11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara Malaysia
19th September 2024

Machine learning for anomaly detection in money services business outlets using data by geolocation

Vincent Lee Wai Seng[1]; Shariff Abu Bakar Sarip Abidinsa[2]

[1] Bank Negara Malaysia, Kuala Lumpur, vincentlee@bmn.gov.my
[2] Bank Negara Malaysia, Kuala Lumpur, shariffbakar@bnm.gov.my

Abstract:
Since 2017, licensed money services business (MSB) operators in Malaysia report transactional data to the Central Bank of Malaysia on a monthly basis. The data allows supervisors to conduct off-site monitoring on the MSB industry; however, due to the increasing size of data and large population of the operators, supervisors face resource challenges to timely identify higher risk patterns, especially at the outlet level of the MSB. The paper proposes a weakly-supervised machine learning approach to detect anomalies in the MSB outlets on a periodic basis by combining transactional data with outlet information, including geolocation data. The test results highlight the benefits of machine learning techniques in facilitating supervisors to focus their resources on MSB outlets with abnormal behaviours in a targeted location.

Keywords:
suptech, money services business transactional data, money-laundering and terrorism financing risk, weakly-supervised, SHAP values

1. Introduction:
Regulators worldwide are increasingly adopting supervisory technology (suptech[1]). By the end of 2023, 54% of financial authorities in developing economies used suptech applications, up from 31% in 2022 (CSL, 2023). One main focus of suptech applications is relating to the supervision of anti-money laundering, counter-terrorism financing and counter-proliferation financing (AML/CFT/CPF).

In Malaysia, the Central Bank oversees the AML/CFT/CPF regime, including financial intelligence, investigation of relevant predicate offences, and supervision of reporting institutions, which includes the money services business (MSB[2]) industry. As at 1Q 2024, there are more than 250 licensed MSBs under the central bank's purview, with more than 80% of them are licensed to conduct currency exchange business at physical outlets. Due to the cash intensive and cross-border nature of MSB activities, the risk exposure of MSB industry stems from money-laundering and terrorism financing (ML/TF) activities. Prior to 2017, central bank supervisors of the MSB sector focused on checklist

[1] CSL (2023) defines suptech as encompassing application of technology and data analytics tools to enhance capability of financial regulator or supervisor to provide oversight on financial industry.
[2] Under the Money Services Business Act 2011, MSB refers to any or all of the following businesses: money-changing business, remittance business, wholesale currency business.

examination approaches to identify and monitor ML/TF risks in the industry. Supervisors are also limited by the high-level aggregated data collected from regulatees, manpower resources relative to the number of MSBs, and outdated risk profiling and assessments.

To address the limitations, a data analytics unit became operational in 2017 to collect transactional data from the licensed MSBs and develop suptech applications. There are three levels of our suptech applications, namely the industry, entity and customer levels (BNM, 2019). Entity level is a key focus as it assists supervisors to identify irregular patterns among the MSBs for the purpose ML/TF risk profiling and calibration of supervisory activities. Based on our engagement with supervisors, a proactive and regular monitoring on hundreds of MSB outlets nationwide is a main challenge due to limited resources. Additionally, differing approaches to peer comparison among MSB outlets in the same area led to inconsistent analysis.

This paper proposes a weakly-supervised machine learning approach to flag out irregular patterns in the MSB physical outlets, using the transactional data with location information. Carcillo et al. (2021) presented the benefits of integrating unsupervised and supervised techniques, improving the accuracy of credit card fraud detection. Barbariol and Susto (2022) also analyses how weak supervision approach can improve the existing anomaly detection algorithm, such as Isolation Forest. Jiang et al. (2023) describes various algorithms that focus on weak supervision due to high cost of annotation, which is also the challenge for this paper.

2. Methodology:

Our weakly-supervised machine learning framework integrates both unsupervised and supervised learning. The framework starts with an unsupervised machine learning on the unlabelled panel MSB dataset, using the Isolation Forest (IF) model for anomaly detection on multiple features generated in the dataset. The anomalous observations churned out by the IF model are then used as training labels for our supervised machine learning. F1 score[3] is used to assess the supervised machine learning models due to an imbalanced distribution of labelled training dataset, of which anomalous labels are of minority population.

A. Data description

The transactional data used in this framework is reported by the licensed MSBs to the Central Bank of Malaysia on a monthly basis, as well as the information relating to the outlet profile of the MSB licensee, including geolocation. The transactional data covers three types of MSB activities: currency exchange business, remittance business, and wholesale currency business. For this paper, the focus is only on outlets that provide currency exchange services as they are more cash-based and conducted at physical outlets as well as the demand is based on geographical factor. For example, MSB outlets locating close to the border tend to conduct more exchange transactions based on the currency of the bordering country.

The currency exchange transactional data contains granular information on each transaction, which includes customer profile information and transaction details. The data also captures where the exchange transaction was conducted, which is the outlet information. Each MSB outlet is uniquely assigned with a unique internally generated 12-digit identifier code, which provides information whether the outlet is the headquarters or

[3] The harmonic means of precision and recall.

branch office of the MSB licensee as well as whether the outlet is a MSB agent to a MSB principal licensee[4].

The transactional data can then be aggregated into a panel dataset based on the MSB outlet identifier code. A time-series dimension is added to the aggregation of the currency exchange transactional data by outlet location, where the data is compiled on a quarterly basis. The time-series dimension accounts for seasonal factors and allows supervisors to monitor MSB outlets regularly.

The aggregated panel dataset comprises of 5,395 summarised observations with 30 features (25 numerical, 5 categorical) derived from millions of currency exchange transactional data. Due to confidentiality concerns, we are unable to disclose all the specific information regarding the features. However, the dataset includes quarterly summary of each MSB outlet's transactions, customers' demographics, and geolocation. As we observed unusual transaction trends in 2020-2021, impacted by the movement control order arising from the COVID-19 pandemic, the data period considered in this project is from the first quarter of 2022 up to the first quarter of 2023, with 2022 data used as training data (n = 4,316) and 2023 data (n = 1,079) used for model test purposes.

B. Unsupervised machine learning (UML)

The benefit of applying a UML approach is that no training labels are required. There are many unsupervised algorithms to consider but there are limitations on some algorithms for the purpose of anomaly detection. For example, K-clustering techniques are sensitive to noise and outliers and more suitable for data with less features. K-nearest neighbours and Kernel Density Estimation algorithms require more computational resources for datasets with higher dimensions. Steinbuss and Bohm (2021) compared few algorithms across multiple fully real data and assessed that Isolation Forest (IF) performed well overall. IF handles outliers and scales of variables well, needing fewer computational resources, but may have biases depending on how the branching cuts are performed.

To reduce the biases to certain features, the IF is applied to identify anomalies from different subsets of the data. First, the features of the dataset are split into three main categories of data subsets: customer, transaction and location. These three categories are selected based on supervisors' own experience in identifying abnormal currency exchange outlet. For customer data category, some example features are percentage of customers identified as higher risk customer type and higher risk nationality. For transaction data category, the main features include transaction value and volume. For location data category, number of peers MSB outlets within 200- and 500-meter radius are some examples of the features.

C. Supervised machine learning (SML)

Currency exchange outlets are labeled as anomalies if the IF model identifies them as such in at least one of the three data categories. Given a 5% contamination rate, an imbalance class of labels is expected.

To fix label imbalances, the paper used Synthetic Minority Over-sampling Technique (SMOTE) – which creates synthetic samples and increases the overall sample size by 57%. In addition to SMOTE, the categorical variables were one-hot encoded. This

[4] In Malaysia, principal licensee refers to a MSB licensee which has obtained the written approval of BNM to appoint MSB agent under section 43 of the MSB Act 2011. Principal licensee is also required to report all MSB transactions conducted by its agents to BNM.

is to ensure data and model compatibility for processing and model performance. Then, as part of the model evaluation process, 80% of the data are then randomly assigned as the training dataset, and the remaining observations to the test dataset. The labelled data are then used as part of training for a classification task using models listed in Table 2 in the Result section.

D. Model explainability

One challenge of applying machine learning models is that it can be difficult for supervisors to understand the model predictions as well as the features that contribute more to the prediction of anomalous outlet. To provide a deeper understanding of the predictions made by our machine learning models, we employ Shapley Additive exPlanations (SHAP). SHAP explains any machine learning model by assigning importance to each feature, enabling both global and local model interpretation.

3. Result:

The isolation forest model produces anomaly scores for each observation where scores approaching 1 has high likelihood of being anomalies and scores close to 0 has low likelihood of being anomalies. As noted in Table 1, 598 out of 4,316 training observations were outliers. About 1% of the outliers were flagged across 3 data categories, indicating that these 1% of currency exchange business outlets behave significantly different to their counterparts. 18% of outliers were abnormal based on 2 categories, while 81% of the outliers were flagged by only 1 of the 3 categories.

Table 1: Number of anomaly predictions on training data by data category

Data Category	Count of anomalies
Customer only	163
Transaction only	147
Location only	173
Customer and Transaction	34
Customer and Location	11
Transaction and Location	64
Customer, Transaction, and Location	6

As for the SML, Table 2 shows that LightGBM records the highest F1 and accuracy score at 74.3% and 93.1%, respectively. Each model listed above goes through a 15-stratified-fold cross-validation, which allows for a similar class distribution as the entire dataset, to ensure data representation and performance.

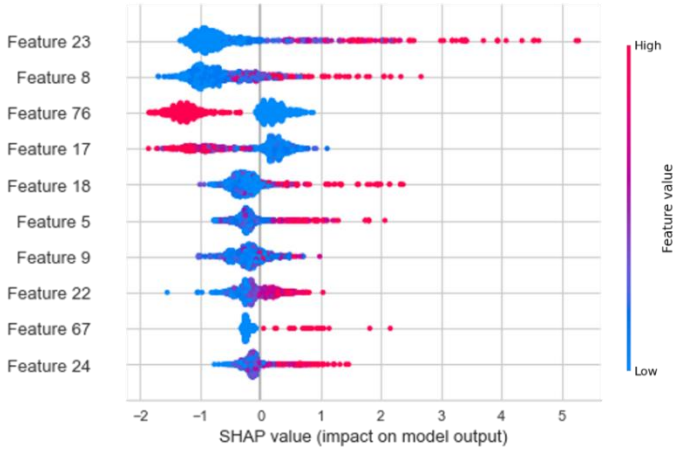
Table 2: Model performance on test set (%)

Model	Accuracy	F1
LightGBM	93.1	74.3
Random Forests	92.3	71.5
Extra Tree	92.4	71.1
Decision Trees	89.7	65.6

Ada Boost	89.6	63.7
Ridge Classifier	85.7	58.5
Logistic Regression	30.0	26.9
SVM	29.6	25.9

Figure 1 displays a beeswarm plot summarizing the SHAP value distribution for the top 10 features, ranked by importance from top to bottom on the y-axis. Based on the SHAP values, the LightGBM model ranks feature 23, related to the customer's profile, as the most influential in identifying anomalies. A high value for feature 23 increases the likelihood of being an anomaly, and a low value decreases it. This insight allows supervisors to focus more on customer profile when conducting second level analysis on the anomalies, while features that rank lower in importance can be given less priority.

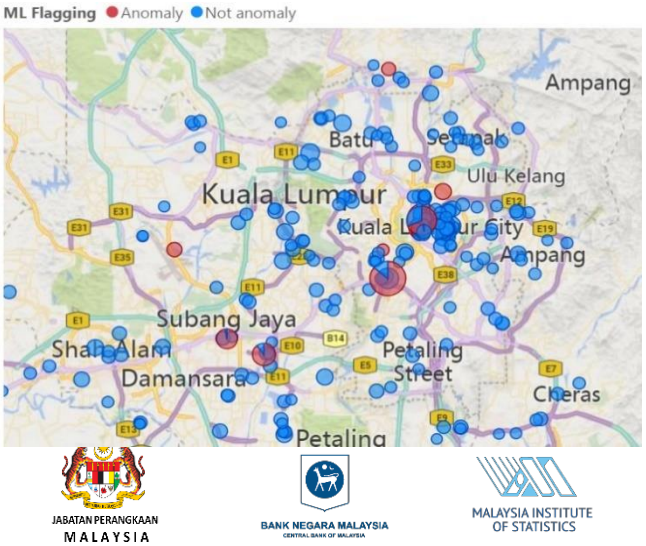
Figure 1: Top 10 feature importance ranking using SHAP value



4. Discussion and Conclusion:

While the model can predict anomalous currency exchange outlets, supervisors need to be able to use this information for their off-site monitoring and supervisory planning activities. To facilitate that, we develop a geospatial dashboard via Microsoft Power BI that maps out all the currency exchange outlets, incorporating colour labels for each outlet based on whether it is predicted as anomalous or not anomalous (Figure 2). As supervisors click on a particular outlet, they can view a table highlighting the top five features contributing to the prediction, based on the SHAP values. Additionally, there will be a table with information about the outlet from the model training dataset.

Figure 2: Snapshot of the MSB Geospatial Analysis Dashboard



In terms of the model performance, predicting true positives of abnormal outlets can be further improved by incorporating more relevant data sources, apart from transactional currency exchange data and outlet profile information. For example, information from financial intelligence or law enforcement agencies regarding illicit activities in a specific location or involving certain currency exchange outlets, if provided to the supervisors, can be useful data points to enhance the machine learning model training performance. Other than inputs from other sources, periodic feedback from supervisors on their validation of the predictions is also vital to ensure low levels of false positives in the model performance.

In conclusion, this paper explores the use of transactional currency exchange data and MSB outlet information to develop a suptech application. We propose a weakly-supervised machine learning approach in identifying anomalous MSB outlets, particularly those providing currency exchange services. The weakly-supervised approach is beneficial due to limited labels of anomalous outlets, which reflects the small number of real-life cases of MSB outlets involving in illicit activities. Supervisors can better understand the model's anomalous predictions using explainable tools such as SHAP. These outputs are then visualised in a geospatial dashboard to ease off-site monitoring by supervisors. This suptech application can be further enhanced with the help of additional information from law enforcement agencies and a periodic feedback loop coming from the supervisors.

References:

1. Bank Negara Malaysia (2019). Promoting Safe and Efficient Payment and Remittance Systems. *BNM Annual Report 2019*. Kuala Lumpur, pp 44-45. Retrieved from https://www.bnm.gov.my/documents/20124/2724769/ar2019_en_full.pdf.
2. Barbariol, T., and Susto, G. A. (2022). Tiws-iforest: Isolation Forest in weakly supervised and tiny ml scenarios. *Information Sciences*, 610, 126–143. Retrieved from https://doi.org/10.1016/j.ins.2022.07.129.
3. Cambridge SupTech Lab (2023). State of SupTech Report 2023. *Cambridge: University of Cambridge*. Retrieved from www.cambridgesuptechlab.org/SOS.
4. Carcillo, F., Borgne, Y. L., Caelen, O., Kessaci, Y., Oble, F., Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557, 317-331. Retrieved from https://doi.org/10.1016/j.ins.2019.05.042
5. Jiang, M., Hou, C., Zheng, A., Hu, X., Han, S., Huang, H., ... & Zhao, Y. (2023). Weakly supervised anomaly detection: A survey. Retrieved from https://doi.org/10.48550/arXiv.2302.04549
6. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422)1. IEEE. https://doi.org/10.1109/ICDM.2008.17
7. Money Services Business Act (2011). Malaysia. Retrieved from https://www.bnm.gov.my/-/money-services-business-act-2011.
8. Steinbuss, G. and Bohm, K. (2021). Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(4), 1-20. Retrieved from https://arxiv.org/pdf/2004.06947