**12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"**

**22-23 August 2024**

# Machine learning for anomaly detection in money services business outlets using data by geolocation[1]

Vincent Lee Wai Seng and Shariff Abu Bakar Sarip Abidinsa,
Central Bank of Malaysia

[1] This contribution was prepared for the conference. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

# Machine learning for anomaly detection in money services business outlets using data by geolocation[1]

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

## 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 allow 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-related 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, outlet geolocation, machine learning, supervision on money services business.

JEL classification: C38, C81, G28.

## Contents



[1] This paper was presented at the 12th biennial IFC Conference on "Statistics and beyond: new data for decision making in central banks" and received the award for the best paper presented by a young statistician.

[2] Respectively Data Analytics Manager, vincentlee@bnm.gov.my, and Supervisor, shariffbakar@bnm.gov.my, at the Payment Services Oversight Department (POD) of the Central Bank of Malaysia. The views expressed herein are those of the authors and do not necessarily reflect those of the Central Bank of Malaysia.

The authors would like to thank Zarifa Izan Zainol Abidin, Mohd. Faizal Abdullah, Siti Hajar Khairul Anwar, Nur Aisyah Mohd Nasir, Muhammad Zulhilmi Shahrolnisham, fellow supervisors of POD and Archana Dilip and Bruno Tissot, BIS IFC, for their insights and comments.

# 1. Introduction

Application of supervisory technology (suptech[3]) has received increasing attention from regulators across the globe. 54% of financial authorities in developing economies have operated one or more suptech applications as at end-2023 as compared to 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).

As for the AML/CFT/CPF regime in Malaysia, Central Bank of Malaysia plays important functions, covering areas such as financial intelligence, investigation of relevant predicate offences, as well as supervision of reporting institutions, which includes the money services business (MSB)[4] industry. As at Q1 2024, there are more than 250 licensed MSBs under the central bank's purview, with more than 80% of them are licensed to conduct money-changing business at physical outlets. Due to the cash intensive and cross-border nature of MSB activities, the risk exposure of MSB industry is mainly coming from money-laundering and terrorism financing (ML/TF) activities. Prior to 2017, central bank supervisors of the MSB sector focused mainly on traditional or checklist examination approaches to identify and monitor ML/TF risks in the industry. Supervisors are also limited by the high-level aggregated data collection from regulated institutions, 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 early 2017 to collect transactional data from the licensed MSBs and then develop suptech applications that can harness risk insights from the transactional data. There are three levels of our suptech applications, namely the industry, entity and customer levels (BNM (2019)). Out of the three, entity level is a key focus as it assists supervisors to identify irregular patterns among the MSBs for the purpose of ML/TF risk profiling and calibration of supervisory activities. Based on our engagement with supervisors, one big challenge for them is to have more proactive and regular monitoring on hundreds of MSB physical outlets located in various parts of Malaysia, despite limited resources. Supervisors also have differing approaches in conducting peer comparison between MSB outlets within a same area over time, resulting in inconsistency in analysis.

This paper proposes a suptech solution, specifically a machine learning approach to flag out irregular patterns in the MSB physical outlets, using the transactional data with location information. This approach allows supervisors to have a more data-driven and regular monitoring in a more automated way. One main challenge in training a robust anomaly detection machine learning model is the lack of precisely labelled data. Hence, a weakly-supervised machine learning approach can address this challenge by having two layers of learning on the data. First is using unsupervised machine learning to generate pseudo-labels. Supervised machine learning models are then trained on the pseudo-labels together. Another reason of having supervised

[3] 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.

[4] Under the Money Services Business Act 2011, MSB refers to any or all of the following businesses: money-changing business, remittance business and wholesale currency business.

machine learning is also to enable supervisors to interpret the model output using SHAP (Shapley Additive exPlanations) values (Wang et al (2024)).

This paper is organised as follows: Section 2 presents an overview of related work. Section 3 describes the dataset used in the machine learning approach and how it is useful to identify anomalies among MSB outlets. The approach for the weakly-supervised machine learning is explained in Section 4, including the models that are selected and ways to explain the model predictions. Section 5 highlights the usage of model outputs by the supervisors and the improvements that can be made to the model. Section 6 summarises the main conclusions.

## 2. Related work

Traditional suptech applications in the area of AML/CFT/CPF supervision has been revolving around detecting unusual networks via analysis on suspicious transaction and threshold reporting as revealed in the overview paper on suptech tools by Broeders and Prenio (2018). More recent applications are also extended to uncovering ML/TF risk triggers from unstructured data using text mining and natural language processing techniques (Appaya et al (2020)). At the same time, some AML/CFT supervisors leverage on suspicious transaction reports (STRs) to train machine learning models to uncover suspicious patterns from transactional data, as highlighted in the paper on suptech applications for AML by Coelho et al (2019). Since STR quality may differ depending on the analysis of reporting institutions, challenge remains for having precise and adequate labelled data.

In relation to the weakly-supervised anomaly detection approach, a paper by 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 analyse how a weak supervision approach can improve the existing anomaly detection algorithm, such as Isolation Forest (IF). Jiang et al (2023) describe various algorithms that focus on weak supervision due to high cost of annotation, which is also a challenge reviewed by this paper.

## 3. Data description and feature engineering

Our solution leverages on the transactional data that were 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 cover 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 and also because the demand is based on geographical factor. For example, MSB outlets located close to the border tend to conduct more exchange transactions based on the currency of the bordering country.

The currency exchange transactional data contain 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 headquarter or branch office of the MSB licensee as well as whether the outlet is a MSB agent to a MSB principal licensee.[5]

The transactional data can then be aggregated into a panel dataset based on the MSB outlet identifier code. At the same time, time-series dimension is added to the aggregation of the currency exchange transactional data by outlet location, where the data are compiled on a quarterly basis. The time-series dimension is important to account for seasonal factors[6] in the performance of the MSB outlets as well as to allow supervisors to monitor MSB outlets on a more regular basis.

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.

An important piece of information relating to the MSB outlet is the full address, allowing us to generate the geolocation, which contributed to the creation of two features. One example feature is the number of peers MSB outlets within 200- and 500-meter radius. Distance between MSB outlets play important role in peer analysis by supervisors to identify unusual patterns across outlets in a particular street or building. Another feature is the region where the outlet is located, eg northern, as we observe unique transaction patterns at regional level.

Other features are relating to the transaction pattern and profile of customers aggregated at the outlet level. This includes MSB outlet's market share of transaction value and volume against all outlets in a postcode, transaction volume by specific time range of the day, and share of higher risk nationality among the customers in the outlet. These features are prepared in consultation with the supervisors and understanding that the MSB outlets within the same area should have similar transaction pattern and customer profile.

[5] 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.

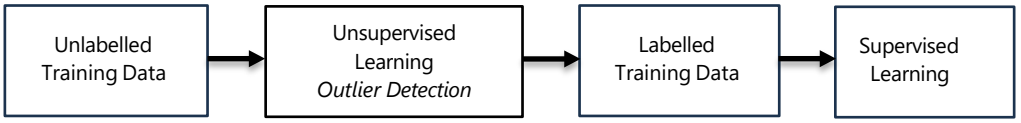
[6] The demand for currency exchange is higher during holiday and festive seasons eg Eid Mubarak, Christmas.

# 4. Methodology

## Weakly-supervised machine learning approach

Our solution adopts the weakly-supervised machine learning framework, which integrates both unsupervised and supervised learning. As a brief overview, the framework visualised in Graph 1 starts with an unsupervised machine learning on the unlabelled panel currency exchange transactional dataset, using the 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[7] is used to assess the supervised machine learning models due to the unbalanced distribution of the labelled training dataset, of which anomalous labels are of minority population.

Weakly-supervised machine learning framework — Graph 1




One benefit of leveraging the weakly-supervised machine learning approach is it allows for noisy or imperfect labels for model training. This fits our use-case of having originally unlabelled panel dataset which then got labelled using the unsupervised learning anomaly detection technique. There are about 800 MSB outlets offering currency exchange services in Malaysia, and it takes up significant human resources to manually label each outlet as "anomaly" or "not anomaly" for each quarter based on historical supervisory findings. Apart from the burden on resources, labelling each outlet based on past supervisory findings would only train our machine learning model to detect anomalies that are similar to that of human findings. Leveraging on machine learning-based anomaly detection models may help supervisors uncover new and emerging anomaly patterns informed by the data.

### A. Unsupervised machine learning

The benefit of applying unsupervised machine learning 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 IF performed well overall. IF is also robust to the presence of outliers in multivariate data and insensitive to the scales of variables.

[7] The harmonic means of precision and recall.

While IF takes less computational resources, it does suffer from 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, currency exchange outlets located in the urban areas of Malaysia may attract more foreign tourists as customers, as compared to outlets based in suburban and rural areas, which may serve more local Malaysian customers who plan to travel abroad. Taking this into account, customer type is one main factor. Based on supervisors' feedback, outlets serving business customers are of higher risk as compared to individuals, due to higher transaction value and the need for MSB licensee to identify the beneficial owner(s) of the business.

Apart from customer type, customer's nationality is also another main factor in differentiating money-changing outlets' behaviour. Specifically, the risks associated with each customer's nationality (such as ML, TF, and PF) is an important differentiating parameter. Money-changing outlets that serve customers primarily from countries with higher risk of financial crimes may be considered as outliers.

In terms of transaction data category, different currency exchange outlets have different transaction patterns based on customer demand. Outlets operating in tourist hotspots may perform more purchase transactions of foreign currencies from foreign travellers in exchange of Malaysian ringgit. In addition, we can also expect more variation in the type of foreign currencies being transacted by currency exchange outlets located in such busy areas. Conversely, currency exchange outlets operating solely in a relatively low-traffic area tend to sell more foreign currencies to meet the demand of Malaysians travelling abroad.

For location data category, currency exchange outlets that are located within the same area tend to behave similarly, such as offering similar exchange rates, since they are more likely to serve similar customer profile demanding for similar currency type. For example, outlets located near the border of neighbouring countries perform more transactions related to the foreign currency of the neighbouring country. Other than exchange rates, there are also limited opportunities for outlets to differentiate their products since physical foreign currency is a homogenous product.

The time period for each of three data categories used as training dataset is from Q1 2022 to Q4 2022. The anomaly detection is then performed on each data category independently using the IF model. A contamination rate of 0.05 was used, which means that only 5% of the total observations were identified as outliers. The low contamination rate is to ensure the false positive rate is reasonably contained.

The IF model identifies anomalies by isolating the observations in the dataset. The isolation process is done by randomly selecting features and split values to construct binary trees until all observations are isolated or a maximum tree height is reached. The path length, $h(xx)$, of an observation $xx$ is the number of edges traversed from the root node to the leaf node that isolates $xx$. Shorter path lengths indicate that the observation was isolated relatively quick, suggesting that it is an anomaly. For an observation $xx$, the average path length $EE(h(xx))$ is calculated across all constructed

trees. The average path length is normalised against the average path length of unsuccessful searches in a binary search tree $cc(nn)$, which depends on the number of observations $nn$ in the data. The anomaly score $ss(xx, nn)$ is then calculated as follows:

$$ss(xx, nn) = 2^{\frac{-EE(h(xx))}{cc(nn)}}$$

where $EE(h(xx))$ is the average of $h(xx)$ from a collection of isolation trees and $cc(nn)$ is the average of $h(xx)$ given $nn$ (Liu et al (2008)).

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

## B. Supervised machine learning

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

| 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 |

Once we have the anomalous observations produced by the IF model, we then label the currency exchange outlets as anomaly if they are considered as an anomaly at least once in the three data categories. This means that the outlets that are not identified an anomaly in any of the three data categories are not labelled as an anomaly. Given that we only introduce a 5% contamination rate, the unbalanced class of labels is expected.

Several data preprocessing steps are conducted to ensure model adequacy. This includes fixing the label imbalances using Synthetic Minority Over-sampling Technique (SMOTE) – which creates synthetic samples and increases the overall sample size by 57%. In addition to SMOTE, one-hot encoding is performed on the categorical variables. This is to ensure that the data and model are compatible for processing as well as for better information gain by the models. Then, as part of the model evaluation process, 80% of the data are then randomly assigned as the training dataset, with the remaining observations assigned as test dataset.

The labelled data are then used as part of training for a classification tasks. We consider the following models and the performance of the models as summarised in Table 2:

- Light Gradient Boosting Machine (LightGBM) – a gradient boosting framework that uses tree-based learning algorithms and excels in handling multi-dimensional data and complex classification tasks.
- Decision Tree – a simple, interpretable model that splits the data into subsets based on feature values, forming a tree structure. This approach provides high interpretability but may overfit easily.
- Random Forest – an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. By averaging the results of many decision trees, random forests reduce overfitting and improve generalisation, making them robust and accurate for classification tasks.
- Logistic Regression – a statistical model that uses a logistic function to model a binary dependent variable. It estimates the probability that a given input belongs to a certain class, making it suitable for binary classification problems. It is easy to implement and interpret but may not perform well with complex relationships.
- Support Vector Machines (SVMs) with Linear Kernel – they work by finding the optimal hyperplane to separate classes in the feature space (by maximising the margin between classes), making it effective for high-dimensional spaces.
- Ada Boost – an ensemble learning method that combines multiple weak classifiers to create a strong classifier. This model adjusts the weights of incorrectly classified instances, focusing more on difficult cases in subsequent iterations. It can significantly improve the performance of weak learners.
- Extra Tree – an ensemble method that builds multiple trees like random forests but with more random splits, which reduces variance and often leads to faster training times. It is robust and can handle complex classification tasks.
- Ridge Classifier – applies Ridge regression (L2 regularisation) for classification tasks, particularly when there are multiple correlated features. It modifies the logistic regression by adding a regularisation term to prevent overfitting. It is useful for handling multicollinearity and ensuring stable solutions.

Based on the results of model performance, 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.

To save computing and time resources, only the best performing model's hyperparameters is optimised for prediction purposes. One of the hyperparameters tuned for the LightGBM model is the learning rate which is an important hyperparameter that determines the step size at each iteration while moving towards the minimum of the loss function. The learning rate determines the contribution weights of each tree when adding it to the existing ensemble of trees. A smaller

learning rate value often leads to better generalisation, while a larger learning rate value may risk overfitting.

Model performance on test dataset

In percentage (%) Table 2

| 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 |

Besides learning rate, n_estimator is another critical hyperparameter that is tuned for optimal model performance. N_estimator is the number of boosting trees in the ensemble, with each tree sequentially attempting to correct the errors of the preceding ones. The optimal n_estimator value was 20, which maximises the model performance without overfitting. Overall, both learning rate and n_estimator are important in developing the best performing model and in ensuring robust and efficient training.

## C. Model explainability

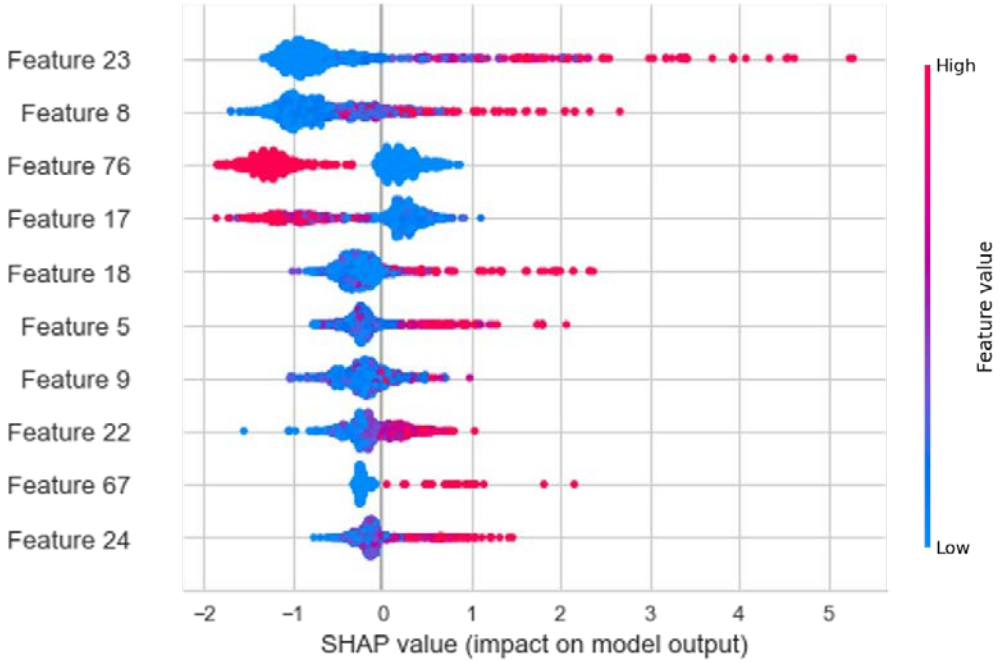
One challenge of applying machine learning models is that it can be difficult for supervisors to understand the model predictions as well as the features that contribute more to the prediction of anomalous outlet. To provide a deeper understanding of the predictions made by our machine learning models, we employ Shapley Additive exPlanations (SHAP). SHAP is a unified approach to explain the output of any machine learning model, based on game theory and Shapley values. It assigns each feature an importance value for a particular prediction, making it possible to interpret the model's behaviour at both global and local levels.

Graph 2 displays a beeswarm plot summarising 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 considered an anomaly, and a low value decreases it. This insight allows supervisors to put more focus on customer profile information when conducting second level analysis on the anomalies, while features that rank lower in importance can be given less priority

Overall, 5 of the top 10 most important features are related to transaction details, 4 features describe location-based information, and 1 feature pertains to the customer's background. This distribution highlights the critical role of transaction and location data in anomaly detection, suggesting that enhancing data quality and analysis in these areas can significantly improve our model's performance. Prioritising

these key features will help us develop more targeted strategies for anomaly detection and prevention.

Top 10 Feature Importance Ranking using SHAP Value Graph 2

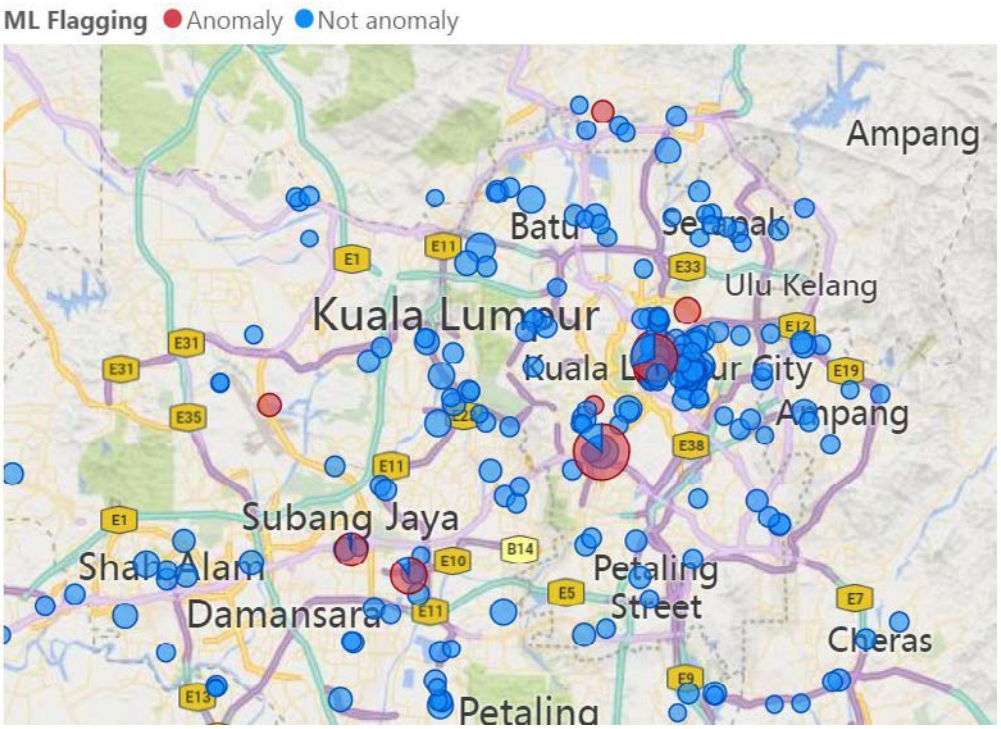



## 5. Usage and further improvements

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 (Graph 3). 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. Apart than that, there will be another table with all the information relating to the outlet, which comes from the dataset used for model training.

Supervisors can also filter the geospatial visual by the time period, which is based on a quarterly period, and observe the quarter(s) for which a particular outlet is predicted as anomaly. Apart from that, supervisors can zoom the geospatial visual into a specific state, city and postcode of interest to narrow down their scope of monitoring. One way that might further improve the dashboard is to provide notification alerts to the specific supervisor when a particular outlet of interest has turned anomalous for the new quarter.

Snapshot of the MSB Geospatial Analysis Dashboard Graph 3




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.

## 6. Conclusion

In this paper, we explore the use of transactional currency exchange data and MSB outlet information to develop a suptech application. In particular, 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 involved in illicit activities. The model predictions of anomalous outlets can only be better understood by supervisors with the help of model explainable tools such as the SHAP. These outputs are then visualised in a geospatial dashboard in order to ease off-site monitoring by supervisors. This suptech application can be further enhanced with the help of

additional valuable intelligence and information from law enforcement agencies as well as periodic feedback loop coming from the supervisors.

# References


Appaya, M, M Dohotaru, B Ahn, T Kliatskova, P Seshan and I Pascaru (2020): "A roadmap to SupTech solutions for low income (IDA) countries (English)", *World Bank Working Paper*, no 153566, http://documents.worldbank.org/curated/en/108411602047902677/A-Roadmap-to-SupTech-Solutions-for-Low-Income-IDA-Countries.

Barbariol, T and G Susto (2022): "Tiws-iforest: isolation forest in weakly supervised and tiny ml scenarios", *Information Sciences*, vol 610, September, pp 126–43, https://doi.org/10.1016/j.ins.2022.07.129.

Broeders, D and J Prenio (2018): "Innovative technology in financial supervision (SupTech)-the experience of early users", *FSI Insights on policy implementation*, no 9, July.

Cambridge SupTech Lab (CSL) (2023): "State of SupTech report 2023", University of Cambridge, www.cambridgesuptechlab.org/SOS.

Carcillo, F, Y Borgne, O Caelen, Y Kessaci, F Oble and G Bontempi (2021): "Combining unsupervised and supervised learning in credit card fraud detection", *Information Sciences*, vol 557, May, pp 317-31, https://doi.org/10.1016/j.ins.2019.05.042.

Central Bank of Malaysia (2019): "Promoting safe and efficient payment and remittance systems", *Bank Negara Malaysia Annual Report 2019*, pp 44-5, https://www.bnm.gov.my/documents/20124/2724769/ar2019_en_full.pdf.

Coelho, R, M Simoni and J Prenio (2019): "Suptech applications for anti-money laundering", *FSI Insights on policy implementation*, no 18, August.

Jiang, M, C Hou, A Zheng, X Hu, S Han, H Huang and Y Zhao (2023): "Weakly supervised anomaly detection: a survey", arXiv preprint arXiv:2302.04549.

Laws of Malaysia (2011): *Money Services Business Act 2011*, https://www.bnm.gov.my/-/money-services-business-act-2011.

Liu, F T, K M Ting and Z-H Zhou (2008): "Isolation forest", in *2008 Eighth IEEE International conference on data mining*, Italy, 15-19 December, pp 413-22, IEEE. https://doi.org/10.1109/ICDM.2008.17.

Steinbuss, G and K Bohm (2021): "Benchmarking unsupervised outlier detection with realistic synthetic data", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol 15, no 4, pp 1-20.

Wang, H, Q Liang, J Hancock et al (2024): "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods", *Journal of Big Data*, vol 11, no 44, March, https://doi.org/10.1186/s40537-024-00905-w.

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

Vincent Lee, Shariff Abu Bakar
Central Bank of Malaysia
Payment Services Oversight Department

(*) The views and conclusions presented in this paper are exclusively those of the author(s) and do not necessarily reflect the position of the Central Bank of Malaysia or of the Board members.

# Data analytics tools to facilitate supervision of large number of regulatees

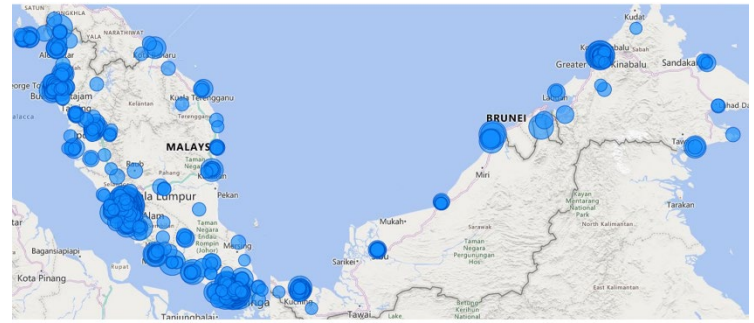
Over 250 money services business (MSB) licensees with close to 800 physical outlets supervised by Central Bank of Malaysia.

MSB industry is considered higher risk for money laundering and terrorism financing in Malaysia, due to the cash intensive and cross border nature of MSB activities.

Prior to 2017, supervisors mainly focus on traditional / checklist approach in examining the MSBs, limited by high-level aggregated data, limited manpower, and outdated risk profiling.

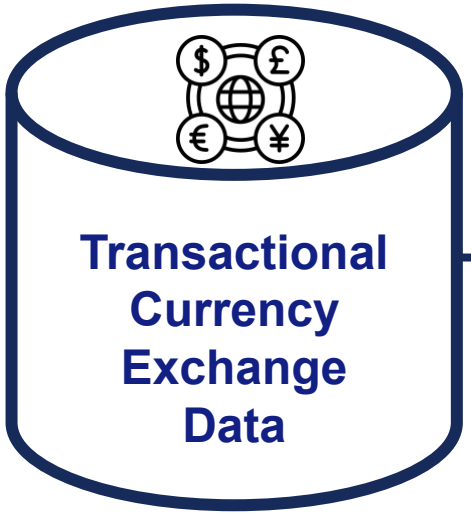
Since then, we collected transactional data to develop data analytics applications to enhance capability of supervisors in providing more proactive and regular oversight on the MSB licensees and their outlets.

The paper focuses on a machine learning approach to flag out irregular patterns in the MSB outlets



# Granular MSB transactional data for anomaly detection of MSB outlets

Training features were derived from past examination findings and engagement with supervisors




Summarised into ***30 features*** by each MSB outlet per quarter

*Period: Q1 2022 - Q1 2023*

## Transaction

- Share of transaction value / volume against outlets in the postcode
- Transaction volume by specific period of the day

## Customer

- Share of higher risk nationality customers
- Share of different customer types

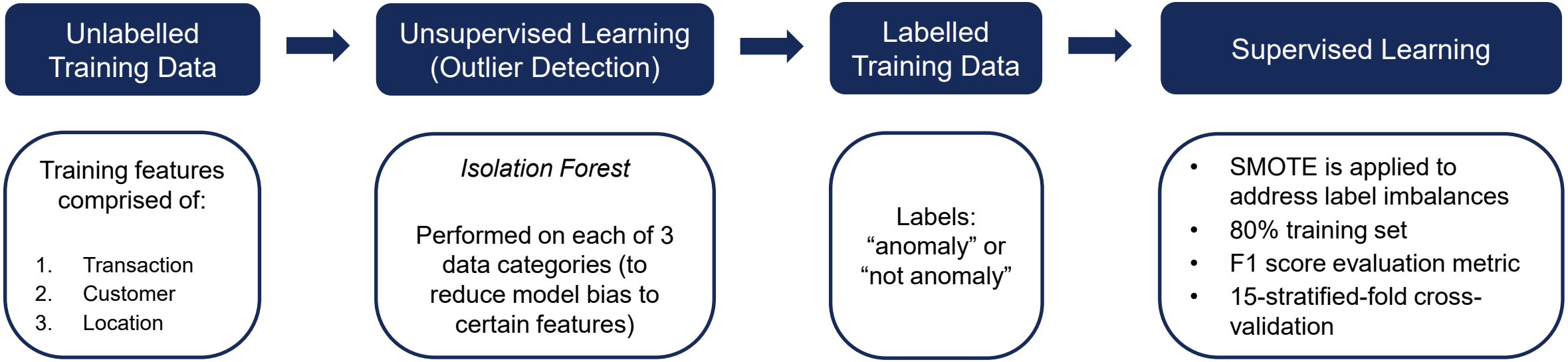
## Location

- Geolocation
- Number of peers within 200-meter radius

# Weak-supervised machine learning approach for anomaly detection

- Allows for silver labels for supervised machine learning
  - Historical findings of anomaly MSB outlets are limited to generate gold labels
  - Solves problem of expensive labelling by subject matter experts

- Assist supervisors to uncover new / emerging anomaly patterns, not found in historical supervisory findings



# Outliers as target labels using unsupervised learning

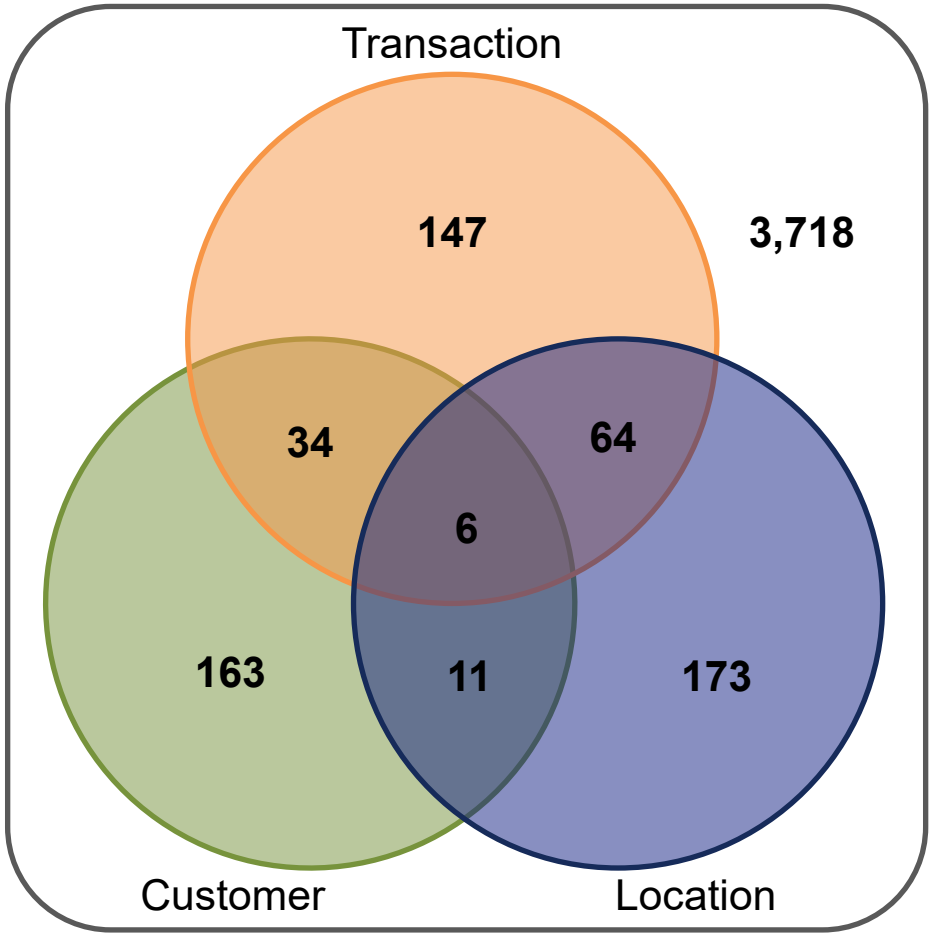
- Isolation Forest (IF) detects anomaly more accurately than other models (Steinbuss & Bohm, 2021).

- The IF model isolates observations by randomly selecting features and split-values. Observations that were isolated quickly (i.e., lesser path lengths) are likely to be considered as anomalies.

- To reduce biases to certain features, IF is applied to three different categories of data:
  - ✓ Transaction
  - ✓ Customer
  - ✓ Location

- Anomaly score:

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}}$$



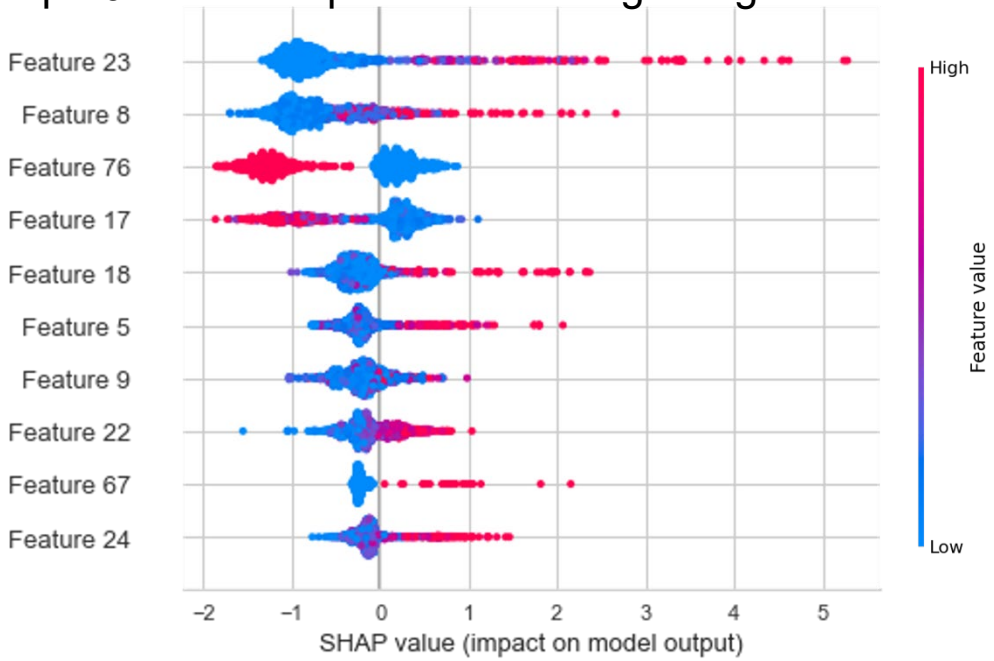
# LightGBM and model explainability

- Best model with highest F1 score:
  - ✓ F1 score as evaluation metric due to class imbalanced in the dataset
  - ✓ F1 score considers both false positive and false negative

| 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 |

- SHapley Additive exPlanations (SHAP)
  - ✓ used to explain machine learning model output, based on game theory and Shapley values
  - ✓ facilitate supervisors to understand model predictors

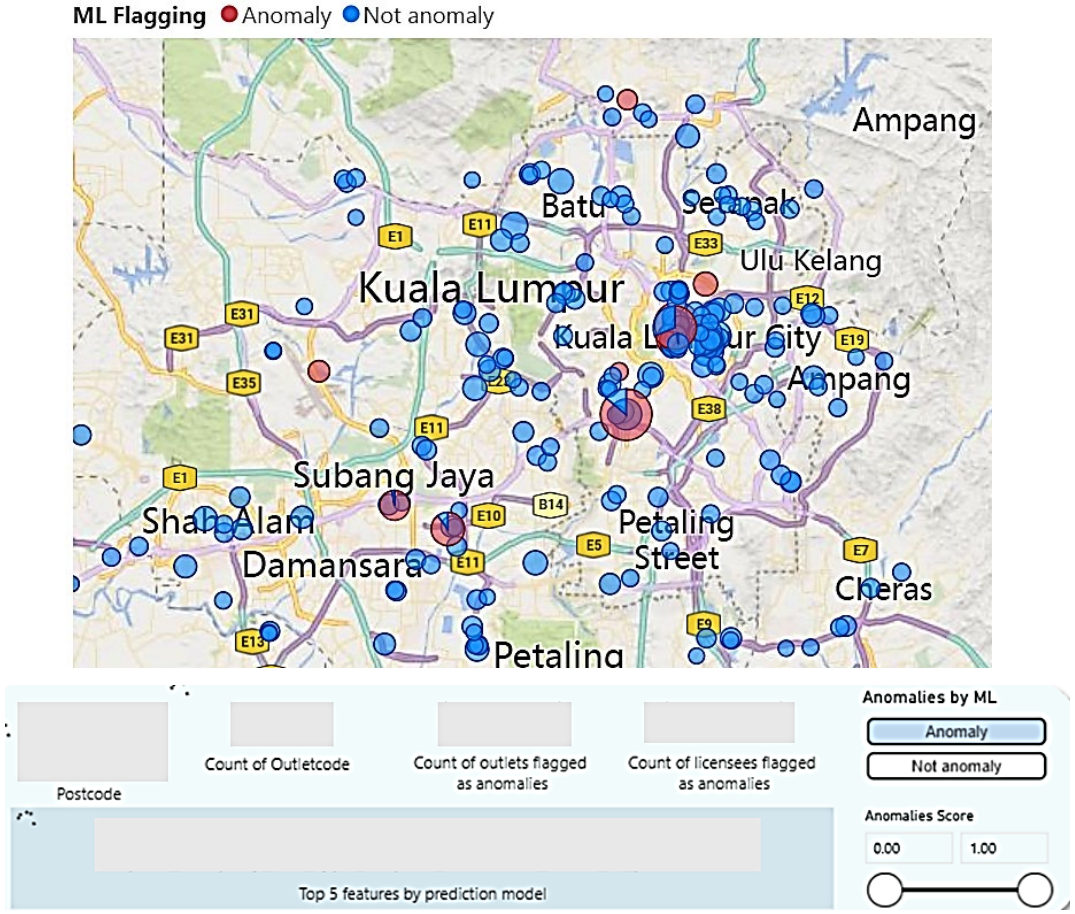
Top 10 feature importance ranking using SHAP value:



# Usage

- Anomalous outlets predicted are visualised in a Geospatial Dashboard for supervisors' monitoring
- Supervisors can filter the geospatial visual by quarterly period, state, city and postcode
- For each outlet, the dashboard highlights the top five features contributing to the prediction (based on SHAP)

Snapshot of the MSB Geospatial Analysis Dashboard



## Future developments

- More relevant data sources to improve model performance
  - ➢ e.g. financial intelligence, findings from law enforcement agencies
- Automated feedback loop: Incorporate periodic feedback from supervisors into model training
  - ➢ To reduce false positives in model prediction

## Conclusion

- Weakly supervised machine learning is beneficial due to limited labels of anomalous outlets
- Model explainability is important for supervisors to understand model predictions
- Visualisation of the model predictions via geospatial dashboard ease supervisors to conduct off-site monitoring on the MSB outlets

# Thank you for your attention!

Vincent Lee vincentlee@bnm.gov.my

Shariff Abu Bakar shariffbakar@bnm.gov.my