The University of
Nottingham

BUSI 4370

ASA COURSEWORK

(Customer Segmentation)

2020-21

JIGNESH MANOCHA

20249953

Section 1

Executive Summary

This report provides an analysis of customer segmentation overtime using customer spending data on various categories of products provided by the company. According to the business case, the aim of this report is to best describe the potential customer segments into which company can examine their spending behaviour and draft policies accordingly.

The methodology of analysis includes feature engineering, data exploration using graphs and statistical summary for better data insights, natural logarithm technique for data normalization. Other techniques include Principal Component Analysis (PCA) for dimensionality reduction using explained variance ratio, K-means algorithm for clustering and average silhouette score metric to measure the cluster's performance.

The data used for this clustering consist of 3000 customers with their transactional data which includes their spending on 20 different product categories, frequency of their visits, total spending, and recency of their transaction.

As a result of this analysis, the customers are segmented into 5 clusters namely High Value, Loyal, Potential, Low value customers and Lookers. Results of cluster analysis reveals that customer segment 2 and 3 are the best potential and profitable clusters which include around 44% of the total customers and majority of them are loyal customers.

The report finds the prospect of the company to focus more on loyal, potential, and high value customers, since acquiring new customers is too expensive as it cost time, money and efforts, than retaining existing ones. And company should plan some promotional campaigns for cluster 1 and 4 as they are high risk churn customers.

Section 2

Feature Description

Initially all features were selected to understand the patterns among them and realize which feature is necessary to cluster the customer based on their spending habits. Also, the three features namely recency, total-spend and frequency were generated as per Chief Data Officer's guidance for better customer behaviour comparison. In order to find the relations between the features, scatter and correlation metrics were used. Based on that it has been found that fruit-veg has a very strong correlation with grocery foods and dairy products, fruit-veg is also strongly correlated with dairy and confectionary, comparatively grocery health pets noted weak correlation with confectionary and grocery food, meat is correlated with grocery food and fruit-veg and so on. To find the more effective results, the features which are strongly correlated should be eliminated as they have tendency to skew the results.

Note: Category columns such as bakery, cashpoint and customer number were not taken into consideration as the customer spending on bakery was nil, for cashpoint it was assumed to be the cash withdrawal which may skew the analysis (lack of information provided by the company) and customer number was not relevant to customer segmentation based on spending, that is why it has been dropped.

- **Data Normalization – Natural Logarithm**

During initial stage of data exploration, it has been found that the data is not normally distributed and most of the data points are left skewed which might be due to presence of outliers. To fix the problem of skewness, data would be normalized, and features were scaled using the natural logarithm scaling technique as it is most common and widely used technique for data normalization. After logging and performing data preprocessing, it has been observed that the data has been normalized and correlation between variables has become quite weaker.

- **Dimensionality Reduction – Principal Component Analysis**

Later, to identify the significant dimensions for further analysis, Principal Component Analysis (PCA) technique of dimensionality reduction has been implemented in order to calculate the dimension with best optimal cumulative variance using explained variance ratio. Basically, PCA is used to reduce data dimensionality by transforming large set of variables into smaller ones. The principal components are just the eigenvectors with eigenvalues which shows the importance of variables into the data, which is computed in descending order, hence first principal component states the most significant component of data and therefore holds out the most essential information of data required for segmentation. It has been identified from the figure that around 53.61% of the variance is illustrated by first three principal components whereas first five principal components explain 66.26% of the total variance in the data.

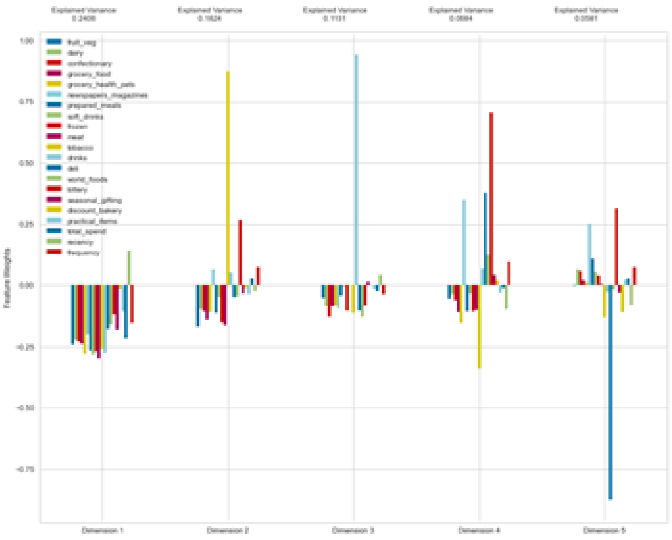

Figure 2.1

- **PC1:** Non- essential items - The rise in principal component 1 is due to the increase in the customer spending on meat, soft-drinks, fruit-veg, drinks and grocery health pets irrespective of negative weights on the graph’s x axis. These features best represent this component and are fully aligned with our initial findings where they are correlated. This component can be called as **“FRESH, DRINKS AND NON-PROCESSED FOODS”**

- **PC2:** Here decrease in PC2 is associated with increase in spending on tobacco, meat, lottery and fruit-veg items. This component can be best represented by these features and called as **"TOBACCO PRODUCTS AND FOOD"**.
- **PC3**: Here fall in PC3 is linked with rise in spending on drinks, confectionary, tobacco, and world foods. These features best represent the PC3. This component can be named as **"CONFECTIONARY-DRINKS AND WORLD FOODS"**
- **PC4:** Here customer spending on lottery, tobacco, drinks, deli, and grocery health pets' results in increase in the PC4 which means these features best represent this component. This component can be called as **"LOTTERY AND NON-ESSENTIAL PRODUCTS"**

It is observed from the principal components that customer spending on tobacco, drinks, meat, and grocery items are the significant features as they are most associated with the data and contributing the most variance and hence therefore should be consider for further analysis for customer segmentation.

Based on the explained variance ratio, 4 dimensions explains the significant amount of variance i.e., 60.45% and therefore selected for further customer segmentation.

Section 3

Customer Base Summary

The patterns were analysed with the help of data exploration using statistical summary of the data features received and generated. It can be noted from summary table below that on an average customer spends maximum amount on tobacco, cashpoint, fruit and veg, grocery-food and drinks. Static increment in the customer spending overtime can be noted from all three percentiles.

Based on RFM features, it can be observed that the average spending of the customers is around 773£, and 9 mean number of days has been passed since last purchase with the frequency of footfall of 65 times. On the other hand, the maximum total spending, recency, and frequency recorded to be the 6589£, 165 days and 374 times respectively which reveals that there are high value purchasing customers exists in the store.

Note: Data has been checked, where no null values has been found but some negative values has been detected in lottery and spend columns of lineitems and category spends sample dataset respectively, which was assumed and converted to zero and all the features with object dtype has been converted into float (original) dtype.

Section 4

Segmentation Methodology

After preprocessing and scaling the data into more normalized form, analysis can be carried forward to segment the customers, based on their similarity. Centroid based clustering algorithm -K-means has been used to segment the customers as it is simple to understand, implement and is widely used technique**.** According to the business case, it can be noted that to determine the number of clusters between 5 and 7 depends on the silhouette score which measures the similarity of the point to its own cluster and the distance between the points using Euclidean distance metric. Silhouette score reveals the quality of clusters that how well the segmentation is and determines the average of silhouette coefficient for each data samples as shown in the figure (4.1).

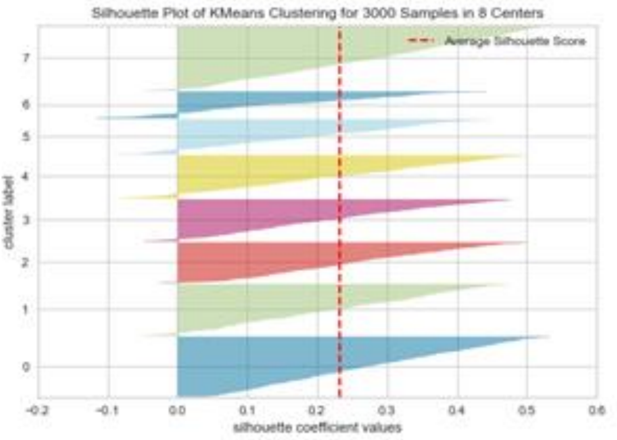

Figure 4.1

According to the business case, Silhouette score has been used to measure the performance of clusters because this score is successfully proven for segmentation and has worldwide applicability thereupon. However, only on the basis of silhouette score metrics it cannot be concluded that 5 clusters are the best as other parameters like interpretability cannot be overlooked.

The average silhouette score is highest when number of clusters to be 5 (chosen best score between 5 and 7) i.e., 23.4, which identify that when there are 5 customer segments, the similarity between their purchase's patterns and categories, and in each individual cluster is quite high. However, the k-means algorithm is very stochastic approach which always shows different clusters, but the pattern always remains same.

Section 5

Results

Below clusters has been examined on the basis of true centre values which are the mean of all data points predicted in the respective clusters.

- **High Value (Cluster 0 -Blue)** – These are the high value spenders in compared to their visits, who spends on an average of 667£ and are the regular visitors to the store (48 times in span of 6 months). There are 609 customers in total, fall under this segment. This cluster's centre reveals that these customers mostly spend on basic food items like fruit-veg, dairy, confectionary, groceries and on non-essentials like newspaper-magazines, deli, world foods and prepared meals.

Recommendation: Since they are high amount spenders but does not come to the stores more frequently, therefore the company should give them reason to visit the store with organizing some promotional campaigns, anticipating their needs, reward them

- **At Risk Customers or Lookers (Cluster 1-Green)** – These customers fall under the low purchase category as they often come to the store but does not purchase high value products or come for 'just looking'. They are second most tobacco and lottery ticket buyers. Around 17% of total customers comes under this cluster. They are also considered as customers at risk as they can churn if careful marketing tactics will not be applied.

  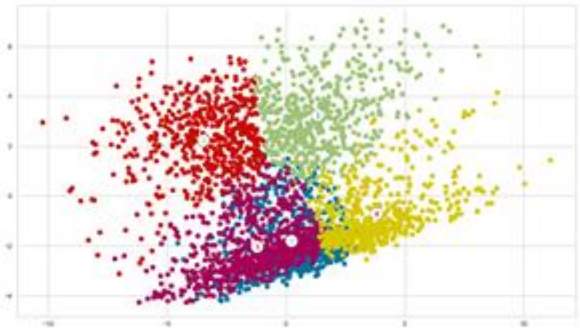

  Figure 5.1

  Recommendation: The company should make strategies to re-engage these customers by providing more details, instructions and market tobacco and lotteries to target these customers with the help of promotional mix tools. Strategical application can turn them into potential customers.

- **Loyal Customers (Cluster 2 -Red)** – Based on total spend, recency and frequency, these are the most revenue generating customers to the company as their purchase value, frequency of visit is highest i.e., 1395£ and 103 times respectively, and only 3 days passed since their last purchase which is the lowest among all clusters. 560 customers out of 3000 lies in this segment. They spend majority on tobacco, dairy products, confectionaries, grocery food and health pets, fruit-veg, lottery, and newspapers-magazines.

  Recommendation: Since they are the major spenders, therefore company should focus on this department with offering great deals, introducing more varieties and along with that company should give VIP treatment, special amenities, pricing, and recognition to them, as they contribute around 80% to the company's revenue.

- **Potential Customers (Cluster 3 -Purple)** – Customers under this segment are considered as the high value frequent buyers as their mean spend is 853£ and the frequency of footfalls are also very high i.e., 52 times. It seems like these customers have healthy diet as, they spend mostly on fruit-veg, grocery items, grocery health pets and apart from that they also likely to spend on meat, prepared meals, frozen, drinks, and seasonal gifting. Quarter of the customers i.e., 25.3% of total, fall under this cluster.

  Recommendation: The company should give them special treatment with great discounts, schemes and offers and conduct some events to build rapport with them. Vigilant implementation can turn them into loyal customers.

- **Low Value Customers (Cluster 4 -Yellow)** – 531 customers fall under this category and are the new customers, that are seen very rarely in the store and when they come, either they buy very cheap products or the products in very less quantity. They mostly buy cheap drinks and lottery items.

Recommendation: Company should try to build relationship with them using surveys, email, feedback, welcome them on-board and give special offers to them, which gradually help to turn them into mid value customers.

Section 6

Summary

Considering the business case in mind, the customers has been segmented with the help of Segmentation Analysis and following the business goal to segment the customers between 5-7 clusters, customers has been put into 5 clusters and their pen profile has been made according to their sharing spending traits.

After careful analysis of Pen Profile of customer segments, it has been concluded that Loyal (Cluster 2) and Potential Customers (Cluster 3) are the most profitable segments for the company, as the customers in these clusters spends high on basics and non-standard groceries irrespective of product value, the customers also come very frequently to the stores with high recency value, which reveals that they are loyal to the company and are fond of company's products. The company should target these customers to retain their trust and loyalty towards company, with lots of attached great discounts, schemes and offers, company should try to conduct promotional campaigns to engage these customers more, to attract new customers.

Company should also organize marketing activities and perform churn analysis in order to retain these customers for longer run.

Recommendation – Further Analysis

More information on customer's demographic data such as age, income, gender etc. could help to perform more robust analysis like for ex, segmentation according to level of income, the age groups for better targeting, and so on.

Note: After analysing and segmenting the data into 5 clusters, the clusters were exported to csv file "output.csv" where clusters and their cluster name columns defines the customer.

References

Google Developers. (2015). *Clustering Algorithms | Clustering in Machine Learning*. [online] Available at: https://developers.google.com/machine-learning/clustering/clustering-algorithms.

Jaadi, Z. (2019). *A Step by Step Explanation of Principal Component Analysis*. [online] Built In. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.