IT1244 Project report

Fraud in Electricity and Gas Consumption

Zhou Lingxuan, Guo Beijia, Duan Shuo, Zhang Dingning
IT1244, NUS

Introduction

This project addresses client-level fraud detection in electricity and gas consumption. Concretely, each customer has demographic attributes and different numbers of transactions; the task is to decide whether the client is fraudulent, not whether any single invoice is anomalous. The dataset is highly imbalanced and requires careful aggregation of invoices per client into informative features. Fraud in utility consumption imposes substantial financial losses and operational risks on providers. Monthly conventional-meter data tends to conceal anomalies within ordinary usage, which complicates learning from weakly correlated inputs under severe class imbalance.

We approach the problem with two complementary supervised models: Light Gradient Boosting Machine (LightGBM) and an Artificial Neural Network (ANN).

Recent Work and Limitations

The first study (Saha et al. 2024) compares ANN and LightGBM on a large Tunisian utility-billing dataset. It engineers features via feature creation, handles class imbalance with Random Oversampling/SMOTE, and reports LightGBM as best, outperforming ANN. A limitation is reliance on supervised learning with resampling and correlation-based selection on aggregated billing data, which may miss rare fraud modes.

The second paper (Taha and Malebary 2020) introduces an "Optimized LightGBM" tuned by Bayesian hyperparameter search and evaluated with 5-fold CV on two real credit-card datasets. It reports strong results (Accuracy 0.9840, AUC 0.9288, Precision 0.9734, F1 0.5695), highlighting the benefit of Bayesian search with LightGBM. A limitation is modest recall/F1 under severe imbalance and the focus on credit-card data, raising transferability questions to energy-consumption fraud.

The third work (Oprea and Bâra 2021) combines classification with anomaly detection on Tunisian utility-billing dataset using extensive feature engineering and a multivariate Gaussian "probability" feature. It reduces leakage by de-duplicating training data, with ensembles like LightGBM outperforming alternatives on unbalanced, weakly correlated inputs. Its processing methodology, creating and piping engineered features, improves classification metrics but remains dependent on careful data hygiene.

Overall, we frame utility-fraud detection as a client-level, imbalanced classification problem with aggregation over sparse monthly invoices.

Dataset and Processing

Data Issues

The principal difficulty is severe class imbalance: the ratio of fraud cases versus non-fraud cases is 4.60%, showing that fraud cases are rare. This misleads the training models toward the majority class and depresses positive-class recall unless addressed through sampling, calibrated thresholds, or cost-sensitive learning. This imbalance pattern motivated our resampling experiments and model selection. Beyond prevalence, the monthly granularity of the data weakens direct signal-to-label correlations; all numerical features show very weak correlation with the target (maximum ~0.07), necessitating careful feature construction and selection to avoid overfitting and computational waste.

Exploratory Data Analysis

Exploratory Data Analysis was conducted on the clients.csv and invoices.csv datasets to understand their structure, completeness, and quality before further analysis. The clients dataset contains 11,741 rows and 6 columns, including a binary target variable (target), while the invoices dataset has 194,374 rows and 12 columns.

Data types were examined, and all columns were found to have consistent types, with no missing values in either dataset. Initial inspections included viewing the first few rows and summarizing column-level statistics to ensure the datasets were properly loaded and interpretable.

Numerical variables in the invoices dataset, particularly the consommation_level columns, were further examined for anomalies. Extremely large values and rare occurrences (appearing ≤2 times) were flagged, resulting in the removal of 2,786 anomalous rows. The remaining 191,588 rows were retained in a cleaned dataset for subsequent analysis.

Data Engineering

Feature Creation

We engineered features to turn irregular, invoice-level histories into a stable, client-level dataset. First, we standardized time and derived basic consumption dynamics such as inter-invoice gaps and index deltas; these capture tempo and change rather than raw timestamps. Next, we condensed each client's sequence into a compact tabular profile by aggregating core numeric signals and a simple regularity indicator (weekday rate).

To avoid spurious seasonality and leakage, we deliberately excluded weak calendar aggregates and unstable transformations of months_number, focusing instead on features that reflect genuine consumption behaviour. Finally, we standardized numerical features and encoded categorical features after the train/test split. The result is a lean, leakage-safe feature set that emphasizes consumption dynamics and sequence structure over raw dates.

Feature Selection and Dimensionality Control

Because aggregation can explode dimensionality, we applied a two-stage filter mirroring a recent study (Saha 2024) on this dataset: an R^2-to-target screen to remove near-zero-signal features, followed by collinearity reduction using correlation matrices to drop one of any highly correlated pair, keeping the member with stronger target association. This approach reduced the feature set substantially while retaining diversity, as advocated in the literature for imbalanced tabular fraud tasks. We then aligned the selected columns across train and test and persisted the curated matrices for modeling.

Visualization and Diagnostics

After performing feature selection, we conducted targeted visual diagnostics to validate that the chosen features were truly informative for predicting fraud. Boxplots of each selected numeric feature against the fraud label were first examined to identify distribution differences between fraudulent (1) and non-fraudulent (0) cases, and where these differences were subtle, KDE plots were used to highlight density separations. For categorical features, bar plots were generated to confirm their predictive relevance and class distribution patterns. Finally, a correlation heatmap among all selected features demonstrated minimal multicollinearity, indicating that the variables provided complementary information without redundancy.

Methods

LightGBM

Our deployed model is LightGBM because tree ensembles are strong on heterogeneous, aggregated tabular data and scale efficiently.

We select LightGBM hyperparameters via Bayesian optimization with 5-fold stratified cross-validation, maximizing ROC–AUC across folds. We keep the optimization objective in mathematical form:

$$\lambda^* = \arg\max_{\lambda \in \Lambda} \frac{1}{5} \sum_{k=1}^{5} AUC\big(\{(p_\lambda(x), y)\}_{(x,y)\in val_k}\big),$$

where $\lambda(x)$ is the model's predicted fraud probability under hyperparameters lambda. After fitting with the best settings, we optimize the decision threshold on the precision–recall curve of the validation split to maximize F1, then report test metrics and feature importances.

ANN

Our final model adopts a Multi-Layer Perceptron (MLP) for binary fraud classification. Fraud patterns often involve complex, nonlinear interactions among features which the ANN can capture through hidden layers. Input features X are standardized, and the network outputs a predicted fraud probability through nonlinear transformations. To address class imbalance, we applied Random Oversampling within each training fold. Hyperparameters, including layer sizes, dropout rates, and learning rates, were optimized via Optuna within 5-fold stratified cross-validation, maximizing ROC–AUC across folds. After fitting with the best configuration, we evaluate the model on the held-out test set and report test metrics (ROC–AUC, precision, recall, F1, and confusion matrix). This pipeline—Standardization → Random Oversampling → 5-fold stratified training with early stopping → test evaluation—

ensures robust learning of complex fraud patterns and reliable performance assessment.

Results and Discussion

LightGBM

Evaluation Methodology

We ran multiple attempts varying imbalance handling, search strategy, and learner family. The selected system is LightGBM (GOSS) inside a pipeline with per-fold RandomOverSampler and BayesSearchCV under 5-fold stratified CV, optimizing ROC–AUC and refitting on the training split. We then evaluated on the held-out test set with a precision–recall-derived F1-optimal threshold and inspected the ROC, PR, F1-vs-threshold, confusion matrix, probability histograms, and gain-based feature importances.

Aggregate Outcomes

Among the six runs, the LGBM with BayesSearchCV and per-fold RandomOverSampler provided the most favorable trade-off between discrimination and minority retrieval. The untuned baseline (Attempt 1) underperformed with a ROC–AUC near 0.658 and an F1 near 0.160. Attempts 5 showed slightly higher ROC–AUC values (0.764), but these improvements did not convert into stronger F1 scores, indicating a less favorable precision–recall balance for our objective. Global oversampling applied before cross-validation (Attempt 4) reduced discrimination (ROC–AUC near 0.667) and yielded a lower F1 (about 0.164). A single decision-tree baseline trained on oversampled data (Attempt 6) recorded the highest nominal accuracy (about 0.914) by predicting predominantly negatives, but it performed worst on discrimination and positive-class F1 (ROC–AUC near 0.557, F1 about 0.144).

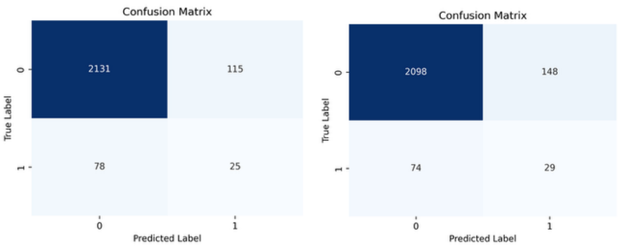

Figure 1. Confusion Matrices for the Best Two Models

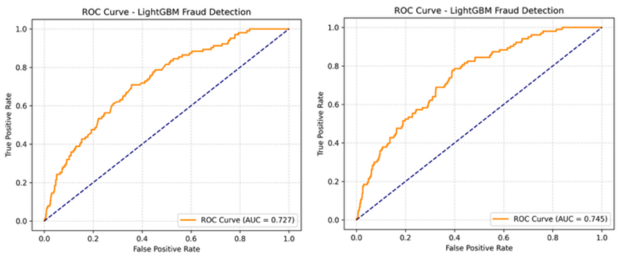

Figure 2. ROC Curve for the Best Two Models

We select the right model because it delivers a slightly improving true positives (29 vs 25) and reducing false negatives (74 vs 78). This means fewer legitimate customers are flagged and more frauds are caught. It also shows a higher ROC–AUC (~0.745 vs ~0.722), indicating better ranking quality across thresholds.

Tuning Insights

Bayesian optimization provided efficient, stable hyperparameter search under noisy 5-fold CV. We constrained depth, leaves, learning rate, and regularization to control capacity. We also enabled GOSS (Gradient-based One-Side Sampling) in LightGBM, which keeps the highest-gradient examples and a small random subset of low-gradient ones. This accelerates training and focuses learning on hard, minority cases, improving discrimination at similar compute. Oversampling was applied per fold inside CV (not globally) to avoid leakage and distributional shift, yielding better minority recall/F1 without inflating false positives. Overall, the combo of Bayesian search + GOSS + per-fold ROS + PR-based thresholding produced stable settings and higher F1 with controlled false alarms.

Model Behavior and Errors

Feature-importance analysis indicates reliance on engineered consumption-dynamics aggregates (e.g., deltas and totals) rather than simple month/year summaries. False negatives occur when usage is low-variance and regular, masking abnormal behavior; false positives cluster around abrupt usage jumps or wide ranges that can also reflect legitimate changes. The F1-optimal threshold chosen for the selected attempt balances these effects while keeping false alarms comparatively low.

ANN

Evaluation Setup

Four successive ANN experiments were conducted, varying the approach to class imbalance handling and model optimization. The first model applied SMOTE for oversampling the minority class. The second model replaced SMOTE with random oversampling. The third model combined random oversampling with Optuna-based hyperparameter tuning to optimize layer widths, dropout rates, and learning rates. The fourth model used random oversampling together with the SciKeras wrapper to integrate Keras models seamlessly with scikit-learn cross-validation and hyperparameter search.

Aggregate Outcomes

Among the four runs, Model 3, the Optuna-tuned ANN, delivered the most balanced and discriminative performance, with ROC–AUC of 0.733 and F1 of 0.189 on the test set. Random oversampling in Model 2 slightly improved recall but introduced more false positives. Model 4 (SciKeras wrapper) produced comparable performance to Model 3, indicating stable integration with scikit-learn cross-validation. Model 1 (SMOTE-only) served as the baseline. Training curves indicated effective early stopping with minimal overfitting, while validation of AUCs remained consistent across folds, suggesting stable convergence.

Tuning Insights

The experiments highlight several key insights for ANN optimization in this imbalanced fraud-detection task. Imbalance handling via random oversampling or SMOTE effectively increased minority-class representation, improving recall and F1 without severely distorting feature distributions. Hyperparameter tuning with Optuna consistently improved ROC–AUC and F1 by finding architectures that balanced expressiveness and regularization. The SciKeras wrapper allows seamless Keras–Scikit-learn integration without affecting performance. Conversely, overly shallow architectures or small networks tended to saturate early, limiting discrimination and reducing AUC, while aggressive oversampling occasionally introduced minor noise, slightly reducing precision.

Model Behavior and Errors

Permutation-based feature importance and output probability histograms revealed that the ANN captured nonlinear dependencies between total consumption dynamics and temporal dispersion features. False negatives occurred mostly among customers with consistent consumption patterns but subtle anomalies, where predicted probabilities hovered near the classification threshold. False positives clustered among profiles with abrupt consumption spikes, likely due to legitimate seasonal or operational variations rather than fraud. Model 3's smoother calibration curve and tighter probability separation suggest it generalizes best and maintains robustness under class imbalance. If operational requirements prioritize recall, the classification threshold can be adjusted downward (~0.45–0.5) to increase sensitivity at a controlled cost to precision.

Model comparison

The comparison of model performance based on F1 score shows that the **LightGBM model (0.2071)** outperformed the **Artificial Neural Network (0.189)**. While both models achieved relatively modest F1 scores, LightGBM demonstrated slightly better balance between precision and recall, indicating stronger overall classification capability for this dataset. This suggests that tree-based methods like LightGBM may be better suited to the current data structure compared to the neural network approach.

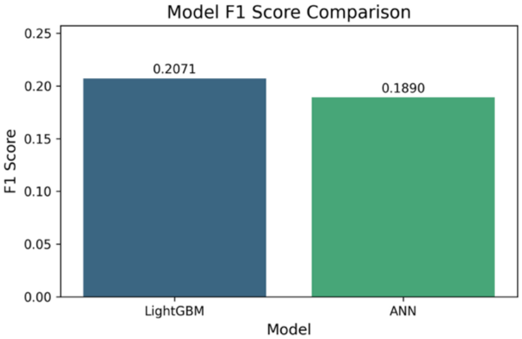

Figure 3. Model Comparison

Human comparison

While experienced analysts can detect certain fraudulent patterns, the volume and complexity of transaction data often exceed human capacity. Our models, ANN and LightGBM, efficiently process large-scale transactions, maintain consistent evaluation criteria, and identify subtle patterns that may be overlooked by human analysts. The system does not need to outperform human experts in every instance to be valuable; even complementary detection provides significant benefit in supporting fraud prevention workflows.

Societal considerations

The system processes sensitive utility records and therefore demands strong privacy controls, minimization, and strict retention policies. For example, fairness risks arise because geographical or tariff attributes can proxy socio-economic factors. With respect to labor, the tool is designed to augment rather than replace investigators by shifting workload from broad manual screening to targeted case analysis.

References

Saha, A., Logofatu, D., & Ray, J. K. (n.d.). *Enhancing fraud detection in utility consumption using neural networks: A comparative study*. Department of Computer Science and Engineering, Frankfurt University of Applied Sciences, Frankfurt am Main, Germany.

Oprea, S.-V., & Bara, A. (n.d.). *Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets*. Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, Bucharest, Romania.

Taha, A. A., & Malebary, S. J. (n.d.). *An intelligent approach to credit card fraud detection using an optimized Light Gradient Boosting Machine*. Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia.

Appendix

LightGBM Full Aggregate Outcomes:

Variant (key change)	ROC – AUC	Accuracy	Precision	Recall	F1
ANN + SMOTE	0.637	0.860	0.090	0.230	0.120
ANN + Random oversampling	**0.632**	**0.870**	**0.110**	**0.270**	**0.150**
ANN + Random oversampling + Optuna-tuned parameters	0.733	0.942	0.242	0.155	0.189
ANN + oversampling + SciKeras wrapper	0.6665	0.9042	0.1325	0.2136	0.1636
ANN + oversampling + add layers	0.7642	0.8825	0.1469	0.3495	0.2069
ANN + oversampling + Class weighting	0.5570	0.9140	0.0591	0.1651	0.1441

ANN Full Aggregate Outcomes:

Variant (key change)	**ROC–AUC**	**Accuracy**	**Precision**	**Recall**	**F1**
ANN + SMOTE	0.637	0.860	0.090	0.230	0.120
ANN + Random oversampling	0.632	0.870	0.110	0.270	0.150
ANN + Random oversampling + Optuna-tuned	0.733	0.942	0.242	0.155	0.189
ANN + Random oversampling + SciKeras	0.732	0.943	0.243	0.155	0.190