Jun 2021

ISSS603-Customer Analytics and Applications

Assignment 2

Bao Le - 01427104

Contents

1. Segmentation based on customers' level of Frequency and Monetary attributes

With the data cleaning process in Assignment 1, we have the cleaned dataset with 391,123 valid records. The Recency, Frequency and Monetary attributes are created using the method described in assignment 1.

The dataset is collected within 2 years, with most purchases occurred during the last 1 year (See Figure 1).

Figure 1: RFM segmentation

For such a short period, we assume that user purchasing behaviour (regarding the products bought) is less dependent on Recency than Frequency and Monetary, i.e., different customers who purchase often and of the same amount will purchase more similar products than different customers whose last purchase is on the same day.

Consequently, we will perform clustering based on Monetary and Frequency attributes, using the K-Means algorithm[1].

First, Monetary_Average and Frequency attributes are scaled to [0,1] interval.

Trying multiple values for the number of clusters give the elbow graph as follows:

[1] If one is to use K-Means clustering to determine which cluster a new customer belongs to, he/she would need to calculate the distance between the new customer's parameter to the cluster's centroid, which is outside the scope of this analysis. This analysis is only working on the existing customer database.

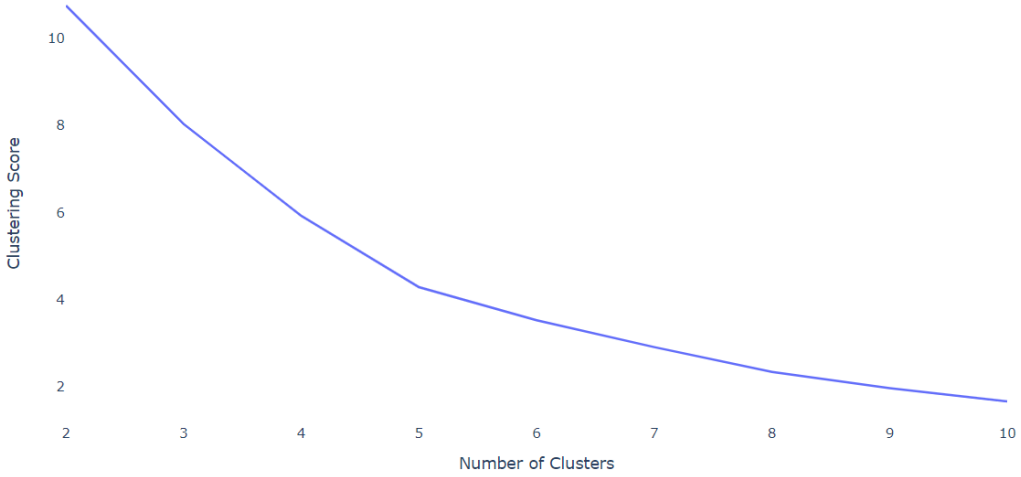

Figure 2: Clustering Score for different number of clusters

Selecting the number of cluster of 5 gives the following result:

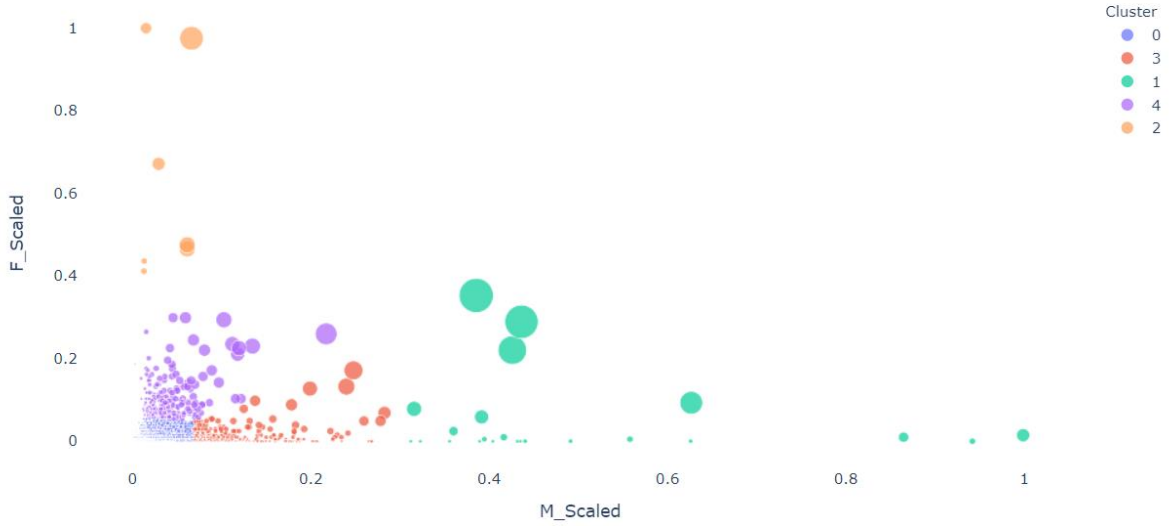

Figure 3: Clustering with n_cluster = 5, based on Scaled Monetary_Average and Scaled Frequency. The bubble size represents the Monetary_Total attribute

We can summarise the attributes of each segment as follows:

- Cluster 0: Customers who purchase infrequently and minimally,
- Cluster 3: Customers who purchase infrequently with a moderate invoice amount,
- Cluster 1: Customers who purchase infrequently with a high invoice amount,
- Cluster 4: Customers who purchase moderately frequently with a low to moderate invoice amount, and
- Cluster 2: Customers who purchase very frequently with a low invoice amount.

The number of customers and number of items purchases for each cluster is tabulated as follow[2]:

Table 1: Cluster characteristics

Cluster	Description (frequency, value)	Number of Customers	Number of Invoices
0	infrequent, minimal	3,737	10,592
1	infrequent, high value	23	210
2	very frequent, low value	7	837
3	infrequent, moderate	340	1,137
4	moderately frequent, low to moderate value)	226	4,370

We could observe that although cluster 2 only has 7 customers, their number of purchases is only one order of magnitude less than that of cluster 0 with 3,737 customers.

The segmentation result is attached in the appendix.

An interesting finding is that the percentage of non-UK customers varies greatly across clusters. The clusters with a high percentage of non-UK customers (clusters 1 and 3) are also the clusters with the higher average invoice value. These customers could be business customers buying in bulk for importing, in contrast to individual customers.

Table 2: Percentage of non-UK customers across segments

Cluster	Number of Customers	Number of non-UK customers	Percentage of non-UK customers
0	3,737	299	8%
1	23	12	52%
2	7	1	14%
3	340	91	27%
4	226	15	7%
Total	**4,333**	**418**	**10%**

When using the MBA technique, different minimum support levels are needed for cluster 1 and cluster 2, as their number of multi-item invoices is much lower than that of the rest.

2. Cross-selling strategy using MBA

Due to the difference in the number of multi-invoices and customers in each cluster, we ought to use different minimum support levels for different clusters.

For clusters 0, 3 and 4, we set the minimum support at 0.03. For clusters 1 and 2, it is easier for a rule to satisfy the minimum support threshold due to the low number of customers

[2] Number of invoices excludes single item invoices, as they do not contribute to the market-basket analysis.

and transactions. As such, we set the minimum support threshold for clusters 1 and 2 at 0.06.

Table 3: Cluster characteristics with Min Support

Cluster	Description (frequency, value)	Number of Customers	Number of Invoices	Min Support
0	infrequent, minimal	3,737	10,592	3%
1	infrequent, high value	23	210	6%
2	very frequent, low value	7	837	6%
3	infrequent, moderate	340	1,137	3%
4	moderately frequent, low to moderate value	226	4,370	3%

Using the purchasing transactions of customers from each cluster, we perform the data engineering steps as follows:

- Aggregate the transactions into invoices (i.e. baskets),
- Eliminate single-item invoices,
- Run the FP growth algorithm[3] to generate a set of rules for each cluster, at min support specified in Table 3 and min lift = 1, and
- Select the top 100 rules based on lift of each cluster for further analysis and visualisation.

The result rule set is included in the appendix. In this report, we will only discuss the highlights from the result.

Table 4: MBA result with original min support

Cluster	Description (frequency, value)	Number of Customers	Number of Invoices	Min Support	Number of Rules
0	infrequent, minimal	3,737	10,592	3%	-
1	infrequent, high value	23	210	6%	914
2	very frequent, low value	7	837	6%	-
3	infrequent, moderate	340	1,137	3%	362
4	moderately frequent, low to moderate value	226	4,370	3%	38

If we set a lower min support threshold, the number of rules is increased as follows:

Table 5: MBA result with relaxed min support

Cluster	Description (frequency, value)	Number of Customers	Number of Invoices	Min Support	Number of Rules
0	infrequent, minimal	3,737	10,592	2%	52
1	infrequent, high value	23	210	4%	27,184
2	very frequent, low value	7	837	4%	2
3	infrequent, moderate	340	1,137	2%	1,246
4	moderately frequent, low to moderate value	226	4,370	2%	224

[3] https://github.com/chonyy/fpgrowth_py

3. Comparing results among segments

Before interpreting the results, we shall keep in mind that the rules in the result set come in pairs, i.e., if rule A=>B is selected, so is rule B=>A, due to the symmetricity of support and lift. We will be using the result from table 5 with relaxed min support.

We notice that the numbers of rules are significantly different across segments.

Segment 2, due to the low number of customers, one pair of rules is detected. Further checking shows that the pair of rules is present in the result for cluster 4, with a similar lift. For all intense and purposes, we can consider segment 2 part of segment 4.

Segment 0 consists of infrequent customers with low purchase value. It makes sense that their purchases are not related, and the rules are few and insignificant (the rules disappear at support threshold = 0.03).

Segment 1 only consists of 23 customers but sees a large number of rules due to their high average purchase value. This is the segment that MBA could bring the most profit.

For segments 1, 3 and 4, due to the large number of rules, a graphical representation is produced as follows. The thickness of the edges represents to lift of the rules.

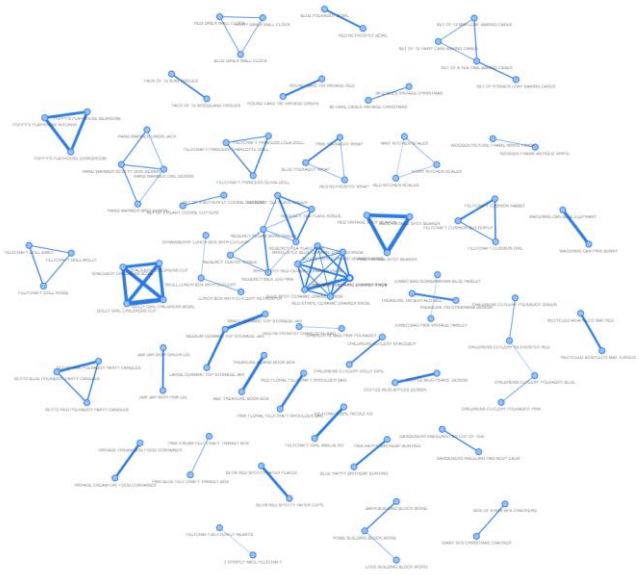

Figure 4: Ruleset of segment 3

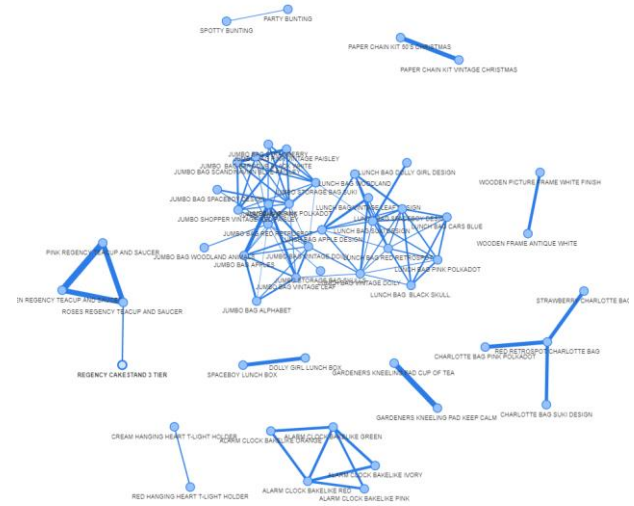

Figure 5: Ruleset of segment 4

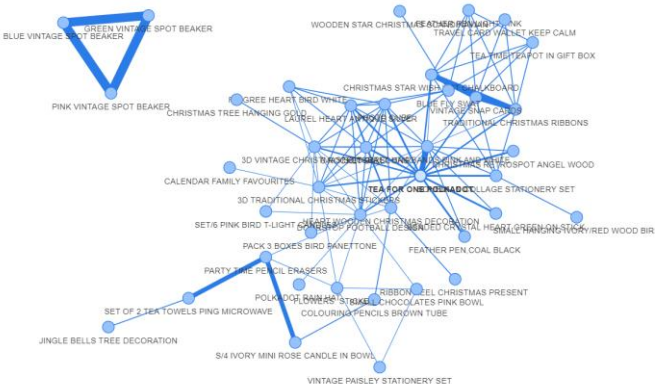

Figure 6: Ruleset of segment 1

The graphs represent the relationship of each segment's top 100 rules (ordered by lift value). We observe that different segments have different graph structures of the ruleset. Segment 3 has many disconnected small spanning trees. Segment 4 has fewer spanning trees, and the rules are more connected. Segment 1 has only 2 disconnected spanning trees, one of which is highly complex.

In simpler terms, customers in segment 3 purchases products from multiple unrelated product types, while customers in segment 1 purchase most products from one or two groups of similar products. Customers in segment 2 is in between the 2 extremes.

4. Recommendation for each segment

At a basic level, the online store can implement one recommendation system that operates the same way across segments: If a customer from a particular segment purchases a product from the segment's ruleset, the recommender will recommend the top 5 related products based on the lift value.

In this report, we will discuss further the recommendation for segment 1, which is the most valuable set of customers.

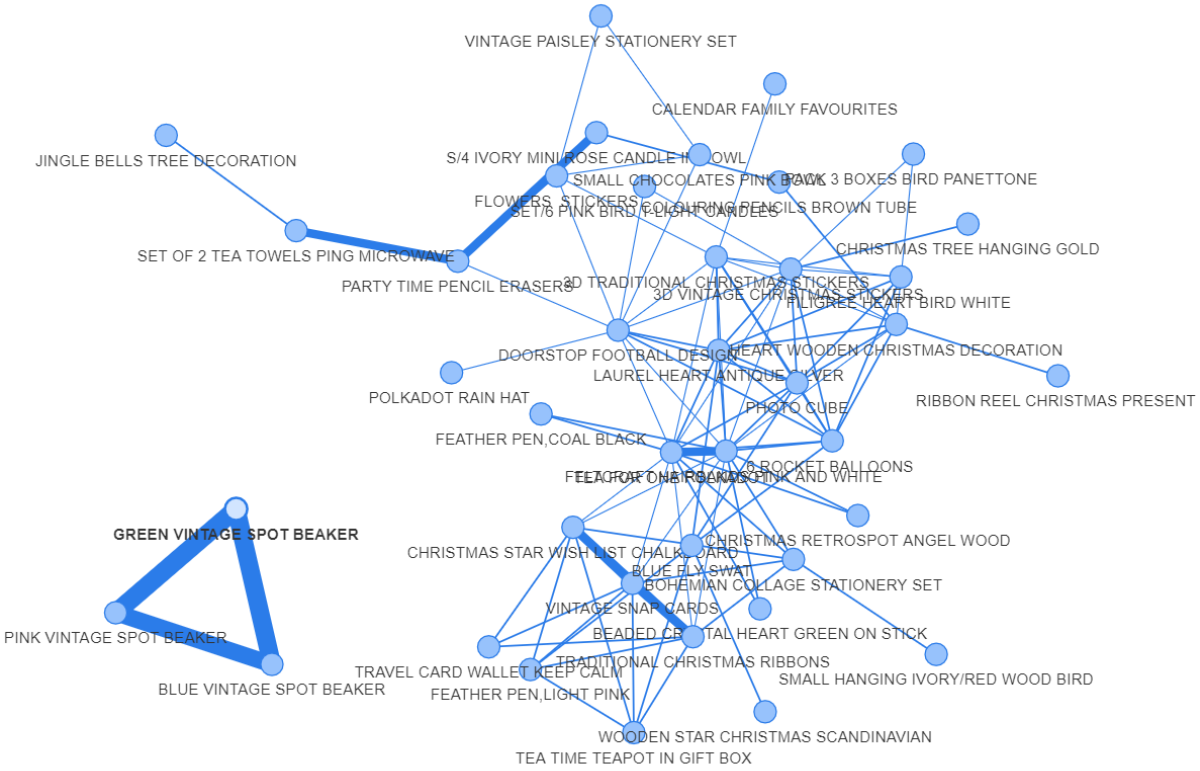

Figure 7: Zoom in of segment 1's rule set

Looking at the larger spanning tree, we observe that most items are Christmas-related products: Christmas decorations or presents. It means that if a new Christmas product is introduced to the store, the store could recommend this product to the customers in this segment, despite it not being included in the ruleset. Similarly, the store could seek other types of beakers to recommend to segment 1's customers.

Using the same logic for segment 4, one could identify that the largest spanning tree of the segment is lunch bag-related products, followed by alarm clocks (Figure 8). However, due to lower purchase value compared to segment 1, the store needs to consider the cost of sourcing new products.

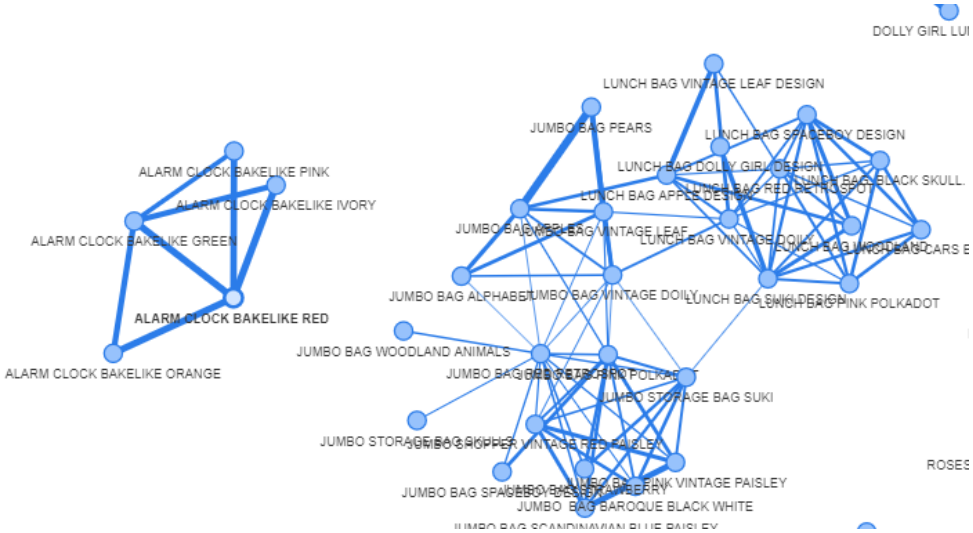

Figure 8: Zoom in of segment 3's ruleset

We do not have any further recommendations for segment 3 due to the fragmented ruleset bringing complexity to any solution and segment 0 due to low customer value.

Appendix

1. Python notebook: Assignment_2_LE_DUC_BAO.ipynb
2. Clustering result: Clustering_result.csv
3. Ruleset result and graphical visualisation using min_support = 3%/6%: MBA result - iteration 1
4. Ruleset result and graphical visualisation using min_support = 2%/4%: MBA result - iteration 2
5. Clustering description: Tables.xlsx