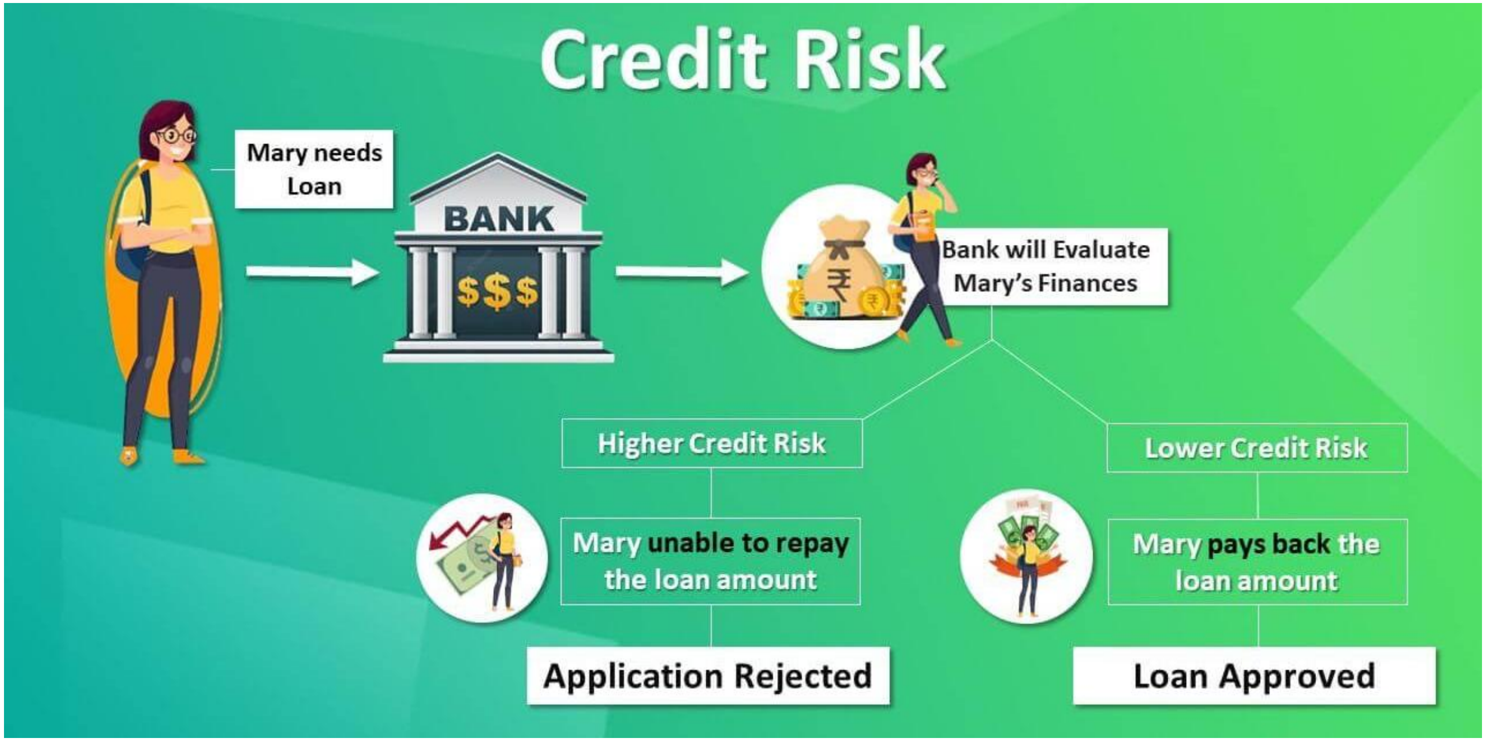
Credit Risk Prediction Using Machine Learning

- Fatima Kssayrawi

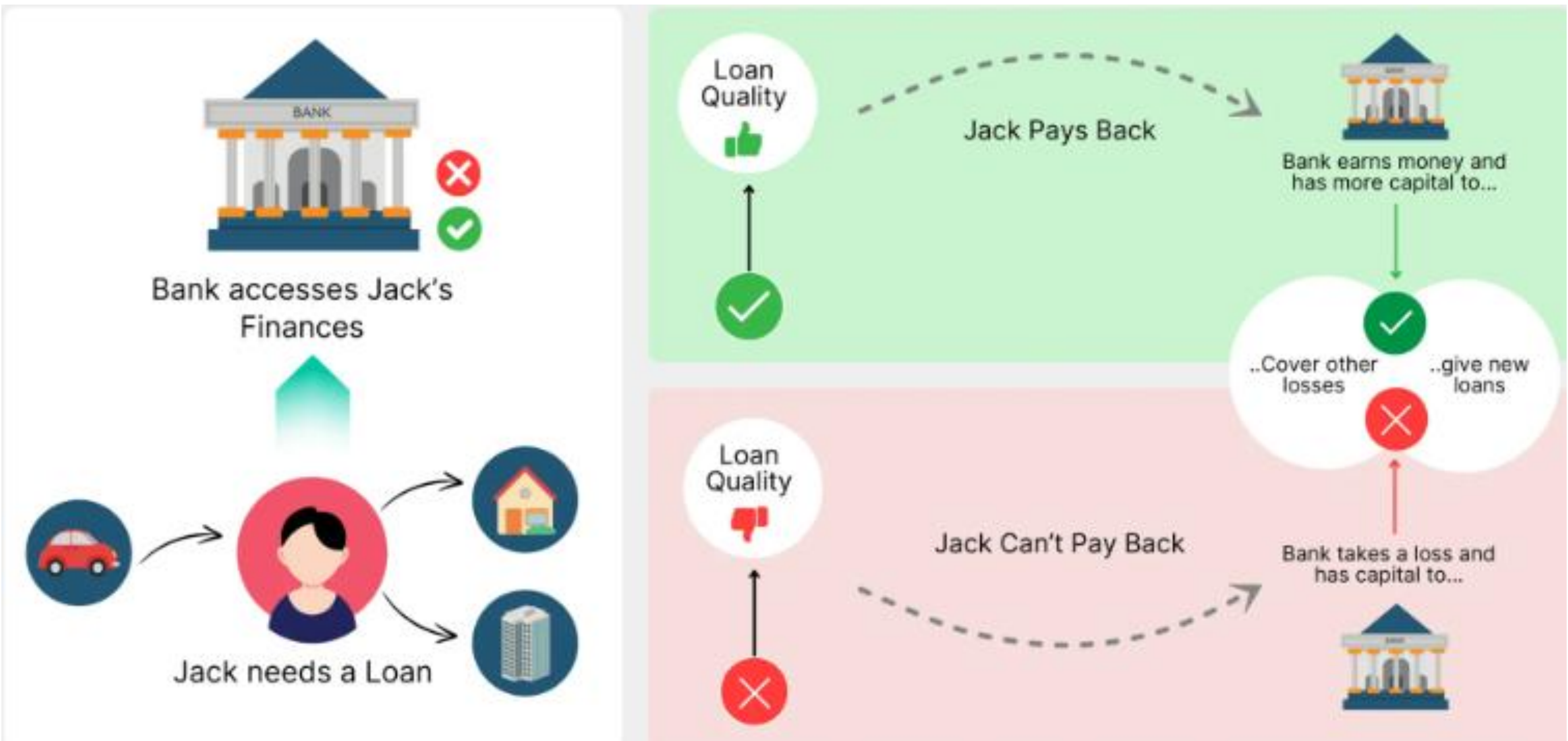
- Artificial Intelligence
- Thursday, April 17, 2025



Problem Statement

What is Credit Risk

It is the possibility that a borrower will **fail to repay a loan or meet contractual debt obligations**. In simple terms, it's the **risk of default**—when a borrower is either unwilling or unable to pay back the money they owe.



Why Is Credit Risk Assessment Important?

Helps Minimize Loan Defaults

- Identifies borrowers likely to default based on credit history, income, and financial behavior
- Enables lenders to approve loans only when repayment likelihood is high

Protects Financial Institutions' Assets

- Reduces the risk of financial loss due to unpaid loans
- Helps determine **loan approval**, **amount**, and **interest rate** based on risk level

Supports Fair and Equitable Lending

- Ensures decisions are based on **creditworthiness**, not biased by race, gender, or background
- Promotes **transparency and fairness** in financial systems

Traditional Approaches to Credit Risk Prediction

1. Credit Scoring Models	• Use statistical methods (e.g., FICO) to assign risk scores based on credit history, income, and other personal attributes.
2. Financial Statement Analysis	• Involves analyzing balance sheets, income statements, and cash flow reports.
3. Expert Judgment	• Credit analysts use industry experience and qualitative insights (e.g., management quality, market outlook) to assess risk.

Project goal, Input & Output

Our Goal

Build a machine learning model that **predicts whether a borrower is high-risk (1) or low-risk (0)** using personal, financial, and behavioral features.

Input

Structured borrower data including:

- Income
- Profession
- House/Car ownership
- City and State
- Experience, Age, etc.

Output

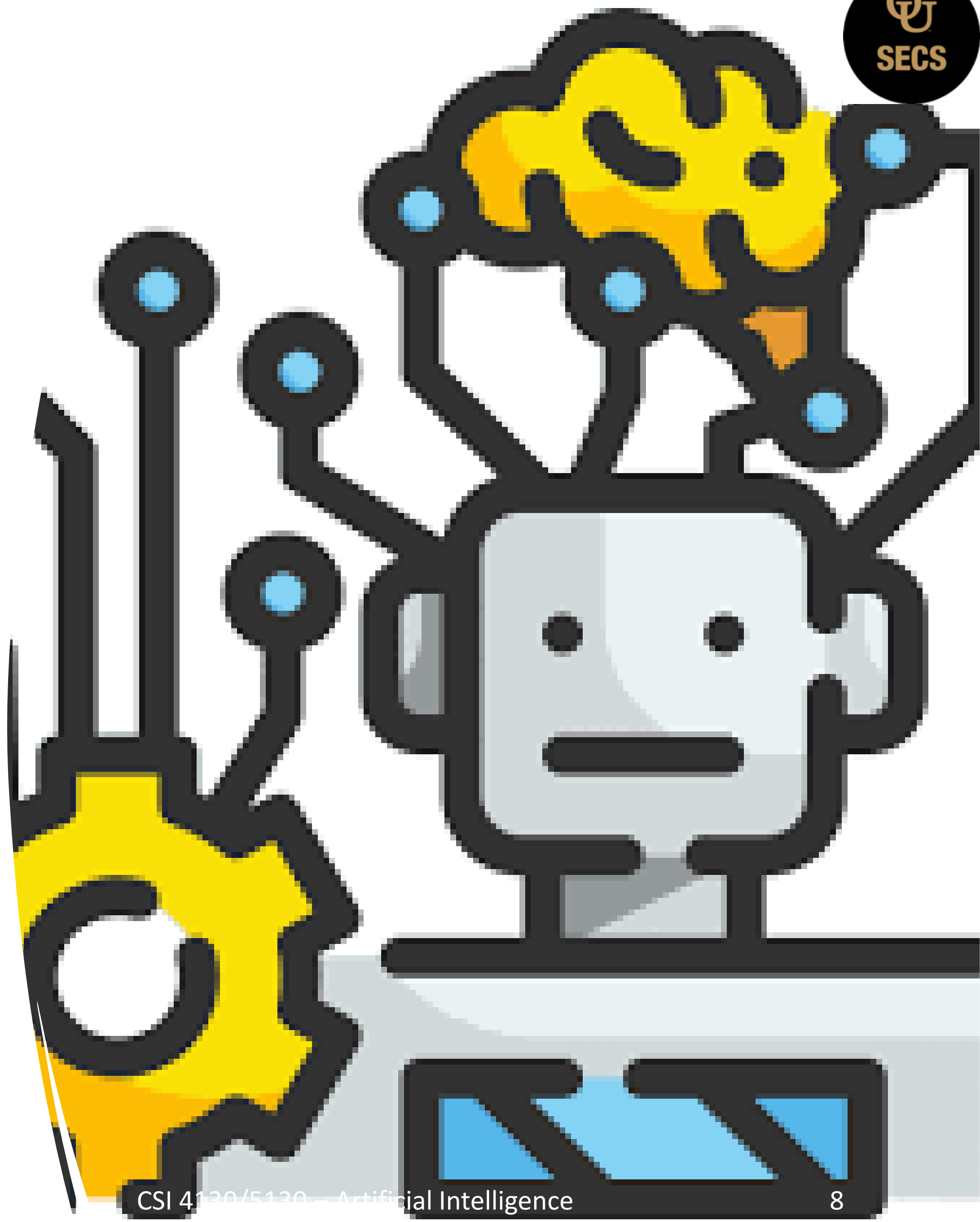
A **binary classification**:

- 1 = High Risk (likely to default)
- 0 = Low Risk (likely to repay)

Technical Challenges

Challenge	Solution
⚠️ **Class Imbalance**	✅ Used **SMOTE** to balance the training data
🧩 **Complex Non-Linear Relationships**	✅ Used **ML models** like Random Forest & GBM
🧠 **High-Cardinality Categorical Features**	✅ Applied **One-Hot Encoding** to transform variables
📏 **Inconsistent Feature Scales** (if relevant)	✅ Applied **StandardScaler** for uniform scaling

My approach: Machine Learning for Credit Risk



📦 Dataset Overview

Source:

Benchmark dataset from Univ.ai with **252,000+ borrower records**

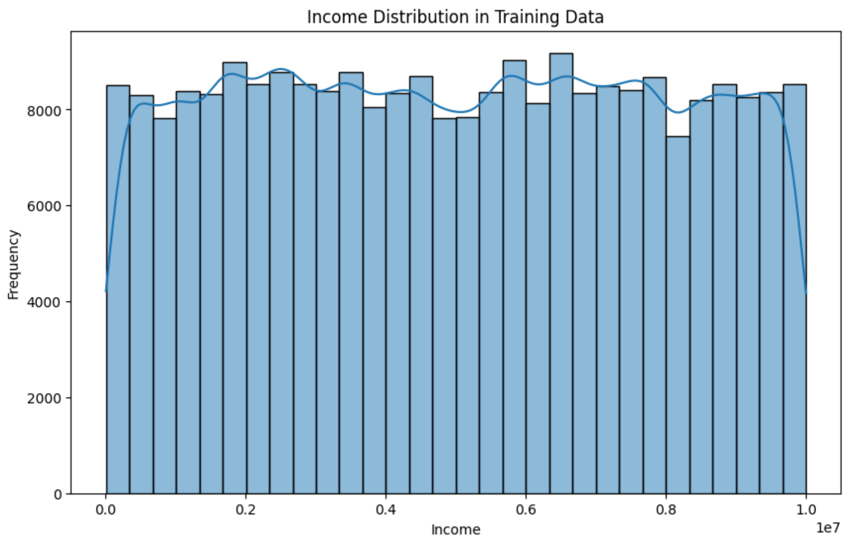
Features include:

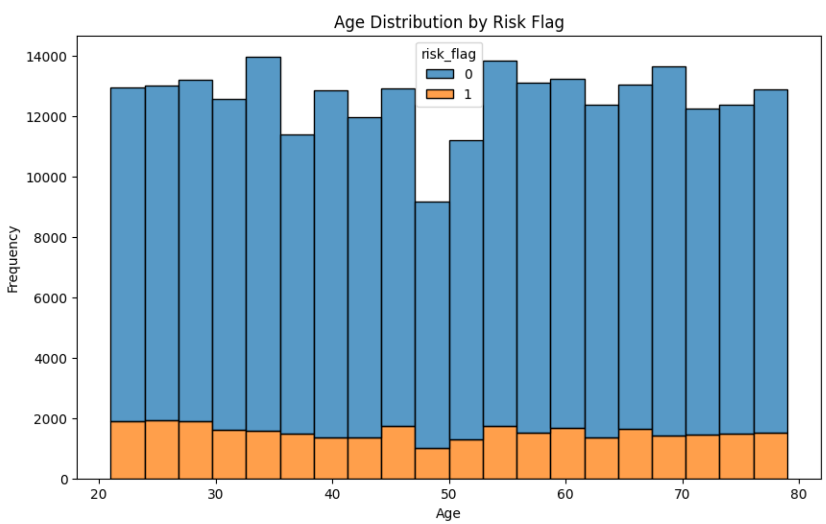
- **Demographics**: age, marital status, city, state
- **Financials**: income, house/car ownership
- **Employment**: profession, years at current job/house

Target variable: *risk_flag*

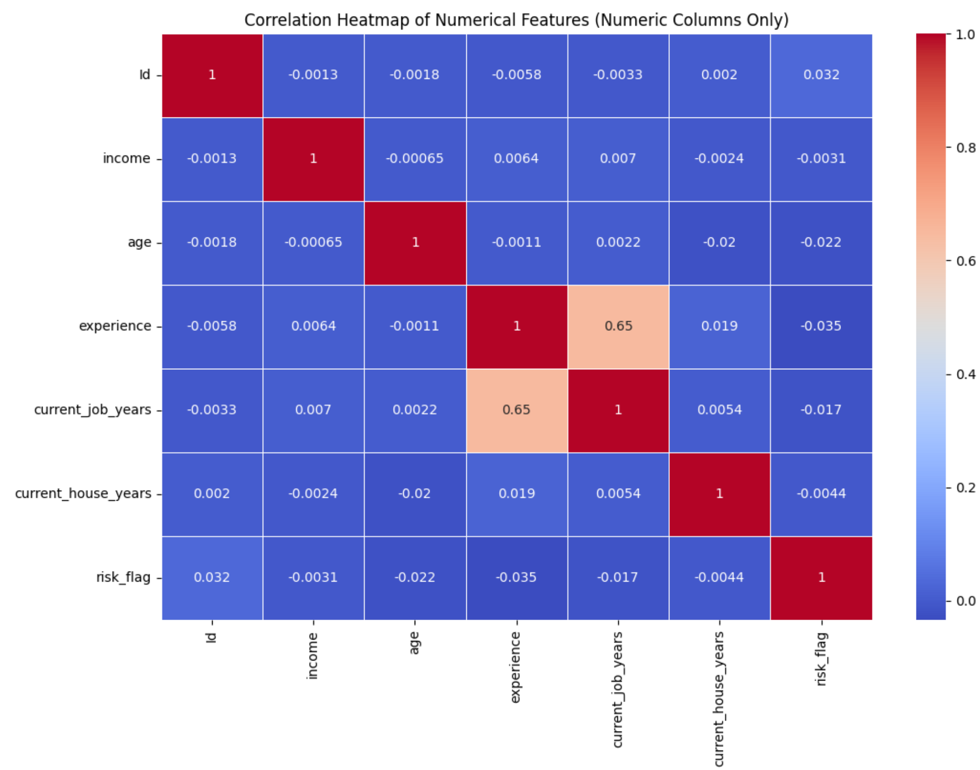
- 1 = **High Risk** (likely to default)
- 0 = **Low Risk** (likely to repay)

Technical Approach (EDA)



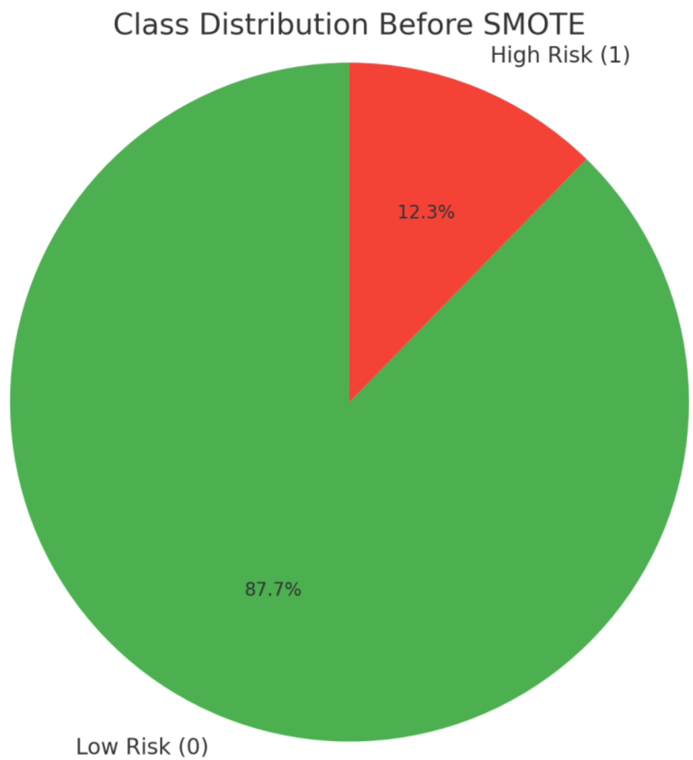


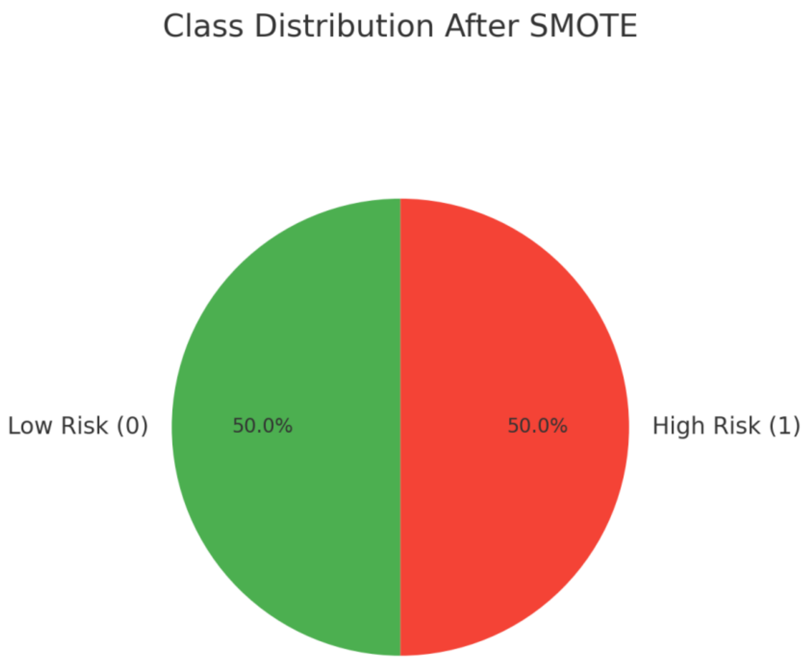
- ✓ Checked feature distributions: age, income, experience, etc.
- ✓ Visualized relationships between features and the target variable (risk_flag)
- ✓ Identified class imbalance between low-risk and high-risk borrowers
- ✓ Detected categorical features (e.g., profession, house ownership) and their unique values
- ✓ Ensured no missing values or duplicates in the dataset



Technical Approach (Data Preprocessing)

- **Preprocessing**:
  - ✓ One-hot encoding
  - ✓ Scaling with StandardScaler
  - ✓ **SMOTE** applied to training data only to address class imbalance





⚙️ Technical Approach (Modeling + Deployment)

Modeling:

- Compared three classifiers:
  - 🌳 Decision Tree
  - 🌲 Random Forest
  - 🚀 Gradient Boosting
- Evaluation Metrics: **Accuracy**, **Precision**, **Recall**, **F1-Score**

Deployment Preparation:

- Model, scaler, and feature list saved (.pkl)
- Custom function developed for real-time predictions

Evaluation & Results

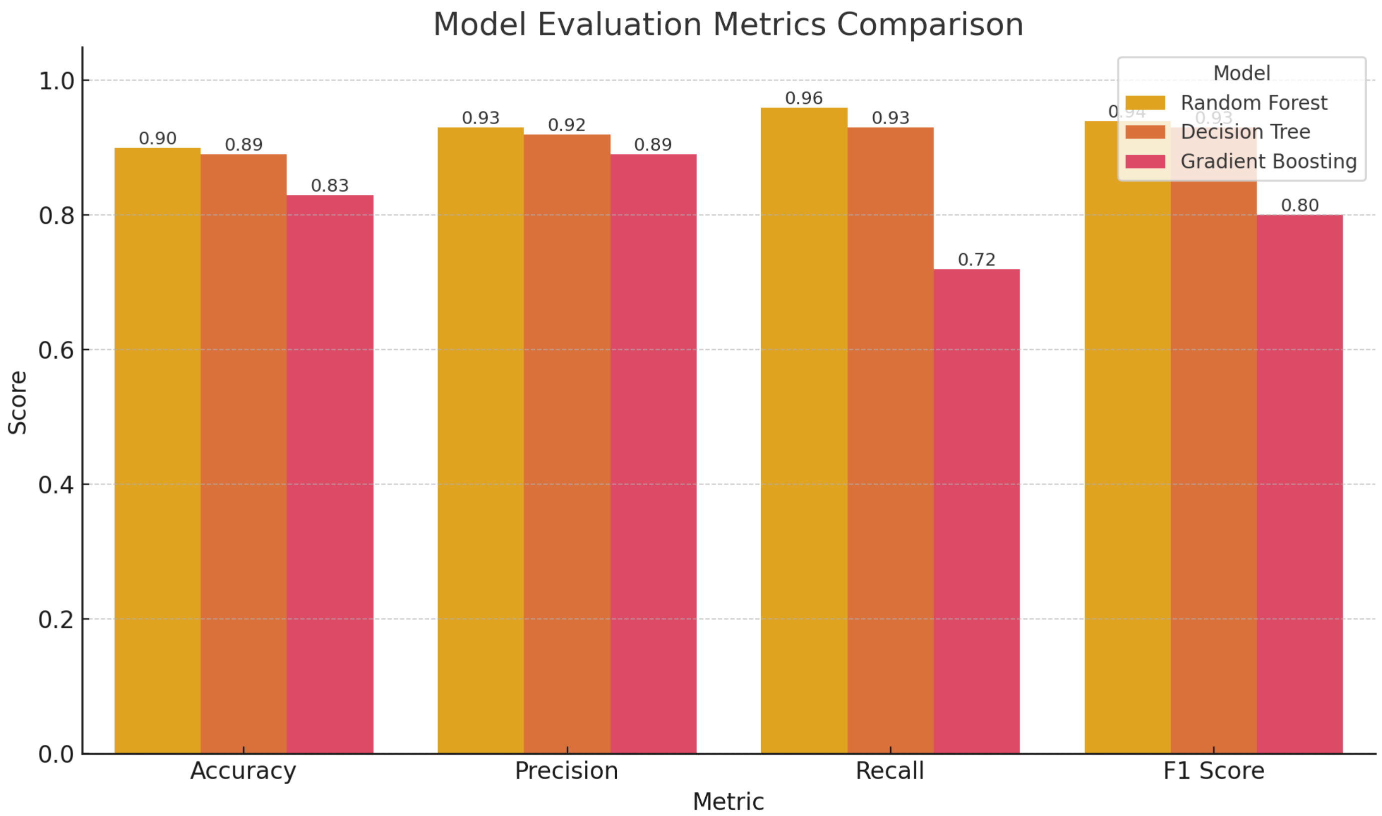
This chart compares **Accuracy**, **Precision**, **Recall**, and **F1-Score** across:

- 🌲 Random Forest
- 🌳 Decision Tree
- 🚀 Gradient Boosting

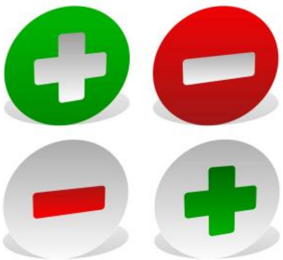
✓ **Random Forest** consistently outperforms the others across all metrics.

✓ **Decision Tree** was close in accuracy but slightly less robust.

✓ **Gradient Boosting** had lower recall, making it less effective at identifying high-risk borrowers.



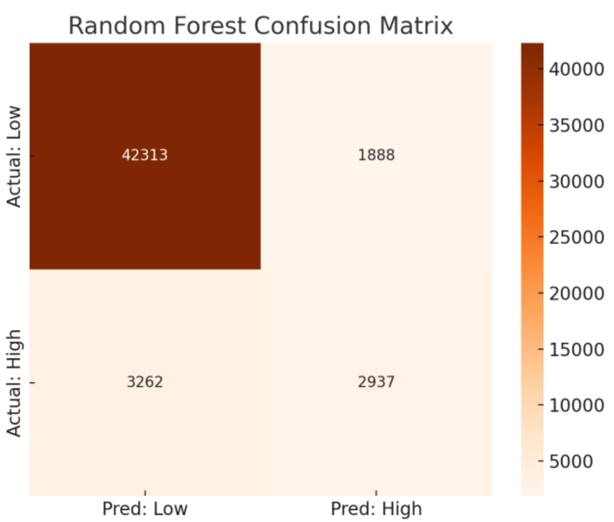
Confusion Matrix Comparison

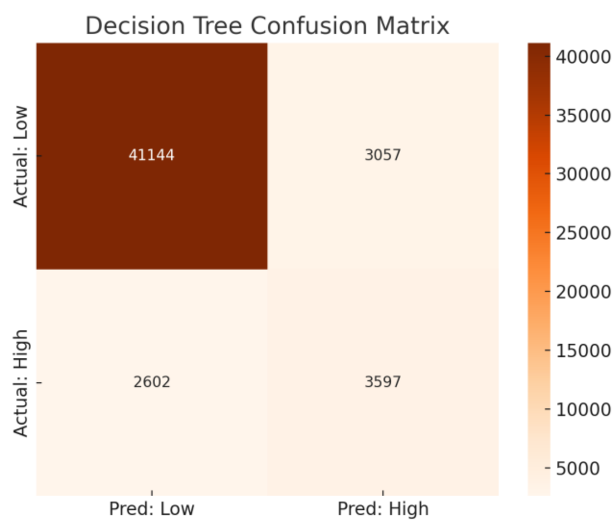


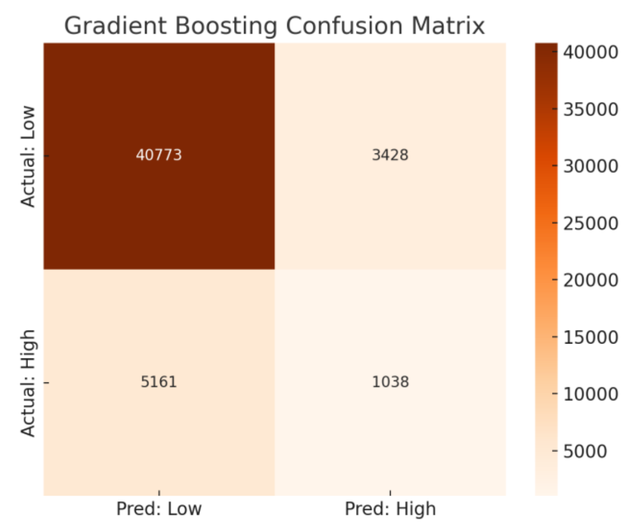
✅ **Random Forest** maintains a **low false positive and false negative rate**, making it highly balanced.

🌳 **Decision Tree** shows a slightly higher false positive rate but better true positive performance than GBM.

🚨 **Gradient Boosting** struggles with **high false negatives**, missing many high-risk cases.







Model Deployment

Final Model Selected: ✅ **Random Forest**

Saved Components:

1.random_forest_model.pkl → Trained classifier

2. scaler.pkl → StandardScaler used during training

3.model_columns.pkl → List of encoded feature columns

🧠 **Prediction Function**

Built a **predict_credit_risk()** function in Python

Accepts **borrower data** as input (e.g., age, income, profession)

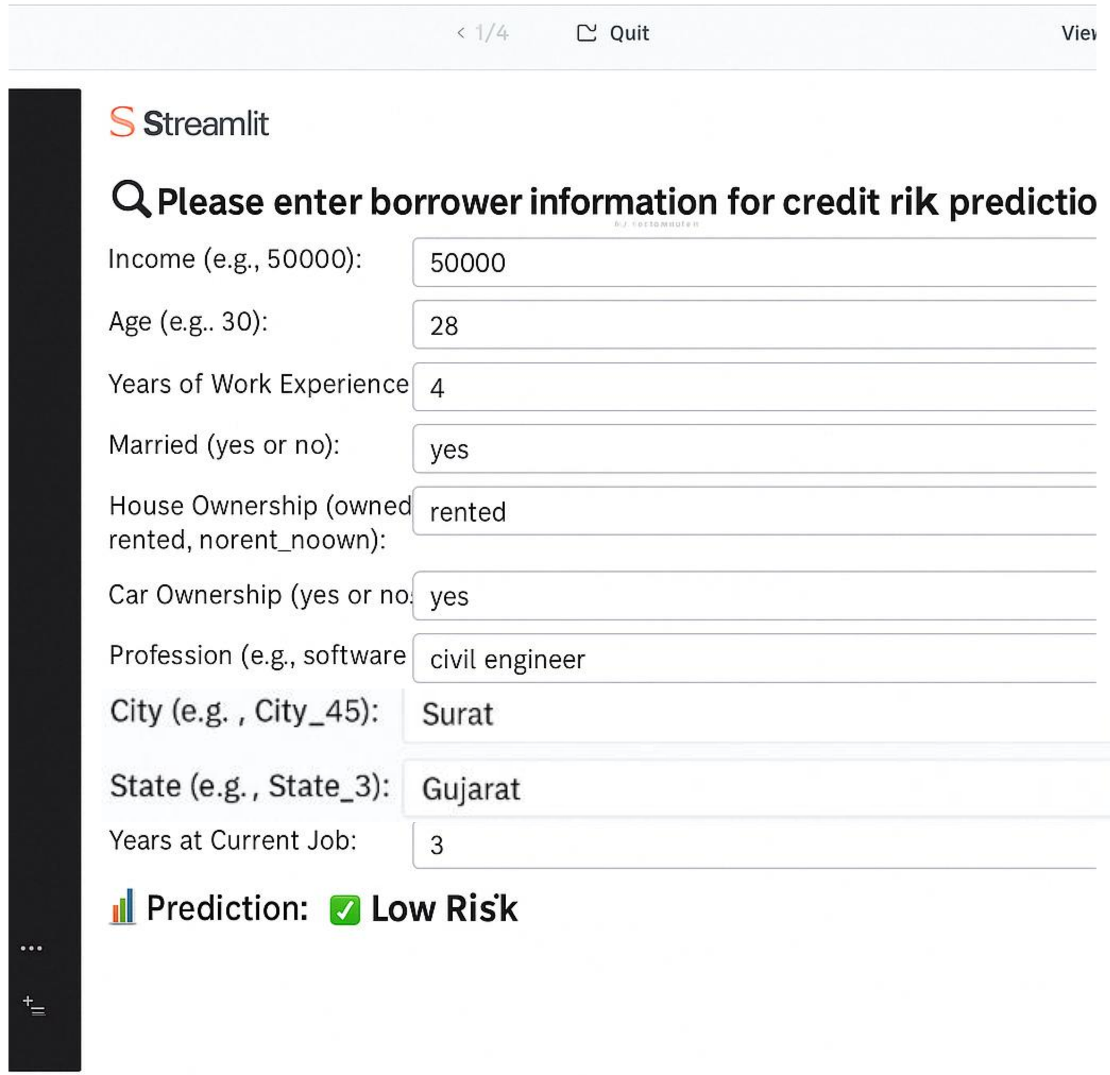
Applies **encoding**, **scaling**, and **model prediction**

Outputs: 1 = High Risk, 0 = Low Risk

Model Deployment

🖥️ **CLI-Based Interaction**

- ✓Users can enter data manually through a command-line interface
- ✓Immediate prediction response



Broader Impact & Limitations

<u>Broader Impact</u>

- ✓ Helps financial institutions make faster, fairer, and data-driven lending decisions.
- ✓ Promotes credit inclusivity by leveraging multiple personal, financial, and behavioral features.
- ✓ Reduces loan default risk, improves portfolio health, and contributes to economic stability.

<u>Limitations</u>

- ✓ Model performance is sensitive to data quality and may degrade with outdated or biased data.
- ✓ High-cardinality categorical variables (e.g., profession, city) may still hide nuances not captured through encoding.

Future Work

1. Build a **Streamlit dashboard** for a production-ready deployment.
2. Integrate **explainability** tools like SHAP to offer transparency in model predictions.
3. Experiment with **deep learning** or hybrid models for even higher predictive power.

Thank You!

https://github.com/fatimakssayrawi9/Credit-Risk-Prediction-Using-Machine-Learning