ONLINE RETAIL CUSTOMER SEGMENTATION

Sayali Kamalapurkar

Data Science Trainee,

AlmaBetter, Bengaluru

ABSTRACT: Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately.

Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

We cleaned the data given to us and then built RFM model to segment the customers based on Recency, Frequency and Monetary. Identified number of clusters using Elbow method and Silhouette score method. Used K-Means clustering on RFM model.

Lastly, did Cohort Analysis which means formed cohorts of customers based on Retention, Acquisition, average purchase amount over the time.

PROBLEM STATEMENT:

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

INTRODUCTION:

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioural patterns play a crucial role in determining the company direction towards addressing the various segments.

Some of the features of the online retail dataset are as follows:

- **InvoiceNo:** It is a 6-digit integer number uniquely assigned to each transaction. If this number starts with ‘C’ it indicates a cancellation.
- **StockCode:** Product code. It is a 5-digit integer number uniquely assigned to each distinct product.
- **Description:** Product name.
- **Quantity:** The quantities of each product per transaction.
- **InvoiceDate :** Invoice date and time.The day and time when each transaction was generated.
- **UnitPrice:** Product price per unit.
- **CustomerID:** Customer identification number. It is a 5-digit integer number uniquely assigned to each customer.

- **Country:** Country name. The name of the country where each customer resides.

STEPS INVOLVED:

1. **Importing libraries:**
   Imported Python libraries such as NumPy, Pandas for data manipulation and Matplotlib, Seaborn for data visualization. Sci-kit Learn and clustering algorithms such as K-Means from Sci-kit learn, also metrics such as silhouette score are imported.

2. **Data Collection:**
   To proceed with the problem dealing first we will load our dataset that is given to us in .xlsx file into a dataframe. Mount the drive and load the excel file into a dataframe.

3. **Explore dataset:**
   Found out different attributes of the dataset, like – shape, number of null and duplicate values in the dataset, statistical information about the numerical columns and categorical columns. Also, found out number of unique values in each column. Dropped cancelled transactions.

4. **Outlier detection and treatment:**
   We plotted box plots for numerical features like – Quantity and UnitPrice. There we saw many outliers, hence we calculated IQR and upper and lower limits. Next, dropped all values which are less than lower limits and higher than upper limits. Finally, we get cleaned data to move onto EDA.

5. **Feature Engineering:**
   Feature engineering is the process of selecting, transforming, extracting, combining, and manipulating raw data to generate the desired variables for analysis or predictive modelling. It is a crucial step in developing Machine Learning models. We created some of the important features such as – Sales_Amount, day_name, Month, Day, Hour, and day_time. Also, datatype of CustomerID column converted to int.

6. **EDA on features:**
   Using visualization libraries in python, did some exploratory data analysis on different features. we did EDA on features to find out highest selling product, most spending customers, days, and months in which highest purchases are done, in which time of the day maximum purchases are done, highest revenue generated w.r.t. products, days, months, etc.

7. **RFM Modelling:**
   RFM analysis is a common approach for understanding customer purchase behaviour. It is quite popular, especially in the retail industry. As its name implies, it involves the calculation and the examination of three KPIs – recency, frequency, and monetary that summarize the corresponding dimensions of the customer relationship with the organization.

   1. Recency (R), the recency value shows the time since the last transaction of the customer's purchase. The smaller the range, the greater the R value.

   2. Frequency (F), the frequency value shows the number of transactions in one period. The more frequency, the greater the F value.

   3. Monetary (M), the monetary value shows the customer value in the form of money spent during the transaction.

8. **K-Means Clustering Algorithm:**
   The K-Means algorithm is a non-hierarchical method that initially takes most of the population components to become the centre of the initial cluster. The K-Means algorithm basically carries out two processes, namely the process of detecting the central location of each cluster and the process of searching for members of each cluster. The workings of the K-Means clustering algorithm are as follow:

1. Determine the value of k as the number of clusters formed.

2. Determine the initial value of the centroid or cluster centre point. At this stage, the centroid value is determined randomly, but for the next stage using the formula below:

$$V_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj}$$

3. Calculate the distance between the centroid point and the point of each object using Euclidean Distance as shown in the formula below:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

4. Grouping data to form clusters with the centroid point of each cluster being the closest centroid point. The determination of cluster members is to consider the minimum distance of objects.

5. Update the centroid value for each cluster.

6. Repeating from step 2 to the end until the value of the centroid point is no longer changed.

9. Elbow Method:

An important thing to remember when using K-means, is that the number of clusters is a hyperparameter, it will be defined before running the model.

In Cluster Analysis, the Elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a dataset.

10. Silhouette Score:

Silhouette coefficient or Silhouette score is a metric used to calculate the goodness of the clustering technique.

It uses compactness of individual clusters (intra cluster distance) and separation amongst clusters (inter cluster distance) to measure an overall representative score of how well our clustering algorithm has performed.

Silhouette score for a datapoint i is given as:

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

where,

b_i is the inter cluster distance defined as the average distance to closest cluster of datapoint I except for that it's part of

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

a_i is the intra cluster distance defined as the average distance to all other points in the cluster to which it's a part of

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Overall Silhouette score for the complete dataset can be calculated as the mean of silhouette score for all data points in the dataset. As can be seen from the formula, Silhouette score would always lie between –1 to 1. 1 representing better clustering.

11. Cohort Analysis:

What is a cohort? In a nutshell, a cohort is simply a subset of users grouped by shared characteristics. In the context of business analytics, a cohort usually refers to a subset of users specifically segmented by acquisition date (i.e., the first time a user visits your website).

A "cohort analysis," then, simply allows you to compare the behaviour and metrics of different cohorts over time. You can then find the highest-performing (or lowest-performing) cohorts, and what factors are driving this performance.

Types of Cohorts –

1. **Time Cohorts:** are customers who signed up for a product or service during a particular time frame. Analysing these cohorts shows the customers' behaviour depending on the time they started using the company's products or services. The time may be monthly or quarterly even daily.
2. **Behaviour cohorts:** are customers who purchased a product or subscribed to a service in the past. It groups customers by the type of product or service they signed up. Customers who signed up for basic level services might have different needs than those who signed up for advanced services. Understanding the needs of the various cohorts can help a company design custom-made services or products for segments.
3. **Size cohorts:** refer to the various sizes of customers who purchase company's products or services. This categorization can be based on the amount of spending in some periodic time after acquisition or the product type that the customer spent most of their order amount in some period.

CONCLUSIONS:

Following are few conclusions which we drew from the whole process.

- Throughout the exercise, we went through various steps to perform customer segmentation. We started with importing data and important libraries. Then, did rigorous data wrangling.
- We have performed RFM Analysis on the data, where we clustered customers based on Recency, Monetary and Frequency aspect. We used Elbow method, Silhouette score method to find appropriate number of clusters. We discovered 4 clusters based on RFM data.
- Further, did cohort analysis to understand how retention and acquisition rate, average amount spends changes over the time.
- However, there can be more modifications on this analysis. One may choose to cluster into more no. depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefers to buy often, finding out customer lifetime value (clv) and much more.