Applications of Predictive Modelling Through Uses of Customer Segmentation, Clustering analysis and ARIMA on Bank Customer Transaction Data

Vidit Sheth
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
vidit.sheth.2023@bristol.ac.uk

Tiancheng Hu
Department of Engineering Mathematics)
University of Bristol
Bristol, United Kingdom
if23971@bristol.ac.uk

Hassan Hassan
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
ze23886@bristol.ac.uk

Shubham Yadav
Department of Engineering Mathematics
University of Bristol
Bristol, United Kingdom
shubham.y.2023@bristol.ac.uk

***Abstract*—Bank transactions, deliver a wide range of customer activities, they serve as a foundational dataset for analyzing financial behaviors and optimizing banking operations. This project, conducted on artificially generated transactional data provided by Lloyds Bank at the University of Bristol, involved comprehensive data analysis of consumer transactions to derive actionable business insights. The study involved 2 datasets, both with different features. The analysis was started by understanding the data and conducting RFM analysis to evaluate customer value based on transaction recency, frequency and monetary value. This analysis helps the bank tailor and optimize their services to each customer. Furthermore, clustering was been implemented to analyze spending patterns of the customers. A time series approach to forecast additional records by using an Autoregressive integrated moving average (ARIMA) model which provided a view of future customer activity and financial trend. Finally, CatBoost was implemented to predict customer churn in the dataset, so as to focus on retention strategies and customer loyalty.**

I. Introduction

With the onset of greater dynamism in the current industrial climate, firms around the world have been faced with the pressures of catering to customers. [1] Customer relationship management is a broadly recognized strategy for acquisition and retention of customers. The main objective of developing customer relations is to make long-lasting and profitable relationships with customers. [2][7] This paper focuses on advanced data analytics to understand customer behavior based on their financial transactions. The goal of this study is to discover trends that can predict future behaviors, and enhance personalized services offered by the bank to maintain and increase customer retention. To deal with the challenges, RFM (Recency, Frequency, Monetary) analysis was employed by us in order to compliment customer segmentation and to inform strategies catered to customer profiles we defined. Further, we made use of an Autoregressive integrated moving average (ARIMA) model to forecast future behaviors of customer transactions to gain an insight into future spending of customers; these could help the bank plan offers on seasonal spending beforehand across the entire year (such as festive spending). Finally, occurrences of customer churn was predictively modelled, using CatBoost, to understand customers drifting away from the bank or switching to competitors.

II. Literature Review

A literature survey was carried out to understand the breadth of methodologies which have been applied in this area, with keener focus given to those applied with success. This started by looking at one paper which focused on the use of customer segmentation on banking datasets, carried out using data mining algorithms to analyze the relationships among respective banks' customers and their transactions. The segmentation helps the bank break down their consumer base in order to offer targeted strategies that focus on individual needs, e.g. risk management. The paper aimed to enhance the effectiveness of customer engagement strategies and lower financial risks associated with banking operations. [3] Another paper dug into deployment of machine learning techniques for fraud detection in banking. They targeted high volume and multifaceted transactions and fraud schemes. The study by Hashemi et al. used class weight-tuning hyperparameters combined with Bayesian optimization to fine-tune machine learning models such as CatBoost, XGBoost, and LightGBM. Their framework enhanced the accuracy and efficiency of fraud detection systems, equipping financial institutions with more robust tools to detect fraudulent activities early and reduce potential losses through more effective and proactive fraud prevention strategies. [4] Another interesting paper by

Zakrzewska et al. focused on application of clustering algorithms for bank customer segmentation, a crucial aspect of knowledge-based marketing in the banking sector. They targeted the challenges of large and multidimensional databases by comparing results of three clustering algorithms: density-based DBSCAN, kmeans, and a two-phase clustering process based on k-means. The research curated evaluation of these algorithms, based on scenarios of high dimensionality with noise aiming, to identify the most suitable approach for segmenting bank customers effectively. [5] Studies related to RFM (recency, frequency, monetary) modeling and rough set theory (RS theory), combined with the K-means algorithm to enhance data mining processes in customer relationship management (CRM) by Cheng and Chen address dividing continuous attributes and using rough set theory to improve the data handling procedure and shorten training times. Their study overcomes the drawbacks of conventional data mining methods. This approach clusters customer value into different classes [3] [5] [7] to evaluate which offers the best accuracy in representing customer loyalty and assists in identifying key customer characteristics that can strengthen CRM strategies. [6]

III. Methodology

This study comprised an analytical approach to treat generated transactional data from Lloyds Bank. The aim of this study to was to identify trends, spending patterns and any possible insight beneficial for the bank to enhance their services for the their customers. The services could range from offering Premium Accounts with benefits for savings rate and rewards on transactions to their best customers, to retaining and establishing customer relations to the customers that have not been engaging enough with the bank. Furthermore, customer churn was analyzed to check and engage with customers before they quit services from the bank. The following make up the analysis done across the study:

A. RFM Analysis

Customer segmentation distinguishes the best customers that have been engaging with the company when compared to the total number of users. RFM analysis is a powerful technique used to categorize customers based on their transactional behavior. The customers are categorized by analyzing three specific aspects, Recency - How recently they have made a transaction, Frequency - How frequent they do transactions, and Monetary - How much they spend in terms of the monetary value during their transactions. These classifications are made by assigning R, F and M scores to individual accounts by setting thresholds. This is done by splitting individuals (identified via their respective account numbers) into quartiles. The scores are allocated in such a way that accounts falling over 75% or the third quartile receive the highest score , i.e. 4, and accounts falling under 25% or the first quartile are given the score 1.

1) Assigning R, F and M scores::

- The R Score is a critical metric that evaluates the time elapsed since a customer's last transaction. In this study, the R score is calculated based on the 'Recency' feature, which measures the number of days since the last transaction was made from a particular account. The R Score depends on the Recency value; the lower the Recency the higher the score.
- The F Score measures how often a customer a made transactions over a particular period. Considering the size of the dataset, this study takes into account the total period when all the transactions were made. The total frequency of transactions for all accounts is defined by the feature, 'Frequency'. The F Score depends on the Frequency; the higher the frequency, the higher the F score.
- The M Score quantifies the total monetary value a customer spent over the period, providing a sum of financial contribution towards the bank. The metric has been calculated based on a feature, 'Monetary'. The data points store in 'Monetary' represent the total sum of amount that an individual account has spent over the entire period. Therefore, the higher the monetary value, the higher the M Score.

The RFM Score is the sum of all the three metrics which is the final output for the analysis. The highest score possible is 12, which indicates accounts with recency, frequency and monetary value each lying in the top quartile of the dataset. These customers have been filtered out in the study as having the "perfect" RFM Score.

2) Customer Segmentation based on RFM metrics: Customer segmentation with RFM Analysis refers to analysing the distribution of Recency, Frequency and Monetary values to categorize customers into distinct groups with respect to their transactional behavior. After careful analysis from the RFM metrics, a total of five segments, representing customer profiles, were selected. The following are the customer segments with their respective threshold values:

- High-Value Customers: Accounts with the highest scores in recency, frequency, and monetary values. All metrics with a score in the top quartile.
- Loyal Customers: Accounts with high frequency and monetary scores , irrespective of their recency score. F Score and M Score which is equal to or above the 3rd Quartile.
- Emerging Customers: Accounts with the highest recency values but lower frequency and monetary values. R Score is above the 3rd Quartile ,but the F Score and the M Score is below the 3rd Quartile.
- Lost Customers: Accounts with low scores across recency, frequency, and monetary. All the metrics are less than or equal to the 2nd Quartile or the mean.
- Need Attention Customers: Accounts with medium scores in all categories. All metrics range from the 2nd Quartile to the 3rd Quartile.

B. Clustering with K-Means and DBSCAN

Clustering is an unsupervised machine learning statistical technique used to group data points which are alike with each other to form a group. These groups are known as clusters. This technique is useful to identify patterns and create segmentations in the data. Some popular clustering techniques include K-Means, K-Medoids, hierarchical clustering, density based spatial clustering, etc. In this study, the dataset has been evaluated for K-Means clustering and DBSCAN.

1) K-Means Clustering: This clustering technique demands a number of clusters that the algorithm should look for before it execution. This value 'K' determines the number of clusters. After preprocessing the data for clustering, selecting the optimum number of clusters is particularly important. A technique coined as Elbow method is used in the study to find the optimum number of clusters required for the dataset. The optimum number of clusters for our dataset was found to be four. After selecting the number of clusters, K-Means algorithm was defined and fitted to the data which gave four clusters. They have been visualized further with the help of Principal Component Analysis which helped in dimension reduction.

2) DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering technique that identifies clusters based on regions of high density separated by regions of low density. Unlike many traditional clustering techniques, DBSCAN does not require the number of clusters to be predetermined. This feature makes it particularly effective for analyzing complex datasets where the number or shape of clusters is not known beforehand. It works by exploring the dataset, classifying any points that do not meet the minimum density criteria as noise, thereby highlighting both the clusters and the gaps or noise between them. Implementing this technique gave out a more than a hundred clusters, therefore the method was not taken ahead for further analysis.

C. ARIMA for forecasting future Volumes

Autoregressive Integrated Moving Average is a statistical model which takes time series data to forecast future values. The model comprised of three components, p - referring to the number of autoregressive terms, d - representing the number of nonseasonal differences required for stationarity and q which is the number of lagged forecasts errors in the prediction equation. This study utilized ARIMA to forecast 30 days' worth of new data aiming to predict future bank transactions. The optimal parameters p, d and q values were selected after trying multiple combinations that would minimize the AIC (Akaike Information Criterion) score. After the model was fitted, it was made sure that residuals were independent and normal. Finally, the model forecasts were projected where balances for the subsequent month were given out for each account present in the dataset.

D. Churn Prediction

The process of analyzing the dataset to identify patterns and signals that could lead to a customer discontinuing the banking service is known as churn prediction. It could be done using various features such as transactional behavior, customer interactions, demographic information, and so on. In this study, the features 'Balance' and 'Amount' have been taken into consideration for training a CatBoost model which is well suited for a challenge like forecasting churn. The model is a good pick as it is robust enough to fight against overfitting, as well as its ability to deal with class imbalance. In the majority of cases for churn predictions, the data points for actual churns are very much low, leading to class imbalance. The forecasts successfully predicted churns with a satisfactory accuracy. The insights could be used to develop strategies to reduce churn and improve customer relationship by highlighting individuals which require attention to improve their levels of customer satisfaction.

IV. Data Description / Preparation

The initial data exploration was done on our first dataset ("dataset one"), where the data was cleaned and checked for outliers. This cleaning allowed for further steps to be implemented. Feature scaling for creating new features was applied firstly to create a new Transaction category which would categorise all the amounts and contain the nature of the transaction. The categories ranged from 'Very small transactions' to 'extreme transactions'. Secondly, a new feature was defined to divide the transactions into various categories of transaction, e.g. shopping, cinema, restaurant, grocery, and so on. Furthermore, statistical metrics such as mean, variance, range of transactions for individual categories were calculated and visualised.

The second dataset contained additional features, 'Balance' and 'Timestamp'. These features were analysed to implement the methodologies used in this study to a major extent. All the steps including cleaning the dataset, feature engineering and general statistics calculation were performed on the second dataset as well. The further preparation to execute all the methodologies are mentioned below:

A. RFM Analysis

There were no additional data cleaning steps involved to perform RFM Analysis on the dataset. However, the feature 'Date', was converted to a datetime object as it was initially a string data type. The further steps involved in conducting the analysis were included in the Methodology section.

B. Clustering with K-Means and DBSCAN

During the pre-processing stage of the dataset with the aim of conducting clustering, it was found that the features 'Balance' and 'Amount' had missing values. These were imputed with the median rather than the mean as the latter is sensitive to outliers and is not always the best fit for financial data. The median values were taken after grouping all the transactions with their respective account numbers. Furthermore, the dataset contained records with null values for transactions that had a single transaction, which were imputed with the overall median for the respective features.

After this, the features were scaled using standard scaler, and the scaled features for 'Balance' and 'Amount' were used for the clustering models. One-hot encoding was used to encode the categorical feature 'category' before moving ahead with the analysis.

C. ARIMA Forecasting

To perform forecasting of new 'Balance' data points from the existing transactions in the dataset there was pre-processing which required the conversion of the 'Date' feature from string to a date format. Time stamp was not required for the analysis, therefore it was dropped. The forecasted dates were calculated by adding thirty days to the last date of the transaction for a particular account. The model was then fitted and run to predict the forecasted values for the 'Balance' feature.

D. Churn Prediction

The pre-processing steps for predicting churn also involved converting the 'Date' feature to a standard date format. The data was imported every time a new methodology was applied so as to avoid any imputed parameter used in any other step. Furthermore, the `Current_date` looks for the most recent date of transaction for an account and then adds a day in order to calculate the number of days since the last transaction. The $mathttDays\_since\_last_{t}ransaction$ then finally subtracts the final transaction date for each account from the `Current_date` to give out the total number of days since the last transaction was recorded. A threshold of more than ninety days was selected to define an account as churn and the target variable in this study. The training variables, 'Balance' and 'Amount' were further imputed for missing values and scaled before the data was split into training and testing. The test size of the data was 20% and the training size was 80%. The CatBoost model was further trained for a thousand iterations with a lower learning rate to achieve better performance. The evaluation metric for the model was set to AUC as it tests the model for binary classification. Finally, class weights were given so that the model pays attention to the minority class, which in this study corresponds to the target variable, churn.

V. Results and Discussion

A. RFM Analysis

The following are the results of the RFM analysis as discussed above. As shown in Figure 1, the individual R, F, and M scores were calculated on the basis of features - Recency, Frequency, and Monetary.

The figure2 is a snippet for the Best Customers who have been classified as the accounts with the highest possible RFM Score.

The figure3 3 is a plot which demonstrates the distribution of the customer segments, as mentioned in the methodology. In order to maximise customer engagement across all segments, 'high-value customers' should continue to get premium services and rewards for their loyalty. Personalised marketing and loyalty initiatives can help keep 'loyal customers'. Personalised suggestions and welcome incentives can help elevate 'emerging customers' and boost their transaction volume. 'Lost customers' should be convinced to return with customised solutions and incentives based on their feedback; and customers who 'need attention' require more personalised offers and promotions to boost their spending and loyalty.

Account No	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
101531259.0	2	123	-716.08	3	1	2	6
104832000.0	4	164	-4565.52	2	2	1	5
105375973.0	4	158	-193.77	2	2	3	7
106601471.0	1	193	10649.00	4	2	4	10
108481285.0	1	368	10038.92	4	4	4	12

Fig. 1. RFM Scores based on Recency, Frequency, and Monetary features

Account No	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
399538448.0	1	379	6965.00	4	4	4	12
719586818.0	1	336	20576.73	4	4	4	12
558119802.0	1	476	37015.18	4	4	4	12
802697323.0	1	425	17227.74	4	4	4	12
930277104.0	1	364	8672.71	4	4	4	12

Fig. 2. Best Customers with the highest RFM Score

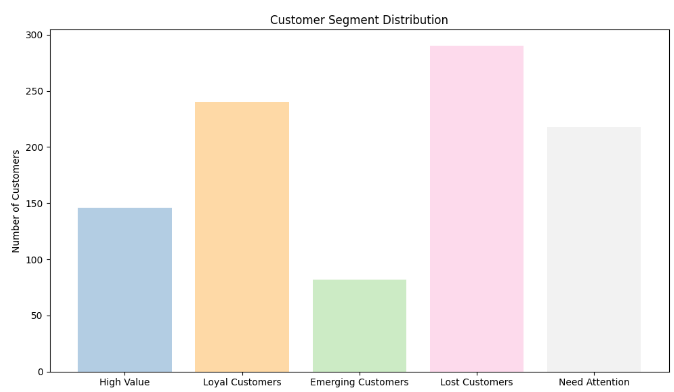

Fig. 3. Customer segmentation based on RFM Analysis with thresholds

B. Clustering

Implementation of K-Means after implementing the Elbow method gave four distinct clusters. DBSCAN on the other hand gave out a total of four hundred and fourteen clusters. Therefore K-Means was the accurate model chosen for this study. Following is the plot for the Elbow method, suggesting that the number of optimum clusters for the dataset is 4:

Furthermore, below is the visualization plot for the clusters with reduced dimensions applying Principal Component Analysis.

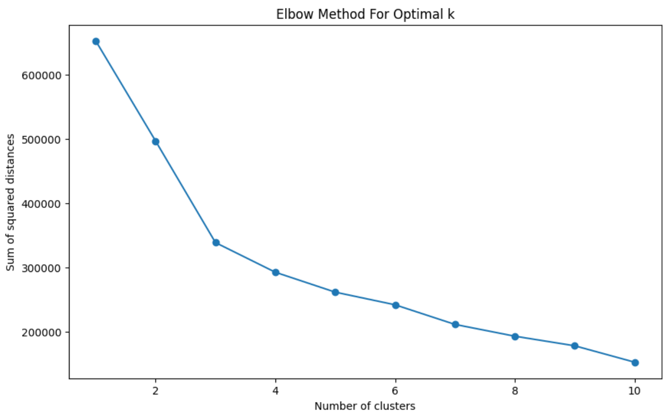

Fig. 4. Elbow plot for K-Means Clustering

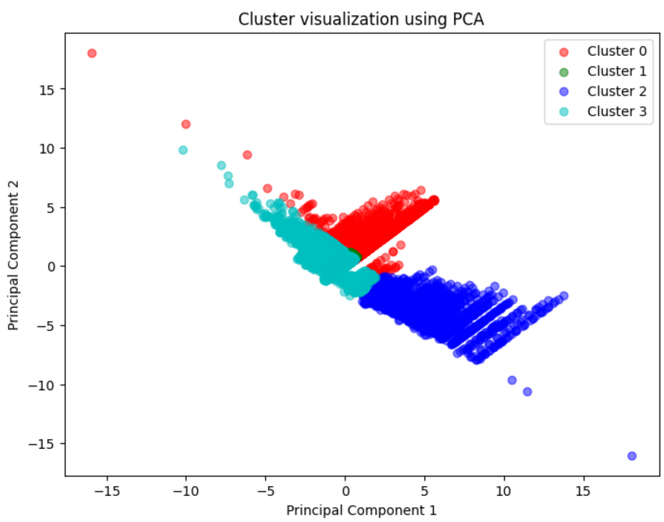

Fig. 5. K-Means with 4 Clusters

C. ARIMA Forecasts

The output from the fitted ARIMA model provided forecasted balances for the next 30 days for each account, which were then saved into a CSV file. The following is depicted in Figure6. Such forecasts can significantly aid institutions like Lloyds Bank, offering insights that facilitate strategic planning and decision-making on a large scale. Strategies like liquidity management for daily operations and loan disbursements could be managed with ease when a model gives out forecasts with good accuracy.

D. Churn Prediction

The rigorously trained CatBoost model was trained with scaled balances and amounts to predict churns. In the following figure, each row corresponds to an individual customer account with the forecasts and the actual churn occurrences. The (1) indicates an actual churn while (0) in the `Predicted_class` indicates the model forecasting a churn as not churn. The model was trained for 1000 iterations although the AUC score of 89.59% was achieved while testing on the

	Account No	Forecast Date	Forecasted Balance
0	101531259.0	2023-12-06	191.548252
1	101531259.0	2023-12-07	168.958249
2	101531259.0	2023-12-08	147.622271
3	101531259.0	2023-12-09	127.470706
4	101531259.0	2023-12-10	108.437803
5	101531259.0	2023-12-11	90.461463

Fig. 6. Head: ARIMA Forecasts with balances for all the accounts for the next 30 days

149th iteration. The forecasts were further saved into a CSV file.

	Account No	Predicted_Class	Churn
89	183546640.0	1	1
210	293191074.0	0	1
276	357651163.0	1	1
433	509627909.0	1	1
488	554771381.0	0	1
500	567499591.0	1	1
525	592961759.0	1	1
623	671824917.0	0	1
649	690941877.0	1	1

Fig. 7. Churn Forecasts by the CatBoost model

VI. Further Work and Improvement

The points of focus for further improvement of this analysis would be to increase the accuracy for the ARIMA and CatBoost model in order to achieve comprehensive results that could contribute to financial institutions in terms of monetary gains. A churn prediction model with a near-perfect accuracy would be highly valuable to high street banks, as well as financial institutions more broadly. Furthermore, this study could be extended to predict anomalies in the transactions. Detecting fraudulent transactions can contribute to a better risk assessment service for the bank and other financial institutions. Anomaly detection can help institutions identify potential threats, improving customer safety. The study could also be extended by using deep learning models that have a higher capacity of learning assisted by hidden layers. These models can be extremely helpful when working with large volumes of data. Recurrent neural network models such as Long-short term memory networks could also be a good option to understand customer behavior and predict churn.

VII. Conclusion

In this work, RFM Analysis has been implemented to comprehend customer segmentation and identify the best customers with the highest Recency, Frequeny and Monetary aspects. The analysis was further extended to segment the customers based on R, F and M scores. Strategies and customer relation improvement services were recommended for each segment of customers.

Next, K-Means and DBSCAN clustering algorithms were used to identify customer behavior and patterns with respect to clusters. The clusters were defined by the elbow curve plotted by analysing the sum of squared distances from the data points to the closest cluster center point.

The ARIMA model was also implemented to forecast balances for a period of 30 days from the last date of transaction for all accounts. The results were saved in a forecasts.csv file that is available on the git-hub repository.

Finally, to detect churn in the dataset, a CatBoost model was developed to predict churns in the dataset that would fall under the category where no transactions were made for a period over 90 days. The model was trained to complete 1000 iterations and gave an AUC testing accuracy of 89.59%, after 149 iterations.

References

[1] Huan-Ming Chuang and Chia-Cheng Shen, "A study on the applications of data mining techniques to enhance customer lifetime value — based on the department store industry," 2008 International Conference on Machine Learning and Cybernetics, Kunming, China, 2008, pp. 168-173, doi: 10.1109/ICMLC.2008.4620398.

[2] Ravi, Vadlamani. (2007). Advances in Banking Technology and Management: Impacts of ICT and CRM. 10.4018/978-1-59904-675-4.

[3] Shokrgozar, Neda. (2016). Customer Segmentation of Bank Based on Discovering of Their Transactional Relation by Using Data ?Mining Algorithms. Modern Applied Science. 10. 283. 10.5539/mas.v10n10p283.

[4] Hashemi Seyedeh , Mirtaheri Seyedeh , Greco Sergio. (2022). Fraud Detection in Banking Data by Machine Learning Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2022.3232287.

[5] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," 5th International Conference on Intelligent Systems Design and Applications (ISDA'05), Warsaw, Poland, 2005, pp. 197-202, doi: 10.1109/ISDA.2005.33.

[6] Cheng Ching-Hsue, Chen You-Shyang. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications. 36. 4176-4184. 10.1016/j.eswa.2008.04.003.

[7] Lloyds Bank – Building Customer Relationships - Business Resource Centre www.lloydsbank.com. https://www.lloydsbank.com/business/resource-centre/business-guides/building-customer-relationships.html (accessed May 01, 2024).

Appendix

The document up to this section should be no more than 8 pages. The appendix section is optional. You can include additional material here, but it will not be marked.