# Credit Risk Default Prediction — Project Report

*Default of Credit Card Clients (UCI), 30,000 clients, 24 variables*

## Project Overview

Predict credit card default using the UCI *Default of Credit Card Clients* dataset (30,000 observations, 24 features). Objectives: (i) build and tune classification models, (ii) handle class imbalance, (iii) evaluate performance beyond accuracy, and (iv) interpret drivers of default for business use.

## Methods

- **Data**: Demographics, credit limit, bill amounts (BILL_AMT1–6), payment status (PAY_0, PAY_2, PAY_3), repayment history.
- **Models**: Logistic Regression, Decision Tree, Neural Network (MLP).
- **Techniques**: SMOTE for class imbalance; GridSearchCV hyperparameter tuning.
- **Evaluation**: Accuracy, Precision/Recall/F1 (default class focus), Log loss, Confusion matrices.

## Key Results

- Neural Network outperformed baselines with **accuracy** $\approx$ **82%**.
- Default-class **precision improved 0.37 $\rightarrow$ 0.65**; **recall** $\sim$ **40%**.
- **Log loss: 0.4446–0.4675**, indicating robust probabilistic calibration.

## Feature Importance

- Logistic Regression coefficients highlighted **recent payment status** (`PAY_0`, `PAY_2`, `PAY_3`) and **bill amounts** (`BILL_AMT1--6`) as strong predictors.
- Decision Tree splits were dominated by `PAY_0` (most recent delinquency signal).
- Interpretation aligns with credit intuition: *recent missed/late payments* + *persistently high balances* $\Rightarrow$ higher default risk.

## Business Implication

- Raising precision from **0.37 to 0.65** improves targeting of high-risk clients by **28 percentage points**, reducing false positives in risk flags.
- Higher recall (~40%) captures more true defaults, strengthening portfolio monitoring and early-stage collections.
- Calibrated probabilities (low log loss) support *risk-based pricing*, *limit management*, and *watchlist* strategies.

## Limitations & Future Work

- Limited features: dataset lacks *income/employment* and *macroeconomic* covariates; results may understate cyclical risk.
- Future: integrate *macro indicators* (unemployment, GDP growth) or *bureau data*; test *Gradient Boosting/XGBoost* and calibrated ensembles.
- Governance: add *threshold tuning* for business KPIs (PD cutoffs by segment) and *stability monitoring* (PSI/CSI) for production use.

## Tech Stack & Files

- **Python**: pandas, numpy, scikit-learn, imbalanced-learn; **Viz**: matplotlib, seaborn.
- Report PDF: `11.pdf` — Code: `11all.py`
- Repo-ready README summary available (overview, methods, results, requirements, run instructions).

## Contact

**Tingyun (Caroline) Kuo** tkuo23@wisc.edu — linkedin.com/in/carolinekuo