Predicting Home Loan Eligibility Using Machine Learning

Group 11 Project

Sakshi Agarwal • Haley Hoang • Mayar Abdelhade • Emelia Appiah • Divya Katamneni

Business Context & Objective

Strengthen Revenue with Smarter Lending

Home loans are a critical revenue stream, but poor approval decisions increase exposure to credit risk.

Manual Reviews Limit Efficiency

Current loan assessments are slow, inconsistent, and difficult to scale as application volumes grow.

Leverage Data to Predict Approvals

Machine-learning models can analyze historical applications to reliably estimate approval likelihood

Reduce Risk with Data-Driven Decisions

Accurate predictions help Easy House focus on strong applicants, lower default rates, and optimize portfolio quality.

Data Overview: Dataset & Key Variables

- ~600 historical loan applications
- **Target:** Loan_Status (1 = Approved, 0 = Not Approved)
- **Predictors include:**
    - Demographics: Gender, Married, Education, Dependents, Self-Employed
    - Financials: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term
    - Risk indicators: Credit_History
    - Property_Area: Urban / Semiurban / Rural

Data Preprocessing: Cleaning and Transformation

01

Impute Missing Values

- Around 20% of values were missing.
- Categorical filled using mode, numerical using median.
- Result: Missing values reduced to 0% with no loss of rows.

02

Remove Loan_ID

- No predictive value, removed for cleaner data.
- Result: Reduced noise and improved model interpretability.

03

One-Hot Encode

- Applied to Gender, Education, Property_Area.
- Result: Expanded feature space with 6 additional encoded variables.

04

Binary Map

- Married and Self-Employed mapped to 0/1.
- Result: Simplified categorical interpretation for linear models.

05

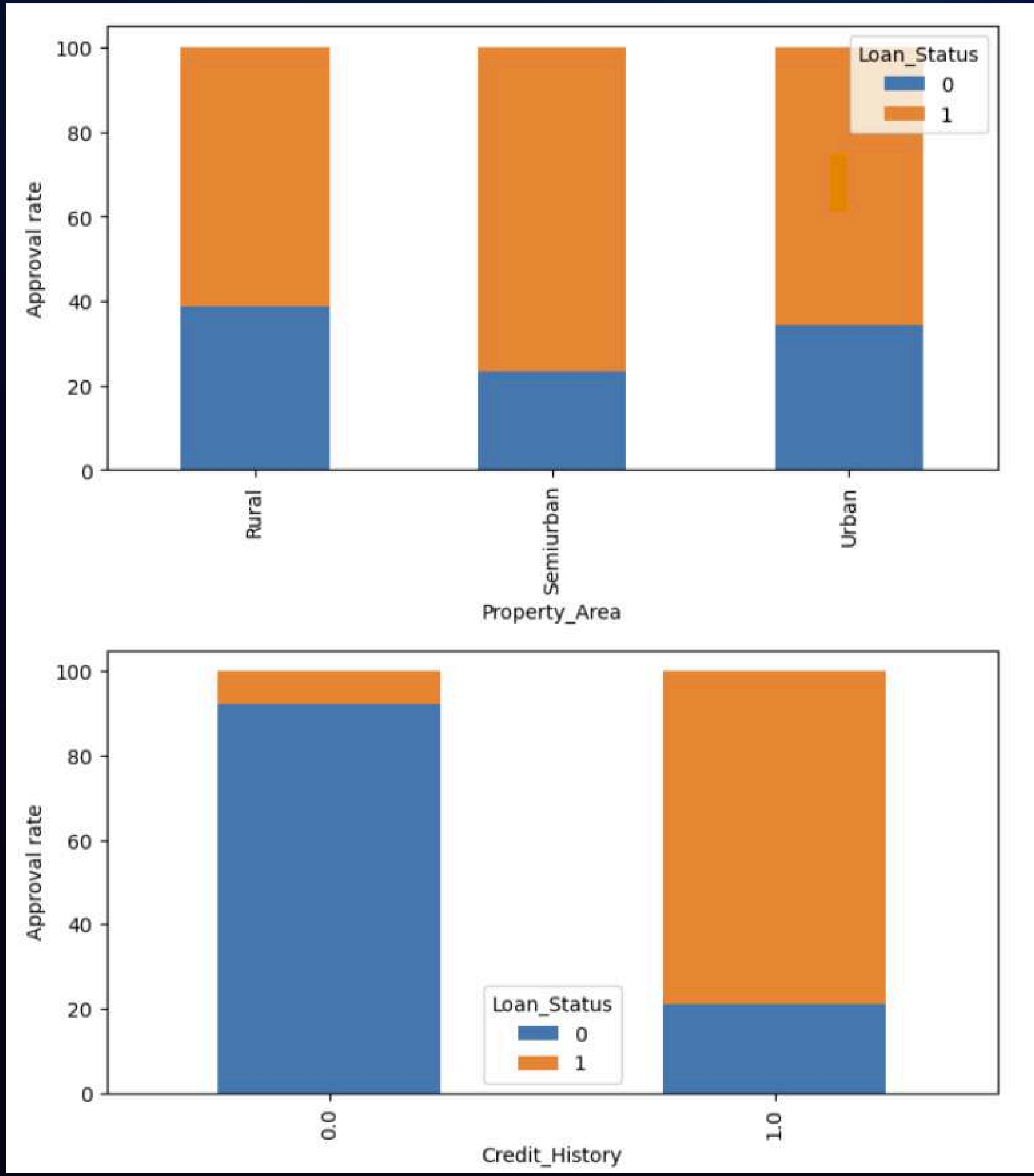
Scale Numerical Features

- StandardScaler is used for consistent numerical ranges.
- Result: Prevented high-magnitude features from dominating Logistic Regression.

Exploratory Analysis – Who Gets Approved?

- **Approval Distribution**: ~68% approved, ~32% not approved → moderate imbalance.
- **Credit History**: Applicants with positive credit history show the highest approval rates.
- **Property Area**: Semiurban applicants have the highest approval rate, followed by Urban.
- **Demographics**: Married and graduate applicants show slightly higher approval likelihood.
- **Income & Loan Patterns**: No strong linear relationship, but income is right-skewed with some outliers

Approval is strongly driven by **credit history and property area**, with smaller effects from demographics



Feature Engineering & Outlier Handling

01 New Features

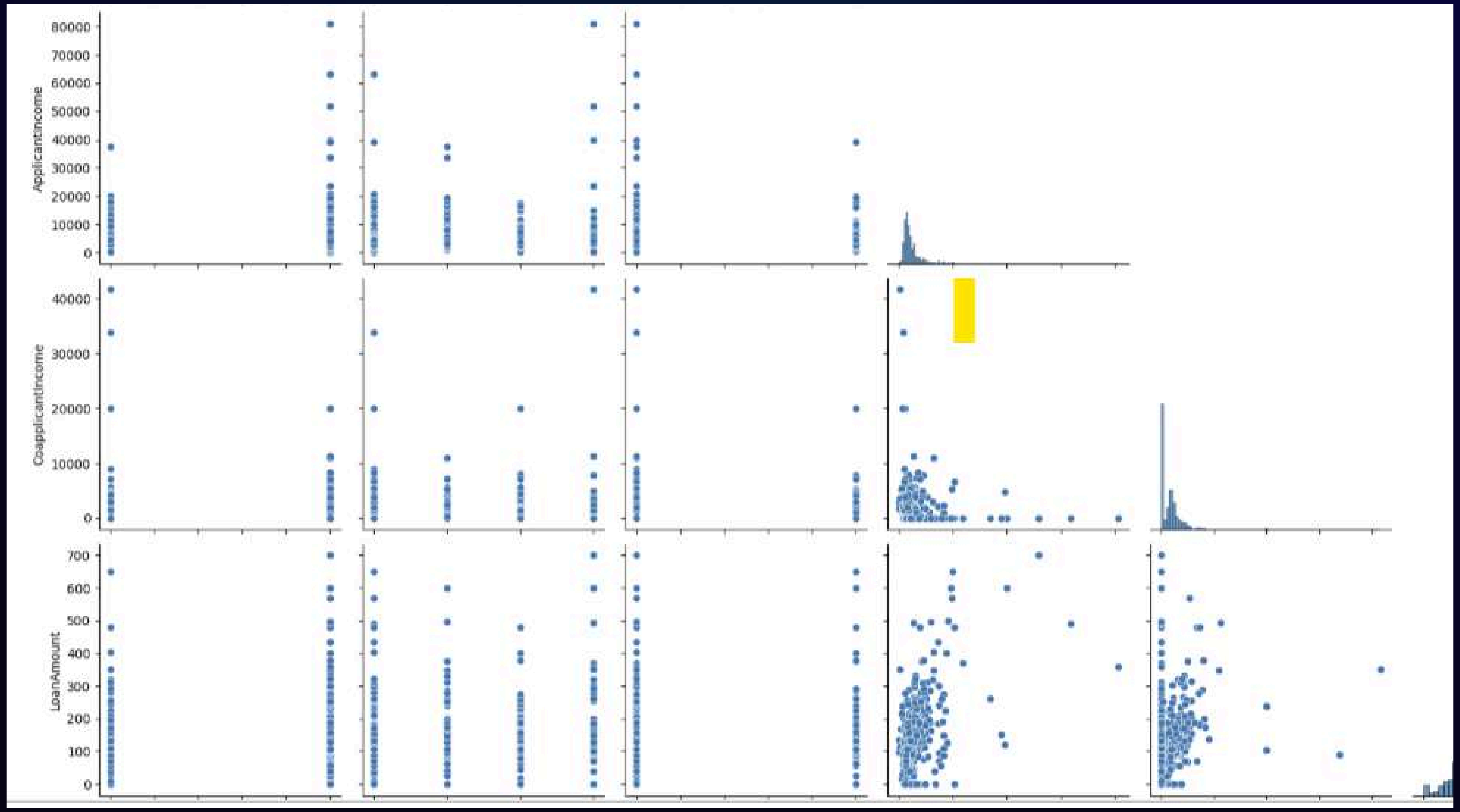
- Engineered financial stability indicators:
- Total_Income (overall earning capacity)
- Balance_Income (disposable income after EMI)
- These features better represent the true repayment ability of applicants.

02 Outlier Removal

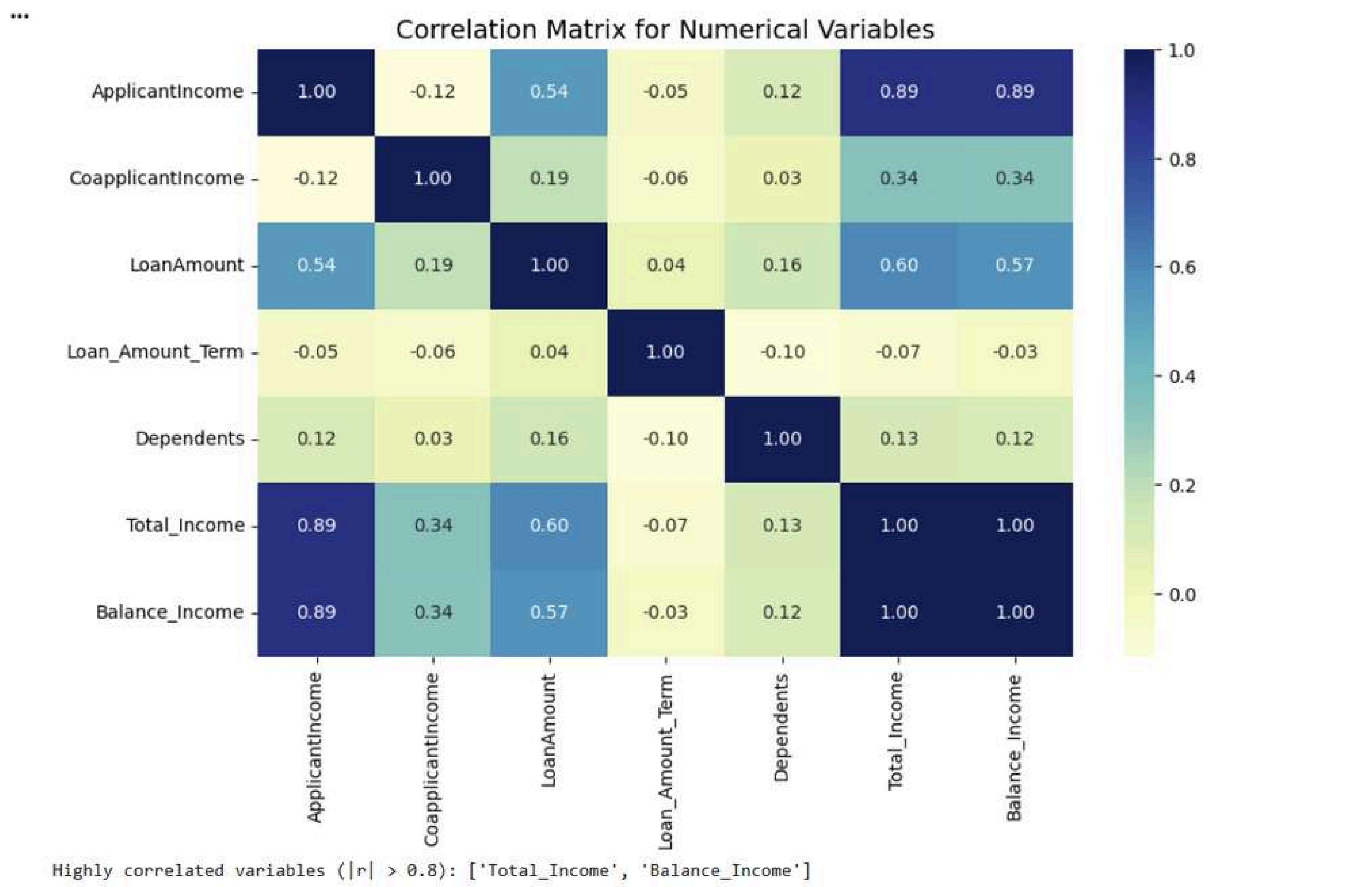
Applied Z-score > 3 to detect and remove extreme values in income and loan amounts.

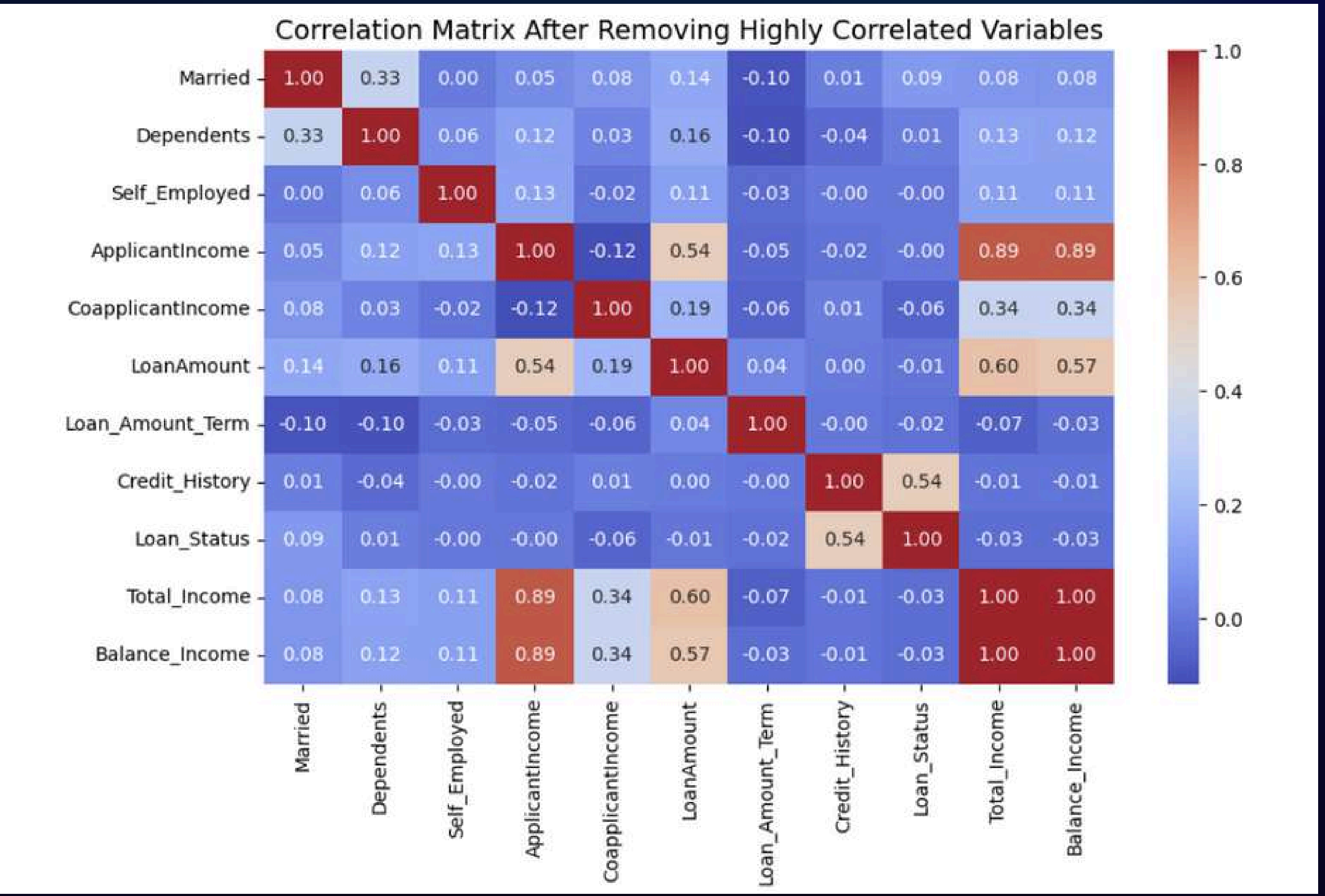
03 Noise Reduction

- Removed ~3.75% extreme records
- Increased model smoothness and reduced variance during training.



Correlation Matrix Analysis





Correlation Findings

- High correlation (r > 0.80) between:
- ApplicantIncome ↔ Total_Income
- Total_Income ↔ Balance_Income

Decision

We kept all three income features because:

- They represent distinct financial meanings
- Tree-based models are robust to multicollinearity
- Removing them would reduce the financial signal

Addressing Class Imbalance with SMOTE

1

Original Target Imbalance

68% approved, 32% not approved. Risk of models over-predicting "Approve."

2

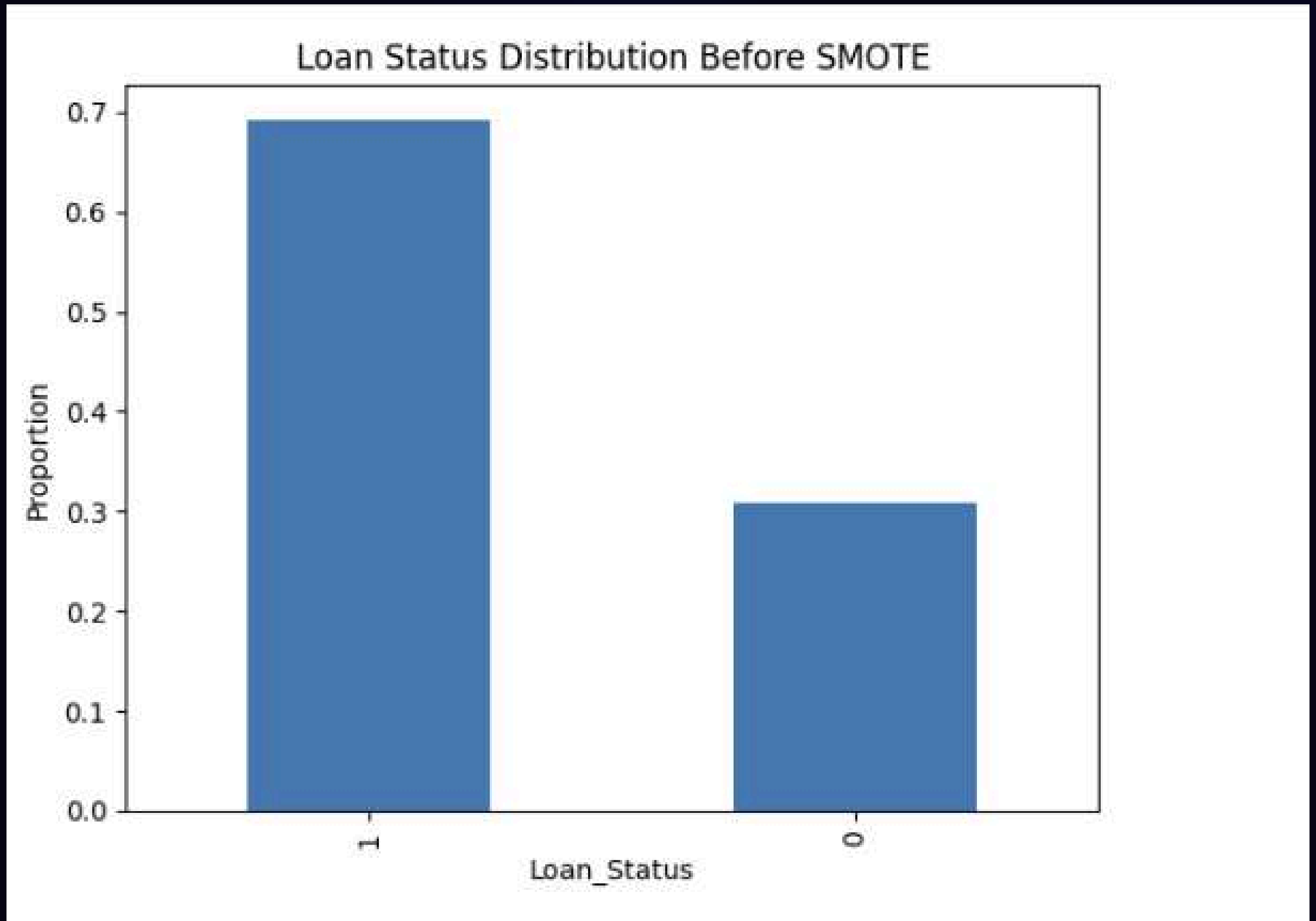
Approach

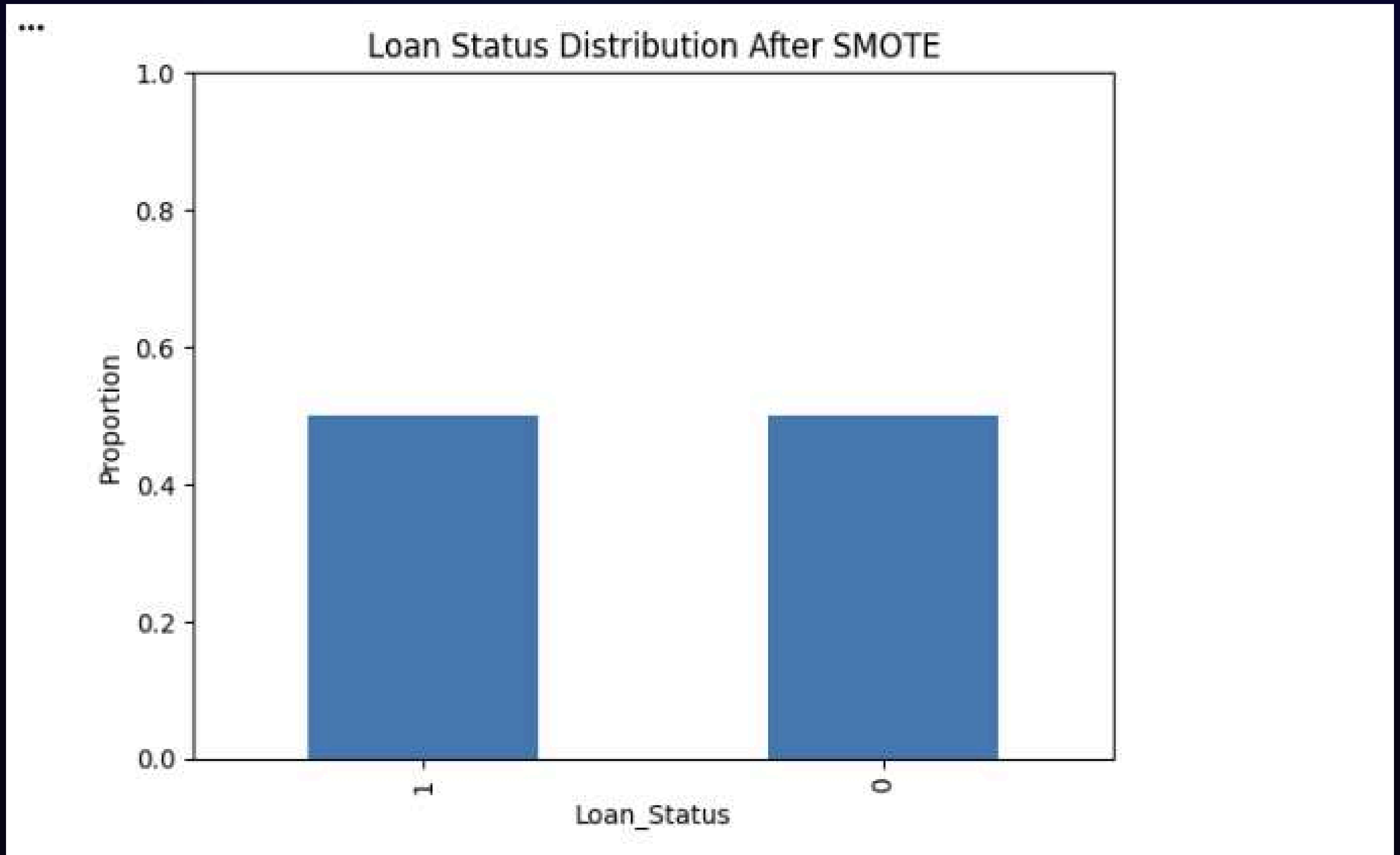
- Train-test split (70:30)
- Applied SMOTE on training data only.

3

Balanced Training Set

Training set balanced to 50% approved / 50% not approved for fair model learning.





Model Selection & Evaluation Metrics

Models Evaluated

Logistic Regression, Decision Tree, Random Forest.

Evaluation Metrics

Accuracy, Precision, Recall, F1, ROC-AUC, 5 fold Cross-Validation.

Business Goal: Avoid Risky Approvals

Crucial to prevent approving borrowers with high default risk.

Business Goal: Qualify Applicants

Equally important to avoid rejecting qualified applicants.

Predictive Modeling: Logistic Regression Baseline

Starting with Logistic Regression offers a transparent, interpretable baseline for binary outcomes like loan approvals.

Model Performance

	precision	recall	f1-score	support
0	0.55	0.57	0.56	58
1	0.79	0.78	0.78	120
accuracy			0.71	178
macro avg	0.67	0.67	0.67	178
weighted avg	0.71	0.71	0.71	178

Precision = 0.79
When the model predicts that a loan will be approved, it is correct 79% of the time

Recall = 0.78
The model successfully captures 78% of all applicants who should receive approval

F1-score = 0.78
Consistent performance between identifying qualified applicants & avoiding incorrect approvals

78%
Training Accuracy

71%
Test Accuracy

The moderate drop in accuracy suggests reasonable generalization, with no severe overfitting, though the model does decrease in performance when applied to unseen data

	Feature	Coefficient	Odds Ratio
7	Credit_History	3.732618	41.788357
0	Married	1.395920	4.038687
2	Self_Employed	0.837317	2.310160
13	Balance_Income	0.795238	2.214968
10	Property_Area_Semiurban	0.044316	1.045312
4	CoapplicantIncome	0.024978	1.025292
5	LoanAmount	-0.119922	0.886990
6	Loan_Amount_Term	-0.166728	0.846430
1	Dependents	-0.177966	0.836971
11	Property_Area_Urban	-0.364252	0.694716
12	Total_Income	-0.449237	0.638115
3	ApplicantIncome	-0.473215	0.622996
9	Education_Not Graduate	-0.593827	0.552210
8	Gender_Male	-0.760972	0.467212

Key Drivers: Odds Ratios

Credit History (41.7x)

Applicants with a clean credit history are 41 times more likely to be approved than those with a poor credit history — by far the strongest predictor in the model.

Married (4.04x)

Married applicants are 4 times more likely to receive approval, holding other factors constant.

Self-Employed (2.31x)

Self-employed applicants have 2.3× higher odds of being approved than non-self-employed applicants.

Balance Income (2.21x)

Higher disposable income after EMI increases approval likelihood — each unit increase doubles the odds of approval.

Model Comparison: Decision Tree

Model Performance: Decision Tree

Captures Non-Linear Patterns

Effective in identifying complex interactions within data.

Accuracy

≈ 66%

Recall (Class 0)

≈ 0.62

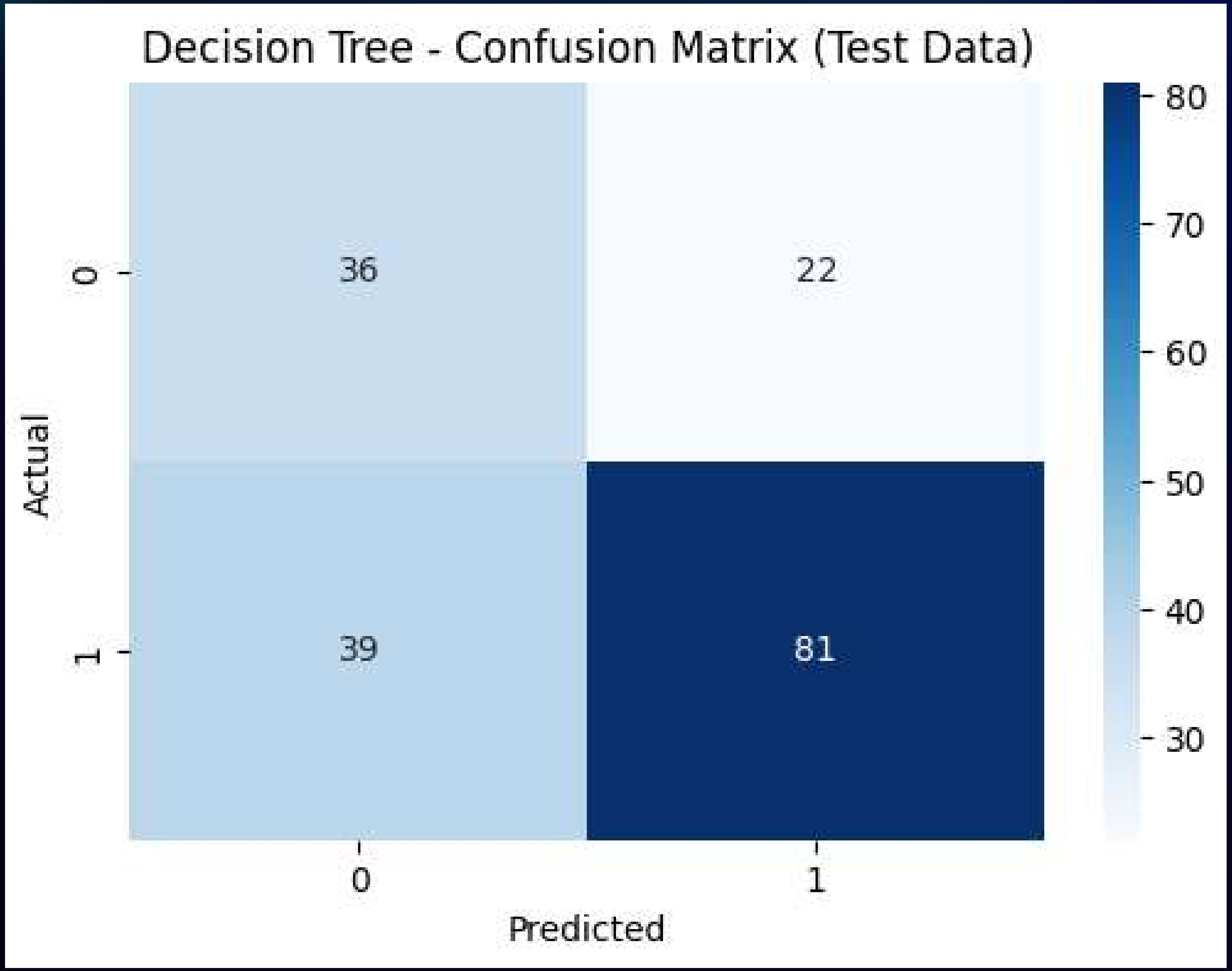
Recall (Class 1)

≈ 0.68

Slightly lower accuracy than Logistic Regression

Key Insights

1. Good at identifying approved applicants
2. Struggles with detecting risky applicants
3. Also misclassifies some safe applicants
4. Indicates overfitting — the tree memorizes patterns but generalizes poorly.
5. Still useful for understanding key drivers such as Credit History, Income, and Loan Amount.



Ensemble Model: Random Forest

Model Performance: Random Forest

Ensemble of many tuned trees for robust prediction. Most robust and highest-performing model overall.

76%

Accuracy

- The model predicts the loan approval outcome correctly for about 3 out of 4 applicants.
- It shows strong overall predictive capability on unseen data.

0.53

Recall (Class 0)

- The model correctly identifies 53% of high-risk applicants, providing moderate ability to catch potentially unsafe loans.
- It detects about 1 out of 2 truly risky cases.

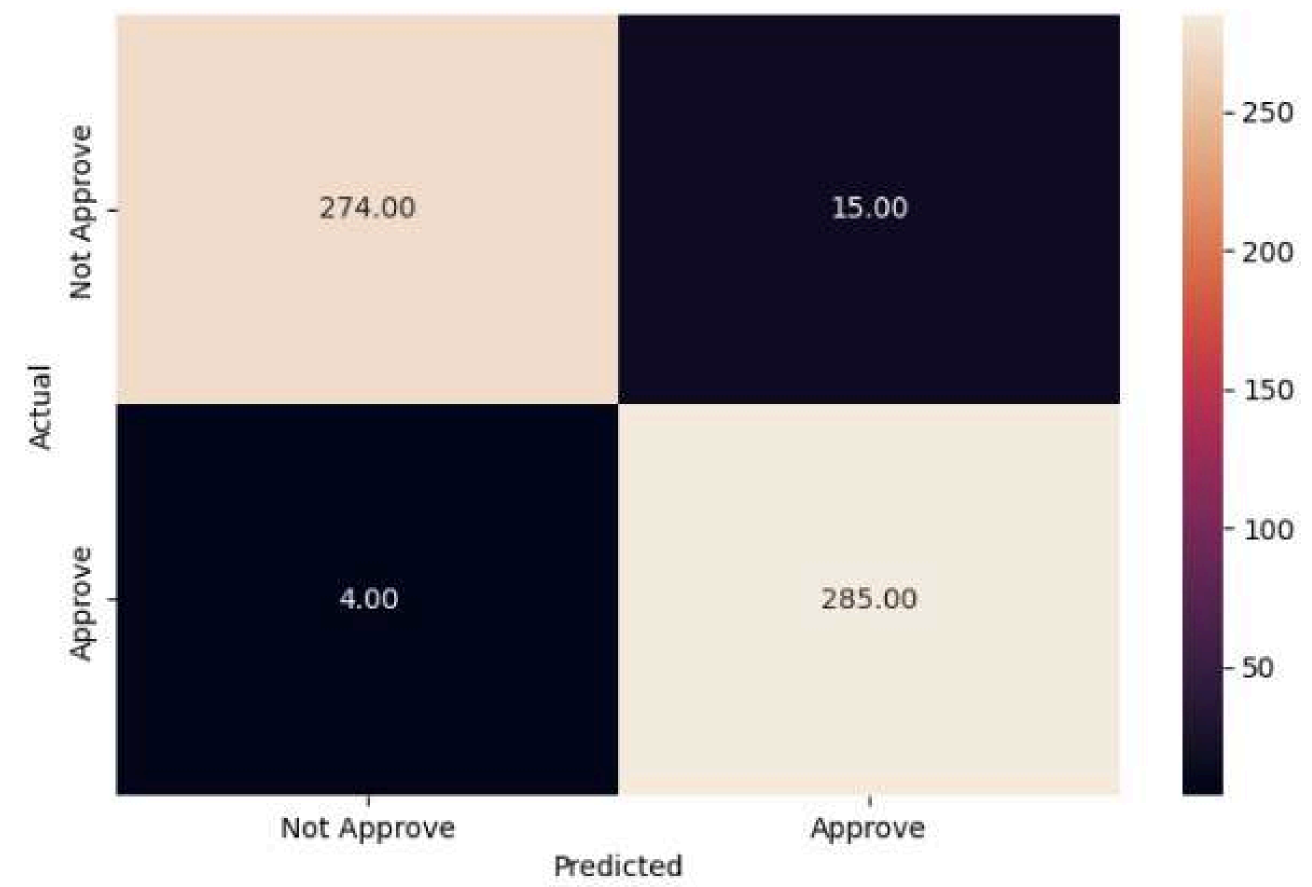
0.88

Recall (Class 1)

- The model correctly identifies 88% of safe applicants, showing a strong ability to approve the right customers.
- It is very strong at catching applicants who are likely to repay.

Confusion Matrix

	precision	recall	f1-score	support
0	0.99	0.95	0.97	289
1	0.95	0.99	0.97	289
accuracy			0.97	578
macro avg	0.97	0.97	0.97	578
weighted avg	0.97	0.97	0.97	578



Training Data

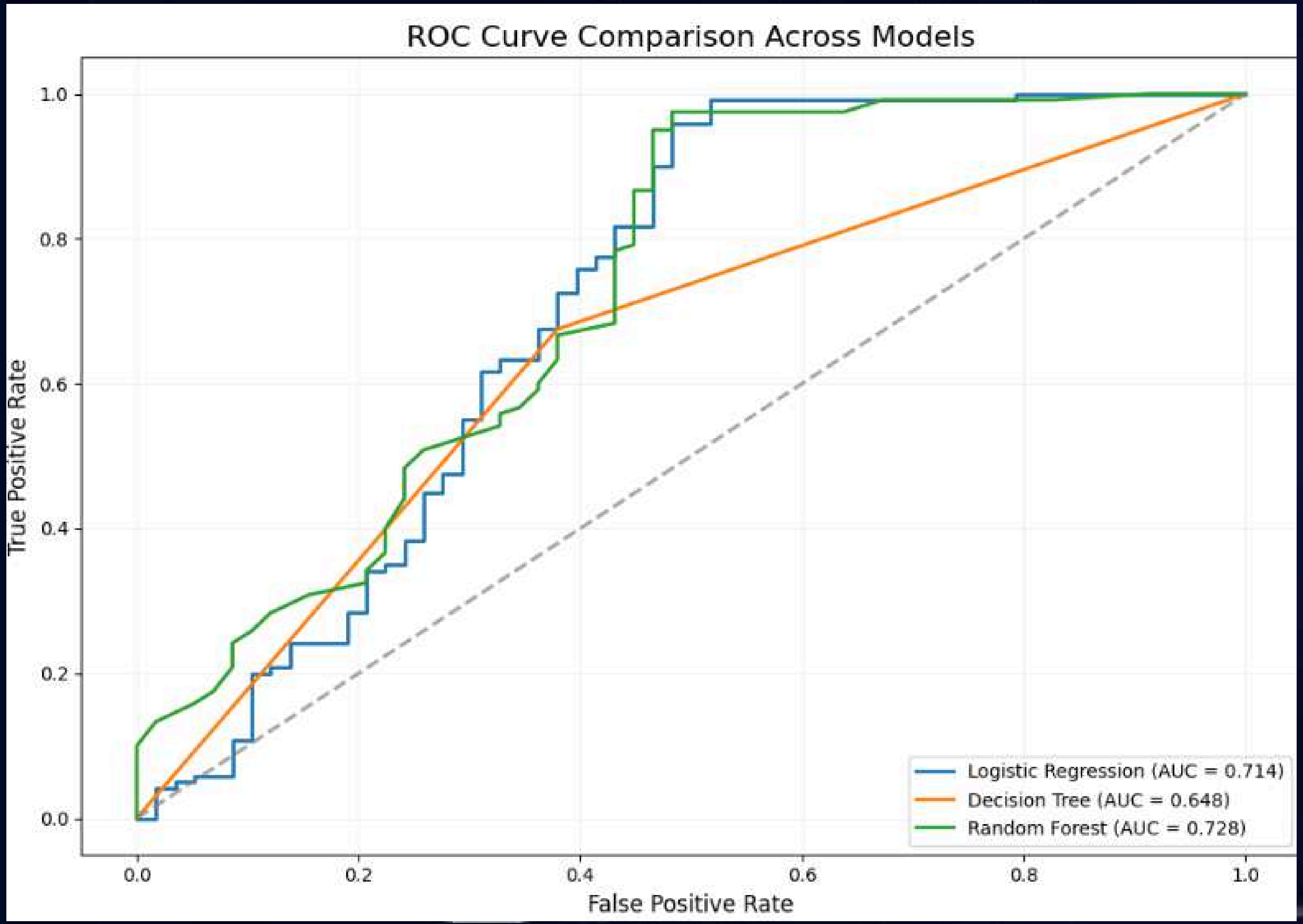
Almost perfect classification on training data — very low errors, suggesting strong learning but possible slight overfitting.

	precision	recall	f1-score	support
0	0.67	0.53	0.60	58
1	0.80	0.88	0.83	120
accuracy			0.76	178
macro avg	0.73	0.70	0.71	178
weighted avg	0.76	0.76	0.76	178

Testing Data

Good performance on unseen data, strong at identifying approved applicants, with room to improve detection of risky cases.

Predictive Model Comparison – Performance & ROC Analysis



- **Random Forest** model achieves highest ROC-AUC (0,728)

Model	Test ROC -AUC	CV ROC-AUC (std)
Random Forest	0.728	0,912
Logistic Regression	0.714	0,837
Decision Tree	0,648	0,756

Key insights

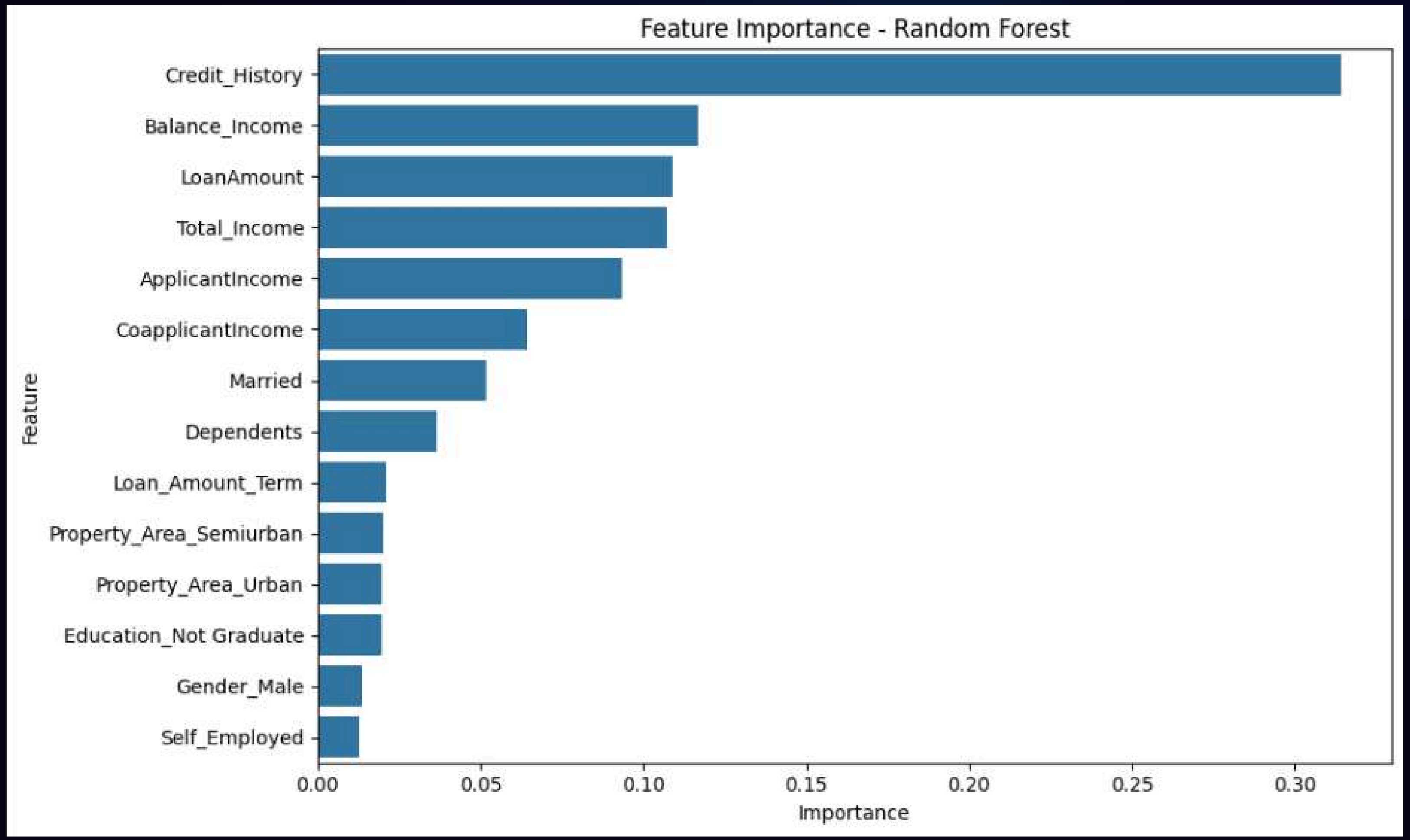
- Random Forest model achieves highest ROC-AUC (0.728)
- The ensemble model also delvers the best accuracy

Overall Model Performance

=== Overall Model Comparison ===

	Model	Test ROC-AUC	CV ROC-AUC (mean)	CV ROC-AUC (std)	Accuracy	Precision (weighted)	Recall (weighted)	F1 (weighted)
0	Random Forest	0.728161	0.911874	0.042392	0.764045	0.756554	0.764045	0.757634
1	Logistic Regression	0.714080	0.836641	0.052333	0.707865	0.710541	0.707865	0.709113
2	Decision Tree	0.647845	0.756171	0.045639	0.657303	0.686567	0.657303	0.666143

Key Drivers of Loan Approval



Credit_History

Strongest driver for approval likelihood.

Financial Health

LoanAmount, ApplicantIncome, Total_Income heavily influence approval.

Coapplicant Income

Contributes meaningfully to approval decisions.

Property Location

Semiurban > Urban > Rural in approval likelihood.

Strategic Business Recommendation

The most important 3 recommendations

1

Strengthen Credit History Evaluation

Strongest predictor of approval and repayment (RF importance = 0.27)

Actions:

- Automate credit bureau checks
- Flagship high-risk applicants early
- Offer credit-building programs

2

Implement Income-to-loan Affordability Metrics

Income explains <24% of approval variation based on the Random Forest feature

Actions:

- Set ITL thresholds (30-40%)
- Encourage joint applicants

Higher repayment sucess
Reduced Delinquency

3

Introduce Loan Amount Risk Tiers

Higher loan sizes reduce approval likelihood
RF importance = 0:14)

Actions:

- Create risk-based loan tiers
- Require collateral for large loans
- Offer alternative products for high risk cases

Supporting Recommendations – Enhancing Fairness & Efficiency

4 Integrate Geographic Risk (Property Area)

Why it matters: Approval rates vary significantly by region (Semi-urban > RuriRural). Adjusting policy by geography improves fairness and portfolio stability.

Actions:
- Adjust interest rates and LTV by region
- Apply additional checks in rural areas

Improved segmentation & stability

5 Remove Low-Impact Demographic Variables

Why it matters: Gender, Education, Married, status, and Self, Employed show near zero predicitive

Actions:
- Exclude low-impact demographit. attributes from the scorring mode.

Impact: Improve compliance and fairness

6 Introduce Conditional Approval Workflow

Why it matters: Borderline cases can still be profitable when structured with appropriate safeguards

Actions:
- Create a medium-risk category
- Include guarantors when applicable

Reliable long-term performance

7 Implement Ongoing Model Monitoring & Governance

Impact:
- Models degrade over time (drift), reducing accuracy and fairness if not regularly updated
- Retrain the model annually

Impact: Sustained long-termperformance

Limitations & Future Enhancements

1

Small Dataset

~600 rows limit generalizability of findings.

2

Historical Bias

Past approvals may contain inherent biases.

3

Missing Other Financial Depth

Lack of data on savings, employment length, collateral.

4

Future Model Upgrade

Explore XGBoost/GBM for improved performance.

5

Dynamic Tuning

Implement cutoff tuning and regular retraining.

THANK YOU