Segmenting Retail Banking Customers Using Unsupervised Machine Learning

Daniel Sutton  
Department of Engineering Mathematics  
University of BristolBristol, United Kingdom  
ee19836@bristol.ac.uk

Hannah Siddiqi  
Department of Engineering Mathematics  
University of BristolBristol, United Kingdom  
pt22769@bristol.ac.uk

Oltion Fazliu  
Department of Engineering Mathematics  
University of BristolBristol, United Kingdom  
xz22993@bristol.ac.uk

*Abstract*—

# Introduction

In modern day retail banking, banks generate huge amounts of data in the form of customer transactions. One of the most important ways banks can use transactional data is for customer segmentation, allowing a bank to identify groups of customers who have similar spending characteristics which can then be targeted with tailored marketing efforts. By accurately segmenting customers in different ways, banks can tailor customer marketing to meet the needs of each individual customer, thus improving customer satisfaction and loyalty to the bank. Incorrect segmentation could lead to customers being ‘bombarded’ with irrelevant offers, causing potential missed revenue opportunities and ultimately the loss of the customer altogether to another bank.

This report investigates how a bank could use different unsupervised machine learning algorithms in order to better cluster customers. The algorithms will be applied alongside the RFM (Recency, Frequency, Monetary Value) model to find optimal customer segmentations for different categories of spending.

For this project, transactional data was provided by Lloyds Banking Group. The data has been artificially produced, and it includes a customer ID, the monetary amount of the transaction, the recipient of the transaction and finally the date of the transaction. In total there are over one million transactions in the dataset, spanning over a 9-month period.

# Literature Review

An overview of related work of similar research in the domain.

# Methodology

The focus of this study was chosen to be on behavioral based segmentation. This involves segmenting customers on purely spending habits; grouping customers who have similar spending characteristics. The main motivation for this is so that banks can offer personalized marketing offers at customers who would be more likely to be interested in taking up such offers. For example, there is no point in a bank offering a premium credit card that for a fee offers luxury bonuses to a customer who has low spending and thus would be less likely to be interested in such an offer.

## RFM

Our main approach for segmenting customers will be using the RFM model. This is a proven successful marketing technique used to quantitively rank and then subsequently group customers based on the recency, frequency and monetary amount of their recent transactions. Groups of customers that fit the required criteria for the target market of a new product can then the subject of targeted marketing. Due to the dataset having no demographic information it was decided that using the RFM model was the best method for grouping customers, as it requires only behavioral information which is all the data set contained.

The RFM model uses the following three factors:

1. **Recency**: The time between the customers last purchase and the most recent date in the table. For this project time was chosen to be days.
2. **Frequency**: The number of transactions made by an individual customer.
3. **Monetary Value**: The amount of money spent by an individual customer.

RFM values will be used alongside unsupervised machine learning algorithms to build 2 different models that a bank can use for customer segmentation.

## K-means Model

A picture containing line, plot, diagram, slope

Description automatically generatedThe first model that will be tested will be a K-means model. A K-means algorithm uses a number k centroids to allocate every data point to the nearest centroid, where a centroid is the center of a cluster. The number of centroids for each test will be found by using the ‘Elbow method’; a commonly used method for finding the optimal k value. To apply the Elbow method, the number of clusters (k) is varied between 1 and 10, where for each value of k the Within-Cluster Sum of Square (WCSS) value is calculated for each value of k. The resulting line graph will be shaped like an elbow, where the point of the elbow is taken as the optimal value of k.

Fig.1. Elbow Method for ‘Dining’ category using RFM

## During the testing of the k-means model, it was decided that k-means should then be performed again on the most valuable group of customers in order to segment this group further. Doing this would provide the bank with a small group of customers that are active, have high frequency and high monetary value, and would thus be customers who would be very likely to be interested in related offers.

## Two Step Clustering Model

The second model that will be tested is a simple two step clustering model. This model will use mean RFM values in order to categories customers into three different clusters as described below:

1. **Low Value**: Customers in this cluster will have lower than average frequency and money values and higher than average recency value.
2. **Potential High Value**: Customers in this cluster will also have lower than average frequency and money values, however they will also have lower than average recency values. Therefore, these are customers who have recently made a purchase in a given segment, and although are currently below average spenders have the potential to become high value customers.
3. **High Value**: Customers in this cluster will have higher than average frequency and money values along with lower than average recency values.
4. **Old High Value**: Customers in this cluster will have higher than average frequency and money values, but will also have higher than average recency values. These are customers who used to be of high value but have become an inactive spender.
5. **Other**: Customers that don’t fit any of the above criteria will be placed into an ‘Other’ group.

One of the main objectives of using this model is to identify customers that fit into the potential high value group. High value customers are very important to a bank as they create a large proportion of revenue, so targeting marketing at potential high value customers can help to increase the number of these customers that do in fact become high value. Identifying old high value customers could also be of use for a bank, as these customers were previously of high value and so targeted marketing here could reactivate, these customers.

It was also decided during testing of this model that the ‘High Value’ group should be iterated on again, where a new group named ‘Extremely High Value’ group can be found by finding customers who have frequency and money values higher than the mean for the ‘High Value’ group, and recency values lower than the mean from that same group. This should provide a very small proportion of total customers who have very high value for the chosen segmentation of spending.

## Normalised RFM model

# Data Description / Preparation

## Data Description

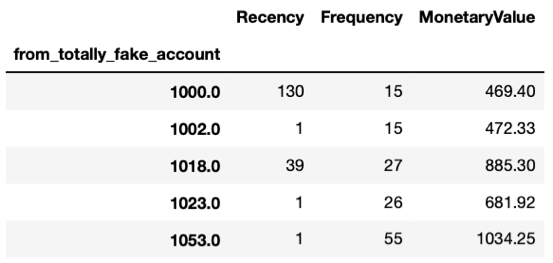
The transactional dataset used to test the different clustering algorithms was the first dataset provided by Lloyds, titled ‘fake\_transactional\_data.csv’. The data had dimension specifications **4x1,048,576**, which although appears extremely large is in reality a greatly scaled down version of the immensely large transactional data large banks handle on a daily basis.

The dataset itself contained four columns. The first, titled ‘from\_totally\_fake\_account’, gave the customer account number of the person making the transaction. In total there were 12,825 individual customers in the dataset. The second column, titled ‘monopoly\_money\_amount’, contained the monetary value of the transaction. The penultimate column, titled ‘to\_random\_generated\_account’, details the receipent of the money from the transaction. Transactions were either being made at retail stores, in which there were a total of 78, or were transactions in the form of bank transfer to other personal banking accounts. The final column in the dataset was ‘not\_happened\_yet\_date’, which gave the date of the transaction. The dataset spanned a period of just over 9 months in total, with the first transactions being made on the 1st January 2025, and the final transactions being on the 2nd October 2025.

## Data Preprocessing

The csv file containing the data was uploaded to a jupyter notebook file in order to be pre-processed. Due to the dataset being completely simulated, there were some inconsistencies within the data and format itself that had to be considered. In all four columns there were numerous examples of missing data, so the decision was made to drop all rows that had missing data in any of the columns.

In order to study customer spending habits in certain store categories it was necessary to group individual stores from the ‘to\_random\_generated\_account’ into several different store categories. Since the data was artificially generated, the store names were not real and so it was down to our discretion to decide which categories to locate certain stores into. Some store names were more ambiguous than others making these more difficult to categorise. In total 10 store categories were created in order to split the 78 stores, with an extra category titled ‘Account Transfer’ [CHANGE PEER 2 PEER] used to categorise all transactions that involved transferring money to another account. Once all stores had been manually assigned to a category, a new column was appended to the data frame titled ‘Category’, which contained the store category for each transaction (or account transfer for these transactions).

Data preprocessing was also required to create different RFM values for each individual model. For the k-means model, raw RFM values were calculated for each individual customer, with values than added into a new pandas data frame that was indexed by each individual account.

*Fig. Example of pandas dataframe with RFM values for each customer appended*.

# Results and Discussion

## Analysis of K-Means Model

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recency** | **Frequency** | **Monetary Value** |
| N | 12825 | 12825 | 12825 |
| Mean. | 1.023 | 903.593 | 16442.472 |
| Std. | 0.195 | 253.374 | 4995.084 |
| Min. | 1.000 | 318.000 | 5172.620 |
| Max. | 6.000 | 1973.000 | 33696.120 |

The first model tested was the k-means model, where k-means was performed twice; firstly for all customers then on purely the group of customers in the ‘top’ cluster. The model was first tested on spending in all categories to identify potential useful clusters for spending in general.

*Fig. RFM statistics for all spending.*

The figure above shows resulting statistics using RFM values for all categories of spending. The table shows how all customers in the dataset are active given the maximum recency value is 6 days and the minimum number of transactions is 318. From this information we can conclude there are no inactive customers, meaning for this study there is no need to create an ‘inacitve’ group of customers that could be targeted to reactivate them.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **R** | **F** | **M** | **N Customers** |
| **0** | 1.0 | 959.2 | 16170.3 | 4092 |
| **1** | 1.0 | 922.3 | 23286.2 | 3626 |
| **2** | 1.0 | 845.7 | 11801.5 | 5107 |

For this test using all spending, it was determined through using the Elbow method that the optimal number of clusters for k-means clustering was 3. The mean values of the 3 clusters are shown in the table below. Given the activeness of all customers, the recency values have little to no relevance for overall spending.

*Fig. RFM mean values and number of customers for each cluster.*

In order to better visualize the calculated cluster, a 3D graph was produced with each distinct colour representing a different cluster. The three axes on the graph are frequency, recency and monetary value.

A picture containing diagram, line, screenshot

Description automatically generated

*Fig. 3D graph showing clusters of customers using RFM values for all spending.*

Green here is cluster 1, and these contain the most valuable customers, with mean values for all 3 variables above average compared with all customers. Cluster 0 contains customers that in fact on average have a greater number of transactions but spend less on average when compared with cluster 1. Cluster 2 contains the least valuable customers, with customers here spending less on average and having fewer than average number of transactions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **New Cluster** | **R** | **F** | **M** | **N Customers** |
| **0** | 1.0 | 851.3 | 21173.0 | 1435 |
| **1** | 1.0 | 927.4 | 23757.2 | 1515 |
| **2** | 1.0 | 1061.7 | 26716.6 | 676 |

In order to segment the most valuable customers, k-means was applied again on cluster 1. Once again through the Elbow method the optimal number of clusters was found to be 3.

*Fig. RFM mean values and number of customers for each new cluster from cluster 1.*

A picture containing screenshot, design

Description automatically generatedThe new clusters for the most valuable customers in all spending categories have been visualized in the 3D graph below.

*Fig. 3D graph showing clusters of most valuable customers using RFM values for all spending.*

The cluster of customers coloured black, labelled cluster 2 in Fig. represent a small group of the most valuable customers. There are 676 customers in this cluster, which is 5.76% of total customers in the whole data set. Average frequency and monetary value in this cluster is significantly greater than overall average, making this group highly valuable. A bank could offer these high, frequent spenders with premium offers.

After this initial testing on spending in all categories, the k-means model was then tested on customer spending in the ‘Dining Out’ store category. This category contained in total 9 different stores, predominantly restaurants.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recency** | **Frequency** | **Monetary Value** |
| **N** | 12825 | 12825 | 12825 |
| **Mean** | 29.420 | 24.929 | 558.963 |
| **Std.** | 32.810 | 16.172 | 251.900 |
| **Min.** | 1.000 | 3.000 | 82.690 |
| **Max.** | 206.000 | 96.000 | 1864.510 |

*Fig. RFM statistics for ‘Dining Out’ spending.*

The table above shows resulting statistics using RFM values for spending in the ‘Dining Out’ category. Compared with overall spending, the recency column in this test now is more useful, as the range of most recent transaction is up to 206 days.

Through using the Elbow method once again, it was determined that the optimal number of clusters for this test was 3 (see fig.1). The mean values of the 3 clusters for the RFM variables are shown in the table below, along with a count of number of customers in each cluster.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **R** | **F** | **M** | **N Customers** |
| **0** | 21.6 | 32.0 | 639.8 | 5068 |
| **1** | 37.9 | 16.5 | 397.4 | 6773 |
| **2** | 10.9 | 46.6 | 1254.6 | 984 |

*Fig. RFM mean values and number of customers for each cluster.*

The graph below shows a 3D representation of the clusters described in the above table.

*A picture containing map, diagram, text

Description automatically generated*

*Fig. 3D graph showing clusters of customers using RFM values for ‘Dining Out’ category spending.*

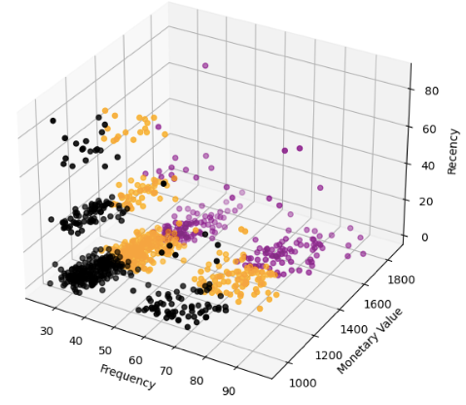
This clustering resulted in two large clusters of customers, clusters 0 and 1, and one much smaller cluster; cluster 2. Cluster 1, the green points on the graph, contaed over half the total number of customers. These are customers that do not regularly dine out, so on average have low frequency and low monetary values and a higher recency value. Cluster 0, the red points on the graph, contains customers that on average have slightly higher spending and frequency values, and lower recency values. The most valuable customers are found in cluster 2, represented by blue points on the graph. Customers here dine out much more than average, thus spending and frequency values are significantly higher than average. Recency average is also much lower than average, highlighting how customers here are very active spenders in this category. The 984 customers in this cluster represent 7.67% of all customers.

Similarly to the method used during testing for all spending, k-means clustering was then performed on this ‘top cluster’ to identify a group of the most valuable customers; very active customers with very high frequency and monetary values.

Once again, the optimal number of clusters was found to be 3. The mean values for these 3 new clusters for the RFM variables are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **New Cluster** | **R** | **F** | **M** | **N Customers** |
| **0** | 10.1 | 46.7 | 1290.0 | 413 |
| **1** | 7.2 | 59.4 | 1544.4 | 195 |
| **2** | 13.7 | 40.0 | 1065.4 | 376 |

*Fig. RFM mean values and number of customers for each new cluster from cluster 2.*

*.*By performing k-means clustering again on the ‘top cluster’ a very small segment of clusters can be identified that contain the extreme end of the most valuable customers. The 3D graph below visualises the resulting clustering.

*Fig. 3D graph showing clusters of most valuable customers using RFM values for ‘Dining Out’ category spending.*

Cluster 1, coloured purple on the graph, contain the most valuable customers. This small group of customers have spent significantly more than average, are the most frequent spenders in this category and are also the most active. The 195 customers here represent 1.52% of all customers. A bank could be highly confident that this small group of customers would be very interested in offers relating to dining out purchases.

## Analysis of Two-Step Clustering Model

The second model tested was the two-step clustering model. It was decided to not to test this model for all spending categories combined (as was done first for the k-means model) due to the importance of the recency column for this model, and for all spending combined all accounts are active with spending within the previous 6 days.

The model was first tested again on the ‘Dining Out’ spending category. Customers were categorised into the five categories outlined in the Methodology. The RFM mean values along with the number of customers per cluster are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **R** | **F** | **M** | **Number of Customers** |
| **High Value** | 3.6 | 50.3 | 901.6 | 2269 |
| **Low Value** | 58.8 | 18.6 | 488.3 | 5940 |
| **Potential High Value** | 4.6 | 15.2 | 422.5 | 2926 |
| **Old High Value** | N/A | N/A | N/A | 0 |
| **Other** | 4.0 | 30.1 | 583.7 | 1690 |

*Fig. RFM mean values and number of customers for each cluster.*

A picture containing diagram, map, text

Description automatically generatedInterestingly, in this test there were found to be no ‘Old High Value’ customers. A potential reason for this could be the fact that the data is artificial.

*Fig. 3D graph showing clusters of customers using Two-Step clustering from RFM values.*

On the 3D graph above the ‘High Value’ customers are shown in green, the ‘Low Value’ customers are shown in red, the ‘Potential High Value’ customers are shown in blue and finally the ‘Other’ customers are shown in purple.

Similar to the method used in the k-means model, it was decided that the ‘High Value’ customer group could be iterated again to find the extremely high value customers. This involved finding customers from the ‘High Value’ group who had frequency and money values higher than the cluster average, and recency values lower than the cluster average. The RFM mean values of this new ‘Extremely High Value’ group along with a new ‘High Value’ group are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **R** | **F** | **M** | **N Customers** |
| **Ex. High Value** | 1.7 | 70.0 | 1280.9 | 192 |
| **High Value** | 3.7 | 48.5 | 866.6 | 2077 |

*Fig. RFM mean values and number of customers for each new cluster.*

A picture containing sketch

Description automatically generatedIterating this way on the ‘High Value’ customer group has created a small ‘Extremely High Value’ group made up of just 192 customers.

*Fig. 3D graph showing ‘Extremely High Value’ customers in orange.*

This small group of customers, making up only 1.50% of all customers, represent a group of very active spenders in the ‘Dining Out’ category who would be highly likely to be interested in offers regarding this category of spending.

Interestingly, after the second round of k-means testing on this category of spending a very similar number of customers were found to be in the most valuable cluster, with 195 in that group compared to 192 in this test.

## Analysis of Normalised RFM Model

# Further Work and Improvement

With access to demographic data on customers, as well as behavioral data, more detailed clustering models could be built that could allow for more precise customer marketing. Having access to more account details, such as monetary amount in the account, could also allow for more precise clustering algorithms to be built.

# Conclusions

A brief summary of the key insights in your report.

##### References

##### 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.