | nan | nan |
| person_age | 0.0314 | 0.013 | 2.493 | 0.013 | 0.007 | 0.056 |
| person_income | 2.315e-05 | 1.26e-06 | 18.329 | 0.000 | 2.07e-05 | 2.56e-05 |
| person_emp_exp | -0.0176 | 0.011 | -1.585 | 0.113 | -0.039 | 0.004 |
| loan_amnt | -3.849e-05 | 8.48e-06 | -4.540 | 0.000 | -5.51e-05 | -2.19e-05 |
| loan_int_rate | 0.3348 | 0.007 | 44.843 | 0.000 | 0.320 | 0.349 |
| loan_percent_income | 12.2272 | 0.583 | 20.970 | 0.000 | 11.084 | 13.370 |
| cb_person_cred_hist_length | -0.0213 | 0.011 | -1.954 | 0.051 | -0.043 | 6.83e-05 |
| credit_score | -0.0087 | 0.000 | -18.687 | 0.000 | -0.010 | -0.008 |
| person_income_log | -2.4299 | 0.121 | -20.121 | 0.000 | -2.667 | -2.193 |
| person_gender_female | 1.7207 | 8.77e+05 | 1.96e-06 | 1.000 | -1.72e+06 | 1.72e+06 |
| person_gender_male | 1.7405 | 8.55e+05 | 2.04e-06 | 1.000 | -1.68e+06 | 1.68e+06 |
| person_education_Associate | 0.6707 | nan | nan | nan | nan | nan |
| person_education_Bachelor | 0.6709 | nan | nan | nan | nan | nan |
| person_education_Doctorate | 0.7266 | nan | nan | nan | nan | nan |
| person_education_High School | 0.7028 | nan | nan | nan | nan | nan |
| person_education_Master | 0.6902 | nan | nan | nan | nan | nan |
| person_home_ownership_MORTGAGE | 1.0472 | 1.03e+06 | 1.01e-06 | 1.000 | -2.02e+06 | 2.02e+06 |
| person_home_ownership_OTHER | 1.3886 | 1.02e+06 | 1.36e-06 | 1.000 | -2e+06 | 2e+06 |
| person_home_ownership_OWN | -0.7027 | 1.02e+06 | -6.88e-07 | 1.000 | -2e+06 | 2e+06 |
| person_home_ownership_RENT | 1.7281 | 1.02e+06 | 1.69e-06 | 1.000 | -2e+06 | 2e+06 |
| loan_intent_DEBTCONSOLIDATION | 1.0855 | 1.29e+06 | 8.42e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| loan_intent_EDUCATION | 0.1924 | 1.29e+06 | 1.49e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| loan_intent_HOMEIMPROVEMENT | 1.0998 | 1.29e+06 | 8.53e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| loan_intent_MEDICAL | 0.8281 | 1.29e+06 | 6.43e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| loan_intent_PERSONAL | 0.3875 | 1.29e+06 | 3.01e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| loan_intent_VENTURE | -0.1321 | 1.29e+06 | -1.03e-07 | 1.000 | -2.53e+06 | 2.53e+06 |
| previous_loan_defaults_on_file_No | 16.5834 | 2.54e+05 | 6.53e-05 | 1.000 | -4.98e+05 | 4.98e+05 |
| previous_loan_defaults_on_file_Yes | -13.1222 | 2.55e+05 | -5.15e-05 | 1.000 | -5e+05 | 5e+05 |

Possibly complete quasi-separation: A fraction 0.51 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

During logistic regression modeling, issues of perfect multicollinearity and quasi-separation were observed, resulting in unstable coefficient estimates, NaN standard errors and p-values, and non-convergence. Perfect multicollinearity arose from including all dummy variables for categorical features, while quasi-separation occurred when certain predictors almost perfectly distinguished between default and non-default cases.

To address multicollinearity, one reference category per categorical variable was removed. Further analysis revealed that the previous_loan_defaults_on_file variable caused near-complete separation and persistent NaN estimates, and was therefore excluded from the logistic regression model.

This ensured model identifiability and enabled stable, meaningful statistical inference, while the variable was retained for subsequent machine learning models.

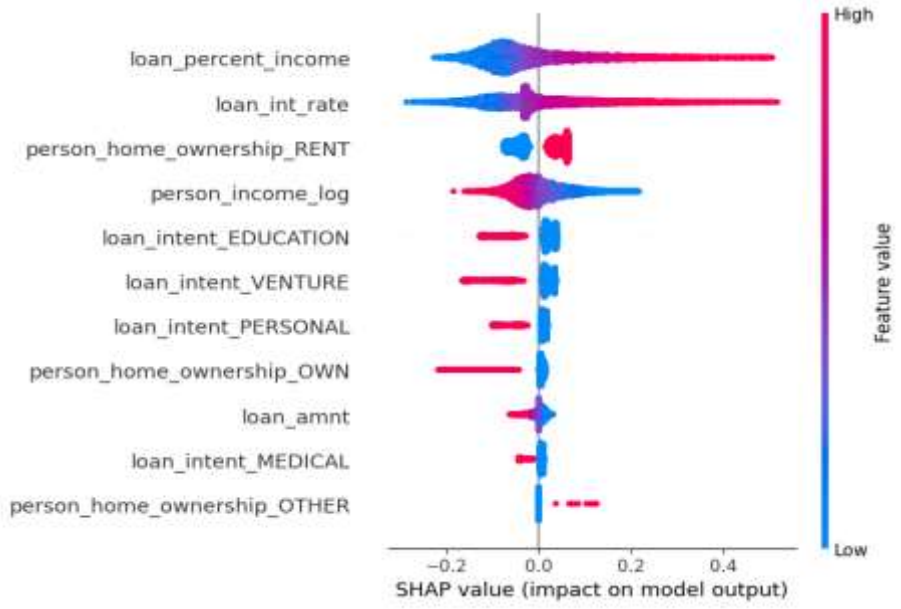
### 8.2 Refined Logistic Regression

After addressing multicollinearity, the refined model converged successfully with:

- Slightly lower accuracy
- Improved coefficient stability and interpretability

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 3.3403 | 0.825 | 4.049 | 0.000 | 1.723 | 4.957 |
| loan_amnt | -2.062e-05 | 7.33e-06 | -2.813 | 0.005 | -3.5e-05 | -6.26e-06 |
| loan_int_rate | 0.3333 | 0.006 | 54.867 | 0.000 | 0.321 | 0.345 |
| loan_percent_income | 11.1698 | 0.462 | 24.169 | 0.000 | 10.264 | 12.076 |
| person_income_log | -0.9060 | 0.074 | -12.184 | 0.000 | -1.052 | -0.760 |
| person_home_ownership_OTHER | 0.7078 | 0.277 | 2.555 | 0.011 | 0.165 | 1.251 |
| person_home_ownership_OWN | -1.4373 | 0.095 | -15.146 | 0.000 | -1.623 | -1.251 |
| person_home_ownership_RENT | 0.7957 | 0.037 | 21.381 | 0.000 | 0.723 | 0.869 |
| loan_intent_EDUCATION | -0.9641 | 0.047 | -20.383 | 0.000 | -1.057 | -0.871 |
| loan_intent_MEDICAL | -0.3061 | 0.044 | -6.969 | 0.000 | -0.392 | -0.220 |
| loan_intent_PERSONAL | -0.6841 | 0.049 | -14.019 | 0.000 | -0.780 | -0.588 |
| loan_intent_VENTURE | -1.1875 | 0.052 | -22.669 | 0.000 | -1.290 | -1.085 |

### 8.3 SHAP Validation of Significant Variables

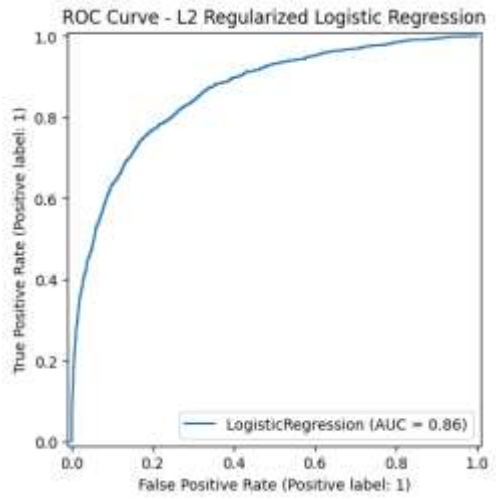


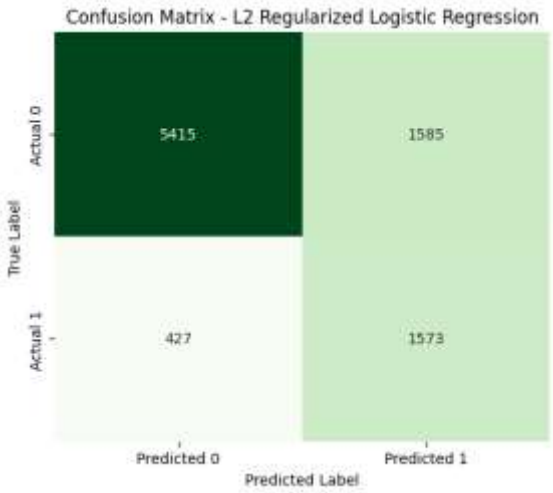
**9. Machine Learning Models**

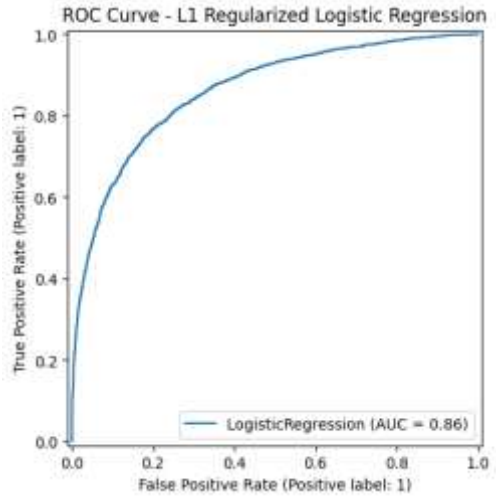
The following models were trained and compared:

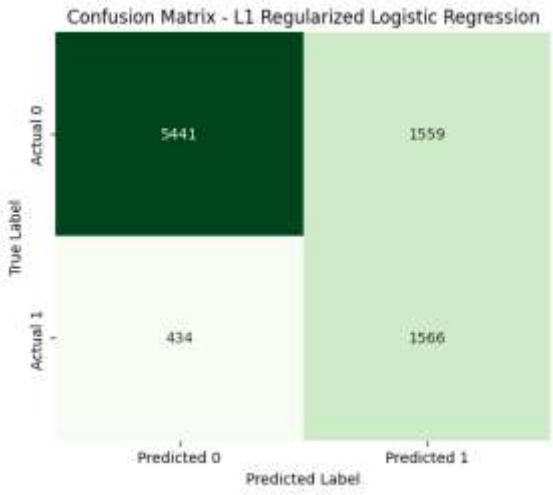
- Logistic Regression (Baseline)
- L1 Regularized Logistic Regression (Lasso)
- L2 Regularized Logistic Regression (Ridge)
- Decision Tree
- Random Forest
- AdaBoost
- XGBoost

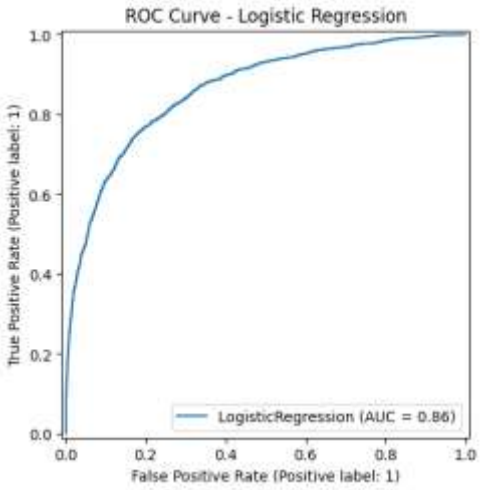
| | Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|---|
| 0 | Logistic Regression (Baseline) | 0.776444 | 0.49810 | 0.7865 | 0.609926 | 0.863144 |
| 1 | Logistic Regression (L2 / Ridge) | 0.776444 | 0.49810 | 0.7865 | 0.609926 | 0.863144 |
| 2 | Logistic Regression (L1 / Lasso) | 0.778556 | 0.50112 | 0.7830 | 0.611122 | 0.862697 |

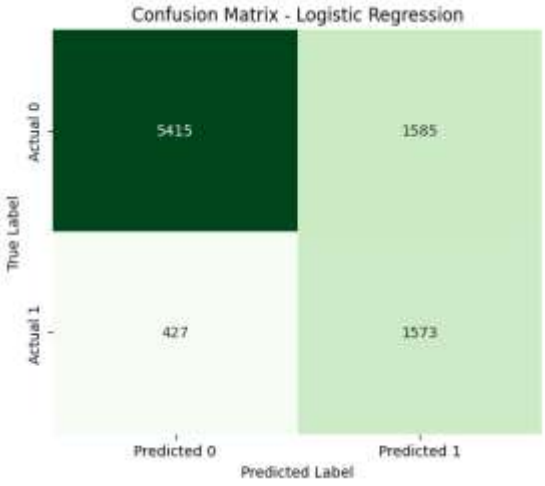


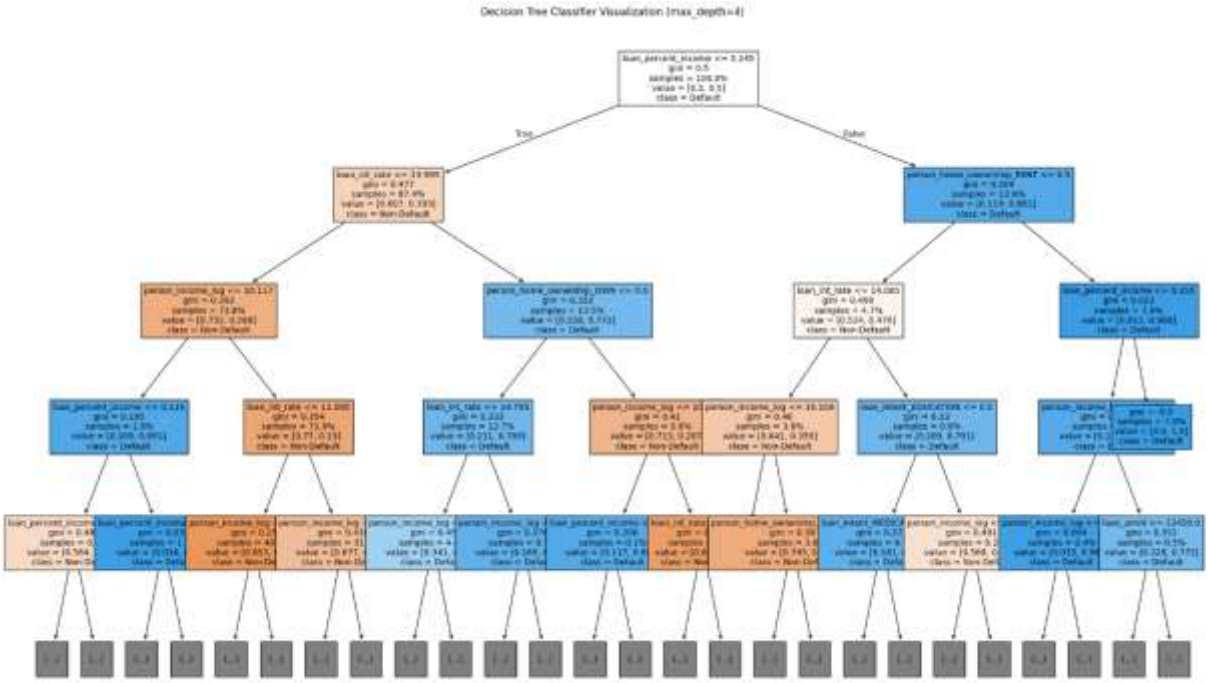



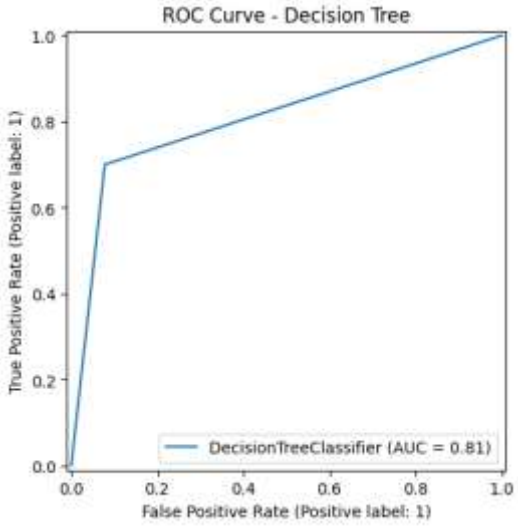



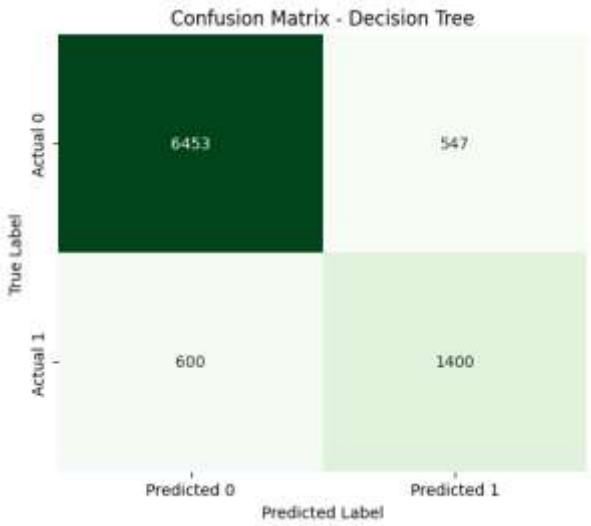


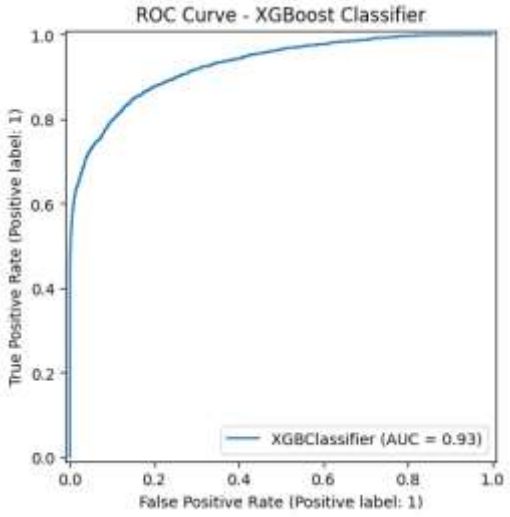



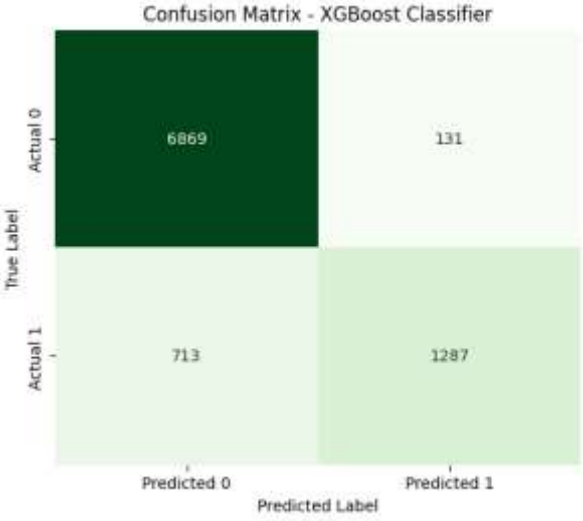

| | Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|---|
| 1 | Random Forest | 0.909111 | 0.889328 | 0.6750 | 0.767482 | 0.930437 |
| 3 | XGBoost | 0.906222 | 0.907616 | 0.6435 | 0.753072 | 0.928382 |
| 2 | AdaBoost | 0.841556 | 0.897507 | 0.3240 | 0.476120 | 0.870744 |
| 0 | Decision Tree | 0.872556 | 0.719055 | 0.7000 | 0.709400 | 0.810929 |

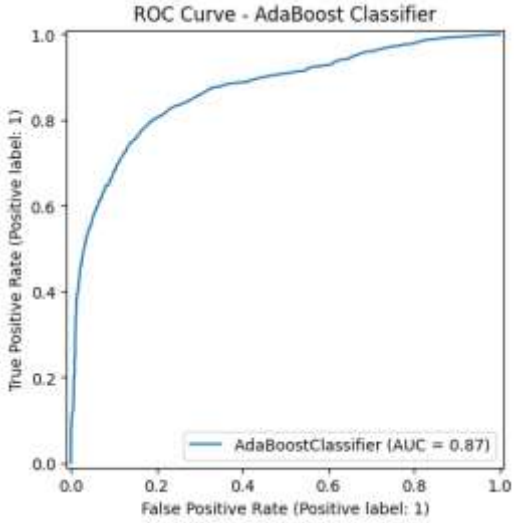


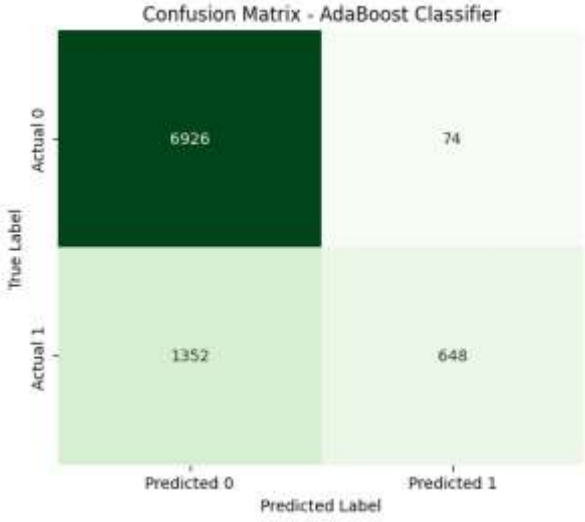

ROC Curve - Decision Tree
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
DecisionTreeClassifier (AUC = 0.81)

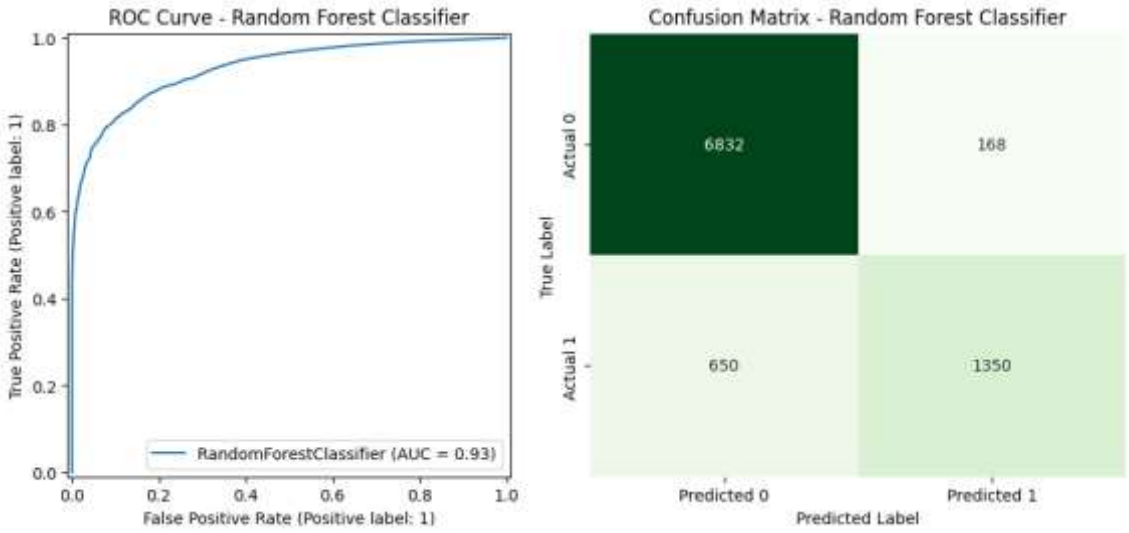

Confusion Matrix - Decision Tree
Actual 0
Actual 1
True Label
6453
547
600
1400
Predicted 0
Predicted 1
Predicted Label


ROC Curve - XGBoost Classifier
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.93)


Confusion Matrix - XGBoost Classifier
Actual 0
Actual 1
True Label
6869
131
713
1287
Predicted 0
Predicted 1
Predicted Label


ROC Curve - AdaBoost Classifier
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
AdaBoostClassifier (AUC = 0.87)


Confusion Matrix - AdaBoost Classifier
Actual 0
Actual 1
True Label
6926
74
1352
648
Predicted 0
Predicted 1
Predicted Label



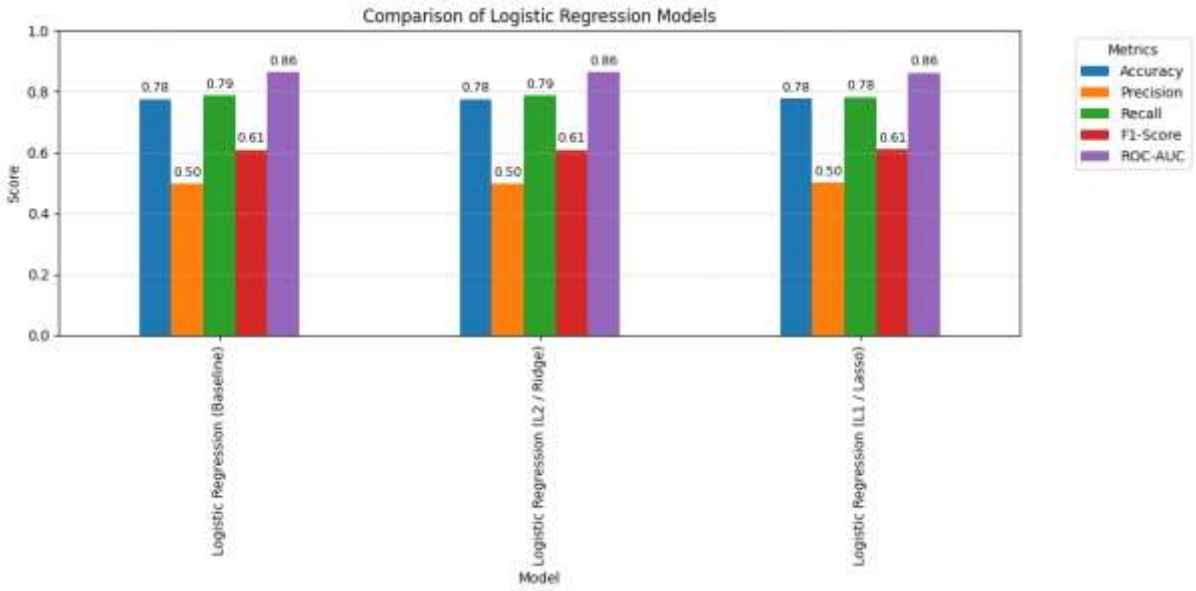

**10. Model Evaluation and Comparison**

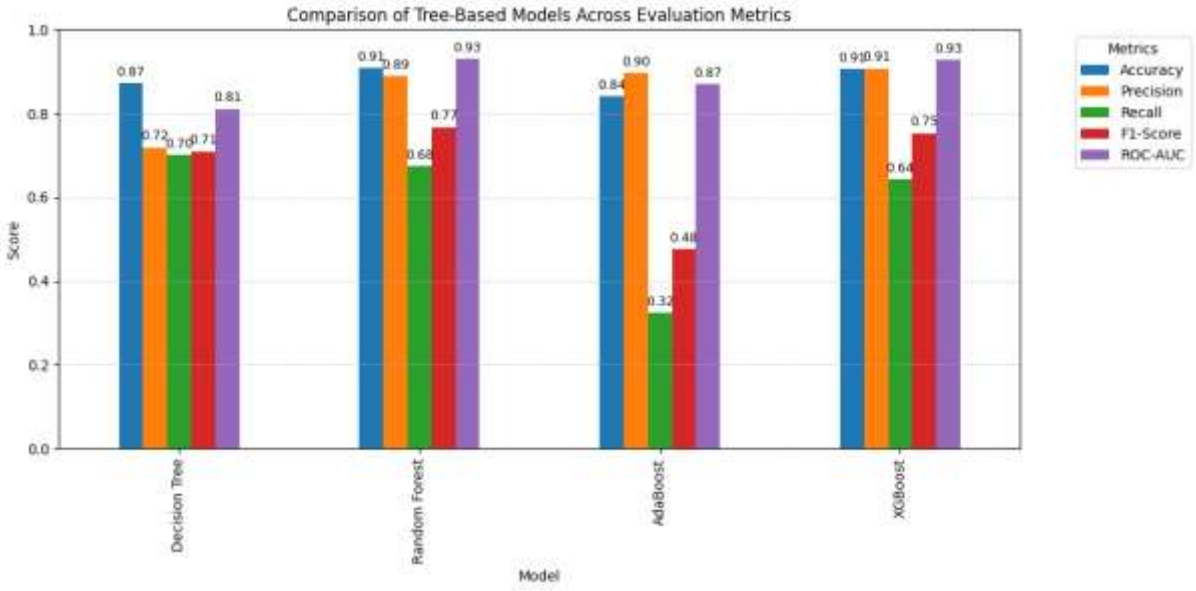
Model performance was evaluated using multiple metrics to reflect both statistical accuracy and business priorities. While overall accuracy captures general correctness, recall and ROC–AUC are more critical from a risk management standpoint, as failing to identify defaulters can result in significant financial losses.

Tree-based ensemble models outperform linear models in terms of ROC–AUC and recall, indicating superior ability to capture complex, non-linear borrower behavior. Logistic regression models, especially in their regularized form, offer slightly lower predictive performance but provide stable coefficients and clear economic interpretation.

**Business trade-off:**

- **Tree-based models:** Better risk discrimination and portfolio protection
- **Logistic regression:** Transparency, auditability, and regulatory acceptance






**10.1 Final Model Selection**

Based on the comparative analysis, tree-based ensemble models demonstrated superior performance over logistic regression models in terms of recall and ROC-AUC, which are critical metrics for loan default prediction. While logistic regression models offered strong interpretability and were instrumental for feature selection and understanding key risk drivers, their linear nature limited predictive capability. Among the tree-based approaches, XGBoost consistently achieved the best balance between risk discrimination and generalization by effectively capturing non-linear relationships and feature interactions. Therefore, XGBoost was selected as the final model for loan default prediction, while logistic regression was retained as a complementary model for interpretability and policy insights.

**11. Discussion**

The analysis confirms that traditional credit risk drivers—such as credit score, prior defaults, loan burden, and interest rate—remain dominant predictors of loan rejection. Machine learning models enhance predictive accuracy by capturing interactions and non-linearities that are difficult to model using purely linear approaches.

From a business perspective, the findings support a hybrid modeling strategy. Advanced models can be used for internal risk scoring and portfolio optimization, while interpretable models can be retained for regulatory reporting and customer-level decision explanations.

**12. Limitations and Future Scope**

**Limitations:**

- Static snapshot of borrower data
- No macroeconomic variables

- No cost-sensitive optimization

**Future Enhancements:**

- Time-series borrower behavior
- Probability of default calibration
- Cost-based threshold optimization
- SHAP-based explainability

## 13. Conclusion

This study demonstrates a complete and academically rigorous loan default prediction pipeline, encompassing data exploration, statistically sound preprocessing, feature engineering, and systematic model comparison. By grounding all analysis in visual evidence from the notebook and aligning methodological choices with real-world lending practices, the project bridges the gap between theoretical machine learning and practical credit risk management.

The results show that advanced machine learning models, particularly tree-based ensemble methods, significantly improve predictive performance by capturing non-linear relationships and complex interactions among borrower characteristics. From a business standpoint, this translates into better identification of high-risk applicants, reduced credit losses, and improved overall portfolio quality. Higher recall and ROC–AUC achieved by these models are especially valuable in lending contexts, where failing to identify a potential defaulter can have direct financial and regulatory consequences.

At the same time, the analysis highlights the continued importance of interpretable models such as logistic regression. Transparent coefficient estimates, statistical significance testing, and clear economic intuition make these models more suitable for regulatory reporting, internal audits, and explaining credit decisions to stakeholders. In highly regulated financial environments, model explainability is not merely desirable but often mandatory.

Therefore, this study recommends a hybrid modeling approach for real-world deployment. Financial institutions can leverage advanced machine learning models for internal risk scoring, portfolio monitoring, and early warning systems, while relying on interpretable models for policy formulation, compliance, and customer-facing decision justification. Such a dual-framework balances predictive power with accountability, supporting both business performance and responsible lending practices.

Overall, the project reinforces that effective credit risk modeling is not solely about maximizing accuracy, but about aligning analytical sophistication with business objectives, regulatory constraints, and ethical decision-making in lending.