**Group Report**

* **Abstract** & **Introduction (10%)**
  + A brief discussion of the problem context, motivation, analysis questions/aims and proposed methods and approaches used.
  + The statement/explanation of the problem and its business context
* **Literature Review (10%)**
  + An overview of related work of similar research in the domain.
  + Discussion of methods/domain relevant prior work with citations to relevant publication
* **Methodology / Data Description/ Preparation (20%)**
  + Includes a discussion of methods applied to address your questions/aims
  + Includes description of data sources, samples and steps for pre-processing if any.
* **Results, insights, and Discussions (30%)**
  + Reporting on the experiments with discussion on insights. Technical challenges are to be discussed here too.
  + Discussion of the results, visualization (figures used), analysis, explanation of insights, relevancy to business
* **Conclusion, Further Work and Improvement (15%)**
  + A brief summary of the key insights in your report
  + Explore what can be done further based on the discussed insights and ways to improve.
* **References**
* **Appendices**
* **Overall (15%)**
  + Structure, writing quality, references format

The Key Question:

Banks generate large amounts of transactional data as a result of day-today operations. How do you think we should use this data?

Our Answer to this Question:

We believe that banks should use their customer data to identify possible user persona’s to best advertise certain financial products and to identify outlier and possibly fraudulent transaction.

**Introduction**

Lloyds Banking Group (LBG), a leading financial-services provider in the UK manages a substantial proportion of the country’s banking transactions, through its various subsidiaries, Lloyds Bank, Halifax and Bank of Scotland. With approximately 30 million customers, everyday LBG generates voluminous data, detailing customer interactions, including the date, time, amount, methods and often the location of the transaction.

Retail banks continue to face threats caused by disruptive FinTech firms, like Paypal, Monzo and Revolut, as well as the rise of independent financial aggregators such as the UK’s MoneySuperMarket.com providing comparison tables, seeking to optimise customers’ financial holdings by analysing purchasing patterns across their customer base. Moreover, large tech companies (Google, Apple) may unsettle traditional banking by integrating financial services into their ecosystem. These disruptions instil motivation for LBG to use their transactional data in the most productive and innovative methods, to increase customer satisfaction, lower costs and churn rates, ensure customers retain a high level of trust, and remaining generally competitive.

There are many possibilities of how to use transactional data to best adhere to the objectives above, for example customer segmenting (through life milestones reached such as marital status or whether they have children, attitudes to spending, preferred methods of transactions, and their network with other bank accounts), fraud detection, lending decision models, predictive analysis, personal finance management, and creating customer value analysis.

However, it is imperative all proposed solutions align with the ethical framework and privacy notice established by LBG, comply with regulatory mandates as stipulated by the Financial Conduct Authority (FCA) and the Prudential Regulation Authority (PRA), and should embody LBG's core values: putting people first, being bold, inclusive, sustainable, trustworthy. In support of this, LBG has provided two artificial datasets generated through agent-based simulations, ensuring that real customer data remains confidential. Moreover, when proposing solutions, we ensure that protecting the customer and their data is the top priority, even if it means forgoing potential profit.

This paper outlines methods in how we believe LBG should best use their data, focussing on two key themes: personal finance and anomaly detection. The personal finance theme explores [fill in- explain personal finance and examples…], helping LBG to remain innovative. We utilise a recency, frequency and monetary (RFM) analysis and then applying unsupervised learning techniques, clustering customers based on spending habits, which can be used to target specific market campaigns, increasing customer personalisation and satisfaction. The anomaly detection topic investigates various methods of identifying abnormal datapoints in the datasets, which can serve as a foundation to finding fraudulent transactions (which requires labelled data to whether a transaction is fraudulent or not). We create anomaly detection models by customer segmentation, by frequency of daily transactions, by unusual hourly activity and by isolation forest, finding that the customer segmentation is the most promising technique.

Note : mention the two datasets in the intro

**Literature Review**

**Personal Finance**

**Segmenting customers via RFM techniques**

# In retail banking, a one-size-fits-all approach to product sales is not very effective. Tailoring services to meet the distinct needs of different customer groups can lead to greater satisfaction and profitability. Calculating an RFM score for each customer can be useful in analysing behaviour to improve marketing strategies. However, several papers found this method limiting, and created more successful models by applying machine-learning algorithms to RFM values. The study “Segmenting Bank Customers via RFM model and Unsupervised Machine Learning” compared three algorithms: K-means, density-based spatial clustering of applications with noise (DBSCAN), and Hierarchical clustering. It concluded that DBSCAN outperformed K-means, while Hierarchical clustering was computationally most intensive. Other papers such as “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market” found applying a Gaussian Mixture Model (GMM) to RMF values was the most valuable out of four algorithms. For this project, we evaluate the effectiveness of K-means, DBSCAN, and GMM algorithms on the enhanced RFM data.

Statistics ISAs:

<https://www.gov.uk/government/statistics/annual-savings-statistics-2023/commentary-for-annual-savings-statistics-june-2023>

Deloitte Insurance Review:  
<https://www2.deloitte.com/content/dam/Deloitte/us/Documents/strategy/us-cons-life-insurance-consumer-study.pdf>

Lloyds package bank account recommendation:  
<https://www.moneysavingexpert.com/banking/best-packaged-bank-accounts/>

Personal Current Account market study update (Covers packaged accounts page 121):

<https://assets.publishing.service.gov.uk/media/53c834c640f0b610aa000009/140717_-_PCA_Review_Full_Report.pdf>

Pet insurance market:  
<https://www.lloydsbankinggroup.com/media/press-releases/2021/halifax/brits-are-three-times-more-likely-to-insure-their-pets-than-themselves.html#:~:text=URL%3A%20https%3A%2F%2Fwww.lloydsbankinggroup.com%2Fmedia%2Fpress,100>

Shows that their a significant complaints around package bank accounts, these stem from people not being able to afford their account, and the offerings of the account being misrepresented and ill-fitting for the people that held the accounts.

Lloyds banking group prides itself on its consumer facing banking practices. When looking through the many services it offered to consumers four main segments stuck out to us. The first was the various current accounts LBG offers, second was the various insurance products LBG offers, third was the different ISA products that LBG offers, and finally was the various debt-based products that LBG offers. This wide array of offering surprised me, I spoke with people who had LBG current accounts, who were completely unaware of these offerings. Many of which I know would have benefitted from such products.

**Current Accounts**

Starting with the current accounts, Lloyds offers four tiers, the first is the standard free to open current account. For three pounds a month a user can upgrade to club Lloyds which offers up to 6.25% interest on the first 5000 pounds within the accounts, and a free yearly benefit such as Disney+, cinema tickets, an annual coffee club subscription, or a magazine subscription. This fee is also waived if the user pays in over 2000 pounds to the bank account. For a further 10 pounds or 21 pounds a month for silver and platinum respectively, Lloyds offers a variety of insurance benefits.

Many of the costs associated with the various accounts seem small considering the benefits on offer, despite this package accounts have a very high complaints rate across various UK institutions. The Personal Current Account Market Study Update written by the Competitions and Market Authority cites that this high number of complaints as a symptom of these accounts being misrepresented and ill-fitting for the people that held these accounts. With the available data the group has, we saw this as an opportunity to change how these accounts could be advertised.

**Insurance**

Lloyds offers five different types of insurance on their website, this includes home, life, critical illness, car, and van. Further, Lloyds has had a rough time with the pet insurance market, where it was forced to both exit due to rising costs and then re-enter when large enough consumer pressure forced them back in the market. It’s clear that it would benefit Lloyd’s banking group to greater understand the insurance market, as this is a point of both struggle and controversy for the firm. When looking for an insurance market analysis, the group found a report from Deloitte which reported who the prime customers for different insurance products are. The reported noted that greater insurance spending was correlated with specific events in people’s lives, mainly getting married, having kids, and buying a house. Further, these customer personas namely couples, parents, and homeowners are the largest market for insurance. Further, over the past couple decades there has been a very fast-growing pet insurance market, Lloyds themselves cites that brits are three times more likely to ensure their pets than themselves. With the data available the group believes that it can create a model that better targets these user personas based on the findings in the Deloitte report. These consumers would not only be the most likely to be looking to buy insurance but also the most likely to already have insurance. Leveraging this would allow LBG to not only acquire new market participants but convert already existing market participants by offering better rates on insurance and making it easy to switch.

**Individual Savings Account**

The individual savings account or ISA is a UK tax advantaged account that allows investors and savers alike to reap the gains of their investments tax free. All dividends, interest earned, and capital gains within the account is entirely shielded from tax. With these benefits it’s no surprise that money within ISA platforms has grown meaningfully, from around 55 billion pounds to 65 billion pounds in the decade prior to the tax year ending 2022. This enormous sum is split roughly between Cash ISA’s and Stocks and Shares ISA with a minority of money being held in the Innovative Finance ISA’s and Lifetime ISA’s. As someone with an interest in personal finance, I heard about the LBG ISA offerings through an advertisement I saw on YouTube prior to the start of this project. Out of interest I looked it up thinking that LBG would follow the same path of their other big 4 peers where the investment products they offer are low return and high cost. I was personally shocked to find that the ISA product offered was not only comparable to the lowest cost options but arguably cheaper.

The standard share dealing ISA on offer has an admin fee of 20gbp every 6 months, with a commission fee of 1.5gbp for funds. The admin fee is free if the subscriber is between 18-25 years old, and the commission fee for funds is also free with a regular investment plan. Within the ISA, subscribers can buy industry leading funds offered by BlackRock iShares which are both globally diversified and low cost. Comparing this to the second cheapest stocks and shares ISA offered by Barclays, they charge a fee of 0.25% of the assets held within the isa and has a 6gbp commission. The most expensive stocks and shares ISA offered by NatWest has a charge of roughly 0.62%.

Given the data available, it would be possible to create targeted advertisements for this product based on the user persona of common ISA subscribers outlined in the UK governments own findings.

**Anomaly Detection**

**Methodology**

**Personal Finance**

The first stage in creating the personal finance solutions was to segment the customers through an RFM analysis. For each unique account in the datasets, this involved calculating:

* Recency values = The number of days between the account’s most recent transaction and the newest transaction in the entire dataset
* Frequency values = sum of transactions the account has made during the time frame of the dataset
* Monetary values = the sum of the total value of the transactions made by the account

These RFM values were used to create models built on 3 different algorithms: K-means, GMM, and DBSCAN. K-means, which is centroid-based, identifies a predefined number of clusters by minimizing the Euclidean distance between points and cluster centres. The optimal number of clusters was established at three for both datasets through the Elbow Method, as illustrated in Figures X and Y. The GMM algorithm, recognising the probability distributions in the data, presumes a Gaussian distribution for points within a cluster, and it was determined that using three components yielded the highest silhouette scores, indicating well-defined clusters. DBSCAN does not require a predetermined number of clusters as an argument but requires trial and testing to find the best epsilon value (specifies the radius of the neighbourhood around a point) and min\_samples value (sets the minimal number of points required to form a dense region). The eps and min\_samples values were chosen as 0.07 and 40 respectively.

A graph with a blue line

Description automatically generatedA graph with a blue line

Description automatically generated

**Anomaly Detection**

User centric model

Isolation forest.

**Data Description and Preparation**

**Data Description**

Dataset 1, characterised by high volume but simpler attributes including 'account number', 'amount', 'third party account', and date, was utilised for preliminary testing of models. Dataset 2 though smaller in scale, offered a higher degree of realism through the addition of 'balance' and 'timestamp' columns, incorporating expenditures and payments, as well as reflecting existing businesses and realistic payment patterns (e.g. monthly income, rent, and subscriptions). Consequently, Dataset 2 was employed to refine the models, ensuring they were adjusted to fit the complexities of transactional data.

**Pre-processing**

For the three unsupervised models used in RFM, the data was scaled using StandardScaler.

**Industry Mapping**

A key part of the different analysis the group accomplished was effectively mapping individual firms to their respective industries. At first, we saw that it may be possible to map companies by their names inferring the industry they may be involved in. After due discussion we decided to take a more informed approach. For the first dataset we grouped companies by their revenues which worked well but was unsatisfactory for the intricacies of the second dataset. To aid this, we employed the use of K means clustering that grouped the businesses based on unique identifiers such as their payment frequency, unique payment days and times, and the value of these payments.

**Results**

**RFM**

* Display the graphs for dataset 1 and 2
* Present the metrics in tables (Silohouette scores, CHI, DBI)

**Insights and Discussion**

**RFM**

* Discuss how it can be used
* Assumptions eg. GMM assumes guassian distribution

Things to do:

Literature Review:

* Anomaly Detection (Diya)

Methodology

* Discuss methodology for Personal Finance (AQ)
  + Discuss methodology used for RFM (JB)
* Discuss methodology for Anomaly Detection (JB)

Data Description/ Preparation:

* Describe first and second dataset (AQ) Note to Amaan: I (Jessie) have started this you might want to add/change bits
* Data preparation for the first and second dataset (AQ)

Results:

* Personal Finance
* Anomaly Detection

Insights and Discussion:

* How do people behave?
* How could this be used to benefit the bank?
* Ethical Discussions

Conclusion:

* Summary of all the insights
* What could be done further: Notably, both datasets are fixed, however in actuality LGB will be receiving an incoming stream of data, meaning our models would need to be adapted to suit a stream