



## Finance Club

### Open Project Summer 2025

**Title:** Building a Behaviour Score: 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.

## 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