Finance Club

Open Project Summer 2025

Title: Credit Card Default Prediction Using Classification and Risk-Based Techniques

Overview:

Bank A aims to improve its credit risk management framework by developing a **forward-looking Behaviour Score** — a classification model that predicts whether a credit card customer will **default in the following month**.

You are provided with anonymized historical behavioral data of over 30,000 credit card customers, with a labeled target variable: default.payment.next.month. This variable indicates whether a customer defaulted on their payment in the **next billing cycle**. The goal is to build a model that can accurately flag potential defaulters **in advance**, allowing the bank to adjust credit exposure, trigger early warning systems, and prioritize risk-based actions.

Variables:

S.no	Column Headers	Explanations
1.	Customer Id	Unique identifier for each customer
2.	marriage	Marital status of the customer (1 = Married, 2 = Single, 3 = Others)
3.	sex	Gender of the customer (1 = Male, 0 = Female)
4.	education	Education level (1 = Graduate School, 2 = University, 3 = High School, 4 = Others)
5.	LIMIT_BAL	Credit limit assigned to the customer (in currency units)
6.	Age	Age of the customer (in years)
7.	Pay_0 to 6	**Repayment status in the past 6 months** **Values:** -2 = No consumption, -1 = Paid in full, 0 = Paid on time, 1–9 = Payment delayed by 1 to 9+ months.
8.	Bill_amt1 to 6	Credit card bill amount for each of the last 6 months (Bill Amount = Previous Balance + New Purchases - Payments Made**)**
9.	Pay_amt1 to 6	Amount paid by the customer in each of the last 6 months
10.	AVG_Bill_amt	Average bill amount over the 6-month period
11.	PAY_TO_BILL_ratio	Ratio of total payment to total bill amount over 6 months
12.	next_month_default	Target variable : 1 if customer defaulted next month , 0 otherwise

Objectives:

- Build a binary classification model to predict customer default (default.payment.next.month: 1 = Default, 0 = No Default).
- Handle class imbalance using appropriate techniques (e.g., SMOTE, class weighting, downsampling).
- Perform exploratory and financial analysis to understand how key behavioral variables influence default risk.
- Engineer features and transformations that are **financially meaningful** and predictive.
- Test and compare multiple classification models such as:
    - Logistic Regression
    - Decision Trees
    - Ensemble Methods (e.g., XGBoost, LightGBM)
- Choose and justify evaluation metrics that reflect **real-world credit risk trade-offs**.
- Set a **classification threshold** aligned with the bank's risk appetite and discuss the business implications of false positives and false negatives.
- Generate production-style predictions on an **unlabeled validation dataset**.
- Ensure that predictions on the validation dataset are generated by maximizing the evaluation metric that best reflects credit risk priorities (e.g., Accuracy, Precision, F1-score, AUC-ROC) through appropriate tuning of the classification threshold.

Deliverables:

1. **Prediction File (CSV):**

Two columns: Customer, next_month_default(1 or 0)

2. **Code:**

A clean, reproducible Jupyter notebook or Collab file or Python script covering:

- Data loading and preprocessing
- Exploratory data analysis (EDA)
- Financial insights from key variables
- Feature engineering and transformations
- Model training and validation
- Final predictions

3. **Report (in notebook or as separate PDF):**

Include a clear and structured summary of your process:

- Overview of your approach and modeling strategy
- EDA findings and visualizations (e.g., variable distributions, correlations)
- Financial analysis of which variables drive default and why (e.g., overdue payments, credit utilization, repayment history)
- Model comparison and justification for final selection
- Evaluation methodology — explain which metric(s) were prioritized and justify their relevance to credit risk.
- Discuss how you selected the classification cutoff
- Business implications
- Summary of findings and key learnings

Data Description

Train Dataset (~25,000 records)

- Features include LIMIT_BAL, age, sex, education, marriage, repayment status (pay_0, pay_2, …), bill amounts, payment amounts, etc.
- Target variable: next_month_default (1 = Default, 0 = No Default)

Validation Dataset (~5,000 records)

- Contains the same feature set without the target
- Your model must predict next_month_default for these records.

Tools and Libraries:

- **Python packages**: pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn, xgboost, lightgbm
- **Optional**: SHAP or LIME for explainability of model predictions

Evaluation:

- **EDA & Financial Insight – 30%**
  Visuals, trends, financial interpretation
- **Class Imbalance & Model Performance – 30%**
  Balancing, tuning, metric evaluation
- **Feature Engineering & Metric Justification – 20%**
  New features, threshold reasoning
- **Code Quality & Report – 20%**
  Clean code, clear summary