Comprehensive Report: Credit Scoring Model

Title: Predicting Loan Defaults Using Machine Learning

1. Introduction

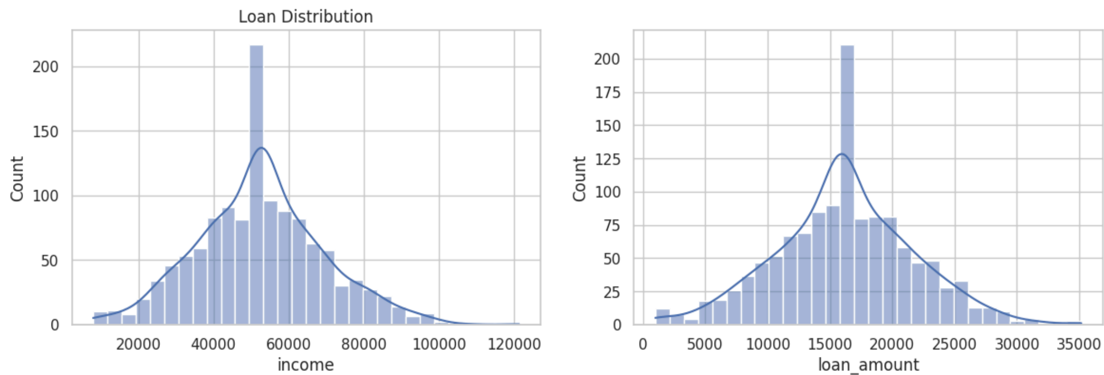
- Objective: Predict loan defaults using income, loan amount, credit history, and term.
- Business Impact: Help banks assess risk and reduce financial losses.
- Dataset: 1,248 records with 5 features (income, loan_amount, term, credit_history, defaulted).

2. Data Preprocessing

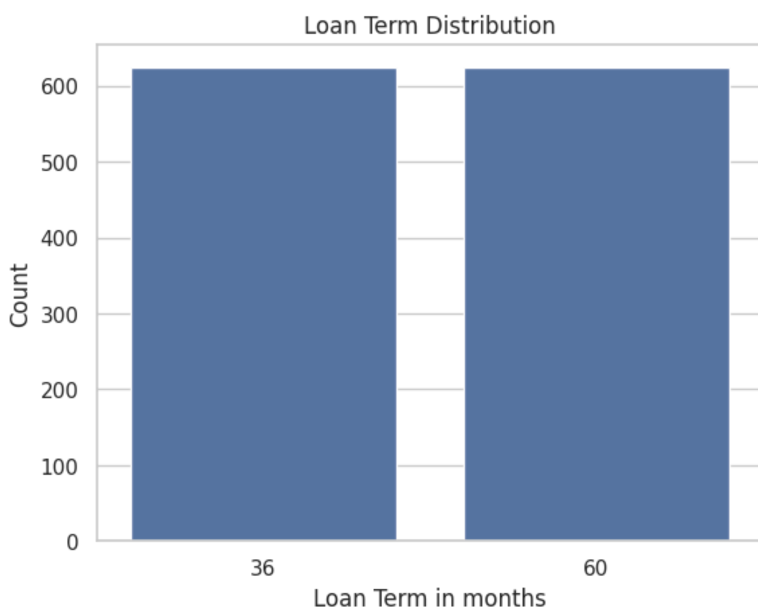
- Missing Values:
  - Filled income and loan_amount with median values.
  - Replaced credit_history missing values with mode (most frequent value).
- Feature Engineering:
  - Created term_in_binary (36 months → 0, 60 months → 1).
  - Applied log transforms to income and loan_amount for normality.
  - Scaled features using StandardScaler.

3. Exploratory Data Analysis (EDA)

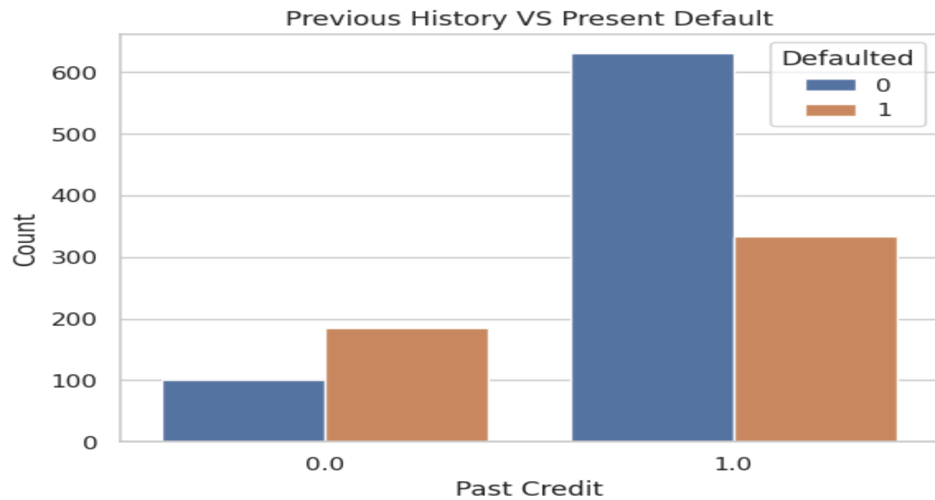
1. Income Distribution & Loan Amount: Right-skewed → Log transform improved normality.



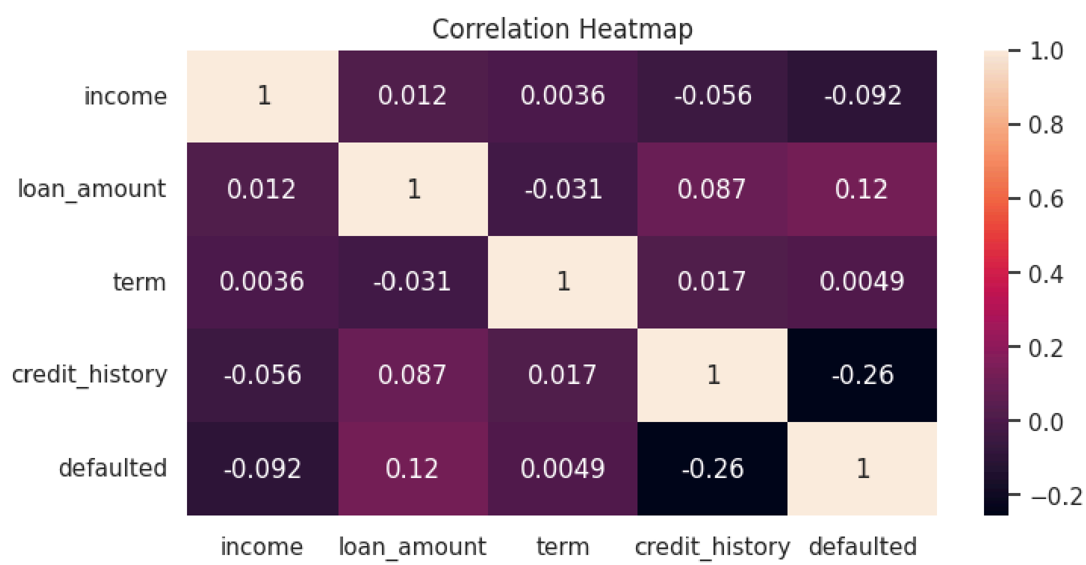
2. Loan Term Distribution: Balanced between 36m (50%) and 60m (50%).



3. Credit History vs. Default: Borrowers with poor credit history (0.0) defaulted more often.



4. Correlation Heatmap: credit_history negatively correlates with defaulted (-0.26).



4. Model Training & Evaluation

- Algorithms Tested:

Model	Accuracy	Precision (Default=1)	Recall (Default=1)
Logistic Regression	65.2%	0.67	0.37
Random Forest	60.0%	0.54	0.44
Decision Tree	55.2%	0.47	0.43

- Key Insight:
    - Logistic Regression performed best but suffers from low recall (missed 63% of defaults).

5. Conclusion & Recommendations

- Conclusion:
    - Credit history is the most influential feature.
    - Models need improvement in detecting defaults (low recall).
- Next Steps:
    - Try XGBoost or Neural Networks for better recall.
    - Deploy as an API for real-time predictions (e.g., using Flask).