University of Bristol MSc in Data Science; DSMP (Data Science Mini Project; EMATM0050), January 2024.

Problem C: Use of Retail Banking Transactional Data
Problem owner: Dr Marie Anderson, Machine Learning Engineer, Chief Data and Analytics Office, Lloyds Banking Group, Bristol.

Lloyds Banking Group (LBG) is one of the UK's largest retail financial-services companies, with over 65,000 employees and over 30 million customers. LBG operates via several distinct brands, including Lloyds Bank, Halifax and Bank of Scotland. With approximately one in five people in the UK holding a LBG current account, we process a huge volume of transactional data. Each transaction provides a record of customer interactions including date, time, amount, method and – where relevant – location. From this we can understand customers' behaviour to find new ways to better support them. For example, we can use segmentation to flag potentially vulnerable customers that may need extra support, such as customers in financial distress who as a result might be at higher risk of falling victim to fraud.

In the Chief Data and Analytics Office (CDAO) department, we work actively with multiple business areas in LBG to develop advanced analytical and machine learning solutions to improve our customer services as well as drive commercially valuable insights and information. Raw current account transactional data provides a high volume and deeply rich source of information we can leverage. As such, we are keen to see what ideas, approaches and insights you, as Data Science MSc students, will show.

Naturally our customer data is treated with absolute confidentiality, and we cannot release even anonymized real data. As such, we have developed agent-based simulations for the generation of artificial transactional data at an individual level. These simulations incorporate a wide range of features and allow you to explore and develop ideas that could realistically be deployed on real data. It also supplies the added advantage that as we know the ground-truths which underlie the synthetic data, we will be able to assess whether your theories are correct – something that is often too difficult, impractical, or unethical to do when working with real customer data.

For teams working on this problem, you will receive data in two stages over the course of the mini-project. The initial release will be a slightly simpler, high volume data set. This should allow you to better understand the nature of the problem statement, test ideas and begin to develop your approaches. We will release a second data set, produced by a different simulation technique, later in the project which will be richer and more realistic, but will be correspondingly more challenging to work with. While you are not required to use both data sets, this will give you an opportunity to strengthen your analysis and approaches.

As we are interested in fostering a wide diversity of potential approaches and possible solutions, we are deliberately stating the problem in very broad terms. Our problem-statement is this: *"Banks generate large amounts of transactional data as a result of day-to-day operations. How do you think we should use this data?"*

We will supply you with the data; we'd like you to supply the ideas, and to develop exploratory proof-of-concept demonstrations of what you think we should be doing with our data. Good luck!

Data link:
Problem C data
(Right-click on the file and Download on your computer. DO NOT open the file using Excel.)