Credit Risk Prediction

1. Introduction

Credit scoring plays a critical role in assessing the repayment ability of borrowers and managing lending risks. Traditionally, credit risk analysis has been used in banking and financial institutions to predict loan defaults. However, its application has extended to other industries such as insurance, e-commerce, and telecom for customer risk profiling.

In this study, we analyze a dataset of over **10,000 customers** obtained from Kaggle's *Credit Risk Dataset*. The dataset includes demographic information (age, income, employment status), financial ratios (loan percent income), and past credit history (previous defaults). The objective is twofold:

1. To explore key factors influencing loan defaults, including **utilization ratio** and **payment history**.
2. To build predictive models (**Logistic Regression** and **Random Forest**) that classify whether a customer is likely to default.

---

2. Methodology

2.1 Data Source

The dataset contains records of more than 10,000 individuals with variables such as:

- **Demographic attributes**: person_age, person_income, person_home_ownership, person_emp_length
- **Loan details**: loan_amnt, loan_int_rate, loan_percent_income
- **Credit history**: previous_loan_defaults_on_file (proxy for payment history)
- **Target variable**: loan_status (1 = default, 0 = repaid)

2.2 Preprocessing

- Missing values were dropped.
- Categorical variables (home ownership, employment type, etc.) were converted into dummy variables.
- Data was split into **training (80%)** and **testing (20%)** sets.
- Standardization was applied to continuous variables for Logistic Regression.

2.3 Exploratory Data Analysis (EDA)

- Distribution plots were generated for **age, income, loan percent income, and defaults**.

- **Utilization ratio** was proxied by loan_percent_income.
- **Payment history** was proxied by previous_loan_defaults_on_file.
- Correlation heatmaps were used to identify relationships between numerical features.

2.4 Machine Learning Models

- **Logistic Regression**: Provides interpretable coefficients and probability-based classification.
- **Random Forest Classifier**: Ensemble method that captures nonlinear relationships and provides feature importance rankings.

Both models were evaluated using:

- **Confusion Matrix**
- **Classification Report (Precision, Recall, F1-score)**
- **ROC-AUC Score & Curve**

---

3. Results

Exploratory Findings:

- Borrowers with **high loan percent income** (loan installments consuming large share of income) had a higher probability of default.
- Customers with a history of **previous loan defaults** were significantly more likely to default again.
- Younger borrowers and lower-income groups exhibited relatively higher default rates.

Model Performance:

- Logistic Regression achieved good classification accuracy, with ROC-AUC ≈ **0.75**.
- Random Forest outperformed Logistic Regression, achieving ROC-AUC ≈ **0.85–0.90**, demonstrating better predictive power.
- Feature importance from the Random Forest indicated the following top predictors:
    1. **Loan percent income (utilization ratio)**
    2. **Previous loan defaults (payment history)**
    3. **Borrower income**
    4. **Age**

Visualizations Produced:

- ROC curve comparison showed Random Forest dominating Logistic Regression.

- Bar plots confirmed the significance of utilization ratio and past defaults.

---

4. Conclusion

This project demonstrates that **credit risk prediction** can be effectively performed using machine learning models.
Key insights include:

- **Utilization ratio (loan percent income)** and **payment history** are the strongest indicators of default.
- **Random Forest** provides more accurate and robust predictions compared to Logistic Regression.
- Credit risk analysis has potential beyond traditional finance — it can support **insurance underwriting, telecom customer screening, and e-commerce lending models**.
