Application of Machine Learning Models to Detect Fraudulent Energy Use

Austin Loh, Celine Liew, Darryl Chan, Lydia Tan

National University of Singapore
e0275395@u.nus.edu, e0775106@u.nus.edu, e0257193@u.nus.edu, e0774533@u.nus.edu

I. Introduction

Fraud has been a pressing issue in many industries. In the energy sector, the industry has suffered a whopping 59.8% increase in median financial loss in fraud within four years, from 2016 to 2020 (Jendruszak 2022). Additionally, the industry loses $96 billion annually to fraud (Northeast Group 2017). To remain profitable, the losses are ultimately passed to legitimate consumers by increased energy prices, thereby going against the customers' interests (Management Solutions 2017). Hence, detecting fraud is important to eradicate this problem.

Fraudulent energy use can happen in two ways: by bypassing the energy meter or by directly tapping into the electricity/gas lines. While some energy losses do happen naturally (called Non-Technical Losses (NTLs)), unusually high NTLs can be an indicator of fraud.

Currently, with energy utility companies undergoing a digital transformation, they have access to vast amounts of customer behavioural data. The data can then be used in various machine learning models to detect fraud.

With extensive research and work with different machine learning models already done in this field with varying degrees of success, this report will focus on evaluating the recall rates of each machine learning model to measure success. Thus, this project is relevant to this module as we are uncovering the most effective model in solving a universal problem.

II. Related Works

For fraud detection in energy consumption, there have been many works published. Dr. Depuru used Support Vector Machines, a machine learning model, to predict energy fraud with 98.4% accuracy for 220 customers (Depuru et al 2011). Dr. Monedero used neural networks and achieved a 50% True positive rate with 13 cases detected as fraud (Monedero et al 2006). Dr. Cody created simple machine learning models with decision trees to detect fraud (Cody et al 2015).

Additionally, we found 2 groups of individuals who have tried to detect fraud in energy consumption with datasets similar to ours. The first group used gradient boosting, a machine learning model, and achieved a 0.885 on its ROC test (Tsepa and Samoshyn 2020). The second group used gradient boosting as well and achieved 0.970 Accuracy (Samaha 2022).

However, both groups did not manage to achieve a high rate of recall - the rate of actual fraud detected - which is a crucial metric in evaluating a model's success.

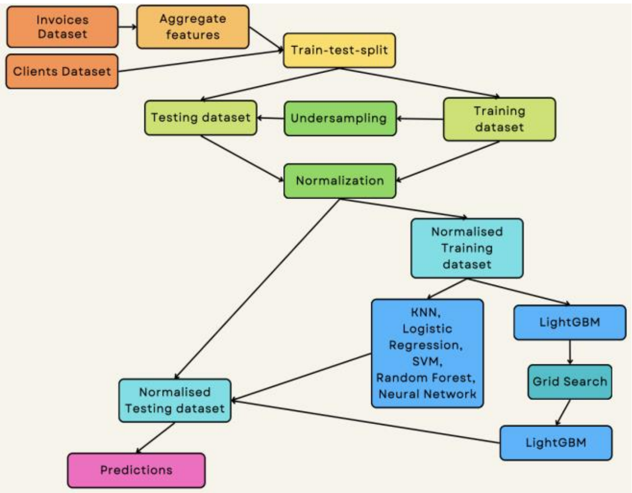
III. Dataset

The dataset we used, the Fraud in Electricity and Gas Consumption dataset consisted of over 135,000 unique customer data. The data was split into two files: invoice and client files. The former had many missing invoices whereas the latter was highly unbalanced, with around 127,000 legitimate to 8,000 fraudulent customers. While this imbalance does not affect the overall accuracy of the models, it affects the recall rate as there is a higher incentive for models to predict customers as legitimate when they are not (Chawla. N. V 2002).

Pre-Processing: Clients who had no invoices were removed, whereas clients with multiple invoices were aggregated by taking the mean of all of their invoices. Afterwhich, the two datasets were combined. For feature selection, we removed the dates of the client's invoices and their ID as they have minimal impact on the outcome of the prediction. Subsequently, we did a random undersample of the legitimate customers, randomly removing some legitimate client data for the model to create balanced data and improve the model's recall rate. We then added the removed data to our test dataset. Finally, as the consumption level had outliers, we normalized their columns to garner equal weights in the model.

IV. Methods

After pre-processing our data, we trained our models with the normalized training dataset as seen in the workflow below.



We viewed the problem as a supervised learning task as the clients dataset included labels on whether a client has committed fraud. As there were only 2 possible labels, we modelled our problem to be a binary classification problem. Hence, we considered several classification models to tackle our problem, namely K-Nearest Neighbours (KNN), Logistic Regression, Support Vector Machines (SVM), Random Forest, LightGBM and neural networks.

We first implemented the base models to test their suitability for our problem. As KNN and Logistic Regression have been discussed in class, we will not delve further into them.

SVM mainly takes the points in the training dataset and outputs a hyperplane that best separates the 2 classes. Points on 1 side of the hyperplane are predicted as 1 class, and vice versa. An optimal hyperplane is one that maximises the distance of each point to the line.

Decision trees can be considered as a combination of multiple if-else statements. Deciding on which feature and value to split on in each node is based on information gained.

Random forest uses multiple decision trees and uses an ensemble method called bagging to combine the outputs of the multiple trees. By splitting the training dataset, each decision tree is trained on a different subset of the training dataset. It then combines the output of all the decision trees using majority voting to get the final prediction.

LightGBM also uses decision trees but utilises gradient boosting instead. Gradient boosting combines the decision trees sequentially and tries to minimise the overall prediction error. LightGBM then uses a histogram-based method, where data is bucketed using a histogram of the distribution. The bins are then used to calculate the gain and split the data (ArcGIS Pro 3.0 2022).

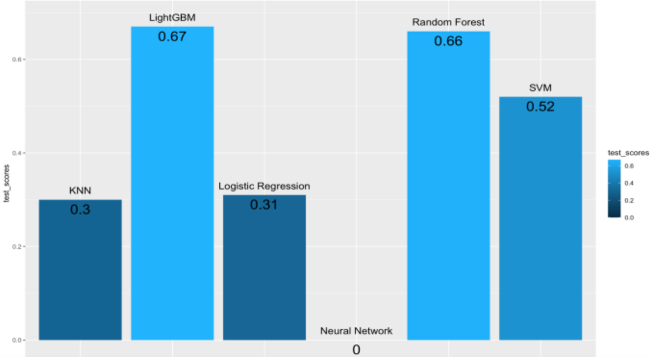
For our dense neural network, we used 3 hidden layers, with ReLU as the activation function for the hidden layers and SoftMax for the output layer.

Testing of all models is done by predicting the class labels with the normalized testing dataset and then comparing the results with the actual class labels.

V. Results

Our group decided that using the recall rate to measure the success of our model is more appropriate as detecting fraud would be more important for companies. While using other measures like Area under Curve or harmonic mean might be a more holistic test, due to limited time we instead focused more on the crucial recall rate.

The recall rates for the base models are shown below.



Particularly, we focused on the LightGBM, Random Forest and SVM due to their relatively high recall score. Due to the need to perform undersampling, cross-validation (CV) would not work. Hence, we split the dataset differently each time and then performed undersampling. By performing our modified 5-fold CV, our Histogram-based Gradient Boosting Classification Tree model (sklearn implementation of LightGBM) showed the best average recall score of 0.66. Actual recall score for each iteration is shown in Appendix A.

We further fine-tuned the hyperparameters of our LightGBM model by running grid search on the hyperparameters. We found that the optimal hyperparameters were 0.12 for learning rate, 22 for max depth, 32 for max leaf nodes, and 47 for min samples leaf as shown in Appendix B.

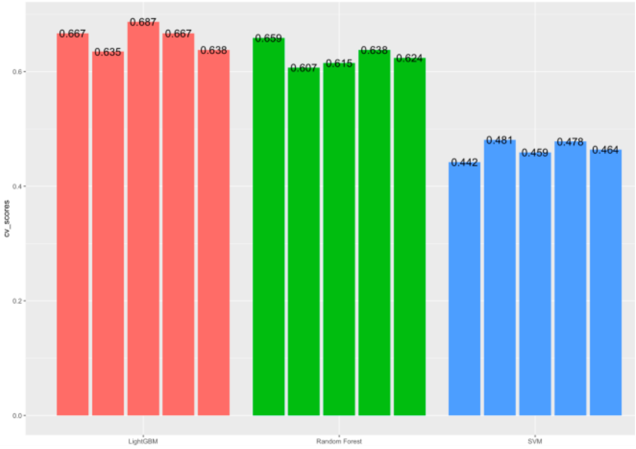
Our tuned model was finally able to reach a recall rate of 0.68, better than the 0.575 (Samaha 2022) and 0.03 (Tsepa and Samoshyn 2020) recall rates achieved by others. This could be due to the recall rates not being optimised for in their models and a slight difference in datasets used.

VI. Conclusion

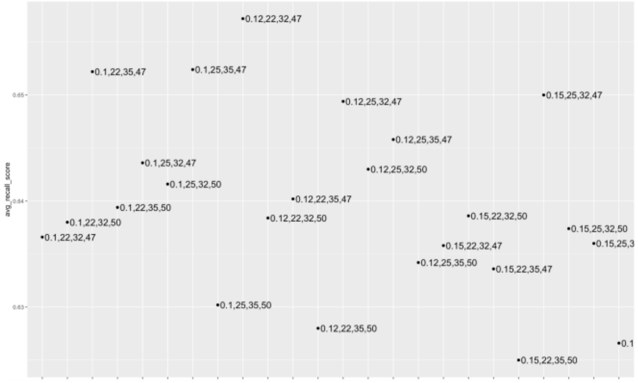
Initially, we believed that the model would perform better if we included more data. However, as there were 15x more legitimate than fraudulent customers, the model had a higher chance of predicting clients as legitimate. This caused the model to incorrectly predict many fraudulent customers as legitimate, which severely affected the recall rate.

To address this, we randomly undersampled the legitimate customers. As a result, fewer legitimate customers were included in the model, creating a more balanced training dataset which greatly improved our previously low recall rate.

VII. Appendix



A: Recall score for modified 5-fold CV for LightGBM, Random Forest and SVM models.



B: Recall scores for running Grid Search for hyperparameter tuning for LightGBM model. (Labels are arranged in following order: Learning Rate, Max Depth, Max Leaf Nodes, Min Samples Leaf)

VIII. References

ArcGIS Pro 3.0 (2022). How LightGBM algorithm works. Retrieved October 26, 2022 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-lightgbm-works.htm

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cody, C., Ford, V., & Siraj, A. (2015). Decision tree learning for fraud detection in consumer energy consumption. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2015.80

Depuru, S. S., Wang, L., & Devabhaktuni, V. (2011). Support Vector Machine based data classification for detection of electricity theft. *2011 IEEE/PES Power Systems Conference and Exposition*. https://doi.org/10.1109/psce.2011.5772466

Electricity Theft and Non-Technical Losses: Global Markets, Solutions, and Vendors. Northeast Group, LLC (2017)

Jendruszak, B. (2022, May 7). *Industry Fraud Index: Which Industries Are Most at Risk?* SEON. Retrieved October 27, 2022, from https://seon.io/resources/industry-fraud-index/#h-loss-to-fraud-increased-the-most

Monedero, Í., Biscarri, F., León, C., Biscarri, J., & Millán, R. (2006). MIDAS: Detection of non-technical losses in electrical consumption using neural networks and statistical techniques. *Computational Science and Its Applications - ICCSA 2006*, 725–734. https://doi.org/10.1007/11751649_80

Samaha, K., (2022). LightGbm | Fraud Detection in ELEC and GAZ. Retrieved October 24, 2022 from https://www.kaggle.com/code/khsamaha/lightgbm-fraud-detection-in-elec-and-gaz/data

Smith, T. B. (2004). Electricity theft: A comparative analysis. *Energy Policy*, *32*(18), 2067–2076. https://doi.org/10.1016/s0301-4215(03)00182-4

Tsepa, A,. & Samoshyn, A., (2020). 4th place in "Fraud Detection" from Zindi. Retrieved October 24, 2022 from https://www.kaggle.com/code/imgremlin/4th-place-in-fraud-detection-from-zindi/notebook