Fraud in Electricity and Gas Consumption

Group 18: Ang See Tien, Lee Hui En, Lim Ze Yan, Loke Ann Chi

Introduction

Fraudulent consumption of electricity and gas remains a critical challenge for utility providers, resulting in substantial financial losses and operational inefficiencies. Traditional rule-based detection systems often fail to adapt to evolving and complex fraudulent behaviours, highlighting the need for more intelligent, data-driven approaches. Machine Learning (ML) offers an effective framework by learning consumption patterns from historical data to identify anomalous or suspicious usage in near-real-time.

This project aims to develop a robust fraud detection system using client and invoice data to identify patterns indicative of fraudulent activity. Given the severe class imbalance inherent in utility fraud datasets, where fraudulent cases typically represent a small fraction of transactions, we investigate approaches to handling imbalanced data. Four ML models are explored: Logistic Regression (LR) as an interpretable baseline; Light Gradient Boosting Machine (LightGBM) and Random Forest (RF) for capturing complex, non-linear relationships; and Support Vector Machine (SVM) for effectively handling high-dimensional classification tasks.

Related Works

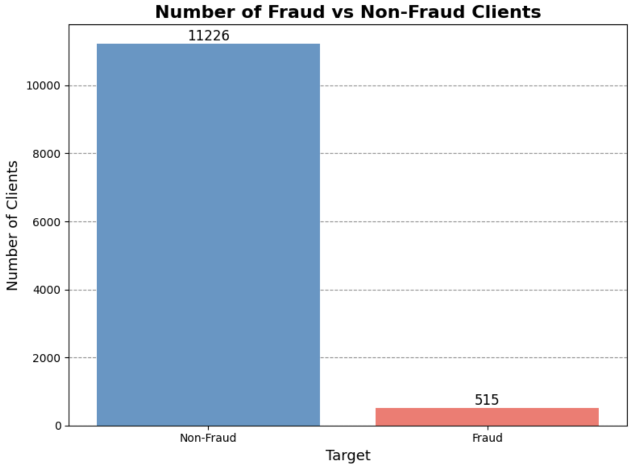
Recent studies on non-technical loss (NTL) detection have explored various machine learning approaches, from classical classifiers to modern ensemble and deep learning techniques.

Proper experimental design is critical for valid performance estimates. Oprea and Bâra (2021) applied feature selection and normalisation to the entire dataset before splitting into train-test sets, a practice that can introduce test-to-train leakage by artificially inflating performance metrics, highlighted by Sasse et al. (2025). This motivates our strict adherence to post-split preprocessing, where all feature engineering and normalisation occur only on training data.

Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) has become a popular approach to address class imbalance in fraud and NTL detection, with recent studies such as Ayub et al. (2022) and Sundaravadivel et al. (2025) routinely applying SMOTE to balance fraud-related datasets. However, there is evidence that in high-dimensional feature spaces, SMOTE may not improve classifier bias toward the majority class, and the synthetic minority samples may not properly reflect the true minority class distribution — potentially introducing noise or over-fitting the model to artefacts of the synthetic generation rather than genuine minority behaviour (Blagus and Lusa 2023). This motivates our decision to compare model performance with other balancing methods in addition to SMOTE and to critically evaluate whether oversampling via SMOTE truly helps in our utility fraud detection task.

Dataset

The dataset itself contains two files: “client.csv” and “invoice.csv”. The client dataset contains records of electricity consumers, serving as the core reference table that links each customer to their corresponding invoice data. It contains six columns, including identifiers such as *id* (unique client identifier), *dis* (district), *region*, *catg* (category) and the *target* variable, which indicates whether the client is involved in electricity fraud (1) or not (0). The invoice dataset stores the electricity consumption and billing information for each client, including variables such as *tariff_type*, *counter_status*, *reading_remarque*, *counter_type* and consumption-related values. After preprocessing, feature engineering and merging, the final dataset contained 11741 samples, comprising 515 fraud and 11226 non-fraud cases, with a total of 33 features, of which 27 are numerical and 5 are categorical, spanning from 1977 to 2019.



Data Quality Issues

Several preprocessing challenges were identified prior to model development. First, both datasets contained date columns stored as strings, which require conversion to datetime format for accurate time-based computations. Second, since each client is associated with multiple invoices (ranging from 5 to 30), the datasets could not be directly joined one-to-one. Instead, invoice-level features were aggregated to derive meaningful client-level statistics, including averages, minima, maxima and standard deviations. Third, several variables, such as *dis*, *catg* and *region*, were originally represented numerically but are actually categorical variables. Hence, they have to be converted to categorical types to prevent misinterpretation as continuous numeric values. However, machine learning models can only work with numerical values so they have to be subsequently encoded. Finally, the target variable exhibited severe class imbalance, with fraud cases forming a very small percentage of the dataset, which could bias model training toward predicting the majority (non-fraud) cases without appropriate correction.

Data Preprocessing

Before analysis, we performed a missing value check using the "isna().sum()" function and confirmed that there were no missing values in either dataset. The date columns were then converted to datetime format for accurate time-based feature computation.

Feature Extraction

Next, we aggregated invoice-level features into client-level summaries by computing various consumption statistics (mean, minimum, maximum, standard deviation, sum, coefficient of variation, maximum-minimum ratio, per-month average consumption pivot). During our research, we observed that irregular month-to-month consumption swings often precede confirmed fraud cases, even when the annual totals appear normal (Nagi et al. 2011). Hence, we derived volatility-related features, such as monthly standard deviation, average consumption change and max/min consumption, to help identify abnormal or inconsistent usage patterns potentially linked to fraud.

Feature Selection

For numerical features, we applied feature selection using the Fisher score, which measures how well each feature separates the two classes by comparing inter-class variance to intra-class variance (GeeksforGeeks 2025). Only the top-ranked features were retained to reduce redundancy and improve model interpretability. For categorical inputs, we applied the chi-square test to evaluate the independence between features and the target variable. Features with p-values greater than or equal to 0.05 were considered independent of the target and thus excluded from further analysis.

Encoding

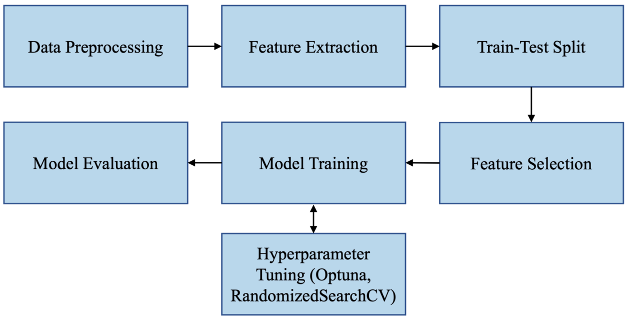
We then performed encoding of categorical variables. Based on the number of unique categorical features, *region* is a high-cardinality feature, *mode_tarif_type* is a medium-cardinality feature and the remaining features (*dis*, *catg*, *mode_reading_remarque*) are low-cardinality features (GeeksforGeeks 2025). Therefore, target encoding was applied to *region* and *mode_tarif_type*, while one-hot encoding was used for *dis*, *catg*, and *mode_reading_remarque*.

Feature Scaling

Finally, we prepare feature scaling to all numerical inputs. For our models, both Logistic Regression and Support Vector Machine (SVM) require scaling as SVM relies on distances between data points, while Logistic Regression relies on Gradient Descent Optimisation, which converges faster with scaled features (Jang 2024). We used StandardScaler, which follows Standard Normal Distribution (SND) and standardises features to follow a standard normal distribution by transforming the data so that the mean becomes 0 and the standard deviation becomes 1. This is ideal for our Logistic Regression and Support Vector Machine models as they assume that the data is normally distributed (GeeksforGeeks 2025).

Methods

Overall Pipeline



Logistic Regression

Logistic Regression is a supervised machine learning and statistical classification model that estimates the probability of a binary outcome — such as whether a client's electricity usage is fraudulent or not — based on input variables. It models the relationship between predictor variables and the likelihood of an event occurring using the logistic (sigmoid) function (IBM n.d.). It is simple to implement and highly interpretable, making it easy to understand how each feature, such as consumption patterns, contributes to the potential electricity theft or anomalies. It also performs effectively on structured datasets, which are common in smart meter or utility billing systems. Moreover, Logistic Regression serves as a strong baseline model, providing a reliable starting point for evaluating performance before exploring more complex machine learning approaches (Riswanto 2025).

Four Logistic Regression model variants were built to compare between the baseline Logistic Regression with three sampling methods designed to address dataset imbalance, including SMOTE-based oversampling, Random Undersampling and a Hybrid method that combines SMOTE and Edited Nearest Neighbours (ENN).

Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework that uses histogram-based decision tree learning algorithms. In each iteration, it learns by fitting the residual errors, effectively correcting the errors made by previous trees. LightGBM uses two key techniques: Gradient-Based One-Side Sampling and Exclusive Feature Bundling. Together, these techniques allow LightGBM to achieve faster computation, lower memory consumption and similar (or higher) accuracy as compared to other gradient boosting decision trees (Ke et al. 2017). LightGBM also has native support for categorical features and can handle class imbalance effectively, making it particularly well-suited for fraud detection on large, high-dimensional and imbalanced datasets.

Three LightGBM models were built to compare different approaches for handling categorical features and class imbalance. One model used LightGBM's native support for categorical features along with its internal class balancing via "scale_pos_weight", while the other two used alternative methods, such as external encoding or SMOTE-based oversampling.

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees trained on random subsets of data and features to reduce noise and solve overfit problems (Soon et al. 2024). Its ability to capture nonlinear relationships and feature interactions makes it suitable for fraud detection, where patterns are often complex and imbalanced. The model also provides feature importance scores, enhancing interpretability by identifying key predictors of fraud. Hyperparameters such as the number of trees, depth and split criteria were optimised using Optuna and RandomizedSearchCV to ensure robust performance. This approach was chosen for its balance between predictive accuracy, interpretability and resistance to overfitting.

Four different Random Forest model variants were developed using two hyperparameter optimisation methods — RandomizedSearchCV and Optuna — each tested with and without SMOTE-based oversampling. This setup enabled comparison of how tuning and resampling techniques affect the model's ability to detect fraudulent clients in an imbalanced dataset.

Support Vector Machine (SVM)

Among various machine learning algorithms, Support Vector Machines (SVMs) have gained popularity due to their strong theoretical foundations and reliable generalisation performance (Xia, J. 2022). SVMs are particularly effective for binary classification, making them suitable for distinguishing fraudulent from non-fraudulent cases in this study. Following the methodology of Nagi et al. (2010) and Abro et al. (2024), who found that the Radial Basis Function (RBF) kernel was the best performing kernel, we evaluate SVMs using the RBF kernel and additionally explore the linear kernel. The linear kernel offers lower computational cost, while the RBF kernel captures non-linear relationships in the data, which may help identify subtle patterns and improve predictive accuracy.

Four SVM models were developed to evaluate the effects of kernel type and class-balancing strategy on fraud detection performance: (1) linear without SMOTE, (2) linear with SMOTE, (3) RBF without SMOTE, and (4) RBF with SMOTE. This setup enables comparison of how kernel choice influences pattern recognition and how SMOTE oversampling affects performance on imbalanced fraudulent and non-fraudulent data.

Training and Evaluation

The dataset was split into 80% training and 20% testing sets. Within the training set, 5-fold cross-validation was applied across all models to ensure robust performance estimates and mitigate overfitting.

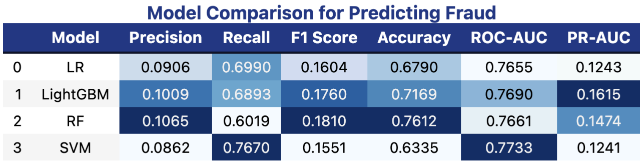
Hyperparameter Tuning

Across all models, hyperparameter tuning was conducted to optimise model performance and ensure consistency across experiments. We primarily used Optuna for automated optimisation, which efficiently searches the hyperparameter space by iteratively updating parameter values based on previous trial results (Akiba et al. 2019). This approach balances exploration and exploitation, achieving faster convergence and improved performance compared to traditional grid or random search methods. For the Random Forest model, we implemented RandomizedSearchCV, which randomly samples from predefined parameter distributions and evaluates performance using ROC-AUC (Scikit-learn 2025). This allowed us to cross-validate and benchmark both tuning methods. The key hyperparameters tuned included the number of estimators, maximum depth, minimum samples per split and leaf, and feature selection strategy. All models were evaluated under identical data splits, random seeds and performance metrics to ensure reproducibility and fairness.

Results and Discussions

We implemented the four models mentioned above to evaluate their effectiveness in detecting fraud clients, training and validating each using identical feature sets and stratified 5-fold cross-validation to ensure a fair comparison. Model performance was assessed using key classification metrics including Precision, Recall, F1-Score, Accuracy, ROC-AUC and PR-AUC.

For each algorithm, the results reflect the best-performing hyperparameter configuration identified during tuning. The base Logistic Regression, LightGBM with external encoding, Optuna-tuned Random Forest and RBF-Kernel SVM variants, all without SMOTE, achieved the highest scores within their respective model families. The results of all models are shown in the table below.

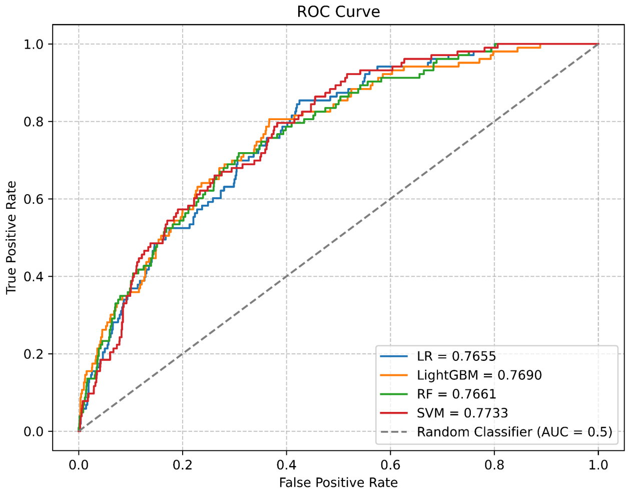


Model Comparison for Predicting Fraud

	Model	Precision	Recall	F1 Score	Accuracy	ROC-AUC	PR-AUC
0	LR	0.0906	0.6990	0.1604	0.6790	0.7655	0.1243
1	LightGBM	0.1009	0.6893	0.1760	0.7169	0.7690	0.1615
2	RF	0.1065	0.6019	0.1810	0.7612	0.7661	0.1474
3	SVM	0.0862	0.7670	0.1551	0.6335	0.7733	0.1241

Ensemble-based models, particularly LightGBM and Random Forest, demonstrated the strongest performance in identifying fraudulent clients from their consumption patterns. Both models effectively captured nonlinear consumption patterns that distinguish genuine users from those with irregular or suspicious behaviour. This aligns with findings that tree-based ensembles handle complex feature interactions and imbalance datasets more efficiently (Soon et al. 2024). LightGBM was ultimately selected as the better model due to its higher ROC-AUC score (0.7690), indicating greater reliability for evaluating model performance on imbalanced datasets (Bhat 2024).

In contrast, Logistic Regression and SVM achieved moderate results. Their comparatively lower recall and F1-scores suggest that these models were less effective at identifying rare fraudulent cases, likely due to their reliance on linear or kernel-based decision boundaries that fail to capture higher-order feature dependencies. Despite this, their stable accuracy indicates that they remain reliable baselines for comparison and interpretability.



AI Against Human Performance

Although the models do not yet surpass humans in contextual reasoning, they can rapidly identify suspicious consumption patterns that humans might overlook in a large dataset, significantly enhancing detection efficiency. In practice, manual fraud detection relies heavily on experience and intuition, which can introduce subjective bias and inconsistency across investigators. By contrast, the LightGBM model achieved high ROC-AUC values, demonstrating reliable discrimination of fraudulent clients under different decision thresholds — a level of quantitative consistency difficult for humans to maintain.

While human analysts remain essential for interpreting nuanced or ambiguous cases, the models serve as a powerful tool to automate the initial screening process. This hybrid approach leverages the efficiency and objectivity of AI with contextual judgement of human experts, resulting in a more balanced and scalable fraud detection system that enhances productivity without displacing human roles.

Societal Impacts

Since the dataset only contained anonymous client IDs and consumption-related attributes, privacy concerns were minimal throughout this project. No personal, financial or demographic information was available, ensuring that model training and evaluation were performed on purely behavioural data. This design also supports fairness, as the models learned patterns from usage behaviour rather than individual characteristics that could introduce bias.

Conclusion

Machine learning can be used to detect fraudulent electricity consumption more efficiently and fairly. By carefully preparing the data and evaluating different models, we built a consistent and transparent approach to comparing their performance. The study also highlights how AI systems can support human decision-making rather than replace it. Overall, our work demonstrates that combining data-driven insights with human expertise can make fraud detection both smarter and more responsible.

References

Abro, S. A.; Hua, L. G.; Laghari, J. A.; Bhayo, M. A.; and Memon, A. A. 2024. Machine learning-based electricity theft detection using support vector machines. International Journal of Electrical and Computer Engineering (IJECE), 14(2):1240–1250. https://doi.org/10.11591/ijece.v14i2.pp1240-1250

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Ayub, N.; Ali, U.; Mustafa, K.; Mohsin, S. M.; and Aslam, S. 2022. Predictive data analytics for electricity fraud detection using tuned CNN ensembler in smart grid. Forecasting, 4(4):936–948. https://doi.org/10.3390/forecast4040051

Bhat, A. 2024. Day 14: Evaluation Metrics for Classification — Precision, Recall, F1-Score, ROC-AUC. Medium. https://medium.com/@bhatadithya54764118/day-14-evaluation-metrics-for-classification-precision-recall-f1-score-roc-auc-7d653695ab52

Blagus, R.; and Lusa, L. 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14:106. https://doi.org/10.1186/1471-2105-14-106

Breiman, L. 2001. Random Forests. Machine Learning 45(1): 5–32. doi.org/10.1023/A:1010933404324

GeeksforGeeks. 2025. Fisher Score for Feature Selection. https://www.geeksforgeeks.org/machine-learning/fisher-score-for-feature-selection/

GeeksforGeeks. 2025. How to fit categorical data types for random forest classification? https://www.geeksforgeeks.org/machine-learning/how-to-fit-categorical-data-types-for-random-forest-classification/

GeeksforGeeks. 2025. StandardScaler, MinMaxScaler and RobustScaler techniques – ML. https://www.geeksforgeeks.org/machine-learning/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

IBM. (n.d.). What is logistic regression? IBM Think. https://www.ibm.com/think/topics/logistic-regression

IBM. (n.d.). What Is Random Forest? IBM Think. https://www.ibm.com/think/topics/random-forest

Jain, A. 2025. Optuna vs GridSearchCV vs RandomSearchCV: Hyperparameter Tuning Techniques. Medium. https://medium.com/@abhishekjainindore24/optuna-vs-gridsearchcv-vs-randomsearchcv-hyperparameter-tuning-techniques-ea8e2ada28d0

Jang, D. 2024. Feature scaling for support vector machines: Unlocking the power of optimized machine learning. Medium. https://medium.com/@jangdaehan1/feature-scaling-for-support-vector-machines-unlocking-the-power-of-optimized-machine-learning-39a9a9024b40

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 3149-3157. Red Hook, NY: Curran Associates Inc.

Li, J. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. arXiv preprint. arXiv:1907.10902.

Nagi, J.; Yap, K. S.; Tiong, S. K.; Ahmed, S. K.; and Mohamad, M. 2010. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. IEEE Transactions on Power Delivery, 25(2):1162-1171. https://doi.org/10.1109/TPWRD.2009.2030890

Oprea, S.-V.; and Bâra, A. 2021. Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets. Computers and Electrical Engineering, 94:107329. https://doi.org/10.1016/j.compeleceng.2021.107329

Riswanto, U. 2025. Building a fraud detection model using logistic regression in R. Medium. https://ujangriswanto08.medium.com/building-a-fraud-detection-model-using-logistic-regression-in-r-0917e2d46b6d

Sasse, L.; Nicolaisen-Sobesky, E.; Dukart, J.; Eickhoff, S. B.; Götz, M.; Hamdan, S.; Komeyer, V.; Kulkarni, A.; Lahnakoski, J. M.; Love, B. C.; Raimondo, F.; and Patil, K. R. 2025. Overview of leakage scenarios in supervised machine learning. Journal of Big Data, 12:135. https://doi.org/10.1186/s40537-025-01087-5

Scikit-learn. 2025. RandomizedSearchCV. scikit-learn Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Soon, H. F.; Amir, A.; Nishizaki, H.; Zahri, N. A. H.; Munirah, L. M.; and Azemi, S. N. N. 2024. Evaluating Tree-Based Ensemble Strategies for Imbalanced Network Attack Classification. International Journal of Advanced Computer Science and Applications (IJACSA) 15(1): 111. https://thesai.org/Downloads/Volume15No1/Paper_111-Evaluating_Tree_based_Ensemble_Strategies_for_Imbalanced_Network.pdf

Xia, J. 2022. Credit Card Fraud Detection Based on Support Vector Machine. Highlights in Science, Engineering and Technology, 23:93–97. https://doi.org/10.54097/hset.v23i.3202