# Client-Level Fraud Detection (Electricity & Gas)

## Overview

This project detects fraudulent clients in electricity and gas consumption data. Each customer has demographic features and a sequence of invoices; the goal is to classify the **client**, not an individual invoice. The dataset is highly imbalanced, so we aggregate invoices into per-client features and use strict validation: an 80/20 stratified split and 5-fold cross-validation on the training portion only. We evaluate a LightGBM model (GOSS + per-fold RandomOverSampler, tuned with Bayesian optimization) against an ANN baseline.

## Project Structure

**Final_Notebook.ipynb** – Main analysis and training workflow.

**Model folder** – Contains trained models and predictions:

- **ANN subfolder**
  - final_model.pkl – Saved ANN model
- **LightGBM subfolder**
  - final_lgb_goss_ros_v2.pkl – Saved LightGBM model
  - lgb_goss_ros_predictions_v2.csv – Test predictions from LightGBM

**Readme.pdf** – Project documentation.

**Project_Report.pdf**

## Reproducing Results

1. Follow the instructions to load the two datasets and install the necessary libraries.

2. Run basic data checks (heads, shapes, dtypes, nulls).

3. Execute feature-engineering cells to aggregate invoices per client.

4. Train LightGBM (GOSS + per-fold ROS) via BayesSearchCV (default n_iter=25).

5. Train ANN model following the code section.

6. Evaluate on the held-out test set; the notebook prints ROC–AUC, F1, precision, recall, and accuracy.

7. Review plots: ROC, PR, F1 vs threshold, confusion matrix, probability histograms, and feature importances.

### Notes and Recommendations

- Random seed is fixed to 42 for repeatability.
- Oversampling is performed **inside each CV fold** to avoid leakage; do not oversample globally before CV.
- The decision threshold is selected from the precision–recall curve to maximize F1; a default 0.5 threshold under-detects positives in imbalanced data.
- GOSS focuses learning on high-gradient samples and improves efficiency on imbalanced, weak-signal problems.

## License and Attribution