Predicting Responsible Credit Card Customers

By: Sukriti Sharma, Prajwala Maharjan, Alex Zheng, Dream M Lopez, Salvador Rangel Huerta

Introduction

The Problem:

- Banks face financial losses when customers default on credit card payments
- Predicting responsible customers is challenging but critical for minimizing risks

Question: How can machine learning models be used to predict responsible credit card customers in order to minimize default risks and financial losses for banks?

Our Objective:

- Develop a machine learning model to predict whether a potential credit card customer is likely to default

Stakeholders:

- banks and financial institutions
- responsible customers

Dataset

	X1	X2	X3	X4	X21	X22	X23	Y
ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
1	20000	2	2	1	0	0	0	1
2	120000	2	2	2	1000	0	2000	1
3	90000	2	2	2	1000	1000	5000	0
4	50000	2	2	1	1100	1069	1000	0
5	50000	1	2	1	9000	689	679	0
6	50000	1	1	2	1000	1000	800	0
7	500000	1	1	2	20239	13750	13770	0
8	100000	2	2	2	581	1687	1542	0
9	140000	2	3	1	1000	1000	1000	0
10	20000	1	3	2	13007	1122	0	0
11	200000	2	3	2	300	3738	66	0
12	260000	2	1	2	22301	0	3640	0
13	630000	2	2	2	6500	2870	0	0

ID: unique identifier

LIMIT_BAL: credit limit for each client

SEX: 1=male, 2=female

EDUCATION: 1=graduate school, 2= university, 3 = high school, 4=others

MARRIAGE: 1=married, 2=single, 3=others

AGE: Age of client

PAY_0 – PAY_6: Repayment status for last 6 months → −1=paid on time, 1= delay by 1 month…

BILL_AMT1 – BILL_AMT6: Bill amount in last 6 months → reflects monthly liability of client

PAY_AMT1 – PAY_AMT6: Amount paid in last 6 months → reflects repayment capacity of client

Default payment next month: 1=default, 0=no default → the output variable we are predicting

Data preprocessing steps: Understand the data, Handle missing vals, Filtering & cleaning, Normalization(standardize numeric columns), Data integration (join/merge multiple tables of different data), Feature engineering (pay & bill amt could be averaged), Feature validation

Methods

Abstract

- For our credit card default prediction task, we evaluated three models: **Logistic Regression, Random Forest, Multi-Layer Perceptron**
- Performance was measured using **precision**, **recall**, **F1-score**, and **ROC-AUC.**
- The goal was to balance precision and recall, focusing on the minority class (defaulters) while maintaining good overall accuracy

Key Metrics

- **Precision:** Proportion of correct positive predictions out of all positive predictions
- **Recall:** Proportion of actual positives correctly identified
- **F1-Score:** Harmonic mean of precision and recall

Modeling Process

MLP Classifier:

- **Preprocessing**
  - One hot encoded categorical features
  - Normalized rest of numerical data
- **Hyperparameter Tuning (Gridsearch)**
  - Best activation function: logistic
  - Best solver: adam
- **Further Potential Improvements**
  - More thorough hyperparameter search
  - Different data preprocessing
  - Greater number and size of hidden layers

```
Classification Report (MLP Classifier):
              precision    recall  f1-score   support

           0       0.85      0.95      0.89      4548
           1       0.67      0.38      0.49      1266

    accuracy                           0.82      5814
   macro avg       0.76      0.66      0.69      5814
weighted avg       0.81      0.82      0.81      5814


Confusion Matrix (MLP Classifier):
[[4313  235]
 [ 784  482]]

ROC-AUC Score (MLP Classifier):
0.7603109399336688
```

Modeling Process Continued

Logistic Regression:

- Utilized as a baseline model for comparison
- Balanced class weights to address the imbalance in the dataset
- Trained using 500 iterations for convergence

Random Forest:

- Ensemble model combining multiple decision trees
- Default hyperparameters with 100 trees and balanced class weights
- Handles non-linear relationships and captures interactions between features

Tuned Random Forest:

- Reduced overfitting by limiting tree depth (max_depth=15).
- Increased min_samples_split=10 and min_samples_leaf=5 for better generalization

Results and Conclusions

INDIVIDUAL WORK SYNTHESIZATION

- Each group member attempted to perform a series of steps to preprocess data, train machine learning models, and evaluate their performance using a credit card default prediction dataset

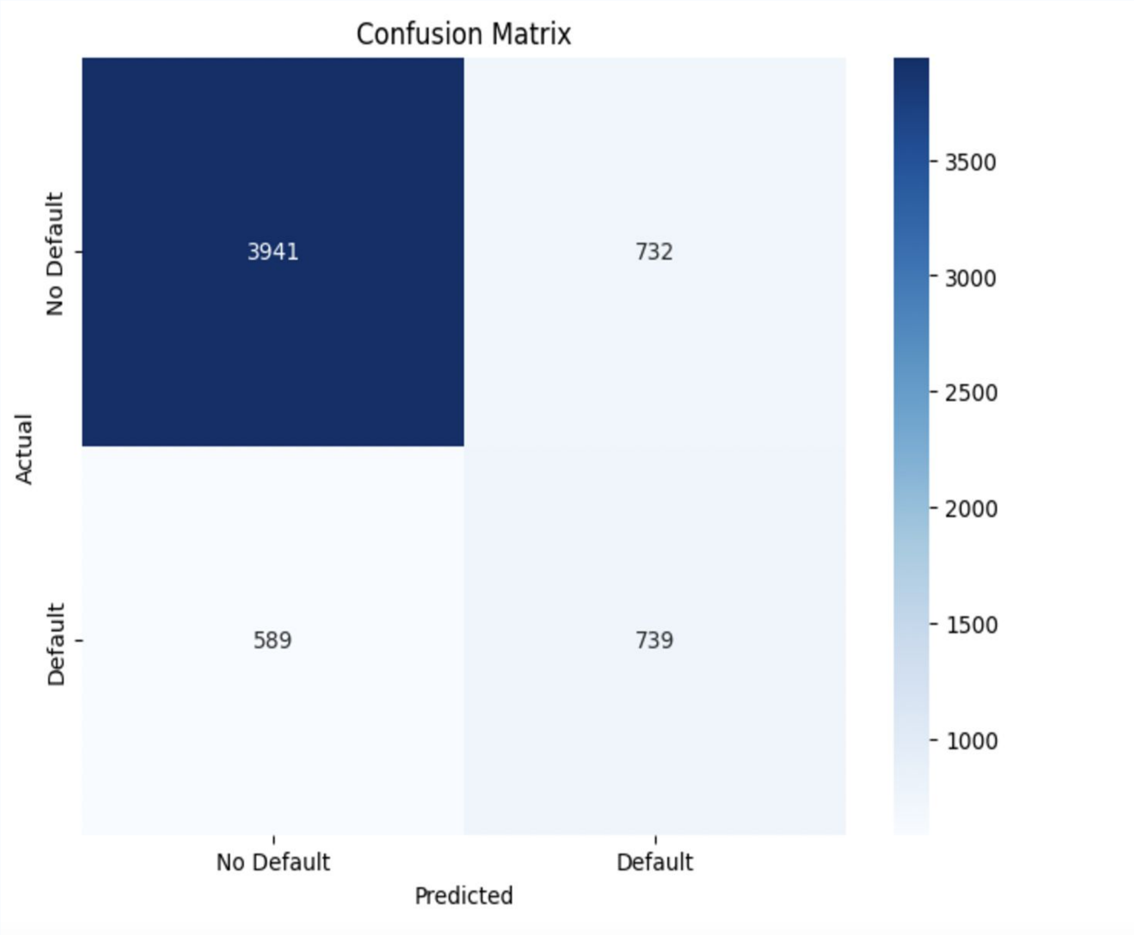
- The following steps were included in the code:
  - Data Loading & Preprocessing, Splitting the data into training and testing sets, Logistic Regression model, Predictions made on the test set, Performance is evaluated using Classification report (precision, recall, F1-score), Confusion matrix, and ROC-AUC score
  - Random Forest model trained on the data, evaluated using Classification report, Accuracy, ROC-AUC score, Confusion matrix (visualized as a heatmap)

```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.84      0.86      4673
           1       0.50      0.56      0.53      1328

    accuracy                           0.78      6001
   macro avg       0.69      0.70      0.69      6001
weighted avg       0.79      0.78      0.78      6001

Accuracy: 0.7799
ROC-AUC Score: 0.7777
```



Findings and Conclusions

Random Forest model: reveal multiple factors that significantly influence whether a client will default on their credit card payment

- These features might include:
  - Credit limit: A lower credit limit could indicate higher financial risk
  - Previous payment history: Clients with a history of late payments are more likely to default
  - Age and education: Demographic factors may also play a role, with younger or less educated clients possibly facing more financial instability
  - Balance utilization: High utilization of available credit may indicate financial strain and a higher likelihood of default

ROC-AUC score: indicates how well the Logistic Regression model, Random Forest model, and MLP model distinguish between clients who default and those who do not

- A higher ROC-AUC score indicates better model performance in identifying high-risk clients who are more likely to default

Implications & Results

Improved Risk Management for Financial Institutions

Identifying high-risk clients early allows financial institutions to take preemptive actions to minimize losses from defaults

Optimizing Credit Policies

- Institutions can adjust credit limits based on the model's predictions
- Clients predicted to be high-risk might be granted lower credit limits, reducing the potential for large defaults and vice versa

Economic Impacts

On a macroeconomic scale, by accurately predicting defaults, financial institutions can reduce the likelihood of large-scale financial crises triggered by widespread defaults, leading to greater economic stability

Model Ready for Real-World?

- Before deploying this model in the real world, the following steps need to be addressed:
  - Fine-tune model hyperparameters and perform cross-validation for robust performance
  - Audit for bias and ensure fairness across different demographic groups
  - Ensure the model complies with regulatory requirements and is explainable for transparency
  - Conduct data quality checks
  - Implement scalable solutions for large datasets and real-time predictions
  - Monitor model performance continuously and retrain as necessary

Any Questions?