TRANSORG ANALYTICS

Prevented customer frauds with 81% accuracy for a leading payments bank

Content

Objective

The objective of the project was to build a anomaly detection model for a leading bank based on account information data and transaction behavior of customers by generating an “anomaly score” to predict the likelihood of a customer being fraudulent

Scope of work:

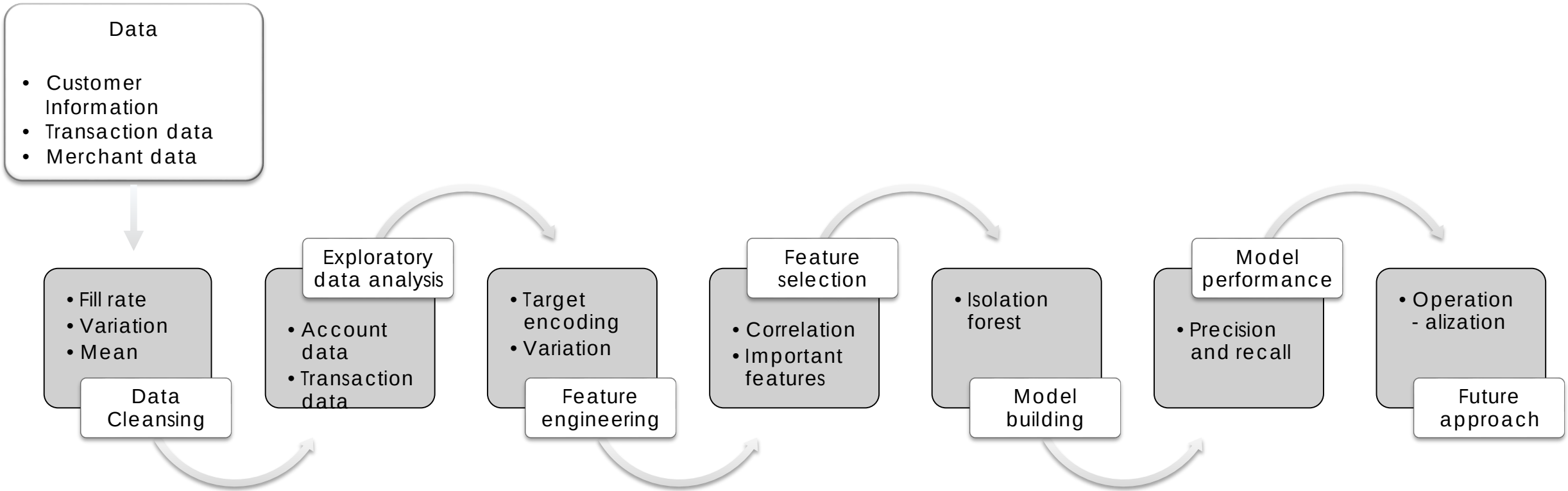
Understand the available data → Perform data exploration → Clean and process the data → Create business variables → Build model to predict the likelihood of fraud

Available data:

- **Customer 360** – Customer account information data
- **Transaction 360** – Customer transaction data
- **Merchant 360** – Merchant data linked to each customer

Approach Overview

Process flow



Methodology (1/2)

Methodology: After understanding all the data fields, performing data exploration and cleaning the data; we followed the below methodology to build the model(s)

1. **Data selection:**
   - Data of active savings account was used as rate of occurrence of fraud is maximum
   - The event rate of fraudulent accounts was ~1.1%

2. **Data pre-processing :**
   - Statistical tests and analysis was done to select list of data variables for model building
   - Selected variables were permanent state, age, occupation code, average transaction time/ volume/ velocity etc.

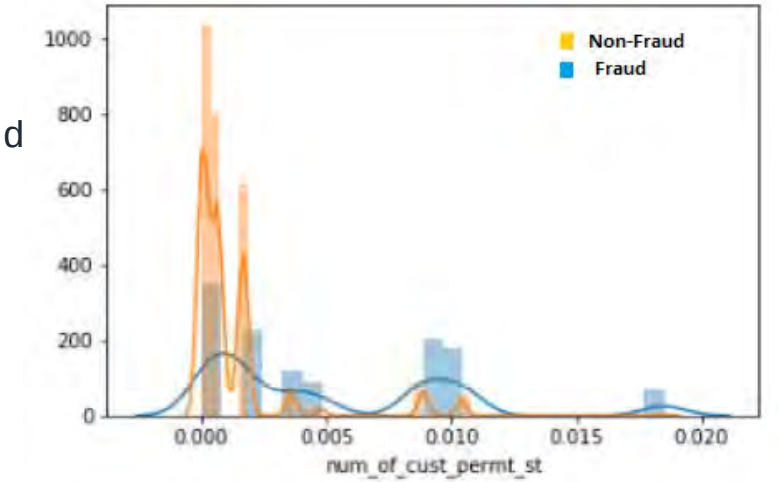
3. **Exploratory data analysis:**
   - Patterns were observed when analysis exclusively on fraudulent accounts were done
   - In some of the fraudulent accounts debit and credit amounts followed a pattern and time gap between transactions were relatively less when compared to a non-fraud accounts

**The numbers mentioned are illustrative.*

Methodology (2/2)

4. Feature engineering :

- As an outlier detection algorithm was used to detect fraud, features were made so that fraud points occur at outlier position in distributions
- Categorical feature such as customer permanent state was engineered based on occurrence of fraud
- Target encoding was performed to obtain numeric continuous values for nominal classes of a category



5. Model building:

- Created machine learning models to predict the likelihood of occurrence of fraud
- The output of the model was an “anomaly score” for every customer
- Model performance was evaluated metrics like precision and recall

Model performance (1/2)

Total population

1.11% of population are fraudulent

Feature Engineering

Isolation forest

Account watchlist for business

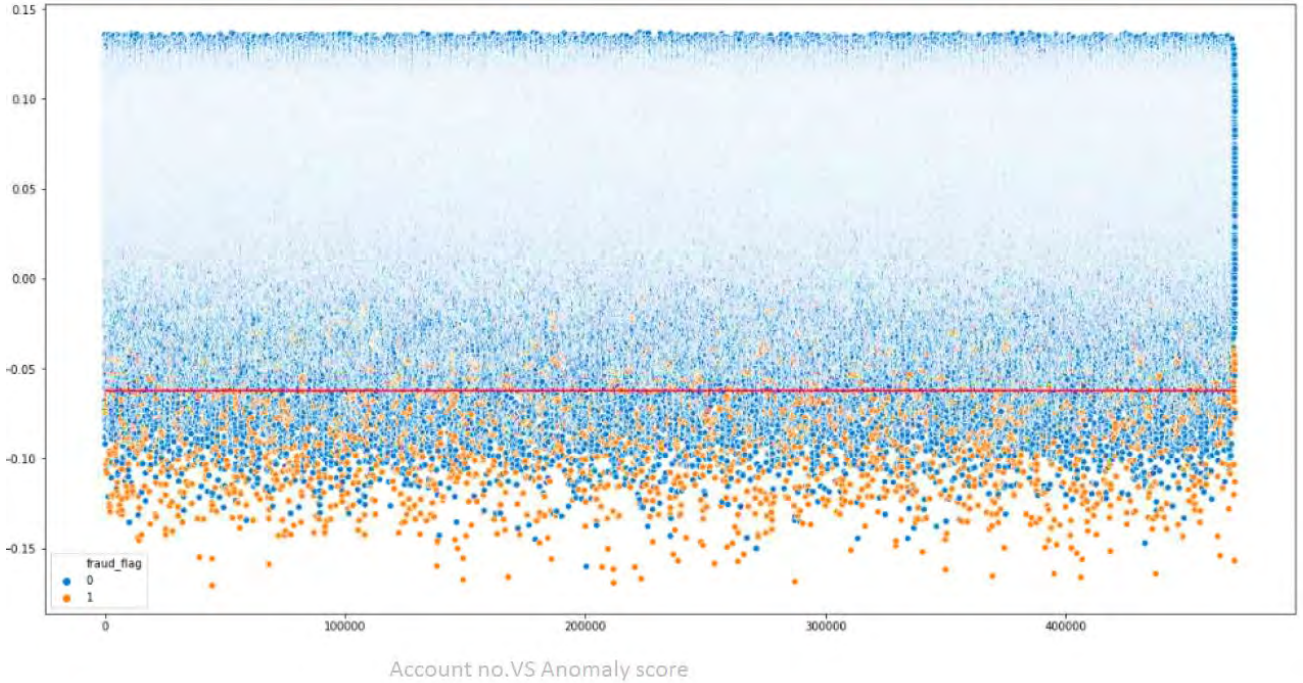
Model generating anomaly score for each customer

Top 1.11% scorers capture 43% frauds

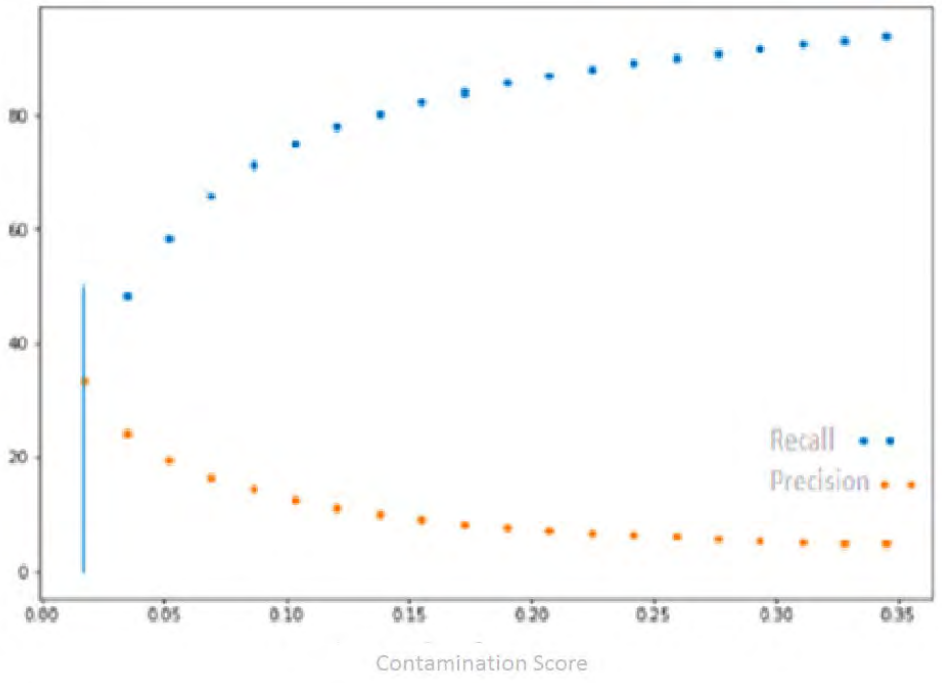
Top 10% scorers capture 81% frauds

**The numbers mentioned are illustrative.*

Model Performance (2/2)



Account no.VS Anomaly score



Contamination Score

Account no. vs Anomaly score

- In top 1.7 percentile, recall and precision is 35 %
- We are covering 43% of account level fraud in top 1.7 percentile
- Out of remaining 57% some accounts are potential fraud

Precision/Recall

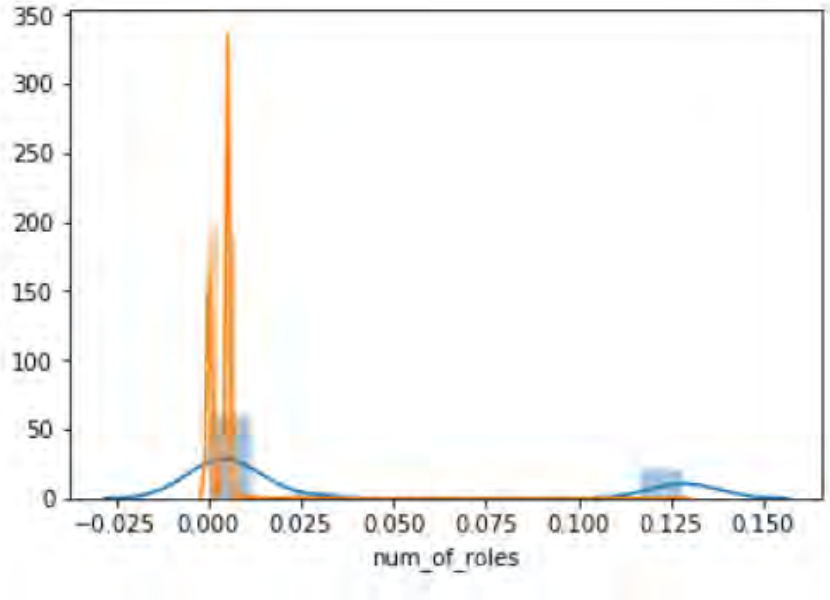
- Contamination score which sets the percentage of points in our data to be anomalous.
- Recall and Precision can be improved by changing contamination score

**The numbers mentioned are illustrative.*

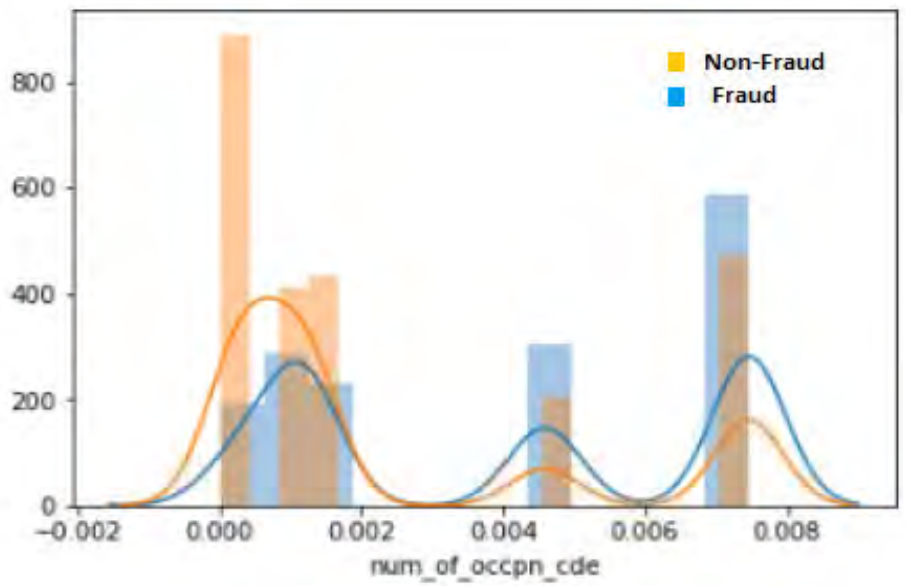
Key Variables and Insights (1/2)

Account Level Features

1. Account opening information:
   - Roles: Medium of account opening
   - Initial fund details



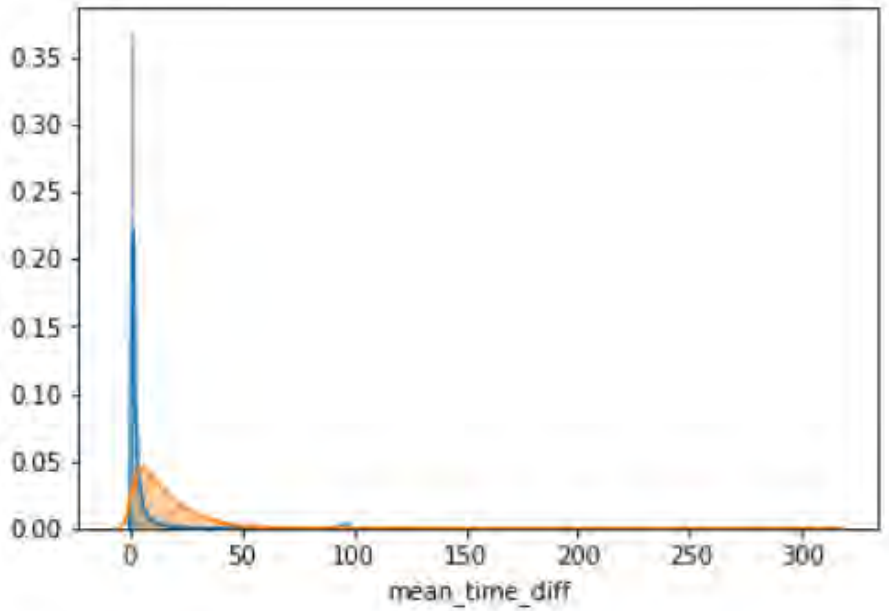
2. Customer personal information
   - Customer permanent state
   - Occupation code

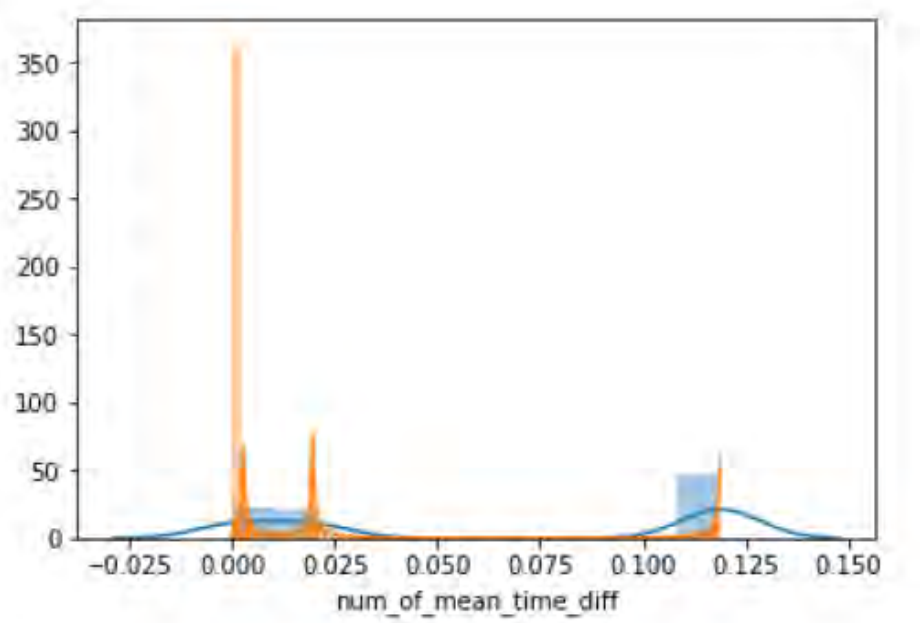


**These are engineered features that were made to fit a scale. We can see that fraud points are occurring at outlier position.*

Key Variables and Insights (2/2)

3. Transaction Features
   - Average time difference between transaction
   - Average transaction volume/velocity
   - Number of time transaction volume crossed a threshold
   - Deviation of transaction volume with mean
   - Transaction count





**These are engineered features that were made to fit a scale. We can see that fraud points are occurring at outlier position.*

Operationalization and Impact

- Tasks performed in improvement
  - The model accuracy was increased after adding more sophisticated features from transactional history
  - Neural Network Auto encoder was also built which yielded better accuracy

- Tasks performed in operationalization

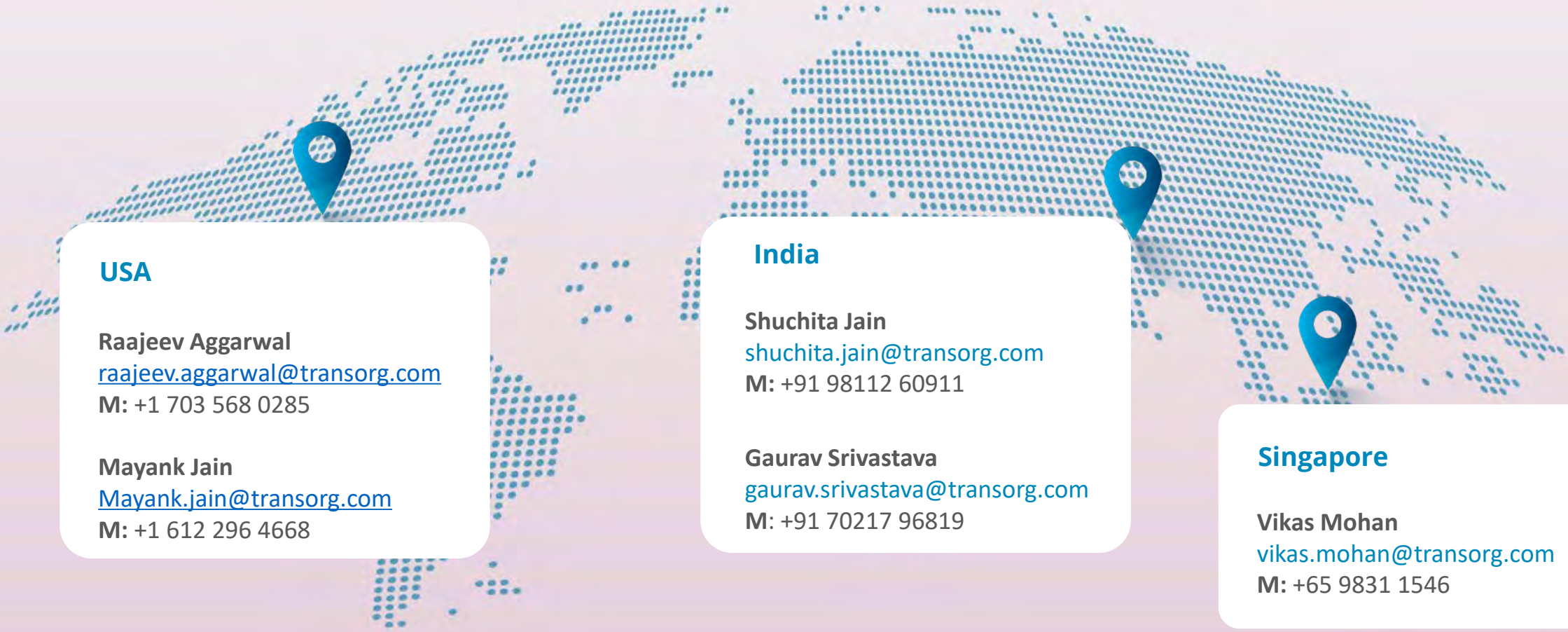
  - Operationalization of the model in the Big Data Architecture was performed using Pyspark and Shell scripting where model was deployed in the Client's production environment

  - A scheduler was created in order to give "anomaly score" at regular intervals.

  - Model accuracy: Top 1.11% scorers capture 43% frauds and Top 10% scorers capture 81% frauds.

  - A list of accounts with high anomaly score in every prediction cycle was shared with Client's Anti Fraud Unit for verification

  - Model performance tracker was also created and retraining was automated, in the case where accuracy dropped down a certain threshold



USA

Raajeev Aggarwal
raajeev.aggarwal@transorg.com
M: +1 703 568 0285

Mayank Jain
Mayank.jain@transorg.com
M: +1 612 296 4668

India

Shuchita Jain
shuchita.jain@transorg.com
M: +91 98112 60911

Gaurav Srivastava
gaurav.srivastava@transorg.com
M: +91 70217 96819

Singapore

Vikas Mohan
vikas.mohan@transorg.com
M: +65 9831 1546

LinkedIn Facebook Twitter YouTube

www.transorg.com | +91-124-4006248

Would you like to connect with us to get advanced analytics solutions for your organization??

Contact Us:

shuchita.jain@transorg.com

gaurav.Srivastava@transorg.com

TransOrg Analytics
www.transorg.com
+91-124-4006248