MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA

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.

11Th MALAYSIA STATISTICS CONFERENCE
"Data and Artificial Intelligence: Empowering the Future"

19Th September 2024

Organized by:

MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

MyStats 202

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on traditional / checklist approach in examining the MSBs, limited by high-level aggregated data, limited manpower, and outdated risk profiling.
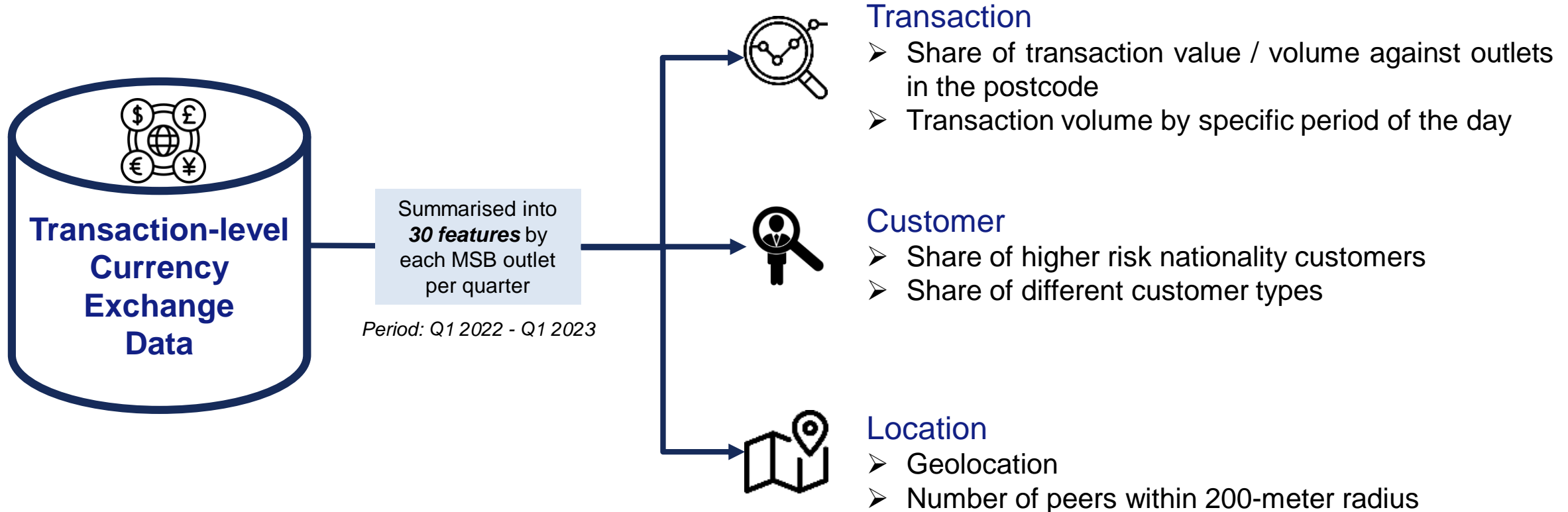
Since then, we collected transactional data to develop data analytics applications to enhance capability of supervisors in providing more proactive and regular oversight on the MSB licensees and their outlets.

The paper focuses on a machine learning approach to flag out irregular patterns in the MSB outlets
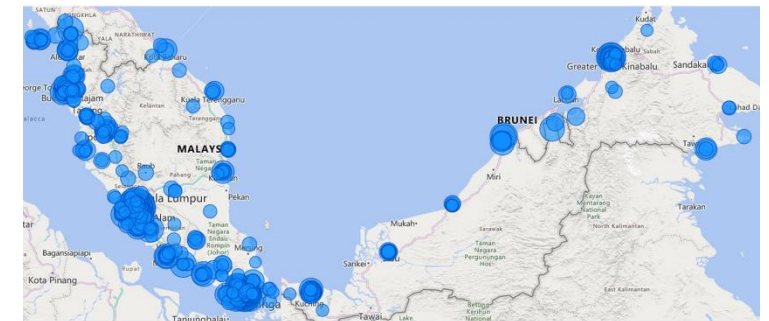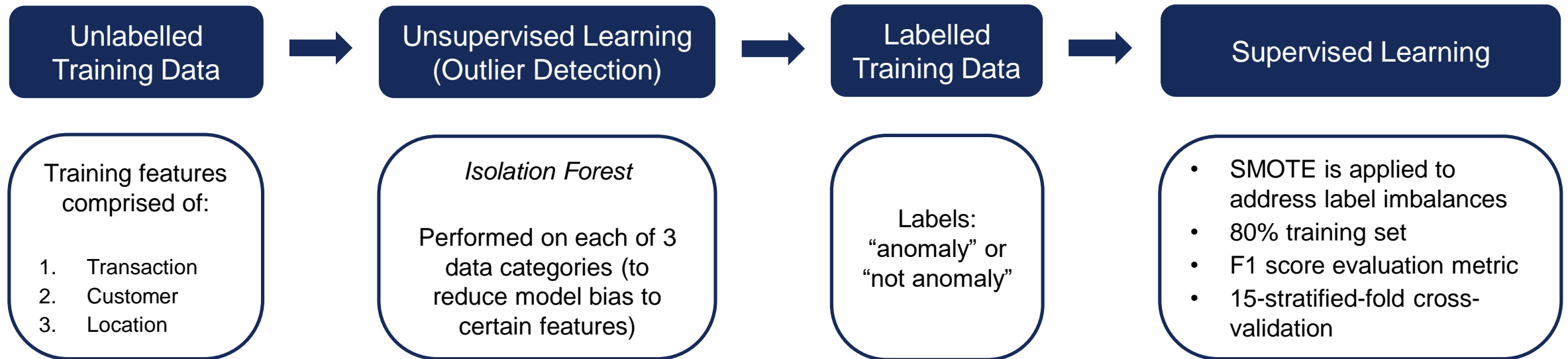
Granular MSB transactional data for anomaly detection of MSB outlets

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

Transaction-level Currency Exchange Data

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

Unlabelled Training Data	→	Unsupervised Learning (Outlier Detection)	→	Labelled Training Data	→	Supervised Learning

Unlabelled Training Data

Training features comprised of:

1. Transaction
2. Customer
3. Location

Unsupervised Learning (Outlier Detection)

Isolation Forest

Performed on each of 3 data categories (to reduce model bias to certain features)

Labelled Training Data

Labels: "anomaly" or "not anomaly"

Supervised Learning

- SMOTE is applied to address label imbalances
- 80% training set
- F1 score evaluation metric
- 15-stratified-fold cross-validation

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:

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

Transaction

147   3,718

34   64

6

163   11   173

Customer   Location
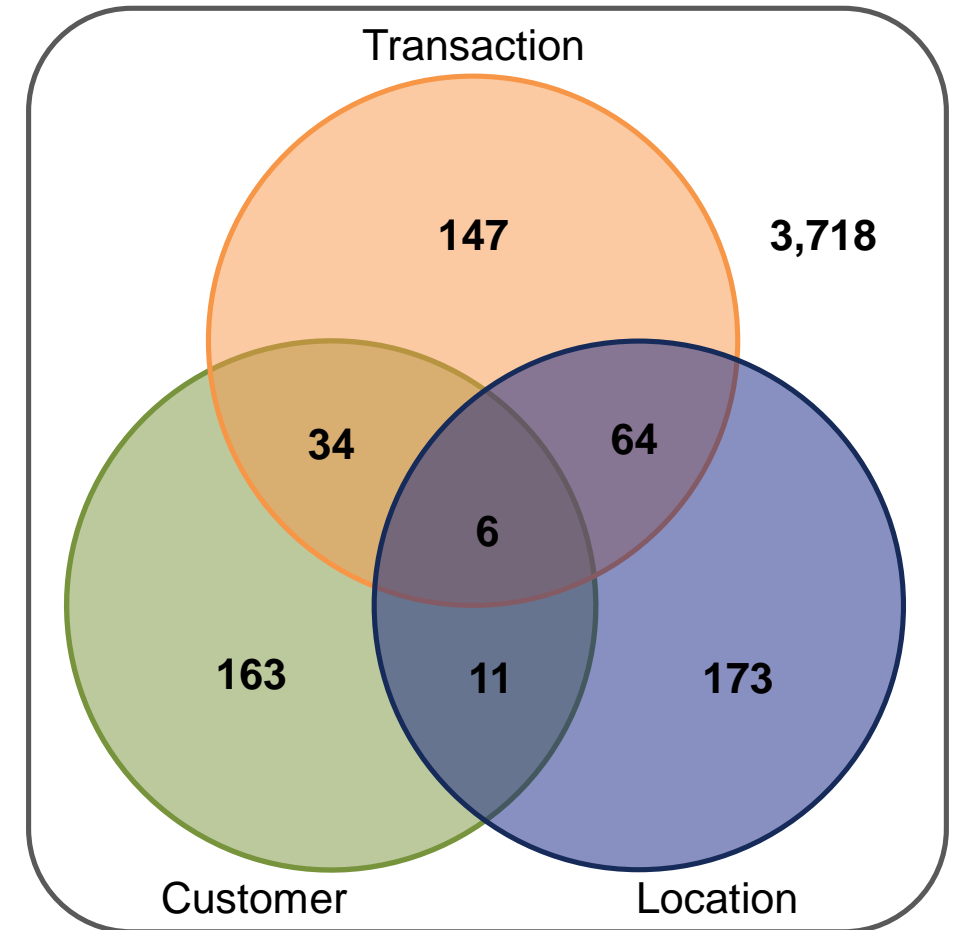
LightGBM and model explainability

- 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:
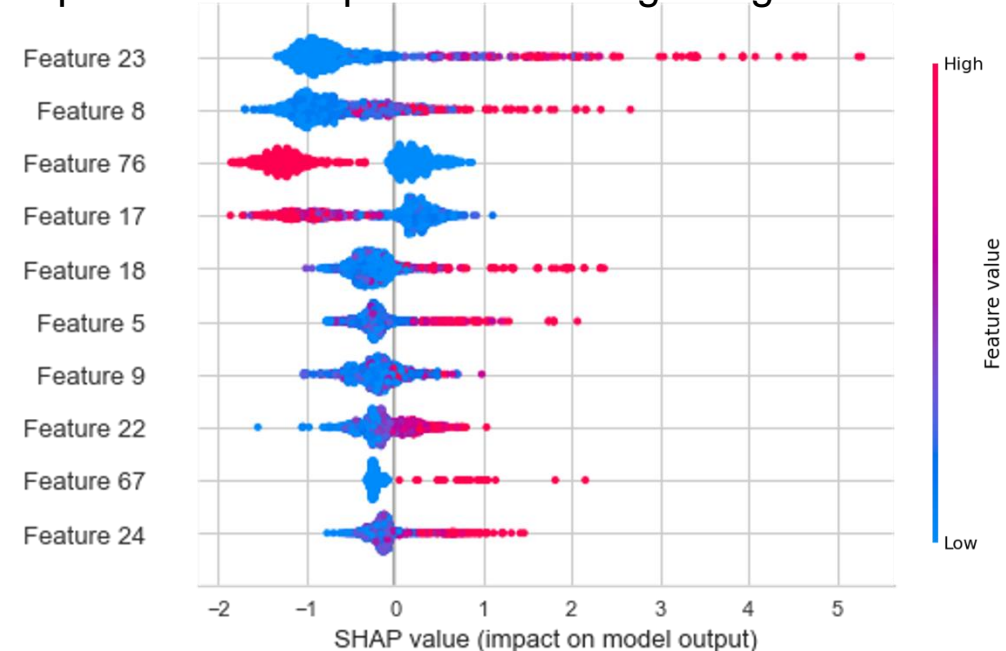
Usage

- 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)

Example cases from supervisors:
- MSB outlets are flagged for having abnormal/lower transaction value than peers in same vicinity

- On-site inspection found non-recording of transactions by the MSB licensees (non-compliance)

- Actions were taken on the MSB licensees, including licence revoked or not renewed

Snapshot of the MSB Geospatial Analysis Dashboard
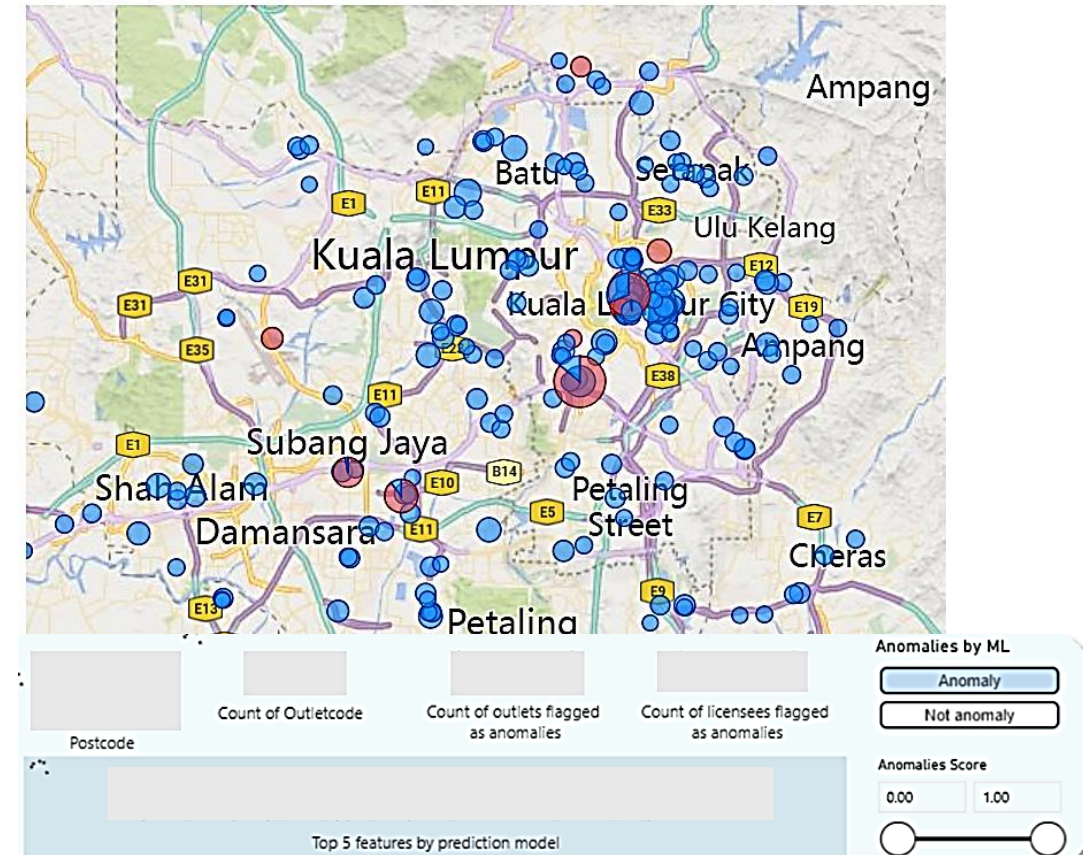
Future developments

- 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
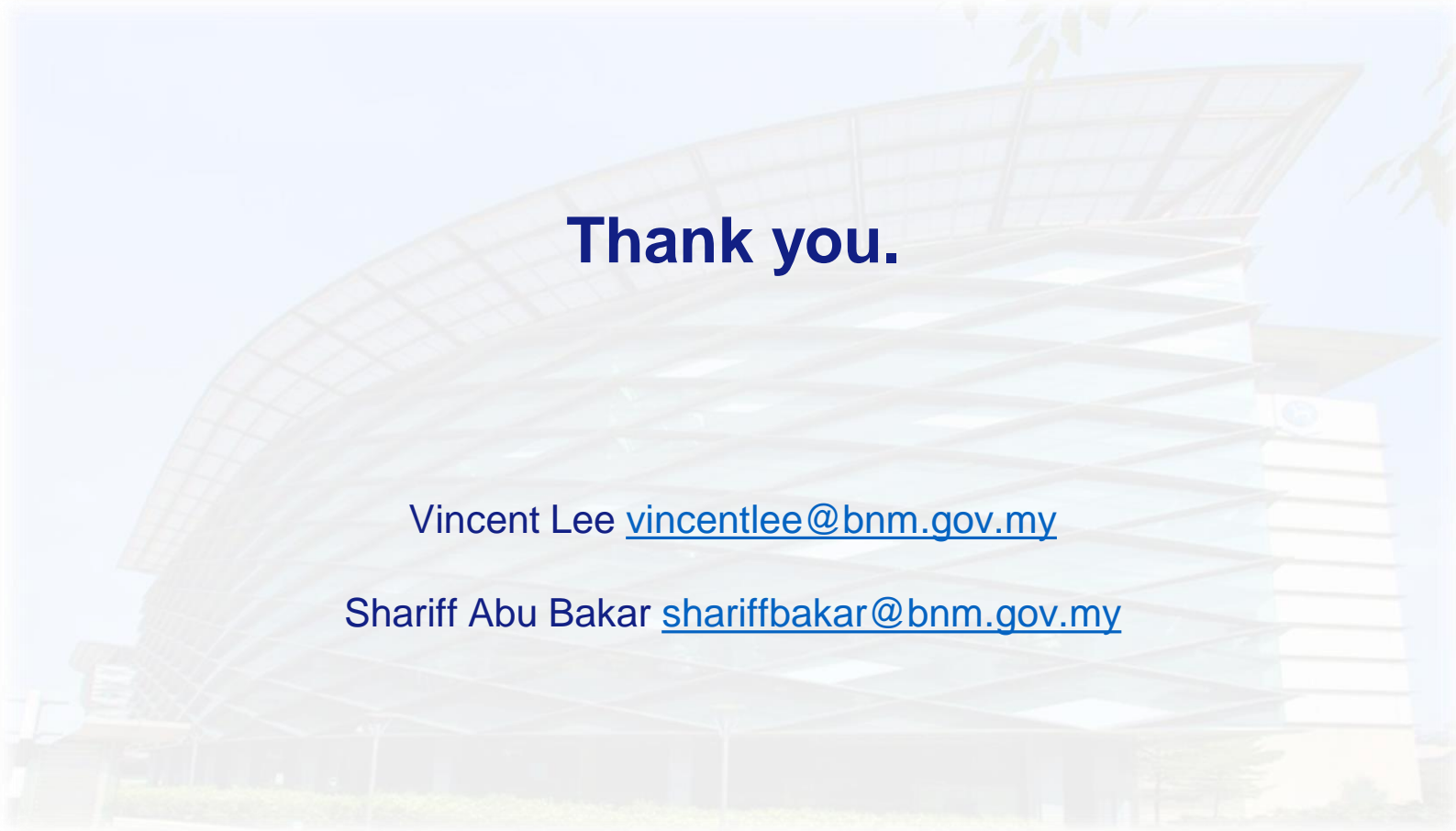 - ➢ To reduce false positives in model prediction

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

11Th **MALAYSIA STATISTICS CONFERENCE**
"Data and Artificial Intelligence: Empowering the Future"

Thank you.

Vincent Lee vincentlee@bnm.gov.my

Shariff Abu Bakar shariffbakar@bnm.gov.my

11Th MALAYSIA STATISTICS CONFERENCE
"Data and Artificial Intelligence: Empowering the Future"

19Th September 2024

Organized by:

MINISTRY OF ECONOMY DEPARTMENT OF STATISTICS MALAYSIA

BANK NEGARA MALAYSIA CENTRAL BANK OF MALAYSIA

MALAYSIA INSTITUTE OF STATISTICS

MyStats 202