Personalised Product Recommendations and Fraud Detection Utilising Unsupervised Learning

Don Issac Joseph
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
ff19734@bristol.ac.uk

Xinyue Zheng
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
fa23233@bristol.ac.uk

Isarapon Prasertstid
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
ws23365@bristol.ac.uk

Rajasree Rajan Unnithan
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
vd23936@bristol.ac.uk

***Abstract*—With the ever-increasing volume of card payments in the UK, commercial banks can access vast quantities of transactional data to enhance customer experiences by personalising services and product recommendations. This report introduces a customer segmentation model that classifies customers into 11 distinct segments from which spending insights from accounts in each segment are used to formulate product recommendations. An XGBoost classifier is then trained to directly predict the customer segment a customer might belong to. This model achieved a 99% F1 score on test data. Furthermore, risk and targeted recommendation models were developed to ensure that products are appropriately and ethically recommended to customers. Finally, an unsupervised fraud detection model was also constructed which flagged 168 anomalous transactions that may relate to fraud.**

I. Introduction

Commercial banks are for-profit organisations that maintain an important role in the UK and other global economies. They provide a critical service to individuals that allows them to store, safeguard and deposit their money whilst also facilitating access to large amounts of capital through loans. The primary source of revenue for commercial banks is the interest earned on loans - which is mainly funded through deposits in customer and business accounts [1]. Interest revenues are however constrained not only by a bank's ability to attract and retain customers - but also by market forces and regulatory limitations that influence lending interest rates and lending capacities. For example, the interest rates a bank can offer are fundamentally limited by the Bank of England and dependent on dynamic macroeconomic conditions. Banks that opt to increase lending by offering more loans face diminishing returns as competition within the market necessitates offering lower interest rates, thereby reducing average profits per loan [1].

Consequently, to preserve business outcomes - it is not only important for commercial banks to expand and maintain their customer base to accumulate funds in deposits for lending but to also diversify revenue streams to ensure robustness to market changes. Many commercial banks such as Lloyds Bank (part of Lloyds Banking Group) gain additional income by offering credit card schemes, selling premium bank accounts with monthly fees, earning commissions from selling insurance products and providing wealth management services [2].

Therefore it can be deduced that enhancing customer loyalty as well as acquiring new customers is paramount for sustaining adequate liquidity to lend money whilst also providing more opportunities for cross-selling financial products for additional revenue. In order to improve customer loyalty, banks can leverage vast amounts of transactional data collected from their customers, to understand consumer spending preferences and latent patterns to personalise services and upsell or cross-sell financial products [3]. With daily card payments in the UK expected to rise from 39.2 million to 60 million by 2026, banks can apply advanced data analytics to an ever-growing volume of transactional data to extract insights to achieve these goals [4].

The main objective of this report is to develop a customer segmentation model utilising feature reduction and unsupervised learning techniques to cluster accounts based on spending trends at specific merchant groups. Insights derived from these clusters will be used to build a predictive pipeline for financial product recommendations, which are tailored to the customer segments and validated through statistical data filtering techniques to evaluate suitability to product recommendations. Additionally, a fraud detection model is developed utilising unsupervised learning techniques to identify anomalous transactions that may potentially indicate fraudulent activity.

II. Literature Review

Customer segmentation is a technique commonly utilised in marketing analytics that seeks to define groups of individuals based on their unique needs or demands [5]. This

is particularly relevant for many industries where consumer preferences vary widely, such as in retail or online shopping. Identifying homogeneous customer segments can enable organisations to tailor product suggestions more effectively or discover emerging needs that could drive product innovation. Lefait and Kechadi utilised a customer segmentation approach called RFM (recency, frequency and monetary) analysis to segment a customer population based on computations of the recency of purchase (R), frequency of purchase (F) and total spend (M) for each account, derived from transactional data [6]. These features can be segmented by calculating quartiles and assigning them ranked scores. Alternatively, clustering algorithms such as K-Means can be applied to identify segments from the data, as demonstrated by the authors. RFM analysis can effectively identify high and frequent spenders but lacks the granularity needed to discern specific customer preferences, such as their choices of merchants for purchases due to its aggregation of total spend per account into a single feature. However, within this report, we aim to use clustering algorithms to identify groups within a broad feature space encompassing spending across various merchant categories, in order to discover more granular customer spending trends.

Clustering algorithms have generally proven effective in customer segmentation tasks. K-Means is a popular choice due to its simple implementation and its efficacy on large datasets as the algorithm's time complexity scales linearly with the number of training data points [7]. More advanced hierarchical methods such as agglomerative clustering are also useful for this task but are less practical due to their quadratic time complexity [7]. Density-based algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are another class of clustering algorithms that cluster data points based on the spatial densities of neighbourhoods of data points within the feature space. DBSCAN was found to be more effective than K-Means in identifying and generalising to clusters of arbitrary shapes for banking customer segmentation tasks, providing its hyperparameters are tuned appropriately [8]. It was also found to be highly scalable to large datasets providing the dimensionality of the dataset is relatively low. Within this report, a variant of the DBSCAN algorithm called HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [9] is utilised as it is able to additionally resolve clusters of varying spatial densities automatically without further hyperparameter tuning.

XGBoost (an abbreviation for eXtreme Gradient Boosting) is a gradient-boosted tree ensemble method introduced by Chen and Guestrin [10] that is also used within this report. Gradient boosting is a technique which uses a series of simple predictive models such as tree-based weak learners which are sequentially trained and incorporated into an ensemble to improve the prediction errors of any existing predecessor trees [11]. XGBoost is designed to be high-performance, interpretable and highly scalable to large datasets, particularly in memory-limited settings [10] which has led to its widespread use in customer/marketing analytics. Ren et al. [12] utilised an XGBoost classifier to predict customer segments (derived from an augmented RFM model) that may churn after being offered retail coupons. It was found that XGBoost achieved a higher test AUC score compared to other gradient-boosted tree methods. However, for this report, an XGBoost model will instead be used to classify accounts into customer segments - which are derived from per-merchant category spending data.

Furthermore, within this report, an anomaly detection model is developed to flag potentially fraudulent transactions. Machine learning-based approaches to anomaly detection are useful when instances of anomalies in data are uncommon. These systems can flag transactions in financial data or other activities that deviate from expected patterns in datasets [13]. Xuetong et al. compared four unsupervised learning models: One-Class Support Vector Machine (SVM), Restricted Boltzmann Machine, Auto-Encoders, and Generative Adversarial Networks [13]. They concluded that all models achieved an Area Under the Receiver Operating Curves (AUROC) higher than 90%. Although the one-class SVM had the lowest performance compared to the other three methods, it can still be used for anomaly detection as it can find decision boundaries that separate data into different classes, allowing this method to identify outliers and distinguish between abnormal and normal data, especially in unsupervised settings where there are no ground-truth fraud labels present [14].

III. Methodology

A. Feature engineering and dimensionality reduction

Dataset 1 is selected over *Dataset 2* (refer to section IV-A for further descriptions on these datasets) for training the customer segmentation model due to its higher volume of spending transactions and due to it featuring a richer diversity of merchant categories with 26 identified unique categories. In contrast, despite *dataset 2* having a total of 31 identified merchant groups, many were observed to be sparse, with 11 of the categories representing the spending of fewer than 1% of the total number of accounts. This left only 20 categories with a reasonable amount of transactions to be useful for segmentation.

To identify customer segments based on per-category spending, a subset of the dataset focusing on commercial payments to merchants is extracted and personal payments (to individual recipient accounts) are ignored. This subset of the data represented the transactions of 8142 individual accounts.

For each account, the sum of payments made in each merchant category is computed. These payments are then normalised to calculate a proportion of total spending in each category per account. This scales the features which can improve the performance of the clustering algorithm whilst also reducing the effects of the class imbalance in transaction counts in different merchant groups. The dataset is then projected down to two dimensions utilising UMAP (Uniform Manifold Approximation and Projection) [15]. UMAP was selected over other popular dimensionality reduction techniques such as PCA (Principal Component Analysis) due to its non-linear mapping enabling it to learn more complex high-dimensional

structures, that can be captured in a lower-dimensional projection. T-SNE (t-Distributed Stochastic Neighbor Embedding) is also a nonlinear data projection algorithm that was considered. However, UMAP has been demonstrated to retain comparable amounts of local structure and more 'global structure' within datasets compared to t-SNE [15]. This is useful in this context as UMAP will be able to discern localised clustering and class separation whilst still capturing higher-level patterns in the per-category spend data. The hyperparameter selection for the UMAP data projection is shown in **Table I**. The *n_neighbours* parameter is set at an arbitrarily low value compared to the total number of data points (8142) to tune the projection onto localised structures within the data to highlight class separations clearer. The *min_dist* is set to a low value close to 0, which forces UMAP to clump similar data points together.

TABLE I: Selected UMAP Hyperparameters for projection.

Hyperparameter	Selected parameter
n_components	2
n_neighbours	25
min_dist	0.10

B. Clustering

The UMAP projection of the dataset is then clustered utilising an HDBSCAN model. Two main hyperparameters are tuned, namely *min_samples* and *min_cluster_size*.

- *min_samples* determines the minimum number of data points within in a neighbourhood for a point to be considered as a core point. This was set to an arbitrarily low value of *min_samples* $= 1$ to lower the density threshold required for clusters to form which allows the model to capture a wider range of customer segments.
- *min_cluster_size* determines the minimum number of data points required to form a core cluster, below this threshold points are designated as noise. This parameter is tuned utilising a combination of visual inspection of the clusters formed in the UMAP projection of the data and the silhouette score which is a common clustering metric in market/customer segmentation tasks due to its interpretability [16].

A series of HDBSCAN models were trained with *min_cluster_sizes* ranging from 100 to 500 (in increments of 50). The results of this tuning study are shown in **Figure 1**.

The optimal value for *min_cluster_size* was selected based on the value that yielded the highest silhouette score, from the plot in **Figure 1**. This hyperparameter value however of *min_cluster_size = 200*, was deemed to be too low a value to cluster the projected data appropriately. This was due to the observation that 1-2 of the visually well-defined clusters in the UMAP projection of the per-category spend data were forced to split into subgroups by the model, which led to the formation of redundant and overlapping clusters. For example, in **Figure 2**, two clusters examined from this model displayed similar spending trends, with both clusters containing accounts that spent proportionally higher at department stores, clothing stores and gaming stores (compared to other categories).

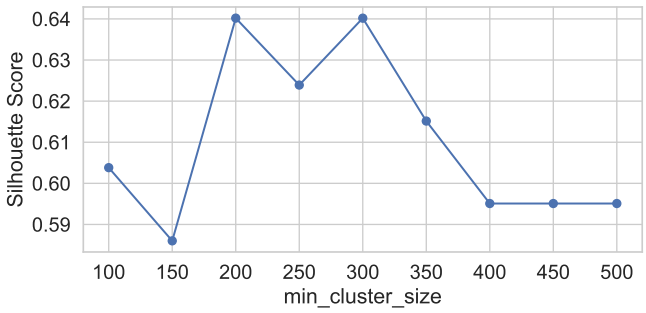
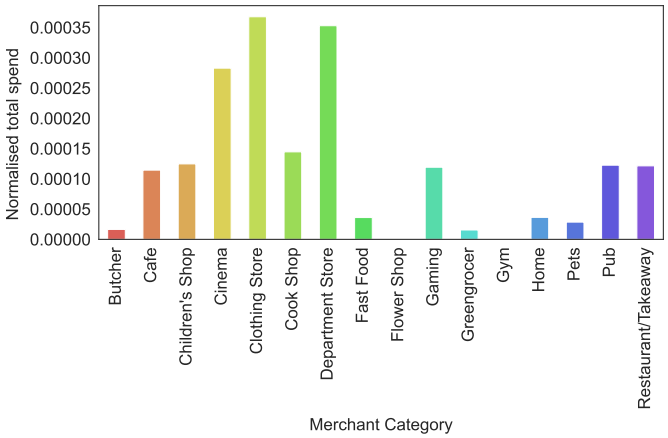
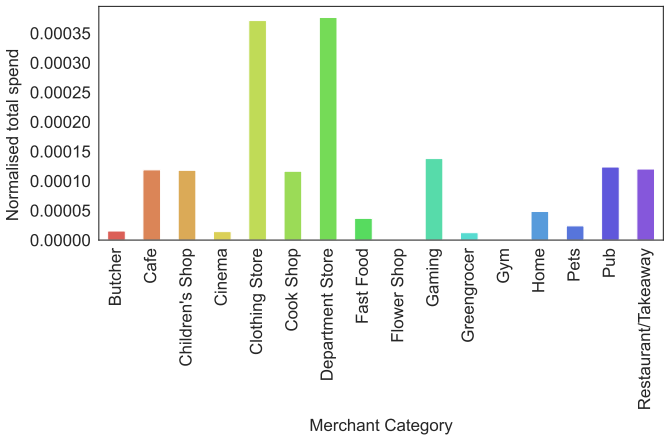


Fig. 1: Plot of silhouette score against *min_cluster_size*.

Following this analysis, the *min_cluster_size* hyperparameter was increased significantly to 350 in order to prevent the HDBSCAN model from splitting well-defined clusters.



(a)



(b)

Fig. 2: Plot (a) and (b) displaying the normalised sum of transaction amounts at different merchants, averaged across accounts in two different clusters.

The final, tuned application of the HDBSCAN model led to the formation of 12 clusters which are visualised in the UMAP projection of the dataset in **Figure 3**.

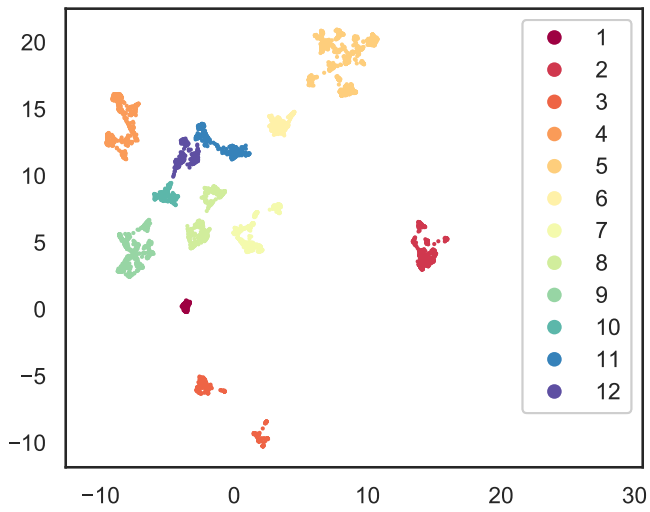


Fig. 3: UMAP projection of the per-category spending data with assigned cluster labels.

C. Product recommendation and prediction

The clustered accounts are then analysed by the proportional total spend across different merchant categories. This analysis enables clusters to be interpreted based on particular merchants where the proportional total spend is significantly higher compared to the other merchants. This enables simple insights to be obtained regarding the personality traits of the account holders within each segment and allows the formulation of different financial product recommendations based on the spending preferences of the accounts. The original per-category spending dataset is then labelled based on these identified customer segments. Utilising these segment labels, an XGBoost classifier is trained to predict the customer segment an account belongs to. This process establishes an automated pipeline that utilises spending data to predict the customer segment an individual likely belongs to. This subsequently corresponds to some predefined and tailored product suggestions to market to those particular account holders.

D. Classification of customer spending relative to total income

Salary estimates were also constructed for a subset of accounts in *Dataset 2*, by pivoting the dataset and filtering for regularly occurring, incoming payments from companies. Total salary estimates could therefore be derived by summing these incoming transactions for each account, as only a single merchant was present per account number. Regular, fixed incoming payments from personal accounts were also included with the salary estimates to create a holistic income estimate for each account. These regular, fixed payments may entail rent standing orders or fixed personal income from self-employment. Regardless, they represent income for individuals and so were included in the income estimation for accounts.

Utilising total income estimates for each account, a ratio could be computed of the total commercial expenditure (payments to merchants) to an individual's total income. The distribution of this ratio allows the application of data filtering techniques to identify individuals who spend high or low amounts relative to their income. As *Dataset 2* also contained merchants which could be categorised, distributions of total spend-to-income across each merchant category for the accounts could also determined. Filtering these distributions allows the identification of *high* and *low* spenders at particular merchant groups.

E. Fraud detection

Dataset 2 is selected to train the model over *Dataset 1* as it includes both online and offline transactions, with account balances that are both positive and negative. It was decided not to utilise merchant categories from *Dataset 2* as features for the model. This decision was based on the observation that infrequent transactions can significantly distort the perceived average spending at certain merchants. Therefore, including such merchants in the same category would result in disproportionate weights given to these anomalies, thus impairing the model's ability to accurately identify anomalous deviations based on transaction values. For example, the average transaction amount at North Face is £270, compared to £50 for other outdoor clothing stores. Therefore, different stores reflect different spending capacities. Continuing to use *Third Party Name* (Merchant name) as a feature - aids in maintaining the model's sensitivity to true anomalies without being misled by statistical values caused by low-frequency, high transaction value outliers.

Our analysis additionally considers factors including transaction timing, weekends, holidays, and salaries to identify key features indicative of user behaviour patterns. Within our approach, we also compute an *anomaly score* for each transaction using the one-class SVM model. Setting an *anomaly score* threshold to flag potentially fraudulent transactions is challenging due to the lack of ground truth fraud labels. Despite this challenge, our understanding of fraud prevalence within transactional data is informed by external research, which indicates that typically, only 0.17% of recorded transactions are fraudulent [17]. This derived value is utilised to set the anomaly threshold for the model. This allows us to position the *anomaly score* with the observed fraud rate, flagging 0.17% of transactions as anomalies. This is assumed to capture a significant portion of fraudulent transactions present in the data. The one-class SVM model flags transactions that might be anomalous. However, this does not imply that all anomalous payments indicate fraud, and so these payments are further categorised. Initially, all transactions flagged by the one-class SVM are classified as *low-risk*. Then, two further criteria are introduced to identify transactions within the *low-risk* category that warrant heightened monitoring, classifying them as *high-risk*. These criteria encompass:

1) *Large* transactions that are flagged as an additional feature and defined where there are transaction amounts exceeding the mean spend at a specific merchant by than ±3 standard deviations.

2) Unusual transaction times are also flagged when individuals make payments between midnight and 5 a.m. Direct debit payments are ignored from this flag.

A transaction is flagged as potentially *high risk* fraud if it is marked as an anomaly from the one-class SVM model and either occurs during unusual hours or involves a large amount. The current logic uses an OR condition, meaning either unusual hours or large amounts can trigger the flag.

one-class SVM hyperparameters are tuned via k-fold cross-validation, which are shown in **Table II**.

TABLE II: Summary of tuned SVM hyperparameters.

Hyperparameter	Selected parameter
γ	Scale
Kernel	Sigmoid
ν	0.1

IV. Data Description / Preparation

A. Dataset overview

Dataset 1, is a simulated transactional dataset provided by Lloyds Banking Group containing instances of 10,148,280 transactions over an annual period. It includes features for sender account number, transaction amount, beneficiary account number or beneficiary business name as well as the date of the transaction. *Dataset 2* is also a simulated transactional dataset that contains instances of 230,596 transactions. It is a lower-volume dataset that contains additional features including post-transaction balance amounts, incoming/outgoing payments from accounts as well as timestamps to the nearest minute for each transaction.

B. Imputing missing values

Within *Dataset 1*, although no explicit missing values or null entries were identified, transactions for one entire day are absent. Since this missing day constitutes approximately only 0.27% of the annual data (based on the daily average number of transactions), imputation is deemed unnecessary given the substantial volume of data that remains.

Dataset 2 contained missing values across most features. Missing *Date* values were imputed using dates from consistent adjacent records or determined based on a *Timestamp* comparison if they differed. For instance, a *Timestamp* gap from 23:08 to 00:14 indicated the start of a new day. Missing *Timestamp* values were imputed using adjacent entries on the same day, resolving all 480 missing cases in these time-based columns. Missing account *Balances* were imputed using the difference between adjacent balance records of the same account. This approach corrected all missing *Balance* entries. Missing values in the *Amounts* column were imputed via consideration of transactions that consistently occurred at either 00:00 or 23:59, typically for regular payments like salaries or subscriptions. The remaining *Amount* values were filled using balance information from adjacent records, resolving 453 cases of missing *Amounts*. Finally, we addressed missing *Third Party Account No* and *Third Party Name* entries by identifying and matching identical monthly transactions, which resolved 69 of the missing values - the remaining of which were dropped. Consequently, *Dataset 2* was reduced from 230,596 rows to 230,195 rows after replacing missing values and dropping any rows that had values which could not be imputed accurately.

C. Merchant categorisation

Merchants are categorised into groups of businesses, based on the semantic meaning of the merchant name as well as similarities in transaction frequencies, the mean spend and variance in spend at the merchants. For example, certain businesses in *Dataset 1* could be identified as food shops by their names such as SANDWICH SHOP or INDIAN RESTAURANT. However, considering the mean and variance in spend as well as the number of transactions, these merchants have been differentiated into restaurant/takeaway and fast food shop categories due to fast food shops (such as SANDWICH SHOP and KEBAB SHOP) having significantly higher transaction frequencies (higher number of visits) and lower average spends (see **Table III**).

TABLE III: Spending statistics at three food merchants.

Spending Statistic	Merchant		
	SANDWICH SHOP	***KEBAB SHOP***	***INDIAN RESTAURANT***
Frequency	111143.000	111339.000	7429.000
Mean (£)	4.400	4.402	35.287
Variance (£)	1.138	1.138	240.814

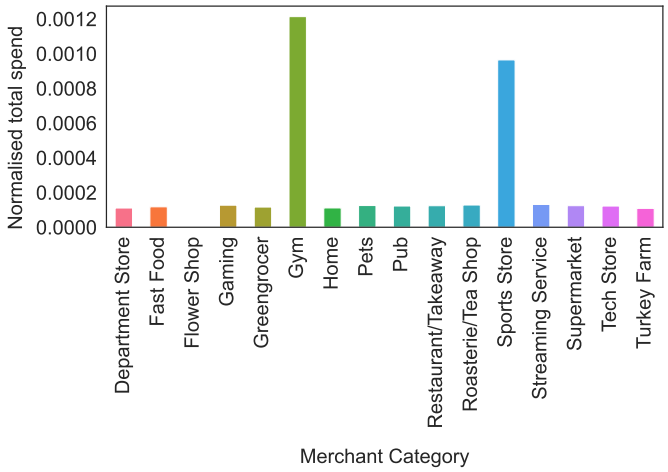
For *Dataset 1*, the commercial transactions (payments to merchants) are categorised mostly utilising the words in the merchant names as well as some manual allocations of merchants. The merchant names are tokenised into words, and a dictionary is created to capture common terms in merchant names and map them to a category. For example, business names containing terms like 'coffee' and 'cafe' are mapped to the merchant category 'Cafe'. This dictionary is then used to map the merchants in the dataset to their respective categories. A similar approach is utilised for *Dataset 2*, except the merchants had to be allocated manually to particular merchant categories as this dataset contained real business names.

V. Results and Discussion

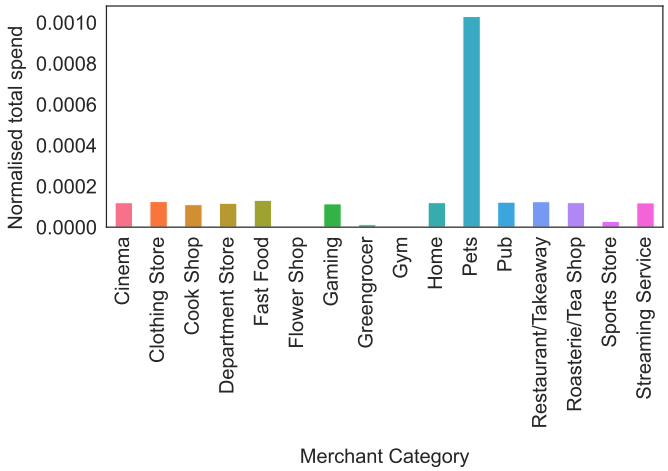
A. Per-merchant category spending insights

After clustering the per-merchant category transactions from *Dataset 1*, the UMAP/HDBSCAN model identified 12 clusters from the spending population of accounts. Plots were then created to analyse the proportions of total spending across the different merchant categories for the accounts in each cluster. For each cluster, 1-2 merchant groups were selected that received the highest proportional spend. For example in **Figure 4a**, it can be observed that accounts in this segment spend significantly higher on gyms and sports stores whilst accounts in another segment displayed in **Figure 4b**, spend significantly higher at pet stores compared to other merchant groups. One cluster was observed to have high spending across the majority of merchant groups and so was ignored. This analysis is

conducted for each customer segment and summarised in **Table IV**. Customer preferences and personality characteristics can be derived from these customer segments based on the spending preferences of different merchants. For example, those individuals belonging to the GS segment can potentially be assumed to be 'fitness' or 'exercise' oriented based on their spending preferences towards gyms and sports stores.



(a) GS customer segment



(b) PE customer segment

Fig. 4: Plot (a) and (b) displaying the normalised sum of payments averaged across accounts in two identified customer segments, plotted against a subset of merchant groups.

Based on these insights, product recommendations can be devised that not only provide benefits to customers but also open up opportunities for Lloyds Bank to cross-sell profitable financial products for the financial benefit of the bank. For example, customers in the CH category, who frequent children's shops are likely to be parents with children. They may for example benefit from a children's savings account to secure funds for their kids' future. This product suggestion may improve the overall customer experience of these individuals as it's relevant and useful to their needs and so may improve their brand loyalty to Lloyds Bank. Furthermore, profitable family-oriented insurance products available in the Club Lloyds Silver Account [18], such as family travel insurance could also be offered - as they may be relevant to these types of individuals. Individuals in the HO customer segment - where spending is concentrated in home stores (such as DIY Stores) may likely be homeowners. They for example benefit from having building and contents insurance products or home improvement loans promoted to them. Additionally, customers in the DM category who frequent cinemas and spend particularly high on streaming services and digital media/content will likely benefit from joining Lloyds Banks' Club Lloyds Silver Account [18] that offers free cinema tickets at *Vue* or free subscriptions to streaming services such as *Disney+*. Due to space limitations, recommendations for other segments cannot be discussed in depth, however a similar approach as outlined in this section can be taken to derive them.

TABLE IV: Identified customer segments and their most spent at (most popular) merchant categories.

Segment Label	Most spent at merchants
CH	Children's stores
GS	Gym and sports stores
AG	Accessory and flower shops
LF	Butchers and greengrocers
PE	Pet stores
CL	Clothing and department stores
FF	Fast food restaurants
TF	Turkey farm
DM	Cinemas and streaming services
TS	Tech stores
HO	Home stores

B. Customer segment and product prediction

An XGBoost classifier is trained using a grid search approach with 10-fold cross-validation, utilising sklearn's *GridSearchCV()* function. This method exhaustively trains and evaluates models across all combinations of pre-defined hyperparameter values (defined in **Table V**) to select an optimal model for testing. The labelled per-category spending dataset is divided into 70%:30% training-test split utilising a random-stratified splitting procedure to maintain approximately the same distribution of classes (customer segments) across the training and test splits. This is done to ensure the training and test sets are representative of the distribution of classes in the overall dataset. This dataset was also scaled utilising the same pre-processing steps conducted prior to training the UMAP/HDBSCAN model outlined in Section III-A. Macro averaged F1-score is utilised as the primary performance metric due to the class imbalance present within the dataset amongst the varying sizes of the customer segments. After training and cross-validation, an optimal model is selected utilising the best hyperparameters defined in **Table V**. This model is evaluated on the unseen testing data split and the results are summarised in **Table VI**

Overall, the model achieved a macro-averaged F1-score of 0.99 on test data, indicating its accuracy in predicting the customer segment an account may belong to. This result provides

TABLE V: Hyperparameter search space for the training of the XGBoost classifier with selected best parameter values.

Hyperparameter	Parameter values	Best parameter
Learning rate	[0.01, 0.05, 0.1, 0.15, 0.2]	0.01
Max depth	[2,5,10,20,50]	5
Number of estimators	[200,600,1000]	600

TABLE VI: XGBoost classifier test results.

Performance Metric	Value
Macro F1 score	0.99
Macro precision	0.99
Macro recall	0.99

confidence that the correct predefined product suggestions for each segment will be recommended for the classified accounts.

C. Risk model and targeted product recommendations

From the distribution of spend-to-income (shown in **Figure 5**), statistical data filtering techniques are utilised to classify accounts to form part of a risk model and a targeted product recommendation system. This distribution is assumed to be approximately Gaussian. Accounts that have a spend-to-income ratio 1 standard deviation above the mean of the distribution are classified as *excessive* spenders. *Excessive* spenders, defined as individuals whose expenditure is particularly high or exceeds their income, may face challenges in financial planning. Marketing financial products beyond their economic means, such as credit cards or loans with interest rates, could lead these individuals into financial hardship. Thus targeting these individuals can be considered highly unethical and as a result, it is advisable to omit them from product recommendation communications and instead offer financial planning and hardship resources/support to these individuals.

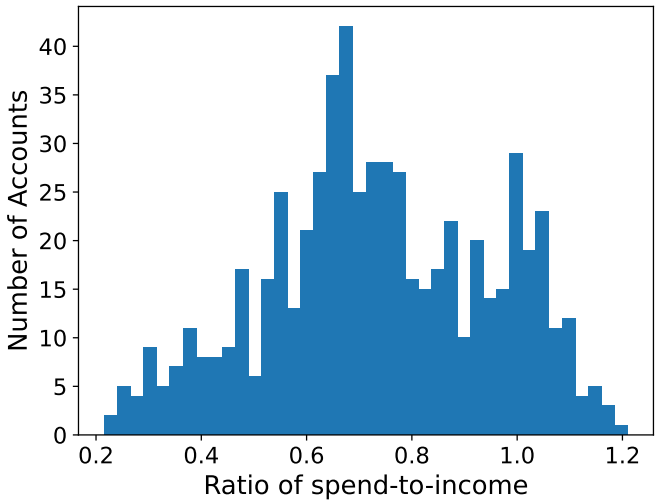


Fig. 5: Histogram displaying the spend-to-income ratio across individual accounts in *Dataset 2*.

However, for the remaining accounts, targeted advertisement of product recommendations based on the particular merchants categories they spend particularly higher at can enable more effective and specific promotion of products. This task was challenging, as most distributions exhibited non-Gaussian properties, with many demonstrating pronounced right-skewness indicative of log-normal distributions. As a result, to estimate the *high*, *low* and *normal* spenders bands from these distributions accurately - they were log-transformed to produce approximately normal distributions. The mean (μ_{log}) and variance (σ^2_{log}) were estimated from these distributions so that upper and lower thresholds could be computed in log space based on values being 1 standard deviation above the mean (see **Equation 1**) and 1 standard deviation below the mean (see **Equation 2**) respectively. These thresholds were then transformed back into real space and used to classify *high* and *low* spenders within each merchant category (see **Equations 3, 4**). *Normal* spenders represent accounts whose spend-to-income ratio lies between these thresholds.

$$UT_{\log} = \mu_{\log} + \sigma_{\log} \quad (1) \qquad LT_{\log} = \mu_{\log} - \sigma_{\log} \quad (2)$$

$$UT = e^{(UT_{\log})} \quad (3) \qquad LT = e^{(LT_{\log})} \quad (4)$$

Where the distributions of spend-to-income ratio at particular merchants are neither approximately Gaussian or log-normal, the upper and lower quartiles were used to assign *high* and *low* spenders. With these new classifications, individuals for example predicted into the DM customer segment, who generally spend relatively more on streaming services and cinemas can be further divided into *high*, *low* and *normal* spenders. *High* spenders at cinemas or streaming services can then be prioritised for DM segment-specific product promotions such as the Club Lloyds Silver premium accounts [18] as they're likely to benefit from signing up and using them more than *normal* or *low* spenders. This can reduce marketing costs compared to advertising unspecifically to an entire segment. Furthermore, it should be noted that the *high* spender thresholds are arbitrarily set and should ideally be adjusted using A/B hypothesis testing to compare how different limits impact product sales metrics and marketing costs - to obtain an optimal threshold.

D. Analysis of anomalous transactions

168 transactions across 5 accounts were identified and labelled as low-risk fraud, but only 28 transactions across 4 accounts within these transactions can considered to be high-risk fraud. It is important to note that unusual spending does not necessarily indicate fraudulent activity. For example, a large money transfer might be legal but could be flagged if it deviates significantly from the user's normal spending patterns. We flag these transactions to bring potentially unrecognized activity to the bank's and user's attention. Low-risk fraudulent transactions occur most often on Sundays, particularly at times like 9 AM and 11 PM. It is important for banks to use experts to review transactions regularly, and any unrecognized activity should be followed by direct contact with the account holders. High-risk fraud, on the other hand, shows a marked increase on Saturday and occurs predominantly between 4 AM and 5

AM. This may indicate more severe fraudulent attempts on weekends and at unusually early times. Such transactions are flagged not only due to their large amounts but also due to their unusual timing, which significantly deviates from the user's normal spending patterns. Unusual timings on weekends may not reliably indicate high-risk fraud, as no external research indicates a correlation between weekend purchases and fraud. To improve this fraud detection system, longer-term behavioural data should be assessed to understand whether transaction times are truly unusual for specific accounts.

E. Data privacy and customer profiling ethics

Banks must be transparent regarding the purposes for which they use customer data collected from account holders, particularly when profiling individual accounts for marketing purposes. Banks can achieve these legitimate interests, by providing additional consent is requested to use customer information in this way to avoid infractions of GDPR (General Data Protection Regulation) and prevent any reputational damage.

VI. Conclusion

In conclusion, within this report, we were able to successfully develop a customer segmentation model that clustered per-merchant category spending data for individual accounts. Through clustering the accounts, we identified 11 unique customer segments that displayed explainable spending preferences at particular merchants. As a result of this, tailored financial product recommendations were then formulated for the segments that are personalised and meet their needs whilst also providing opportunities to cross-sell products for additional income. A predictive XGBoost classifier was also successfully trained on this dataset to predict the customer segment a particular account may belong to, creating an automated pipeline for segment-specific product predictions. This model achieved a high F1 score of 99% on test data. Income estimates were also derived from *Dataset 2* that enabled the development of a spend-to-income risk model that blacklists individuals from product recommendations as well as enabling targeted recommendations towards *high* spenders in particular merchant categories. Finally, a fraud detection model was developed utilising one-class SVMs which identified 168 anomalous transactions with a subset of 28 transactions specifically flagged as indicating a *high-risk* of fraudulent activity.

VII. Further Work and Improvement

A. Model improvement

The SVM-based fraud detection model could be improved by utilising more accurate estimates of the proportions of fraudulent payments in transactional datasets. These can likely be obtained by leveraging the internal knowledge of fraudulent activity within a commercial bank such as Lloyds Bank. Furthermore, product recommendations targeted at particular customer segments derived from the HDBSCAN model should be continuously reviewed. If customer feedback or sales metrics are below expectations, the customer segment definitions should be revised.

B. Bias quantification and reduction

Both **Dataset 1** and **Dataset 2** lack demographic information such as the age, ethnicity, and gender identity of individual account holders. This creates uncertainties as to whether the identified customer segments are representative of the diversity of the spending population. This can potentially lead to biased product recommendations. For instance, if *high* spenders within a merchant category predominantly belong to a specific demographic group - the marketing strategies that target these segments could inadvertently bias economic opportunities against other groups by excluding them from relevant advertising for useful banking products. Incorporating personal demographic data into the datasets may help quantify these disparities and aid in tuning or redefining models to capture these groups more equitably.

References

[1] M. McLeay, A. Radia, and R. Thomas, "Money creation in the modern economy," *Bank of England quarterly bulletin*, p. Q1, 2014.

[2] Lloyds Bank, "Products and services," 2024. Available online: https://www.lloydsbank.com/products-and-services.html [Accessed: 19/04/2024].

[3] D. Koteshov, "Big data in banking." https://startups.epam.com/blog/big-data-in-banking, March 2024. [Accessed: 20/04/2024].

[4] Statista, "Card payments per day in the united kingdom (uk) in 2006 and 2016, with a forecast for 2026," 2017. [Accessed: 20/04/2024].

[5] M. McDonald, "The role of marketing in creating customer value," *Engineering Science & Education Journal*, vol. 6, no. 4, pp. 1–111, 1997.

[6] G. Lefait and T. Kechadi, "Customer segmentation architecture based on clustering techniques," in *2010 Fourth International Conference on Digital Society*, pp. 243–248, IEEE, 2010.

[7] F. Giannotti, C. Gozzi, and G. Manco, "Clustering transactional data," in *Principles of Data Mining and Knowledge Discovery: 6th European Conference, PKDD 2002 Helsinki, Finland, August 19–23, 2002 Proceedings 6*, pp. 175–187, Springer, 2002.

[8] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 197–202, 2005.

[9] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[11] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[12] Y. Ren, P. Fu, and W. Yu, "Prediction of coupon usage behavior based on customer segmentation and xgboost algorithm," in *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*, pp. 42–47, 2021.

[13] X. Niu, L. Wang, and X. Yang, "A comparison study of credit card fraud detection: Supervised versus unsupervised," *arXiv*, 2019.

[14] Activeloop AI, "What is one-class svm," 2023. Available online: https://www.activeloop.ai/resources/glossary/one-class-svm/ [Accessed: 20/04/2024].

[15] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[16] J. Korczak, M. Pondel, and W. Sroka, "Discovery of customer communities–evaluation aspects," in *Conference on Advanced Information Technologies for Management*, pp. 177–191, Springer, 2019.

[17] A. Nadim, I. M. Sayem, A. Mutsuddy, and M. Chowdhury, "Analysis of machine learning techniques for credit card fraud detection," pp. 42–47, 12 2019.

[18] Lloyds Bank, "Club Lloyds Silver Account," 2024. Available online: https://www.lloydsbank.com/current-accounts/all-accounts/club-silver-account.html [Accessed: 19/04/2024].

APPENDIX

DSMP Group 38 - GitHub Repository URL: [https://github.com/UoB-DSMP-2023-24/dsmp-2024-group-38]