IV. DATA DESCRIPTION/ PREPARATION

1. Data Description

The dataset used in this project is from Dr. Marie Anderson, Machine Learning Engineer, Chief Data and Analytics Office, Lloyds Banking Group, Bristol. The raw transaction data provides a record of customer interactions, including date, time, amount, mode, and associated location. For reasons of confidentiality, the raw transaction data is subjected to agent-based simulations to generate artificial transactional data at the individual level. Our research is based on this dataset.

In the initial release of the dataset, there are 1,048,575 rows of transaction data and four variables: 'from\_totally\_fake\_account', 'monopoly\_money\_amount', 'to\_randomly\_generated\_account', and 'not\_happened\_yet\_date'. The following is a detailed description of the variables:

**from\_totally\_fake\_account**: virtual transfer out account. Character variable representing the trading account.

**monopoly\_money\_amount**: virtual transaction amount. Numeric variable representing the spend/transfer amount.

**to\_randomly\_generated\_account**: virtual to account. Character variable representing the transaction merchant/virtual personal account.

**not\_happened\_yet\_date**: virtual transaction date. Date type variable from 1 January 2025 to 31 December 2025.

本项目中使用的数据集来自Marie博士。原始交易数据提供了客户互动记录，包括日期、时间、金额、方式和相关地点。出于保密性的原因，原始交易数据经过基于代理的模拟，生成个人层面的人工交易数据。我们的研究是基于这个数据集展开的。

在最初发布的数据集中，共有1048575行交易数据，和四个变量，分别是from\_totally\_fake\_account，monopoly\_money\_amount，to\_randomly\_generated\_account，not\_happened\_yet\_date。下面是变量的详细说明：

from\_totally\_fake\_account：虚拟的转出账户。字符型变量，代表了交易账户。

monopoly\_money\_amount：虚拟的交易金额。数值型变量，代表了消费/转账金额。

to\_randomly\_generated\_account：虚拟的收款账户。字符型变量，代表了交易商户/虚拟的个人账户

not\_happened\_yet\_date：虚拟的交易日期。日期型变量，从01/01/2025到31/12/2025。

1. Data Quality