Customer Segmentation of Online Retail

Gaurav Yogeshwar

ABSTRACT

The customer base is usually quite small and individually targetable. But, as a business grows in size, it will not be possible for the business to have an intuition about each and every customer. At such a stage, human judgments about which customers to pursue will not work and the business will have to use a data-driven approach to build a proper strategy.

For a medium to large size retail store, it is also imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their 'best' or high-valued customers. Since the resources that a company has are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, companies use the technique of customer segmentation.

PROBLEM STATEMENT

Customer segmentation has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers. For example, a company can make a high-end product, a budget product, or a cheap alternative product, depending upon whether the product is intended for its most high yield customers, frequent purchasers or for the low-value customer segment. It may also fine-tune the features of the product for fulfilling the specific needs of its customers.

Customer segmentation can also help a company to understand how its customers are alike, what is important to them, and what is not. Often such information can be used to develop personalized relevant content for different customer bases. Many studies have found that customers appreciate such individual attention and are more likely to respond and buy the product. They also come to respect the brand and feel connected with it. This is likely to give the company a big advantage over its competitors. In a world where everyone has hundreds of emails, push notifications, messages, and ads dropping into their content stream, no one has time for irrelevant content.

Finally, this technique can also be used by companies to test the pricing of their different products, improve customer service, and upsell and cross-sell other products or services.

ATTRIBUTE INFORMATION

- **InvoiceNo**
  Invoice number. Nominal, digit integral number uniquely assigned to each transact ion. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**
  Product (item) code. Nominal 5 digit integral number uniquely assigned to each distinct product.
- **Description**
  Product (item) name
- **Quantity**
  The quantities of each product (item) per t ransaction.
- **InvoiceDate**
  Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice**
  Unit price. Numeric, Product price per unit in sterling.

- **CustomerID**
  Customer number. Nominal, a 5 digit integral number uniquely assigned to each customer.
- **Country**
  Country name. Nominal, the name of the c ountry where each customer resides.

INTRODUCTION

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable. Insights from customer segmentation are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning. A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the company philosophy and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy.

- **Zero segments**
  This means that the company is treating all of its customers in a similar manner. In other words, there is no differentiated strategy and all of the customer base is being reached out by a single mass marketing campaign.

- **One segment**
  This means that the company is targeting a particular group or niche of customers in a tightly defined target market.

- **Two or more segments**
  This means that the company is targeting 2 or more groups within its customer base and is making specific marketing strategies for each segment.

- **Thousands of segments**
  This means that the company is treating each customer as unique and is making a customized offer for each one of them.

Once the company has identified its customer base and the number of segments it aims to focus upon, it needs to decide the factors on whose basis it will decide to segment its customers.*F*actors for segmentation for a business to consumer marketing company:

- **Demographic**
  Age,Gender,Education, Ethnicity, Income, Employment, hobbies, etc.

- **Recency, Frequency and Monetary**
  Time period of the last transaction, the frequency with which the customer transacts, and the total monetary value of trade.

- **Behavioral**
  Previous purchasing behavior, brand preferences, life events, etc.

- **Psychographic**
  Beliefs,personality,lifestyle,personal interest, motivation, priorities, etc.

- **Geographical**
  Country, zip code, climatic conditions, urban/rural areal differentiation, accessibility to markets, etc.

STEPS INVOLVED

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.
4. Removing future warnings in seaborn plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Checking the Unique values in the column (if any)
8. Converting the data types to similar objects as the Analysis Demands.
9. Formatting the "size" column into a single column in the dataset.
10. Eradicating special characters from the dataset columns.

- **EXPLORATORY DATA ANALYSIS**

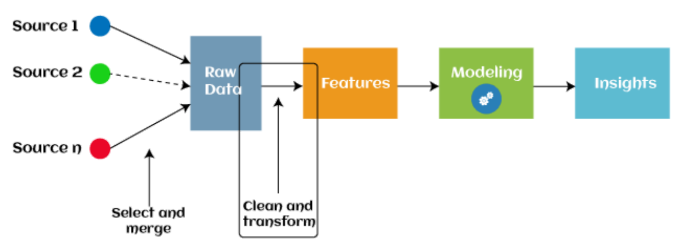
  Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it..

- **EXAMINING NULL VALUES**

  The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analysing the first thing which we will do is to examine the null or missing values on the Dataset. It is the first step that will make the results "more" accurate should be handled before it affects the performance of the models that predict the outcome.

- **Feature Engineering and Data preparation**

  Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improves the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.



- **Cohort Retention Analysis**

  Cohort analysis is a subset of behavioral analytics that takes the data from a given eCommerce platform, web application, or online game and rather than looking at all users as one unit, it breaks them into related groups for analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined time-span.

It is a tool to measure user engagement over time. It helps to know whether user engagement is actually getting better over time or is only appearing to improve because of growth.

App Launched ↓ % Active users after App Launches →

Cohort	Users	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Jan 25	1,098	100%	33.9%	23.5%	18.7%	15.9%	16.3%	14.2%	14.5%	Retention over user lifetime		12.1%
Jan 26	1,358	100%	31.1%	18.6%	14.3%	16.0%	14.9%	13.2%	12.9%			
Jan 27	1,257	100%	27.2%	19.6%	14.5%	12.9%	13.4%	13.0%	10.8%	11.4%		
Jan 28	1,587	100%	26.6%	17.9%	14.6%	14.8%	14.9%	13.7%	11.9%			
Jan 29	1,758	100%	26.2%	20.4%	16.9%	14.3%	12.7%	12.5%				
Jan 30	1,624	100%	26.4%	18.1%	13.7%	15.4%	11.8%					
Jan 31	1,541	100%	23.9%	19.6%	15.0%	14.8%						
Feb 01	868	100%	24.7%	16.9%	15.8%							
Feb 02	1,143	Retention over product lifetime		18.5%								
Feb 03	1,253											
All Users	**13,487**	100%	**27.0%**	**19.2%**	**15.4%**	**14.9%**	**14.0%**	**13.3%**	**12.5%**	**13.1%**	**12.2%**	**12.1%**

It proves to be valuable because it helps to separate growth metrics from engagement metrics as growth can easily mask engagement problems. In reality, the lack of activity of the old users is being hidden by the impressive growth numbers of new users, which results in concealing the lack of engagement from a small number of people.

- **Transforming Skewed Data for Machine Learning**

From my point of view, when a model is trained whether they are linear regression or some Decision Tree (robust to outlier), skew data makes a model difficult to find a proper pattern in the data is the reason we have to make a skew data into normal or Gaussian one.

Standard functions used for such conversions include Normalization, Log, Square root , Cube Root, Box-Cox and Yeo–Johnsonand It all depends on what one is trying to accomplish.

- **RFM Segmentation**

RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).

RFM analysis helps to answer the following questions: Who was our most recent customer? How many times has he purchased items from our shop? And what is the total value of his trade? All this information can be critical to understanding how good or bad a customer is to the company.

After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assign the required order. For example, suppose that we divide each metric into 4 cuts. For the recency metric, the highest value, 4, will be assigned to the customers with the least recency value (since they are the most recent customers). For the frequency and monetary metric, the highest value, 4, will be assigned to the customers with the Top 25% frequency and monetary values, respectively. After dividing the metrics into quartiles, we can collate the metrics into a single column (like a string of characters {*like ‘213’*}) to create classes of RFM values for our customers. We can divide the RFM metrics into lesser or more cuts depending on our requirements.

- **STANDARDIZATION OF FEATURES**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

$$z = \frac{x_i - \mu}{\sigma}$$

ALGORITHM

K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what K-means clustering algorithm is, how the algorithm works, along with the Python implementation of k-means clustering.

What is the K-Means Algorithm?

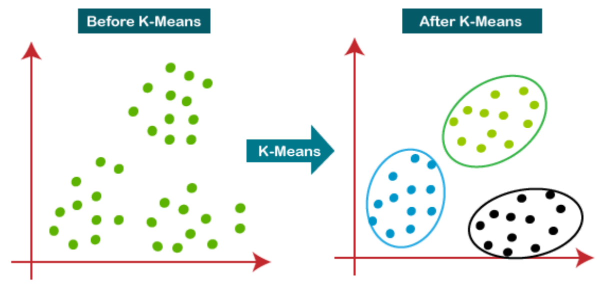
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

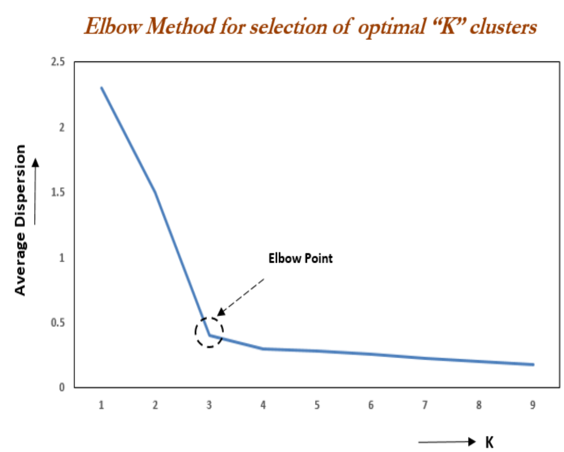
The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to the particular k-center, create a cluster.Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



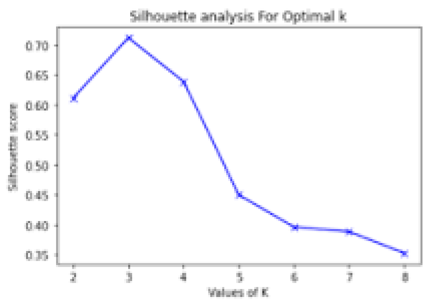
Elbow method to find optimum k value:



The Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

Silhouette score method:

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:



Line plot between K and Silhouette score

1. Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by **a**.
2. Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by **b**.

The Silhouette Coefficient for a sample is,

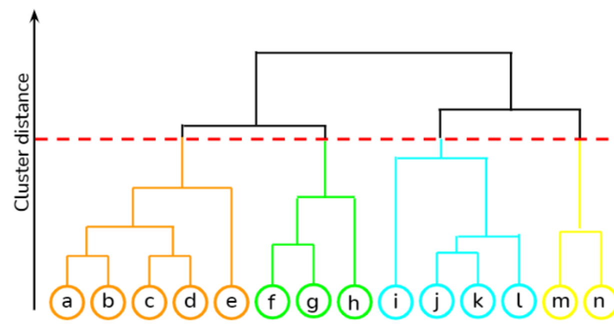
$$s = \frac{b-a}{\max(a,b)}$$

Hierarchical Clustering

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work.



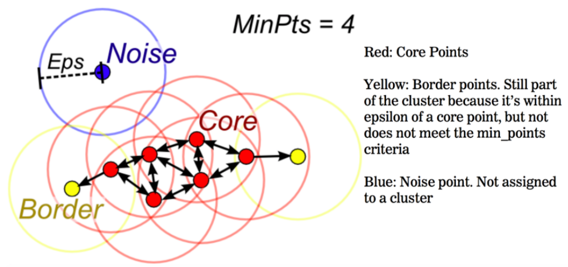
The hierarchical clustering technique has two approaches:

1. **Agglomerative:** Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

Density-based spatial clustering of applications with noise (DBSCAN)

Fundamentally, all clustering methods use the same approach i.e. first we calculate similarities and then we use it to cluster the data points into groups or batches. Here we will focus on Density-based spatial clustering of applications with noise (DBSCAN) clustering methods.

Clusters are dense regions in the data space, separated by regions of the lower density of points. The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise". The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.



CONCLUSION

- This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.
- Using a recency, frequency and monetary (RFM) analysis, the customers have been segmented into various clusters and got a silhouette score of 0.39 for two clusters
- By applying different clustering algorithms to our dataset, we get the optimal number of clusters is equal to 2.
- Since this is an unsupervised learning approach, there is no 100% correct answer, the number of clusters will vary depending on the company's requirement and understanding of domain knowledge.
- The business can focus on these different clusters and provide customers with services of each sector in a different way, which would not only benefit the customers but also the business at large.