Finance Club

Open Project Summer 2025

Title: Credit Card Behaviour Score 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.

The goal is to go beyond prediction and create a **financially interpretable** model that helps the bank understand default **patterns**, take early action, and manage credit exposure.

Click here to download the Dataset (Project 2)

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_m(Pay_0 to 6)	PAY_0 represents the payment status in the most recent month (Month 0), PAY_2 represents the payment status one month before that, and so on.Values: • -2 = No credit consumption (no bill) in month $m-1$ • -1 = Bill generated and fully paid in month $m-1$ (same month payment) • 0= Partial or minimum payment made (revolving credit) • ≥1 = Payment delayed by that many months (1 = 1 month overdue, etc.)
8.	Bill_amt_m(Bill_amt1 to 6)	Total bill amount at the end of month *m* (1 = last month, 6 = six months ago). • > 0 means customer owes money (paid less than previous bill) • = 0 no spending • < 0 customer overpaid previous bill (credit carried)
9.	Pay_amt_m(Pay_amt1 to 6)	Payment amount made in month m towards the bill generated in month m-1.
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

How these link per month (example for month 5):

- **`Bill_amt5`** = bill generated at end of month 5 (say, May).
- **`pay_amt5`** = payment made in month 5 to cover bill from month 4 (April).
- **`pay_5`** = payment status in month 5, showing whether April's bill was paid:
  - `-1` means fully paid on time
  - `0` means partial/minimum payment made
  - ≥1 means payment overdue by that many months

Clarification:

1.For Payment Status = 0

It indicates that the customer did not pay the full amount due, but at least paid some minimum required amount to avoid being classified as overdue.

2.Small bill amounts with a payment status of -2 can occur due to residual charges, billing timing differences, or carryover balances from previous months.

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.

- Go beyond basic EDA — analyze **behavioral trends** like payment delays, repayment consistency, and utilization
- Engineer features and transformations that are **financially meaningful** and predictive like credit utilization ratio, delinquency streaks etc.
- 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).
Please name your CSV file exactly like this:
submission_<YourEnrollmentNumber>.csv

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 insights and 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.
- Metrics result on train dataset (especially accuracy , F1 score, recall and F2 score)
- 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

Here's how we will assess:

Step 1: Effort & Deliverables Check

We will first verify that:

All required deliverables (prediction CSV, code, and report) are submitted.

Your work reflects actual implementation of what's outlined in the project document — no copy-pasting or generic solutions.

Your code and analysis show that you understood and worked through the problem.

Only submissions that meet these standards will proceed to the next evaluation stage.

Step 2: Evaluation by F2 Score

Since we’re expecting more submissions than initially planned, we will further evaluate qualifying submissions based on the F2 Score calculated using your predictions on the validation dataset.

Please ensure all work submitted is your own original effort and not copied from any external source. We will review the submissions carefully to check for any plagiarism or copying.