**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 : make aims and objectives more clear

**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>

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: current accounts, insurance products, ISA’s, and debt offerings. This array of offerings was surprising as although the group was aware of LBG as a brand, they were unaware of the offerings.

**Current Accounts**

LBG offers four tiers of current accounts, the standard offering, for three pounds more a month a club Lloyds account, for ten pounds more a month a silver account, and for twenty-one pounds more a platinum account. The club Lloyds account gives consumers one of a selection of benefits that include Disney+, cinema tickets, and an annual coffee club subscription, alongside greater interest on savings with the further upgrades give access to insurance. Despite the variety of benefits that these accounts offer at below market rate, these types of package accounts receive many complaints. The Personal Current Account Market Study Update written by the Competitions and Market Authority cites that this is due to these products being misrepresented and ill-fitting for those that held the account. With the data provided LBG has the power to change this and shape the industry for the better.

**Insurance**

Lloyds advertises five different types of insurance on their website, this includes home, life, critical illness, car, van, and has had negative previous experience with the pet insurance market. This was due to LBG misunderstanding the costs associated with pet insurance, exemplifying the need for the bank to greater understand this industry and possible buyers. A report discussing the insurance market released by Deloitte

(Work ended here)

Lloyds advertises 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**

Anomaly detection is a significant task in the banking industry to detect any outliers and mark them as anomalies to prevent fraudulent transactions. Implementing various methods to analyse and understand the historical transactions is a key factor in maintaining the bank’s overall efficiency. There are numerous methods to perform anomaly detection and many studies compare the efficiency as well as the results of these techniques.

In An Evaluation of Unsupervised Outlier Detection Methods for Univariate Time Series Data in Financial Transactions, various methods are performed to compare and conclude which is effective in the finance industry. The Interquartile Range (IQR) was used as a statistical method to identify outliers. This method relies on the assumption that the data distribution is close to normal, which may not always hold true in real-world scenarios and could be considered more effective if used in conjunction with other techniques. Isolation Forest was highlighted as an efficient algorithm for anomaly detection in high-dimensional data. It isolates anomalies instead of profiling normal data points, making it effective for datasets with a mixture of normal and anomalous points but can struggle with complex dependencies in time series datasets.

In the study On the Detecting Anomalies within the Clickstream Data: Case Study for Financial Data Analysis Websites, the IQR method was employed to label the anomalies. It was performed by calculating the IQR for the data, defined as the difference between the third quartile (Q3) and the first quartile (Q1) of the dataset. Followed by labelling sessions as anomalous based on the number of outlier features detected in each session. This method provides a straightforward statistical technique to detect values that deviate significantly from the central tendency of the data, which further helps in finding unusual user behaviours that could indicate potential threats on financial websites.

**Methodology**

**RFM**

The first stage in creating personal finance solutions was segmenting 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

The RFM values were used to create models built on 3 different algorithms: K-means, GMM, and DBSCAN. The optimal number of clusters for K-means and GMM was established using the Elbow Method at either three or four as illustrated in Figures X and Y, which were both tested. 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 0.18 and 85 respectively.

It was decided to omit ‘Recency’ from the three clustering models, based on histogram analyses in Appendix A, revealing in datasets 1 and 2, all customers were recent spenders, having made transactions within the last 5 or 7 days of the observed period, respectively (apart from one anomalous account in dataset 2). This lack of variation implied that 'Recency' would not contribute meaningful segmentation among customers. In the models applied to dataset 2, the median 'Balance' was introduced as an additional variable, providing more nuanced customer segmentation.

A graph with a blue line

Description automatically generatedA graph with a blue line

Description automatically generated

**Personal Finance**

To leverage the complex user transactional data and transform its complex nature into manageable chunks, rigorous clustering and filtering were initiated which made use of these painstakingly generated datasets. Users were categorised according to their unique spending pattern, the frequencies of their transactions, and their tendencies to patronise particular businesses by the utilisation of sophisticated clustering and filtering techniques. This fine-grained segmentation proved instrumental in identifying each person’s financial routines and revealed commonalities with other groups of users. These insights helped in the construction of recommendation engines that were entrusted to recommend the best-suited banking product–be they savings accounts, current accounts, Individual Savings Accounts (ISA), or different insurance products. The data was then used to determine the relationship between spending patterns and the best financial solutions to the user's need, allowing financial institutions to tailor their offering to meet the demands of specific customers. For instance, individual users who spent small amounts and with high frequency may have been presented with different banking accounts and product possibilities than those who spent larger amounts and with less frequencies. Similarly, the engine was able to identify possible candidates for ISAs or insurance policies that matched their financial aims and based on lifestyle and frequency of interactions with different company sectors recommend the most appropriate products and services. (closing stanza)

**Anomaly Detection**

*User centric model*

One method of creating an anomaly detection model was to build dataframes containing thresholds of spending personally tailored to each LBG account. Each LBG account had three ‘Overall Thresholds’, three thresholds for specific third-party accounts they interacted with, and three thresholds for industries (dataset 1) or businesses (dataset 2) they interacted with. The model worked so when a new transaction was processed, if the LBG account had previously interacted with that third party account or business, if the amount was below threshold 1 it would be classified as ‘No risk’, if it was between thresholds 1 and 2- ‘Low risk’, between thresholds 2 and 3- ‘Medium risk’, and if it exceeded threshold 3 it would be classified as ‘High risk’. If the account had not previously interacted with the third-party account or business, then the ‘Overall Thresholds’ would be used. Thresholds X were calculated using Equation X, where Q3 is the third quartile, the multiplier was either 1.5, 3, 5 depending on if X =1,2,3, and IQR is interquartile range of either their overall spending, spending for that specific account or business they interacted with.

ThresholdX= Q3 + multiplier \* IQR

This method was initially tested on the first dataset (allowing for experimentation such as tweaking multipliers) and then further adapted to fit the complexities of the second dataset. For instance, in Dataset 2 the presence of distinct expenditure and payment patterns for each LBG account made it infeasible to use a single dataframe. Consequently, separate dataframes were created for payments and expenditures for each account to accommodate the unique data characteristics, and the model had to be iterated, assessing whether a transaction was an expenditure or a payment.

*Isolation forest*

The isolation forest technique is a recently well-known algorithm to detect outliers. The method uses binary trees to find anomalies, which leads to a linear time complexity and low memory consumption, making it ideal for handling big datasets. Randomly subsampled data is processed in a tree structure using randomly chosen features in an isolation forest. Since it took more cuts to separate the samples farther into the tree, they are less likely to be anomalies. Comparably, samples that end up on shorter branches also point to anomalies because the tree could more easily distinguish them from other observations.

The first dataset was used to extract all the business side of transaction that is all transactions happening between user and a business. The data was pre-processed to extract unique business names and create a dictionary to hold individual business dataframes. Following this, the isolation forest was applied to the business dataframes with contamination of 0.50. The resulting dataframes had a column with the anomaly score and another column which indicated whether the transaction was anomaly where ‘1’ meant it is not an anomaly and ‘-1’ as anomaly. A horizontal bar chart was plotted to visualize the anomalous transactions resulting from this model. This method was used as a base to adapt it to create a user-centric model with the second dataset.

The second dataset varied from the first and initial steps involved preparing the dataset for analysis by modifying certain columns to a specific format. After preprocessing, data was split into training set of 80% and test set of 20%. The transactions were then categorized into two types namely payments that had positive transactions and expenditures where it had negative amounts which was converted to positive for further steps. This was followed by feature engineering to provide a multi-dimensional view of the transactional behaviour. This included extracting the hour, day of the week, day of the month and month of the year depending on the transactional timing. Summarising of the transactions were performed by aggregating transactions by account number and the third party names. The isolation forest model was trained for each unique account number, setting the contamination factor to 0.01 which flags 1% of the data as anomalous. Then the model was executed on the test set to identify anomalies. In the results, a prediction column was appended to the dataframe where ‘1’ indicated the transaction was not an anomaly and ‘-1’ as anomalous transaction. This was further categorized into low risk, medium risk and high risk anomalous transactions. The risk categorization was performed by computing the Interquartile Range (IQR) for each business. By incorporating the risk categorization system, the model has enhanced the anomaly detection framework and providing valuable insights into transactional data.

**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.

The entries of the first dataset are designed such as to indicate their use cases are suitable for prioritising creativity and engagement rather than data authenticity, with entries that span over various monetary transactions directed to quirky entities like ‘HIPSTER\_COFFEE\_SHOP’ and ‘TOTALLY\_A\_REAL\_COFFEE\_SHOP’, and all of them being timestamped to a future date of ‘01/01/202. The comprehensive nature and significant size indicate its application is more logical towards testing scenarios. On the other hand, the second dataset comes across as being more realistic and structured by its nature. It is successful in depicting a transaction log which is typically found in various banking sectors. It includes over 230,000 entries and despite also being simulated, the data is presented with realistic account numbers, third-party names like ‘Westport Care Home’, and date stamps that start from ‘01/01/2023’, which hints at a more serious intention for simulation, such as to be utilised for data analysis and software testing.

**Pre-processing**

*Cleaning*

Dataset 1-

Dataset 2: For the second dataset, there was a total of 1402 unique rows with null values (0.6% of the dataset) for Date, Timestamp, Account No, Balance and Amount. These are crucial pieces of information required for further analysis, and since they represent a small proportion of the dataset, it is acceptable to remove them. These missing values were saved to a separate dataframe, as it can be useful to investigate why these rows contain null data, potentially due to systemic issues. Moreover, the Date and Time column were parsed into a Datetime format and combined to create a ’Datetime’ column.

The initial analytical framework’s underlying datasets were carefully prepared to fulfil a range of analytical and machine learning objectives. Originally, diverse datasets were derived with a specific primary focus to enable contrasting model training and testing purposes along with in-depth exploration of data. Among various datasets that were devised one such dataset featured user-to-user transactions, which meticulously captured the peer-to-peer financial interactions, and this was a substantial aspect of understanding the social and communal characteristics of monetary exchange. Another dataset zoomed in on user-to-industry transactions which detailed the myriad ways an individual engages oneself with various sectors, this illuminated patterns of spending and an overlook of consumer behaviour concerning the businesses.

The devised dataset that was leveraged extensively was crafted using user-to-user transaction data and it represented a culmination of data cleaning and preparation efforts. It was subjugated to a rigorous process of data filtering to weed out irrelevant entries to ensure smooth analysis. This ministered dataset then underwent initial analysis which included breaking down the transactional behaviours, spotting key trends and also creating a baseline statistic for a range of variables which included transactional volumes, frequencies and values. This analytical stage was an essential part of developing the foundation for more sophisticated analysis as well as for developing preliminary hypotheses for user behaviour.

The core of this dataset's utility was its application to clustering algorithms which was particularly for generating recommendations for various account types offered in the banking industry based on spending history and financial behaviour and was crucial for the RFM analysis {add more..}. It aided in identifying user segments based on patterns of transactional activity by converting the raw transactional data into actionable insights. Application of clustering methods like K-means was made possible by the pre-processed dataset which had been denoised and normalised as required. This enabled the customization and marketing campaigns of banking products for these specific user groups on a minute level.

Moreover, with extensive exploratory data analysis (EDA) complete on these datasets incorporating visual representation of data was made possible, employing statistical graphs, plotting and other methods helped to uncover underlying structures, spot outliers and to test fundamental assumptions about the nature of the market. The heavily data-driven approach for focused user interaction and then the creation of more advanced predictive models later was greatly influenced by this exploratory step.

The creation of these datasets was a meticulous and strategic endeavour and the primary goal in sight was the optimization of the value of data contained in the transactional records. To guarantee that the datasets reached the analytical objectives whether they were for clustering, recommendation systems, or in-depth transactional studies, it involved immense manual labour and hence paved the way for useful findings and robust financial models.

*RFM*

For the models created using RFM values, the data was normalised using StandardScaler, ensuring no single feature dominates others due to its scale, which is important for distance-based algorithms such as K-means.

*Anomaly Detection*

For the *user centric anomaly detection model*the datasets were split at random into training and testing sets with a 80:20 split. The training data was used to create a dataframe of each user and their typical spending habits, and the testing data was used as ‘new transactions’, which were then classified using the model, as either No, Low, Medium or High risk transactions.

**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 - Dataset 1**

Figures A, B and C visualise the clusters generated using K-means, GMM an DBSCAN on the monetary and frequency values, with each point representing a customer, colour-coded by the cluster it belongs to, indicating groups of customers with similar ‘M’ and ‘F’ values. Each model was evaluated using three metrics: Silhouette Scores (how similar an object is to its own cluster compared to others, with a higher score indicating better defined clusters), Calinski-Harabasz Index (evaluates clusters based on the mean between-cluster variance divided by the within-cluster variance, with higher values indicating better cluster separation and more defined clusters) and Davies-Bouldin Index (Determines the average 'similarity' between clusters, where lower values signify clusters that are further apart and less similar, which is desirable for well-separated clusters). Table X indicates a four-cluster solution yielded the most effective results for both K-means and GMM, with K-means slightly outperforming GMM. Conversely, DBSCAN optimal parameters generated three clusters, with its performance on dataset 1 being significantly inferior to the other two algorithms.

**RFM - Dataset 2**

Figures X and Y show 3D scatter plots visualising clusters resulting from K-Means and GMM analysis, using Frequency, Monetary and Balance values for each customer. Table X indicates that segmenting the customers into three clusters creates slightly better models for both K-Means and GMM, with K-Means outperforming GMM in all three metrics.

A diagram of a diagram

Description automatically generated with medium confidence

**A screen shot of a graph

Description automatically generatedA graph of data on a white grid

Description automatically generated with medium confidence**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Dataset 1** | | | **Dataset 2** | | |
|  |  | **Silhouette Score** | **Calinski-Harabasz Index** | **Davies-Bouldin Index** | **Silhouette Score** | **Calinski-Harabasz Index** | **Davies-Bouldin Index** |
| **K-Means** | **3 clusters** | 0.47 | 7908.38 | 0.76 | 0.43 | 898.36 | 0.85 |
| **4 clusters** | 0.48 | 8642.27 | 0.70 | 0.43 | 821.46 | 0.96 |
| **GMM** | **3 clusters** | 0.47 | 7603.15 | 0.76 | 0.38 | 721.37 | 0.92 |
| **4 clusters** | 0.47 | 8315.67 | 0.69 | 0.36 | 608.36 | 1.19 |
| **DBSCAN** | **3 clusters** | 0.20 | 1846.02 | 1.48 | N/A | N/A | N/A |

**Anomaly Detection**

*User centric*

A graph of a number of long columns

Description automatically generated with medium confidence

**A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated**

*Isolation Forest*

A green rectangular bar with red and yellow text

Description automatically generatedA graph with numbers and lines

Description automatically generatedA graph with a line and a blue line

Description automatically generated

**Discussion**

**RFM**

In both Figures A and B, the K-means and GMM algorithms similarly segment customers into four groups (Table X) with only slight variations in borderline classifications. DBSCAN (Figure C), forms three clusters, with green representing customers who spend moderately frequently but at a higher amount, blue showing customers who spend moderately frequently at lower amounts, and yellow illustrating customers who spend most frequently, at moderate amounts. However, DBSCAN is limited, with 37.71% of the datapoints marked as ‘noise’, which is not useful if LBG use this model for customer segmenting, as over a third of their customers would be excluded. However, these ‘noise’ datapoints may be helpful for identifying unusual spending pattens. Since DBSCAN model’s performance was lower than the other models, it was not applied to the second dataset. Figures X and Y again show K-means and GMM closely aligning in their clustering of customer segments, which are detailed in Table Y.

Dataset 1

|  |  |  |
| --- | --- | --- |
| **Colour of cluster in K-Means** | **Colour of cluster in GMM** | **Description of customer segment** |
| Blue | Yellow | Spend less frequently, and spend low amounts |
| Yellow | Blue | Spend more frequently but a lower amount |
| Green | Purple | Spend less frequently but higher amounts |
| Purple | Green | Spend the most often and the highest amounts |

Dataset 2

|  |  |  |
| --- | --- | --- |
| **Colour of cluster in K-Means** | **Colour of cluster in GMM** | **Description of customer segment** |
| Blue | Green | Spend least frequently, the least amount, and have the lowest bank balance |
| Green | Blue | Spend moderately frequently, a high amount, with a moderately low bank balance |
| Orange | Orange | Spend the most frequently, a high amount with the highest bank balance |

For LBG, segmenting customers using clustering algorithms like K-means (which proved most effective across all three metrics for both datasets) is key for targeted marketing. For instance, premium products could be tailored for higher, frequent, and most recent spenders as well as the accounts keeping more of their money with LLoyds. Excluding 'Recency' from models was appropriate here due to synthetic patterns of the data, but LBG could incorporate it to their models with real transactions and accounts. Furthermore, given the voluminous nature of transactional data, the models’ computational complexity needs to be considered. Given the number of data points (n), the number of clusters (k), the number of dimensions (d), and the number of iterations required for convergence (i), K-means tends to be the most computationally efficient algorithm *O*(*n*⋅*k*⋅*d*⋅*i*), especially when the number of clusters (k) is much smaller than the number of datapoints. DBSCAN can also be computationally effective, with an average complexity of *O*(*n*log*n*), and GMM is generally the least computationally effective of *O*(*n*⋅*k*2⋅*d*+*k*3⋅*d*3) for each iteration. Moreover, both datasets are fixed in nature, yet as LBG will be receiving an incoming stream of data, they must continuously refresh customer segments for relevant marketing.

When clustering RFM values, several ethical considerations must considered. For example, LBG must ensure customer data is collected and used in compliance with privacy laws such as GDPR or CCPA, and customers should be informed and give their consent about what data is collected and how it will be used. Moreover, there must be transparency about the use of data analytics in tailored marketing, and it should not exploit vulnerable customers or encourage irresponsible spending.

**Anomaly Detection**

Banks are vulnerable to fraud and attacks, and therefore implementing anomaly detection- identifying datapoints that deviate from expected transactional patterns- is the first line of defence. Anomaly detection benefits LBG as it protects customers, fosters their trust, and can boost retention rates. Our results show tailoring models to individual accounts enhances effectiveness of the models, recognising that a £500 spend at Gap Kids may be anomalous for one account but normal for another.

Moreover, having a granular system of classifying transactions at different risk levels helps to LBG prioritise which transactions needs the Bank’s attention, as well aggravating the customers less by not freezing their accounts when a transaction only slightly deviates from their usual spending. Table X suggests subsequent actions LBG can initiate, following a transaction being flagged at a certain risk level.

Table X

|  |  |
| --- | --- |
| **Risk Level** | **Action** |
| No Risk | No Action Required |
| Low Risk | Monitor transactions closely. Block contactless and request PIN. Notify customer of unusual activity via secure message or email, to confirm it was them |
| Medium Risk | Temporary hold on transaction. Direct customer contact |
| High Risk | Account freeze. Full investigation. Customer support and resolution. |

Implementing the model required striking a balance between leniency and fraud prevention. With customer safety as LBG’s main priority, our design opted for a cautious approach, flagging about 5% of transactions and designed to capture more false positives, rather than overlooking true fraud. However, LBG can modify these settings (such as the multipliers) or explore alternative metrics such as deviations from the mean, to best align with their objectives. The model faced difficulties with recurring transactions, like monthly salaries, subscriptions, and fixed bills, which have consistent amounts and hence, an IQR of zero. This caused them to be inaccurately flagged as anomalies. To resolve this, a revised threshold calculation excluding IQR was used (Equation Y) using a percentage increment which varied depending on the threshold X, ensuring regular payments were not mistakenly flagged. Upon evaluating the model used on Dataset 1, it emerged that smaller transactions, like coffee purchases, had a low IQR and were being flagged as high-risk regardless of the multiplier used. To address this, the model was refined to include a post-risk assessment, adjusting the risk level based on the transaction's absolute value—setting benchmarks at £15 for low risk, £50 for medium, and £100 for high. Lastly, anomaly detection models must adapt to evolving customer spending habits. Thus, transactions deemed no-risk or falsely flagged can be updated into the account's profile of typical transactions, continually refining the threshold dataframe.

Equation Y: Threshold X= Q3 \* (1+ percent\_increment)

Using the isolation forest in the first dataset was performed by setting generic parameters which lead to a broader detection of anomalies, giving rise to more false positives. While creating the user centric model in the second dataset a more refined approach was taken to yield better results. The parameters were fine-tuned, potentially reducing the false positives, and increasing the number of true anomalies being detected. Furthermore, the risk categorization provided a more nuanced view of the transactions and gave more insights into the anomalous transactions.

**Conclusion**

Things to do:

Intro (JB)

Literature Review:

* RFm (JB)
* Anomaly Detection (DM)
* Personal Finance

Methodology

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

Data Description/ Preparation:

* Describe first and second dataset (AQ)
* Data preparation for the first and second dataset (AQ)

Results:

* Personal Finance
* RFM (JB)
* Anomaly Detection (waiting on results)

Insights and Discussion:

RFM (JB), personal finance, anomaly detection user centric (JB), anomaly detection isolation forest

* 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:

**Appendix**

**A green and black graph

Description automatically generated**

**A group of different colored bars

Description automatically generated with medium confidence**

In Fig. \ref{fig:rfmresultsdataset1}, the K-means and GMM algorithms similarly segment customers into four groups (Table \ref{table:dataset1\_segments}) with only slight variations in borderline classifications. The DBSCAN model (Fig. \ref{fig:rfmresultsdataset1}), forms three clusters, with green representing customers who spend moderately frequently but at a higher amount, blue showing customers who spend moderately frequently at lower amounts, and yellow illustrating customers who spend most frequently, at moderate amounts. However, the DBSCAN model is limited, with 37.71\% of the data points marked as ‘noise’ (purple coloured), which is not useful if LBG uses this model for customer segmenting, as over a third of their customers would be excluded. However, these ‘noise’ data points may help identify unusual spending patterns. Since the DBSCAN model’s performance was lower than the other models, it was not applied to the second dataset. Figs. \ref{fig:rfmkmeans} and \ref{fig:rfmgmm} again show K-means and GMM closely aligning in their customer segment clustering, in Dataset 2, detailed in Table \ref{table:dataset2\_segments}.

For LBG, segmenting customers using clustering algorithms like K-means (which proved most effective across all three metrics for both datasets) is key for targeted marketing. For instance, premium products could be tailored for higher, frequent, and most recent spenders as well as the accounts keeping more of their money with LLoyds. Excluding 'Recency' from models was appropriate here due to synthetic patterns of the data, but LBG could incorporate it into their models for real transaction data where the recency values vary more. Moreover, Datasets 1 and 2 are static, yet in reality, LBG will be receiving an incoming stream of data. Therefore, LBG must continuously refresh clustering customer segments to maintain relevant targetted marketing.

Given the voluminous nature of transactional data, the models’ computational complexity needs to be considered. For the number of data points (\( n \)), the number of clusters (\( k \)), the number of dimensions (\( d \)), and the number of iterations required for convergence (\( i \)), K-means tends to be the most computationally efficient algorithm \( O(n \cdot k \cdot d \cdot i) \). DBSCAN can also be computationally effective, with an average complexity of \( O(n \log n) \), and GMM is generally the least computationally effective with \( O(n \cdot k^2 \cdot d + k^3 \cdot d^3) \) for each iteration.

When clustering RFM values, several ethical considerations must considered. For example, LBG must ensure customer data is collected and used in compliance with privacy laws such as GDPR or CCPA, and customers should be informed and give their consent about what data is collected and how it will be used. Moreover, there must be transparency about the use of data analytics in tailored marketing, and it should not exploit vulnerable customers or encourage irresponsible spending.t