Optimisation model for investigating fraudulent transactions in the banking sector

Mathematical, Modelling and Consulting Skills – MSc Operational Research 2023/2024

Contents

1. Introduction

This project is designed to present a small-scale simulation of the UK banking system, exploring optimal solutions for minimising the multi-bank fraud activity through individual or combined optimisation objectives. The solutions are supposed to be scalable (subject to computational resources), hence algorithmic approach is sought, in the form of an end-to-end optimisation model.

The main purpose of the project is to deliver a systematic way to protect banks' customers who may be targets of fraudulent activities. Thus, the focus is on the outgoing transactions – identifying the scammers is not in scope but any interesting ideas are welcome in that regard. The bank account type is also irrelevant, only current accounts and credit cards transactions are included for simplicity.

2. Key facts about the transactional data

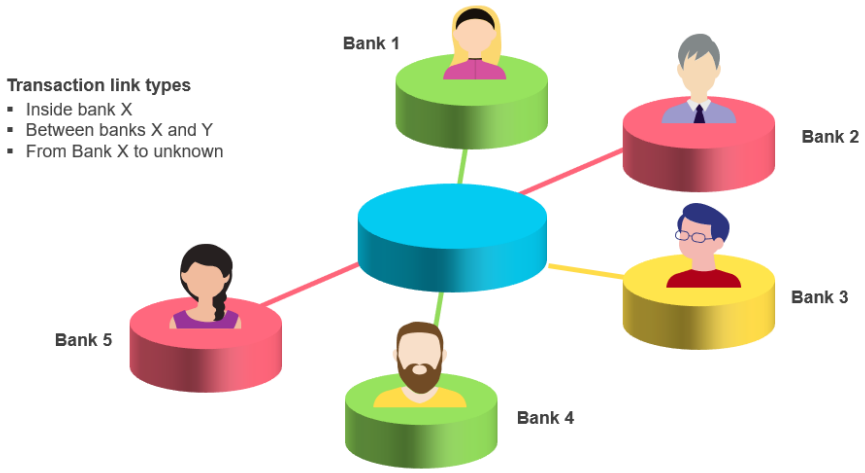
First and foremost, this data is simulated by Sopra Steria and does not include any sensitive information about customer-centric or financial features. However, it is a good representation of the 'real world' banking data, using experienced consultants' subject matter expertise to generate it.

Furthermore, no 'real world' data is perfect. Any inconsistencies in the data are possible – in fact, the provided data's quality is closer to the standards of the digital challenger banks, which have become

popular in the past several years, due to their data-driven excellence and data quality (mainly because they started from scratch and designed modern data pipelines). In comparison, some traditional banks are likely to own multiple legacy data systems, sometimes struggling to reconcile identical financial figures from different sources and often dealing with numerous data quality issues (for example, 300-year-old customers or missing payments on active loans for years).

Five banks participate in the simulation and all transactions in the input dataset originate from one of them. There are 3 types of transaction directions:

- Inside bank X (where X is between 1 and 5), i.e. both the sender's and receiver's bank account belong to the same bank;
- Between banks X and Y (where both X and Y are between 1 and 5);
- From Bank X (where X is between 1 and 5) to unknown international bank.



Cash withdrawals will obviously not point to any receiving bank and it is important to mention that as part of this project the banks have no responsibility when the funds are withdrawn at an ATM (for simplicity we do not consider stolen cards and assume all withdrawals are genuine).

Furthermore, most of the transactions listed in the dataset are outgoing (debit transactions). There are some paid in transactions as well (such as regular income, transfers from savings accounts into current account and interest received from savings). These can be useful in the analytical stage of the project, providing meaningful insights and summary dashboards. However, credit transactions (paid in) cannot be flagged as fraudulent in the current setup.

3. Fraud probabilities

- **Customer probability** - historical information about scammed customers is available in banks' books and each customer is assigned a customer probability. Based on internal historical models (which can be trusted), this probability shows the respective bank's expectation of the likelihood that the customer can be scammed. Note that higher probabilities do not necessarily mean these customers will certainly have fraud transaction – it is just an indication.
- **Transactional and description probabilities** – similarly, but at transactional level, each bank's historical data has driven calculations of how likely each transaction is to be fraudulent, based on *strictly confidential* logical conditions related to the financial or descriptive component of the transaction, respectively. Any guesses of what the internal banks' probability models take into account are more than welcome.

4. Investigators and case priorities

Each bank has their own in-house team of investigators. Initial research has shown that Bank B has a slightly larger size than the others and is much more vulnerable to scam. The other four are of similar size but Bank A is strongest in their fraud detection modelling. The banks' teams of investigators are of the following size:

Bank	Investigation team size
A	8
B	12
C	10
D	10
E	10

The main decision to be made in the model is whether a particular transaction should be investigated or not. It is up to the modelling team to come up with their criteria – it can be fully based on the transactional, description and customer probabilities or can use any other classification methods applicable to the existing data and its categories.

Note if the transaction is between two of the UK banks, they can share investigation resource, subject to availability. Otherwise, the investigation should be undertaken within the customer's home bank.

Furthermore, there is an independent third-party company which specialises in fraud investigation. Each bank is allowed to hire external investigators for individual cases with the following associated cost:

Transaction Priority	Investigation time (in days)	External investigator hire cost
1	0.25	£40
2	0.5	£60
3	1	£100
4	2	£150

Priority 1 transactions are investigated for ¼ day, Priority 2 – ½ day, hence one investigators can handle multiple of these cases per day. Priority 3 transaction require a full day of investigation and priority 4 – two days (only international ones can be priority 4 but not all of them). Remember, for priority 1 to 3 the investigation time can be split between the two banks if it is a cross-bank transaction within the UK banking system.

5. Timeline and iterative approach

The input data covers the period from 01 October 2023 until 31 July 2024. The optimisation model should run daily, with the goal to 'learn' from the effectiveness of the investigation outcomes in the previous days.

At the end of each day, the investigation outcome should be determined as follows:

		Fraud/Scam:	
		Yes	No
Investigated?	Yes	True Positive	False Positive
	No	False Negative	True Negative

It is up to the modelling team to decide how the model should 'learn' from these outcomes. The supplied data contains a dataset with all fraud transactions and the different type of scam associated with each of them. If a transaction is investigated and it does not appear in the fraud database, this is considered to be false positive, hence the investigator's time will be wasted. On the contrary, if the decision is to not investigate a particular transaction and it is included in the fraud database, this means that the funds are lost as undetected fraud. True positive and true negative cases represent success in the decision making, i.e. the investigator's time is planned efficiently.

Remember that the model runs each day but does not have access to the fraud cases data for that day prior to the end of the run.

Once the model's outcome is checked against the fraud database at the end of each day (and only the snapshot for that day), the model can 'learn' by adjusting one or all probabilities (transactional probability, description probability and customer probability) for the future days accordingly. This can be done through random variate generation following a distribution of choice and amending the probabilities as forecasted for the next day. Note that a customer can have multiple transactions per day and there is no guarantee that the one with highest probabilities would be fraudulent (if any).

6. Objectives and constraints

The ultimate goal of this project is to build a model which will 'learn' efficiently over the time and will be able to maximise the accuracy of which transactions should be investigated for fraud, based on iteratively adjusted probabilities, amount, categorisation, description, common fraud types and so on.

Here are the key objectives and these can be combined or complemented with additional ideas:

- What is the feasible solution that minimises the fraud cost for the entire banking system? The cost consists of:
  - Lost money from 'false negatives', i.e. fraud transactions that have not been investigated;
  - Additional costs for hiring extra investigators.

- What is the feasible solution that maximises the 'true positives' and saves money to the customers? Also, can you define a dual objective in combination with the previous one?

- If any of the banks is disadvantaged or advantaged in comparison to the others, can you achieve a more balanced solution where all banks are in fair competition? Note that 'fair' can be subject to interpretation by the modelling team. It does not necessarily mean equal number of cases, it could be similar proportion of detected cases, similar total funds which are successfully investigated etc.

7. Additional questions

1. What is the best approach to 'learn' from the investigation outcome on a daily basis? Can you identify any stochastic approach which takes into account transaction description and/or category, as well as the customer level probability?
2. Can you perform sensitivity analysis on the number of investigators employed by each bank? For this exercise assume that the cost of having them is a normally distributed random variate with a mean of £90 per day and a standard deviation of £5. Would it be worth for the banks to use the fixed costs external investigators and keep fewer permanent employees in these teams? Also, try to perform a sensitivity analysis of the number of external investigators in case their headcount is not sufficient, with the aim to determine the ideal number of available experts.
3. Can you identify and summarise any common differences between fraud and non-fraud transactions?
4. Is there any 'seasonal' trend in the existence and/or identification of fraud transactions?
5. How do the international transactions compare to the domestic ones?
6. Is there any difference between the investigation success rate of the internal bank transactions vs between banks and international?
7. Are any banks' customers more vulnerable to scam than others?
8. How do the transactional outliers compare between fraud and non-fraud?
9. What is the trade-off between investigating transactions of a higher priority vs these with lower?
10. How would the model change if the fraud database would not contain information about 'false negatives'? In the real-life scenario banks would only see exact information about cases that have been investigated.
11. Once you reach the last day of the timeline, can you use the trained model and start again from day 1 and achieve the following targets for 'true positives'? Try this with and without looking at the fraud cases data retrospectively.

Transaction Priority	Investigation success target
1	50%
2	60%
3	70%
4	80%

Appendix: Project Database:

Table 1: 231013_Transactions_Input.xlsx

Name	Format	Range
transaction_id	**integer**	**1 - 315182**
description	char	as seen on the bank statement
Amount	numeric	
category	char	Pre-processed: Income, Online Shopping, Utilities, Housing, Shopping, Transfers, Transportation, Dining Out, Cash Withdrawal, Groceries, Electronics, Streaming Services, Credit Card Payment, Healthcare, Interest, Bank Fees, Charity, Home Improvement, Loan Payment, Holiday, Investment, Entertainment, Personal care
date	DD/MM/YY	01/10/2023 - 31/07/2024
month	char	October - July
customer_id	**integer**	**10003 - 19951**
type	char	income/spending
In_or_Out	char	paid_in/paid_out
bank_to	char	Bank A to Bank E, Intrnl (international) and blank (cash withdrawals)
bank_from	char	Bank A to Bank E
transac_prob	numeric	0.1 - 0.8 or blank
description_prob	numeric	0.1 - 0.9 or blank
priority	integer	1-3

Table 2: 231013_Customer_Base.xlsx

Name	Format	Range
customer_id	**integer**	**10003 - 19951**
home_bank	char	Bank A to Bank E
customer_prob	numeric	0.2 - 0.9

Table 3: 231013_Fraud_Cases.xlsx

Name	Format	Range
transaction_id	**integer**	**1 - 315182**
is_scam_transaction	binary	always 1 for these cases
fraud_type	char	AdvanceFee, Impersonation, Purchase, Investment, Romance, InvoiceMandate
case_id	integer	1001576 - 8992744 (internal index for the investigators)