Customer Segmentation in Retail Using K-Means and Markov Chains

Shreya Babare
Faculty of Engineering
MSc Data Science
University of Bristol
Bristol, United Kingdom
kn24855@bristol.ac.uk

Tanya Sharan
Faculty of Engineering
MSc Data Science
University of Bristol
Bristol, United Kingdom
xe24084@bristol.ac.uk

Yupeng Liu
Faculty of Engineering
MSc Data Science
University of Bristol
Bristol, United Kingdom
rf23153@bristol.ac.uk

Jiaqi Zhang
Faculty of Engineering
MSc Data Science
University of Bristol
Bristol, United Kingdom
vq24786@bristol.ac.uk

***Abstract*—Can spending patterns reveal more than just habits, can they anticipate behavior? In today's data-driven world, banks are no longer just custodians of money, they are stewards of behavioral insight. Every payment, transfer, and balance shift offers a glimpse into a customer's financial health, priorities, and potential future decisions.**

We began with the first quarter of the year, leveraging Recency, Frequency, and Monetary (RFM) metrics to develop monthly behavioral profiles. These profiles were then clustered using K-Means to identify meaningful customer segments. To understand how customer behavior changes over time, we implemented a Markov Chain model that used Q1 patterns to predict how customers might transition between segments in the subsequent month. To build on this, we explored credit-to-debit ratios as a proxy for financial balance and laid the groundwork for future overdraft risk modeling. Our approach demonstrates the feasibility of transforming raw transaction logs into predictive tools that can help banks identify at-risk customers, personalize support, and enhance overall financial well-being.

Keywords: Behavioral Segmentation, RFM Modeling, Unsupervised Learning, K-Means Clustering, Markov Chain Model, Retail Banking Analytics, Financial Behavior Prediction, Credit-to-Debit Ratio, Overdraft Detection.

I. Introduction

Lloyds Banking Group serves over 30 million customers in the UK. With one in five adults holding a Lloyds current account, the bank processes millions of transactions every day from contactless payments at high-street shops to direct debits for bills, subscriptions, and everything in between. Even in the synthetic data we worked with, there were familiar spending patterns: purchases at places like Boots or Mountain Warehouse, followed by recurring payments to coffee shops or art supply merchants. Although the data was synthetic, each of these transactions holds valuable information on a customer's financial behavior, lifestyle, and potential future needs.

The main focus of this project was on how transactional data can be used to extract more information, not just to understand what customers are doing now but to anticipate what they might need next. If we can identify patterns in how people spend, save, and move money, we can start to support them more proactively: offering the right services at the right time, identifying early signs of financial difficulty, or even preventing issues like overdrafts before they occur.

To do this, we worked with a synthetic dataset modeled on real customer behavior. We began by looking at the first quarter of the year - January to March, and calculated key behavioral indicators using the Recency, Frequency, and Monetary (RFM) model. This allowed us to group customers into different segments based on how recently they transacted, how often, and how much they spent.

We then used a Markov Chain model to track how customers moved between these segments from month to month. This gave us a way to predict what segment a customer was likely to fall into in the following month, forecasting changes in financial behavior. We tested the model's predictions against real (simulated) data from April to see how accurate our approach was. In addition, we looked at credit-to-debit ratios, a simple way to understand whether customers are consistently spending more than they're receiving. This opens the door to identifying potential financial stress, especially when combined with overdraft data, which we plan to explore further in future work.

By combining customer segmentation with predictive modeling, this project demonstrates how existing data can be used to build a more responsive and supportive relationship with customers, which would adapt to their financial behavior in real time.

II. Literature Review

Customer segmentation plays a crucial role in customer relationship management (CRM), marketing optimization, and business strategy. Several models and algorithms have been proposed to analyze and classify customer behavior based on purchase patterns, demographic traits, and lifecycle value.

The RFM (Recency, Frequency, Monetary) model is one of the most widely adopted frameworks for customer segmentation. Chen et al. [1] applied RFM-based segmentation in the context of online retail, demonstrating the model's effectiveness when coupled with clustering techniques for actionable targeting. Enhancements to this approach were introduced by Cheng and Chen [14], who combined the RFM model with Rough Set (RS) theory to manage uncertainty in data and improve classification granularity. Aggelis and Christodoulakis [15] also utilized the RFM framework in retail analytics, highlighting its compatibility with customer clustering for value-driven marketing.

Jiang and Tuzhilin [2] introduced an optimization-centric segmentation technique, framing customer base division as a performance maximization problem. This method emphasized personalized solutions over rule-based grouping, aligning segmentation directly with marketing utility. Wong and Wei [3] took a predictive perspective, proposing a data analytics model that integrates segmentation with customized service prediction to enhance online shopping experiences. Vela and García [10] offered a domain-specific application by profiling budget air travelers using clustering techniques, contributing insight into traveler preferences and behavior.

Clustering algorithms are foundational to customer segmentation. Traditional K-Means, while popular, is limited to numerical data. Huang [4] proposed extensions to the algorithm, resulting in the K-Modes and K-Prototypes variants that support categorical and mixed-type data. Ahmad and Dey [6] developed a K-Means-based approach specifically for mixed numerical and categorical attributes, improving clustering precision in real-world datasets. Cao et al. [5] addressed initialization sensitivity by introducing a novel seeding technique for categorical data clustering, enhancing both speed and accuracy. Determining the correct number of clusters is another critical step; Marutho et al. [7] employed the Elbow Method and purity evaluation to identify optimal cluster count in news headline analysis. Liu and Deng [8] further advanced this technique by integrating fuzzy logic to determine unknown targets in open-world environments, which has broad applications in emerging customer pattern detection.

Customer lifecycle modeling often utilizes Markovian approaches to track behavioral transitions over time. Pfeifer and Carraway [9] modeled customer interactions as a Markov chain to predict future behavior and calculate customer lifetime value (CLV). Cheng et al. [12] applied a Markov chain-based data mining model in the automotive maintenance industry, using behavioral states to estimate customer longevity and value. Netzer et al. [16] developed a Hidden Markov Model (HMM) to uncover latent customer relationship states, enabling a more dynamic view of engagement, churn, and reactivation. These probabilistic frameworks are mathematically supported by Freedman [13], who provided a convergence theorem for finite Markov chains, ensuring long-term modeling reliability. Hwang [11] proposed a stochastic model that values customers based on probabilistic transitions, offering a data-driven case study in lifecycle-based valuation.

Collectively, these studies highlight the evolution of customer segmentation from simple rule-based heuristics to sophisticated data-driven methods. The integration of clustering, predictive modeling, and Markov-based behavioral analysis reflects a comprehensive approach to understanding and managing customer relationships. Algorithms are increasingly tailored to accommodate mixed data types, uncertainty, and real-time adaptability, which are critical for practical CRM deployment in dynamic business environments.

III. Methodology

Since the aim of this project was to explore how customer transactional data can be used to group individuals based on behavioral patterns, and to predict how those behaviors may change over time. To implement this, the project was divided into two main phases: customer segmentation and behavioral prediction.

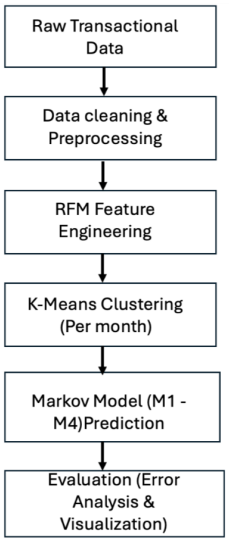

Fig. 1. Overview of the RFM-based customer segmentation and prediction process.

The first phase involved pre-processing and cleaning the data, followed by the calculation of Recency, Frequency, and Monetary (RFM). These metrics were used to quantify customer activity, and the basis for segmenting customers using K-Means clustering, resulting in clearly defined customer behavior segments. In the second phase of the project, we analyzed how customer behavior changed over time by tracking movement between segments. Based on the clusters from the first quarter (Q1: January to March), we implemented a Markov Chain model to estimate the likelihood of customers

transitioning from one behavioral cluster to another. This approach helped to forecast the distribution of customer segments for the following month (April) and compare the results with the actual customer behavior.

A. Data Cleaning & Pre-processing

The dataset "simulated_fake_transactions_dataset" consisted of individual transactional activities made by Lloyd's Banking Group customers. Each row in the dataset represented a single transaction made by a customer and included details such as the date and time of the transaction (Date and Timestamp), the amount of money (Amount), the resulting account balance (Balance), and the recipient (Third Party Account No or Third Party Name). We started by cleaning the dataset to fix errors, fill in missing details, and make sure it was consistent and balanced.

First step involved standardizing the date and time features/ columns by merging them into a single `DateTime` column, allowing straightforward time-based aggregation and sorting. Then moved to the inconsistencies in the third party transaction fields. When the merchant name (`Third Party Name`) was available and the corresponding account number was missing, the transaction was assumed to be a payment to a business and labeled as a merchant transaction (`MTrx`). Similarly, if the `Third Party Account No` was present but the name was missing, the transaction was tagged as a peer-to-peer transfer (`P2P`), indicating money sent to another individual. In cases where both fields were missing, the transaction lacked any identification of the recipient and was therefore labeled as Unknown.

TABLE I
EXAMPLE OF TRANSACTION LABELING BASED ON THIRD-PARTY INFORMATION

Third Party Name	Third Party Account No	Label
Boots	–	MTrx
–	123456789	P2P
–	–	Unknown
CoffeeWorld	–	MTrx

In addition to resolving inconsistencies, further cleaning steps involved addressing missing values for the columns `Account No`, `Amount`, and `Balance`. Rows missing all of these fields were dropped, as they would not have helped provide valuable insights. We also standardized data types across key fields to ensure consistency. Both `Amount` and `Balance` were converted to floating numbers to support reliable numerical operations, while `Account No` was stored as a string to avoid formatting issues caused by automatic float conversion (e.g., 123456789.0). The merged `DateTime` column was converted to a proper datetime object, allowing for consistent and accurate time-based filtering, grouping, and trend analysis.

B. Exploring Transaction Patterns and Engagement Trends

Customer behavioral characteristics were derived by analyzing patterns in transaction frequency, monetary value, and transaction type over time. These attributes were chosen to help guide feature selection and improve the overall segmentation and modeling pipeline.

Transaction frequency was calculated as the total number of transactions made by each customer within monthly intervals, capturing overall engagement. Monetary behavior was measured by aggregating debit transactions to represent total monthly spending. Monthly granularity was maintained to ensure consistency with RFM feature construction and the subsequent Markov transition analysis.

Transaction categories were defined using the `merchant_category_group` variable provided in the dataset, which grouped transactions into high-level labels such as Retail & Fashion, Supermarkets, Entertainment & Gaming, and Financial Services. To simplify interpretation and reduce noise from rare or unclear entries, low-frequency or ambiguous categories were grouped under a general "Other" label. For each customer, the proportion of transactions within each category was calculated to highlight dominant spending preferences.

Looking at daily spending patterns, recurring spikes in transaction activity were observed at the beginning of each month—likely corresponding to simulated salary deposits, rent payments, or monthly subscriptions. Between these peaks, daily activity remained relatively steady, indicating consistent but lower engagement. These patterns reinforced the value of incorporating features such as `Recency` and `Frequency` into the customer segmentation process.

C. RFM Calculation

For a deeper understanding of customer engagement and financial behavior, we applied the RFM (Recency, Frequency, Monetary) model. Which is a widely used approach in behavioral analytics. This method allowed us to assign scores to each customer based on how recently they transacted (Recency), how often they transacted (Frequency), and how much they spent (Monetary), where spending was calculated using only negative transaction amounts to reflect outflows.

Initially, we calculated RFM values using the entire year's worth of transaction data. This provided a holistic view of each customer's behavior across the full 12-month period. However, since Markov models rely on historical data, monthly RFM calculations were necessary to predict behavior in Month 4 (April). We then narrowed our focus to monthly RFM calculations for the first quarter (January to March). This allowed us to observe how customer behavior evolved over these three months and provided the month-by-month detail needed to track how customer behavior changed over time.

For each month in Q1, Recency was calculated as the number of days since a customer's last transaction, Frequency represented the total number of transactions made during that month, and Monetary reflected the sum of all spending (i.e., total value of negative transactions) by that customer. These monthly RFM values were then used as input features for clustering, helping us track and predict changes in customer behavior over time.

TABLE II
RFM VALUES FOR MONTH 1 (RECENCY, FREQUENCY, MONETARY) FOR SAMPLE CUSTOMERS

Account No	Recency (M1)	Frequency (M1)	Monetary (M1)
100100738	2	17	-1192.90
100837224	4	3	-288.05
101348775	0	29	19228.36
103439190	0	30	30373.01
104009728	0	33	-11420.88

D. Customer Segmentation Using Clustering

Once we had calculated the RFM values for each customer, the next step was to identify patterns in customer behavior by grouping similar profiles together. To carry out clustering, we applied K-Means, an unsupervised ML algorithm that separates data into distinct groups based on feature similarity. In our case, the input features were the `Recency`, `Frequency`, and `Monetary` values derived for each customer during the first quarter.

Before clustering, we normalized the RFM features using Min-Max scaling to ensure that each metric contributed equally to the distance calculations. This technique transforms each feature to a common scale, which ranges between 0 and 1, using the formula:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

In this case, X represents the raw value of a customer's Recency, Frequency, or Monetary score. X_{min} and X_{max} refer to the minimum and maximum values of that feature across all customers. For example, if one customer had a Frequency of 20, and the minimum and maximum values for Frequency were 1 and 50 respectively, the scaled value would be:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} = \frac{20 - 1}{50 - 1} = \frac{19}{49} \approx 0.39$$

Without this step, the `Monetary` values could dominate the Euclidean distance calculations in the K-Means algorithm and skew the clustering results. Reason being, they are numerically larger than Recency or Frequency.

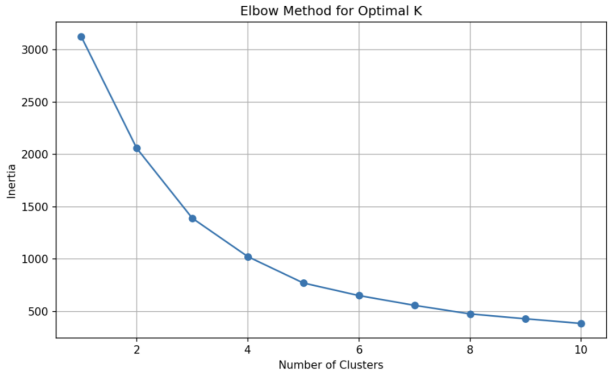

Fig. 2. Elbow plot indicating optimal number of clusters at $K = 6$.

Now to determine the number of clusters, we used the Elbow Method (Fig.2), which plots the Within-Cluster Sum of Squares (WCSS) against different values of K. The ideal number of clusters is often found at the point where adding more clusters no longer significantly reduces the WCSS. In our case, the elbow was observed at K = 6.

TABLE III
CLUSTER INTERPRETATION BASED ON RFM SEGMENTATION ($K = 6$)

Cluster	Customer Profile	Profile Description
0	Steady Savers	Moderate recency, frequency, and monetary; consistent average engagement
1	Re-Engagement Opportunity	High recency with low frequency and spending; minimal recent activity
2	Routine Users	Low recency and high frequency; engaged with moderate monetary value
3	Premium Contributors	Recent but infrequent; high monetary spenders (e.g., large transactions)
4	Cautious Transactors	Mid-recency and frequency; low overall spend but regular activity
5	Unique Financial Profiles	Behavioral outliers with atypical or extreme RFM characteristics

Each resulting cluster represented a distinct customer profile. Some clusters captured customers who transacted frequently and recently but with low spending, while others contained customers who transacted less often but made larger transactions. Since we had a 3D feature space, we applied PCA, a dimensionality reduction tool, to visually validate our results. PCA projected the three RFM features into two principal components, allowing us to plot and assess how well-separated the clusters were in a 2D space.

E. Customer Transition Modeling Using Markov Chains

After segmenting customers into clusters using K-Means, we observed how these customers moved between clusters over time, particularly across the first three months of the year. To model these transitions, we used the Markov Chain Model which is a stochastic model that represents transitions between different states with a set of probabilities. In our case, each "state" is a behavioral cluster, and we are interested in estimating the likelihood that a customer moves from one cluster to another between months.

Consider a scenario in which customers are grouped into clusters based on their spending behavior—Cluster 0, Cluster 1, and so forth. At the end of each month, customers may remain in their assigned cluster or transition to a different one; for example, a customer categorized as "Low Engagement" in one month may shift to "Routine User" the next. Figure 3 captures these transitions by illustrating how the relative size of each cluster changes over time. The horizontal axis represents sequential months, while the vertical axis denotes the proportion of customers within each cluster. Notably, Cluster 4 increases in prominence during the second month, and Cluster 2 expands sharply in the third. These structural shifts in behavioral segments form the foundation for Markov Chain modeling, which estimates the probability of customers

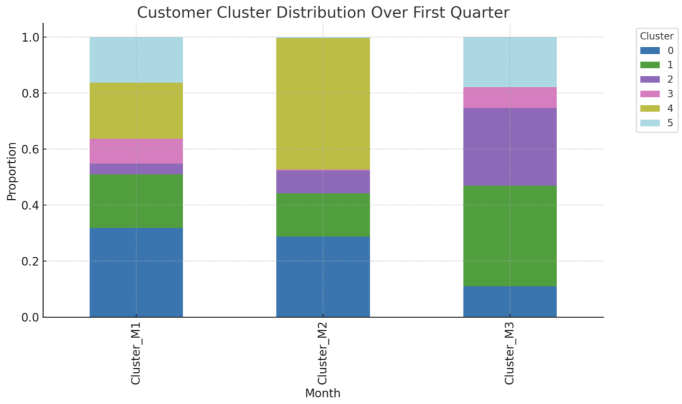

Fig. 3. Cluster distribution over Q1, indicating behavioral shifts.

transitioning between clusters based solely on their most recent state.

So we constructed a transition matrix $\mathbf{P}$ using the actual cluster labels from January to February and February to March. Each element P_{ij} of the matrix represents the probability of a customer transitioning from Cluster i to Cluster j:

$$P_{ij} = \frac{\text{Customers transitioning from Cluster } i \text{ to } j}{\text{Total customers in Cluster } i}$$

This gives us a matrix of the form:

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & \dots & P_{05} \\ P_{10} & P_{11} & \dots & P_{15} \\ \vdots & \vdots & \ddots & \vdots \\ P_{50} & P_{51} & \dots & P_{55} \end{bmatrix}$$

Each row of the matrix sums to 1, since every customer must transition to some cluster, including the possibility of remaining in the same one.

To apply the model, we treated the cluster distribution in March (Month 3) as a state vector $\mathbf{x}_t$, where each element represents the proportion of customers in a given cluster.

To predict the distribution in April (Month 4), we used the following expression:

$$\mathbf{x}_{t+1} = \mathbf{x}_t \cdot \mathbf{P}$$

By multiplying the current cluster distribution with the transition matrix, we can estimate where customers are likely to move next, helping us forecast behavior based on previous trends.

IV. Results and Discussion

A. Behavioral Trends in Transaction Data

The exploratory data analysis gave us insights into how customers interact with their accounts, highlighting patterns in transaction frequency, spending categories, and balance fluctuations. All of which helped feature selection and modeling approach.

The difference in how frequently customers interacted with their accounts was quite noticeable. Some displayed regular activity, such as daily or weekly transactions, while others used their accounts more sporadically. These differences suggested the presence of distinct behavioral profiles, ranging from highly engaged customers with consistent routines to occasional users with more ad hoc financial habits. Spending behavior also varied significantly. Majority of transactions were concentrated in categories such as Retail & Fashion, Financial Services, and Supermarkets followed by smaller, less frequent purchases in areas like Books & Arts, Food & Beverage, and Entertainment & Gaming.

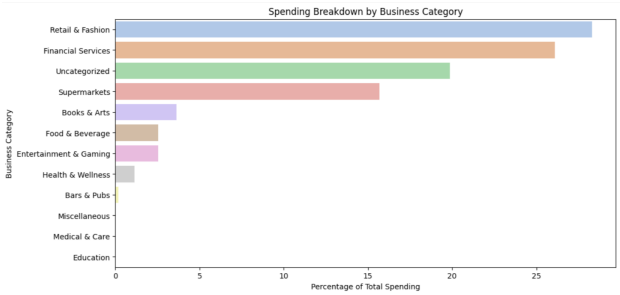

Fig. 4. Spending Breakdown by Business Category.

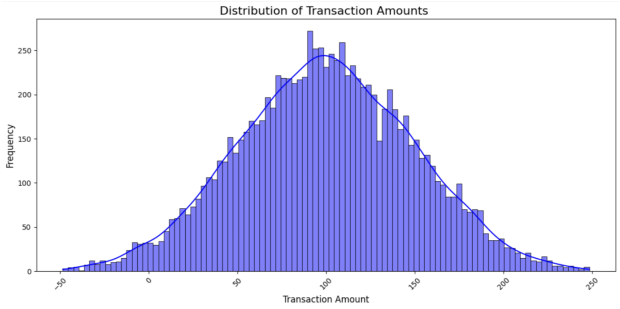

Fig. 5. Distribution of transaction amounts showing a near-normal spending pattern.

The distribution plot (Fig.5.) for transactional amounts indicated a normal spread of spending, with most transactions falling within a low to mid-range bracket. This reflected a tendency toward routine financial activity and supported the assumption that the data was designed to simulate everyday banking behavior. While few extreme transactional values were present, they were minimal and did not significantly skew the overall distribution. Likewise, the correlation between `Amount` and `Balance` helped establish outliers, specifically in accounts with highly variable transactions. Some customers maintained stable balances throughout the year, while others experienced noticeable fluctuations. Based on these observations, we decided to flag overdraft accounts, which offers a starting point for future financial vulnerability analysis.

The patterns observed in daily spending trends, there were recurring spikes in transaction activity at the beginning of each month, likely reflecting simulated salary deposits, rent payments, or monthly subscriptions. Between these peaks, daily

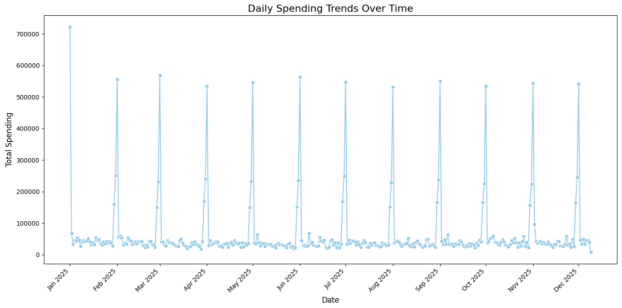

Fig. 6. Daily spending trends highlighting periodic spikes at the start of each month.

activity remained relatively steady, indicating consistent but lower engagement. These trends highlighted the importance of incorporating features such as `Recency` and `Frequency` into our customer segmentation approach.

B. Behavioral Transition Modeling Using Markov Chains

To evaluate the performance of the Markov Chain model, we compared its predicted distribution of customer clusters for April with the actual observed distribution.

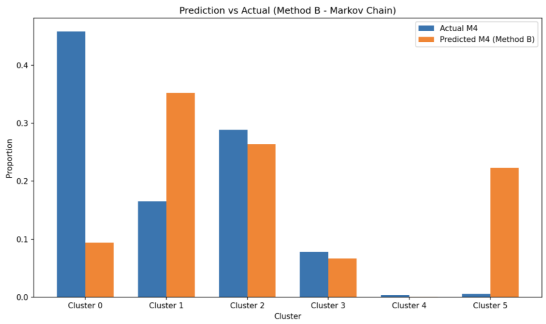

Fig. 7. Predicted vs actual cluster distribution for April using the Markov Chain model. Predictions are based on Q1 transition data.

The results revealed distinct behavioral shifts across clusters. Cluster 0, customers with consistent and moderate transaction activity, showed the largest discrepancy. The model projected that only 9.2% of users would remain in Cluster 0, whereas the actual observed proportion was significantly higher at 46.8%, resulting in a prediction error of over 37%.

In contrast, the model predicted that 35.6% of customers would fall into Cluster 1 (disengaged users), while the actual proportion was only 16.4%, resulting in an overestimation of approximately 19.2%.

Whereas, clusters 3 and 4 demonstrated relatively accurate predictions by the model. For Cluster 3, which comprised of high-value but infrequent transactors, the predicted proportion was 6.1%, while the actual share stood at 7.4% — a difference of just 1.3 percentage points. Similarly, for Cluster 4, representing low-spending but steady users, the prediction was 0.3% compared to an actual value of 0.4%, with a marginal error of 0.1 percentage points.

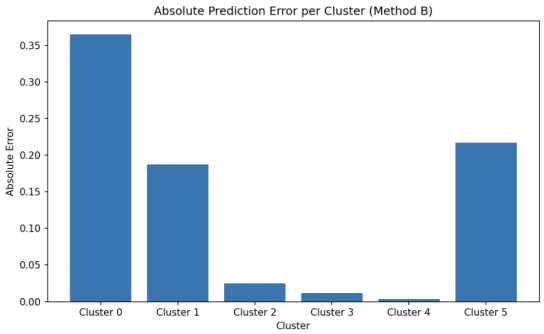

Fig. 8. Absolute prediction error for each cluster using the Markov Chain model.

The prediction errors observed in the Markov Chain model point to two distinct challenges. Cluster 0, stable and consistently engaged users, was significantly underestimated. Its lower representation in Month 3 led the model to assign it less weight during training, causing it to overlook the cluster's reappearance in Month 4. This indicates a broader tendency of the model to under-predict clusters that are less frequent in the training data. In contrast, the model failed to capture the sharp rise in Cluster 5 during Month 4, as this behavior had not been previously observed in the training data. Because the Markov model relies entirely on historical transitions, it lacks the flexibility to adapt to sudden behavioral shifts not previously observed in the data.

These results highlight the key limitations of using a first-order Markov chain to predict customer behavior. Since this method relies on the immediate previous state when estimating transitions, it cannot account for a longer behavioral history or external variables that may influence customer actions. As a result, the model tends to misrepresent segments where outliers play a significant role.

Future versions of the model could benefit from incorporating longer behavioral timelines and more detailed financial signals, such as income frequency, fixed expenses, and average monthly spend. Introducing higher-order Markov Chains or Hidden Markov Models (HMMs) would allow for more context-aware transitions between segments. With added features such as transaction categories, recurring salary patterns, and credit score trends, the model would more closely reflect real-world banking behavior. To quantify the financial stability of customers, the standard deviation of account balance over time can be computed.

$$B_i = \sigma\left(\text{Balance}_i(t)\right)$$

where $\text{Balance}_i(t)$ is the balance at time t, and σ denotes the standard deviation. These future enhancements could enable Lloyds to anticipate customer needs more accurately and deliver personalized support.

V. Further Work and Improvement

While the current project focused on segmenting and modeling customer behavior using transactional features like Re-

cency, Frequency, and Monetary value, there are several areas where the analysis could be extended to gain deeper insights and help design more tailored financial support for customers.

Looking more closely at credit-to-debit ratios could offer informative insights into financial stability. Customers consistently spending more than they receive may be at higher risk of financial strain or potential overdraft, while those with healthy ratios may indicate a stronger ability to manage finances. Although we explored this ratio early on, it was not included in the clustering model because it was highly variable and influenced by individual financial contexts (e.g., salary schedules or loan payments). If we had more detailed income data, we could use this feature more reliably to better understand customer behavior and spot potential risks.

Similarly, we added an overdraft flag to mark when a customer's balance dropped below zero which is a simple way to spot possible signs of financial difficulty. While this flag was helpful in identifying at-risk behavior, we did not include it in the prediction model because it did not follow a clear pattern and was rather ambiguous. In future, this could be explored by looking at how often overdrafts happen, how long they last, and how quickly customers recover. These patterns could be analyzed over time using tools like time-series models or anomaly detection to help catch early signs of financial stress. Adding features like credit scores, overdraft history, and income patterns could make future models more accurate and insightful.

VI. Conclusion

This study offered a glimpse into the behavioral patterns that quietly shape how individuals interact with their finances. By applying the RFM model to monthly customer activity, we were able to segment individuals into distinct behavioral clusters using K-Means. These clusters highlighted variations in spending patterns, engagement levels, and transaction frequency, offering a practical framework for understanding customer needs and tailoring services accordingly.

To move beyond static segmentation, we modeled customer transitions between segments over time using a first-order Markov Chain. While the model was able to capture general directional trends in behavior, it revealed limitations in capturing more complex or persistent patterns. Particularly in the case of stable users and disengaged customers.

A key takeaway is that, this work sets the foundation for future developments. Metrics like credit-to-debit ratio and overdraft frequency, though initially explored, could be further developed into meaningful features with the right supporting data, such as income history or credit scores. Expanding the modeling approach to include higher-order or probabilistic models could provide greater accuracy and context, helping banks like Lloyds identify risk, personalize support, and build more resilient financial relationships with their customers.

References

[1] D. Chen, S. Sain and K. Guo,"Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 3, pp. 197–208, 2012.

[2] T. Jiang and A. Tuzhilin,"Improving personalization solutions through optimal segmentation of customer bases," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 3, pp. 305–320, 2008.

[3] E. Wong and Y. Wei,"Customer online shopping experience data analytics: Integrated customer segmentation and customized services prediction model," International Journal of Retail & Distribution Management, vol. 46, no. 4, pp. 406–420, 2018.

[4] Z. Huang,"Extensions to the K-Means algorithm for clustering," Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283–304, 1998.

[5] F. Cao, J. Liang and B. Liang,"A new initialization method for categorical data clustering," Expert Systems with Applications, vol. 36, no. 7, pp. 10223–10228, 2009.

[6] A. Ahmad and L. Dey,"A K-mean clustering algorithm for mixed numeric and categorical data," Data and Knowledge Engineering, vol. 63, no. 2, pp. 503–527, 2007.

[7] D. Marutho, S. H. Handaka, E. Wijaya and Muljono,"The determination of cluster number at K-mean using elbow method and purity evaluation on headline news," in Int. Seminar on Application for Technology of Information and Communication, Semarang, Indonesia, pp. 533–538, 2018.

[8] F. Liu and Y. Deng,"Determine the number of unknown targets in open world based on elbow method," IEEE Transactions on Fuzzy Systems, vol. 29, no. 5, pp. 986–995, 2021.

[9] P. Pfeifer and R. Carraway,"Modeling customer relationships as Markov chains," Journal of Interactive Marketing, vol. 14, no. 2, pp. 43–55, 2000.

[10] M. R. Vela and E. M. García,"A segmentation analysis and segments profile of budget air travelers," Cuadernos de Turismo, vol. 26, pp. 235–253, 2010.

[11] H. S. Hwang,"A stochastic approach for valuing customers: A case study," International Journal of Software Engineering and Its Applications, vol. 10, no. 3, pp. 67–82, 2016.

[12] C. J. Cheng, S. W. Chiu, C. B. Cheng and J. Y. Wu,"Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan," Scientia Iranica, vol. 19, no. 3, pp. 849–855, 2012.

[13] Ari Freedman,"Convergence theorem for finite Markov chains," in Proc. REU, Chicago, IL, USA, 2017.

[14] C. H. Cheng and Y. S. Chen,"Classifying the segmentation of customer value via RFM model and RS theory," Expert Systems with Applications, vol. 36, no. 3, pp. 4176–4184, 2009.

[15] V. Aggelis and D. Christodoulakis,"Customer clustering using RFM analysis," in Proc. 9th WSEAS Int. Conf. on Computers, Wisconsin, WI, USA, pp. 1–5, 2005.

[16] O. Netzer, J. M. Lattin and V. Srinivasan,"A hidden Markov model of customer relationship dynamics," Marketing Science, vol. 27, no. 2, pp. 185–204, 2008.

GitHub URL: https://github.com/EMATM0050-2024/dsmp-2024-groupm17