NOTTINGHAM UNIVERSITY BUSINESS SCHOOL

Academic Year: 2020 – 2021

ANALYTICS SPECIALISATIONS & APPLICATIONS

1st Coursework

CUSTOMER SEGMENTATION USING RETAIL TRANSACTIONAL DATA

Student ID: 20243144

Apr 2021

CONTENT

A EXECUTIVE SUMMARY

The key business objective, as stated in Chief Data Officer message, is to define different customer segments that best informs marketing activities.

Given the resource constraints and the common understanding that different clustering techniques:

- require different methodologies of features transformation to deliver optimal inputs for modelling,
- overall are not comparable via one common parameter, as each of them uses a distinctive strategy in identifying segments,

The technical approach was to predefine an appropriate segmentation algorithm while focusing on **exploratory** (data summary and feature selection), **feature transformation & extraction**, **model's local optimisation** and especially **cluster interpretation** to deliver actionable insights. Due to its intuitiveness and interpretability, k-Means was selected as the clustering model for this project.

Throughout the process, customer inputs for further transformation and modelling were selected including their **overall loyalty** (spending, visits, recency), **variety (**number of products, categories), **products purchased** (types, value), the **features of their shopping trips** (basket size, purchase time) and **change in their engagement** (spending) over the period.

The model output provided 5 distinctive segments, namely **Super Customers**, **Bustling Homemakers**, **One-stop Shoppers**, **Economical Neighbours** and **Potential Lapsers**, with the first two suggested as the most attractive target, contributing over 60% of revenue while only accounting for 45% of the population. The differences between segments also helped to derive recommendation on customised strategy for product and communication.

B DATA EXPLORATARY

Data summary

The provided material included four data tables. One contains original transaction records ('lineitem_sample.csv'), including customer number, transaction time, product code, category, quantity, spending, and was used to compute three other 3 tables.
A data cleaning step, therefore, was conducted for 'lineitem_sample.csv', in which 865 records with negative value for 'quantity' and 'spend' were found, considered error and removed (more details in Data Cleaning – Support Document).

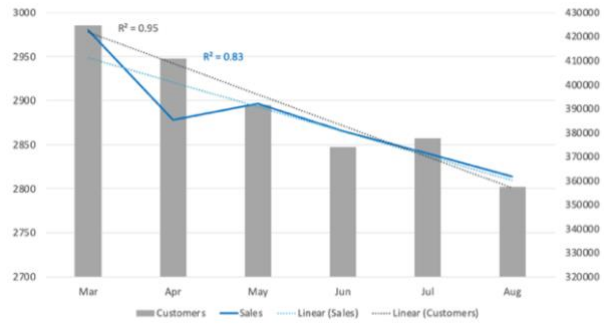


Figure 1. Monthly tracking for total sales and number of unique customers

Exploring the data, it can be seen that the store had reached maturity stage when most of current customers (2985/3000) shopped in the first recorded month.

Over the period, the store witnessed declines in both Total sales and Customer base (SQL query 6 – Support Document), putting minor objective for the clustering task in defining the potential and vulnerable segments.

Features' generation

With the objective to provide business a detailed description for each cluster, the strategy was to create **all relevant behaviour variables**, investigating their relationship, selecting the best for modelling while reserving others in clusters' analysis.
The groups of features are briefly described in Table 2 below. The generation process included a step of retrieving data from SQL Server (SQL Query 7,8 – generating '%cate_clean_uncross.csv' & 'all_except_cate_final.csv'), followed by data merging using Python Pandas.

SHOPPING BEHAVIOUR (RFM)	**Customer loyalty**	Spending (total, weekly), Visit (total, weekly), Days from last purchase
	Variety	Product quantity, Number of categories, unique products, Products per category
	Shopping trip features	Basket size (value, number of unique products per basket), Visit time (percentage of weekend visits/total)
	Change in loyalty	Change in spending (computed by divide individual spend in last 3 months for individual spend in first half)
SHOPPING MISSION	**Product purchased**	Cost per item, Proportion of individual spending allocated for 20 categories

Table 2. Generated features and their indication

Customer base and features' selection

Based on summary statistics of selected features , the average shopper in total spent £771 and £30 weekly, with nearly 3 visits a week (43% happened during weekend). Their norm baskets valued at £15 and consisted of 5 different products, with a unit cost of £1.4. There was a slight decrease in aggerated spend change (0.95) that in line with finding from revenue's monthly tracking.
The most-selling categories were Tobacco (12%), Dairy (9.2%), Fruit & Vegetables (9%).

An investigation on the correlation of created features showed that there was no significant relationship between **behaviour (RFM)** and **shopping mission** attributes.

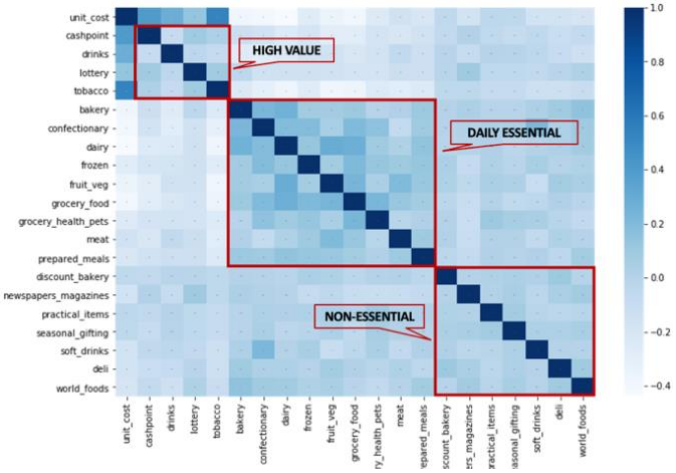


Figure 3. Shopping mission features correlation map

Among RFM features, beside the apparent relationship between spending, visits, product numbers and days from last purchase, it is interesting to observe a slight negative impact between total visits and basket size (value: - 0.37, product: -0.5), hinting that there could be certain clusters visiting less frequent with larger shopping cart.

Among the shopping mission features, 'high value' products, especially 'tobacco', showed medium negative correlation with essential/ cooking-aid products, indicating that there were customer segments having different shopping missions and spent mostly on either one of these two categories.

This suggested an approach for dimensional reduction by grouping similar categories into bigger mega-categories that better explain the shopping missions, including **'High Value'**, **'Daily Essential'** and **'Non-Essential'** (Figure 3).
The highly correlated features (correlation rate >=0.7) were then shortened before applying Feature Extraction, as similar-direction vectors can cause the algorithm to overemphasize their contribution. The 11 remained features and their manifestation are included in Table 15 – Support document.

C FEATURE EXTRACTION

Features' transformation

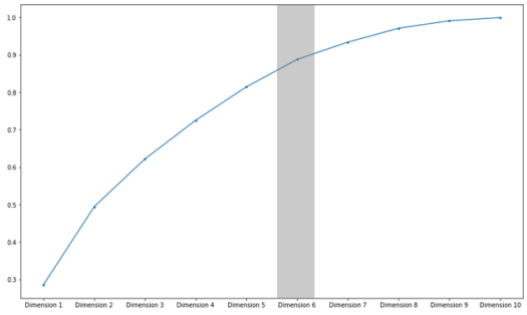
Investigating these features' histogram (Figure 16 – Support document), it can be seen that none of them followed normal distribution, and most had left-skewed pattern.
The data, therefore, was first altered using **Power Transformation** (to maximize the effect when applying to full value range), followed by Standardization algorithm to balance the unit between variables. **"Yeo-Johnson"** transformation was applied due to its robustness in unskewing the items in comparison with Logarithm.

Features' extraction

As remained features still showed moderate correlation, a Principal Component Analysis was applied to further reduce dimensions and maximised clusters' separation.

Using the 'rule of thumb', 3 break points were identified (the points after which witnessed noticeable decrease in variance accumulation), including 2, 6, 8 dimensions. Among them, the option of 6 components was the best one, as it reserved 89% of total features' variance.

Figure 5. Variance accumulation by number of Principal Components



The aggerated components and their variance contribution are illustrated in Figure 6.

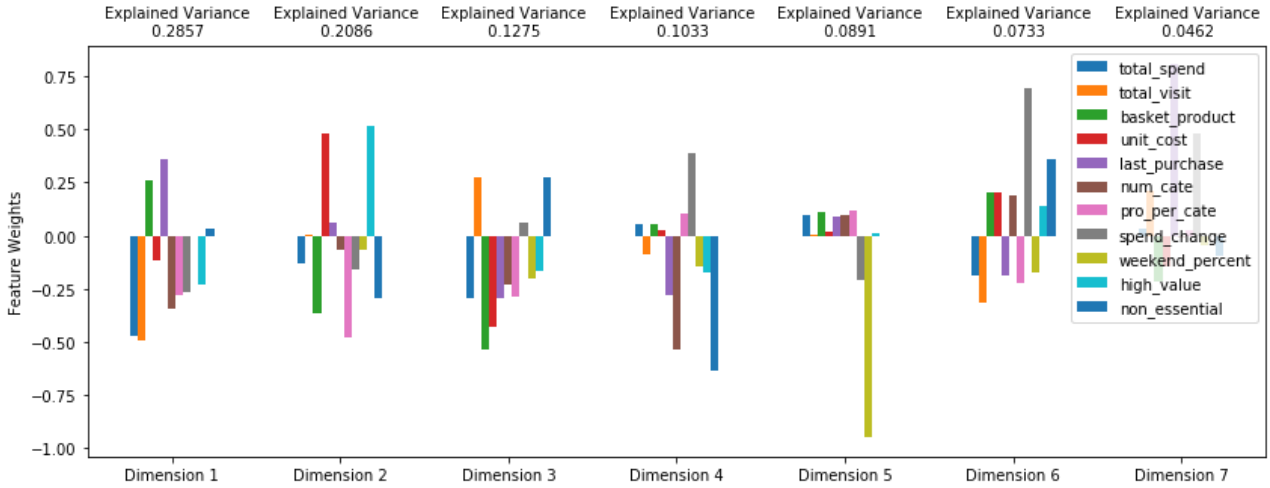


Figure 6. Principal Components' variance contribution and features association

- **The first component** having high negative association with the total spending, number of visits, and positively related to days from last purchase, represented the Loyalty of customers towards the store.
- Mostly derived from unit cost and high value categories, **the second feature** indicated the value of purchased products.
- **The third variable** described time utilisation behaviour, creating mainly by the basket value input.
- Positively related to the increase in spending, but negatively measured by Non-essential products, products number, **the fourth component** explained engagement level of shopping trips.
- Similarly**, the fifth component** also described visit feature but focusing on whether the trip happened during weekend or not.
- **The last component** represented changes in customer spending/ engagement over the period.

D MODEL SELECTION

Between different techniques

As different clustering algorithms require different preparation techniques, and their performances, especially interpretabilities are difficult to compared**, k-Means** was selected due to its intuitiveness and speed in implementation. Regarding other common models:

- **Connectivity and Distribution based clustering** are overall computationally expensive.
- **Density-based clustering** requires tuning two hyper-parameters, hence is more complicated in clusters selection process. Also, its advantage in reducing noise cannot be applied due to the necessity to fit each customer to one cluster.

All in all, their superiority over k-Means in not assuming globular clusters became less appealing in this case as the features had been pre-processed to achieve overall 'spherical' shapes.

Between different clusters' number

To pick up the most efficient result generated by k-Means, the extracted components were fed into a model iteration with different number of segments (from 3 to 10) and compared using below indexes:

- **Within-Cluster-Sum of Squared Errors** (WSS or Elbow Method) calculates the sum of square distance from a point to it assigned centre, showing how dense the clusters are. The smaller the index the better.
- **Silhouette score**, taking into account the distance between points in same cluster in comparison with instances from other clusters, informs on the separation between clusters. High Silhouette Score is desirable for better clustering.

Figure 7. WSS/ Elbow method by number of clusters

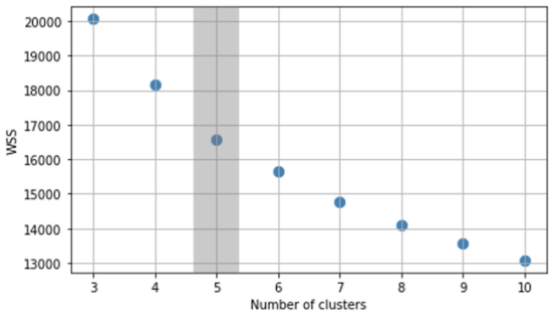
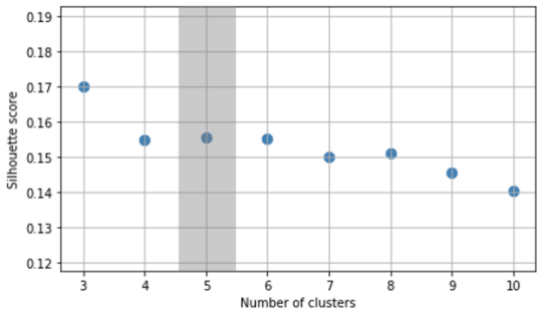


Figure 8. Silhouette score by number of clusters



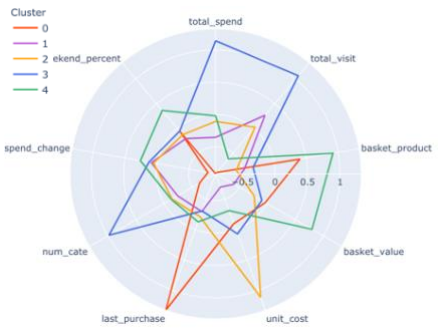
Results from Figure 7 and 8 showed that:
With the WSS method, 5-cluster option was the **'plateau point'** for the elbow, meaning that after this point the decrease in the distance between instances in one cluster to their centroid was slowed down.
Similarly, the Silhouette score suggested that the 5-cluster model had slightly higher performance in terms of both separation between clusters and sample size distribution (Figure 17 – Support document), thus, being the most optimal option.

E RESULTS & RECOMMENDATION

Comparative analysis

The segments were formed by applying k-Means (k=5) to the reduced data by PCA (n_components = 6) as explained above.
To provide an intuitive comparison, the **average values** of key features were **standardised** and grouped to respective clusters. Polar charts were generated illustrating the shopping behaviour (Figure 9) and categories spending allocation (Figure 11). The detailed and unstandardised average stats of each Cluster was presented in Table 18 – Support document.

Figure 9. Shopping behaviour features by cluster (standardised value)



In terms of **spending**, **visits** and **variety,** Cluster 3 (hereinafter referred as the **Super Customers**) possessed the highest indicators (spend = £1,404, visits = 116 times, categories = 17), while in contrast, we had Cluster 0 (**Potential Lapsers**) with the lowest performance (spend = £263, visits = 23 times, categories = 14).
For **days from last purchase**, **Potential Lapsers** showed an extremely high index (35 days). **Bustling Homemakers**, though far from that, also had a slightly higher indicator versus other segments (5 days). Despite that fact, they had overall higher **spending increase** and proportion of **weekend visits** (50%).
On **basket size**, **Bustling Homemakers** was the winner (value = £24, products = 9), while Cluster 1 (**Economical Neighbours**) owned the smallest basket (value = £8, products = 3). Interestingly, for Cluster 2 (**One-stop shoppers**), their medium spending per visit (value = £12) only accounted for on average 2 unique products. They also had the highest **unit cost** (£2) in contrast with the **Economical Neighbours** (£1).

Figure 10. Mega-Categories spending (%) by cluster (unstandardised value)

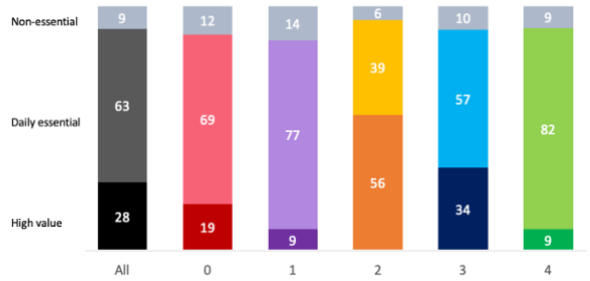
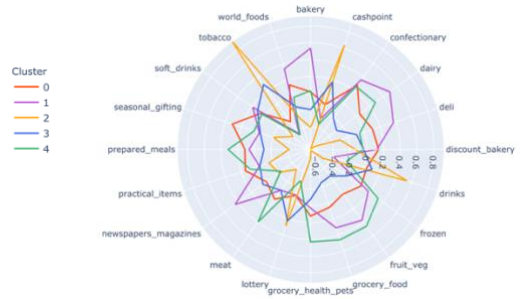


Figure 11. Categories spending (%) by cluster (standardised value)



Regarding shopping mission, the baskets of **Bustling Homemakers** and **Economical Neighbours** were mostly filled with **Daily Essential products** (82% and 77% respectively). While **Bustling Homemakers** skewed towards **cooking-aid ingredients**, **Economical Neighbours** favoured **Ready to Eat** foods. In contrast, **One-stop shoppers** mainly spent for **High value** categories (56%).
The other two segments had their spending allocation more in line with the categories' contribution in total sales, though **Potential Lapsers** showed special interest towards **seasonal gifting**.
For an easier view, each segment' pen portrait is included in a full-page section (**Meet the segments -** page 8) and should be reviewed before below Recommendation section.

Business recommendation

The two most attractive segments suggested for business are **Super Customers** and **Bustling Homemakers**, due to:

- **Their large contribution** to business when accounting for over 60% of total sales while only taking up 45% of customer base, which helps the company to **centralise the resources.**
- Their **stable performance and consistent loyalty** over the period (SQL Query 10), with **Bustling Homemakers** showed even a slight increase in individual spending in the last 3 months versus the first half (cannot observe in trend line as 'spend_change' was aggerated and did not count in individual spending weight).

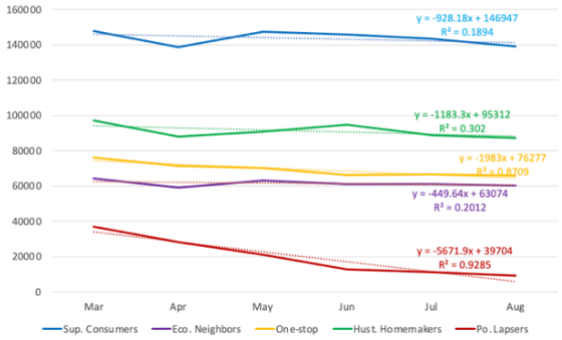


Figure 12. Monthly tracking for total sales by cluster

For **Super Customers**, who purchased a wide range of products, the strategy should be providing them the variety they looked for. Bundling or sampling could be a great way to not only benefit the business but also satisfy customers.
For **Bustling Homemakers**, it is important to meet their demand for convenient shopping. With their portfolio highly skewed towards essential supplies, the offer for online shopping and fresh foods delivered could be a good choice.

Beside the attention towards most attractive customers, it is suggested that the company consider a more comprehensive approach to reduce the decreasing trend in sales overtime.

As each of the segments showed different product demand, a simple **Recommender System** based on clusters can be deployed to suggest suitable items for customers via current communication channels.

The information on **Potential Lapsers** can be useful to identify vulnerable customers, especially days from last purchase, but at the same time can provide signal for business to tackle the 'critical point' in losing engagement**. Communication once a week**, for example, could help remind the customers to come back and encouraging their shopping habit with the store.

Meet the segments

THE SUPER CUSTOMERS

CLUSTER SIZE: **20%**

VALUE SIZE: **37%**

Being the most valuable segment to business, the **Super Customers** contributed nearly 40% of total sales, despite only accounting for one fifth of the population.

They owned high frequency of visit, while actively seeking for wide ranges of products and categories.

With dominant influence on revenue, their basket arrangement highly resembles the company product portfolio.

THE ECONOMICAL NEIGHBOURS

CLUSTER SIZE: **21%**

VALUE SIZE: **16%**

Despite having modest individual spending, the **Economical Neighbours** can still be considered as a highly engaged customer group since their frequency and recency indexes were very positive.

Unlike the Hustling Homemakers, they did not mind on purchasing in small basket and visiting the store frequently.

The unit cost derived from this group was the lowest, reflecting their interest in ready to eat foods and low value items.

THE ONE-STOP SHOPPERS

CLUSTER SIZE: **20%**

VALUE SIZE: **18%**

The **One-stop shoppers** paid their attention to a limited range of products and seemingly possessed a purposive shopping attitude.

While this led to smaller size of baskets regarding quantity, with their superior cost spent per item, they still possessed a shopping cart of considerable value, which is understandable as the categories they seek for were mostly high in value (‘tobacco’, ‘drink’, ‘cashpoint’).

THE BUSTLING HOMEMAKERS

CLUSTER SIZE: **24%**

VALUE SIZE: **24%**

The **Hustling Homemakers** appeared to be convenience seekers: they purchased in large basket in compensation for lower visit frequency versus average.

Their time utilisation also showed in the fact that half of their visits was during weekends.

Furthermore, the average spending change was the highest for this group, indicating their growing loyalty towards the store.

Their key category was Daily Essential / Cooking-aid products (‘frozen’, ‘fruit_veg’, groceries).

THE POTENTIAL LAPSERS

CLUSTER SIZE: **15%**

VALUE SIZE: **5%**

Accounting for 15% of population, but only contributing to 5% of revenue, the **Potential Lapsers** no doubt was the least profitable group.

This conclusion can also be drawn out from their inferior parameters in total visits, number of products and change in spending through time.

They can be detected by long absent period of around 1 month, and a relatively higher spend proportion for non-essential items such as practical items and seasonal gifting.