Technical Documentation

Online Retail Customer Segmentation

Tushar Wagh, AdityaSingh Thakur, Meenakshi Singh

Abstract

Many small online retailers and new entrants to the online retail sector are keen to practice consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so. In this project we are using unsupervised machine learning techniques to segment customers for an online retailer.

On the basis of the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using the k-means clustering algorithm and Hierarchical clustering the main characteristics of the consumers in each segment have been clearly identified.

Keywords:- *Retail , RFM Analysis , K-Means Clustering , Elbow , Silhouette , Hierarchical.*

1. Problem Statement:-

The main purpose of this project is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively.

In this project, our task is to identify major customer segments on a transactional data set which contains all the transactions 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.

2. Introduction:-

In recent years, middle class income has been grown significantly around 53 million people by 2010 (66% growth). Internet become an important component in today's business life. Many organizations decide to extent their business model by exploiting online strategy to reach superior growth, profit, reputation and matching customer needs. It is not only record purchase data but also potentially record location, demographic and psychographic of online customer. The abundant of online data make it possible to implement big data analytics using certain algorithm to gain customer-centric insight. Internet is commonly used to sale and advertise products or services of multiple genre. There are benefits of using internet as media advertising such as cheaper, easily accessible, use as per convenience, price comparisons, etc. The birth of online retailing using internet started with the launch of two big online retailing websites, eBay and Amazon. Customer buying behavior is a buying behavior of customer for personal consumption, it could be individuals or household consumption. This buying behavior shows how customer purchase goods and services. Comprehensive understanding of buying patterns will benefit the company for strategic marketing, segmentation, distribution, and promotion.

Customer Segmentation is considered an effective method for managing customers while developing diverse marketing strategies, it is the process of dividing customers into homogeneous and distinct groups. Segmentation could be done according to customer characteristics, which

are tracked online helped by certain algorithm. Company need focusing the target customer then gaining maximize profit with win-win situation for company-customer. Customer segmentation is one of solution to optimize the result of win-win situation.

3. Customer Segmentation:-

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.

Why segment your customers?

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.

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.

4. Data Overview:-

The data used in this project is a transactional dataset 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.

The transaction dataset of this online retail store has 8 variables as shown below, and it contains all the transactions occurring in years 2010 and 2011.

Attributes Information:-

1. InvoiceNo: Invoice number. If this code starts with letter ‘c’, it indicates a cancellation.

2. StockCode: Product (item) code, a 5-digit integral number uniquely assigned to each distinct product.

3. Description: Product (item) name.

4. Quantity: The quantities of each product (item) per transaction.

5. InvoiceDate: Invoice Date and time, the day and time when each transaction was generated.

6. UnitPrice: Unit price, Product price per unit in sterling.

7. CustomerID: Customer number, a 5-digit integral number uniquely assigned to each customer.

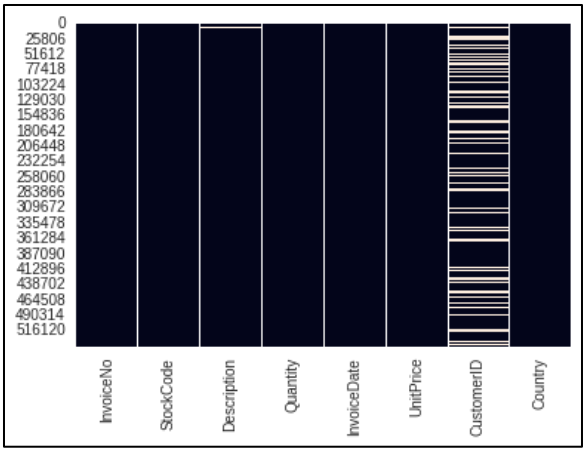
8. Country: Country name.

5. Steps Involved:-

5.1 Data Exploration:-

At the very first we started with loading the dataset and importing the required libraries. Next step was inspecting the data by looking at the top and bottom rows, checking its shape, columns and basic info of the dataset. The dataset contained 541909 rows which are the transactions and 8 columns. Looking at the basic information, we can see there are some missing values present in the dataset.

Before diving into insights from the data, we removed duplicate entries from the data. The data contained 5268 duplicate entries. Then we checked for missing values and observed the null values in Description and CustomerID columns as shown in below figure. Imputing missing CustomerID values was not possible and our main task was customer segmentation so we dropped missing CustomerID values which are around 25%. We also dropped other missing values without losing much data.



5.2 Exploratory Data Analysis:-

Explored the columns one by one and looked at the total number of products, transactions, and customers in the data, which correspond to the total unique stock codes, invoice number, and customer IDs present in the data. There were 4372 customer records present in the dataset having 22190 transactions in total. We observed the minimum value for quantity is negative, this type of transactions is of canceled orders and we saw in the description that these transactions have C at the start of InvoiceNo. There were 8872 canceled orders in our dataset which are huge. The number of products and the number of descriptions didn't match. We can say that some of the products might have more than one description. We checked the number of transactions from each country in the data. We observed around 88% of orders are coming from the United Kingdom.

Next, we tried to get some insights from the data by plotting graphs and charts. We saw top customers who have purchased the maximum quantity. The average quantity purchased by the customer is 250 per order.

Interestingly, customers have placed orders 4 or 5 times on average and the maximum number of orders being 146 from a single customer. We saw not only maximum transactions come from the UK but also most customers are located in the United Kingdom.

5.3 RFM Analysis:-

RFM stands for Recency, Frequency, and Monetary. **Recency, frequency, monetary** value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:

- **Recency:** How recently a customer has made a purchase
- **Frequency:** How often a customer makes a purchase
- **Monetary:** How much money a customer spends on purchases

RFM analysis is a marketing technique in analyzing customer behavior such as how recently a customer has purchased, how often the customer purchases, and how much the customer spends. It could improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions. The advantage is that the customers' behavior can be captured by using a relatively small number of features. The RFM variables are appropriate for capturing the specifics of the customer's purchase behavior.

In order to conduct the RFM analysis, the original dataset needs to be filtered. Considering only the UK based customers data for maximum impact and not to form clustering on geographical conditions. Also filtered the canceled orders and created a total cost column. The prepared dataset had 349227 rows and 10 columns including the date column as well which we created in EDA.

Started with calculating Recency, first fixed a reference date as 2011-12-10 as the last date of transaction in our dataset is 2011-12-09. Then calculated the day's difference between the most recent transaction carried out by the customer and reference date. Next calculated the frequency to know how many times a customer has purchased. Then calculated Monetary value for all customers and combined the recency, frequency and monetary to form a single dataset.

After getting the RFM values, created 'quartiles' on each of the metrics and assigned the required order. We divided each metric into 4 cuts. For the recency metric, the highest value, 4, is 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, is assigned to the customers with the top 25% frequency and monetary values, respectively. After dividing the metrics into quartiles, we collated the metrics into a single column (like '213') to create classes of RFM values for our customers. The RFM score is calculated by summing up the RFM quartile metrics. Finally, we created segments within this score range of RFM score 3-12, by manually creating categories in our data.

5.4 RFM Based Clustering Models:-

Data Preprocessing:-

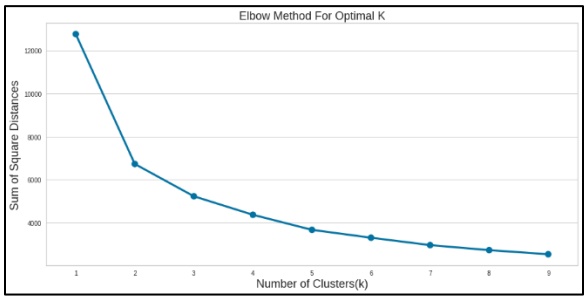
Next step is clustering the RFM model using different clustering algorithms such as K-means clustering and Hierarchical clustering. Before that, plotted the distribution of recency, frequency and monetary and saw their distributions were skewed. As K-Means require normally distributed data so we applied log transformation to reduce skewness. Scaled the data using Standard scaler.

5.4.1 K-Means Clustering:-

K-means is a well-known clustering algorithm that is frequently used for unsupervised learning tasks.

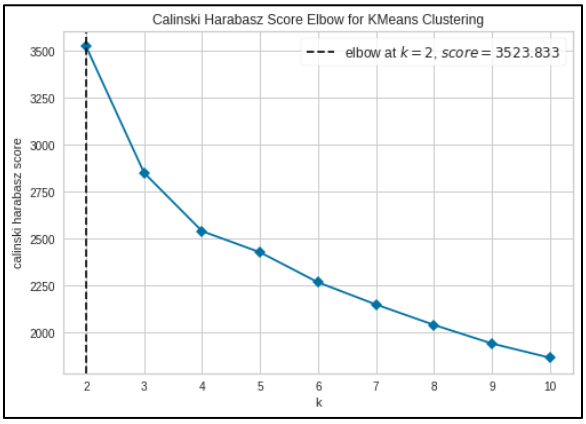
1) Elbow Method:-

In this section, we built multiple clusters upon our normalized RFM data and tried to find out the optimal number of clusters in our data using the elbow method. For each cluster, we have also extracted information about the sum of squared distances through which we built the elbow plot to find the desired number of clusters in our data.



As the number of clusters increases, the variance (within-cluster sum of squared distances) decreases. Here, we cannot see a very distinct elbow point. One might infer the optimal value of K to be 2, 3 , 4 and 5.

Taking Calinski Harabasz score as the metric, we get the following elbow plot for our data:



From the above plot, we can see that the optimal number of clusters is 2.

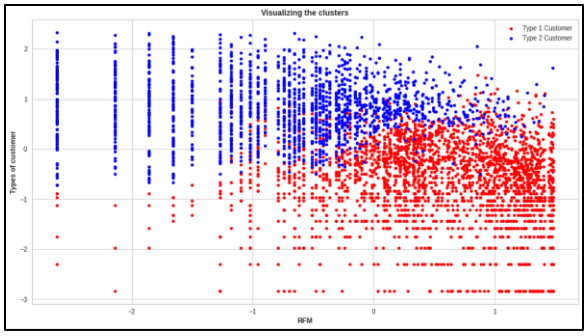
2) Silhouette Analysis:-

Silhouette analysis can be used to study the separation distance between the resulting clusters. We checked the silhouette score for the number of clusters ranging from 2 to 9. The best silhouette score obtained is when the number of clusters is 2.

```
For n_clusters = 2 The average silhouette_score is : 0.39408103379054493
For n_clusters = 3 The average silhouette_score is : 0.29475936365115435
For n_clusters = 4 The average silhouette_score is : 0.2975051811313832
For n_clusters = 5 The average silhouette_score is : 0.28291048922517165
```

5.4.2 Hierarchical Clustering:-

We used an Agglomerative Hierarchical Clustering algorithm but before that drew the dendrogram to help us decide the number of clusters. We got 2 clusters from dendrogram and then applied hierarchical clustering on our RFM data using clusters k=2.



We can see that; Customers are well separated when we use Hierarchical clustering and the number of clusters equal to 2.

6. Results:-

We performed RFM analysis on the online retail dataset and then used RMF data to perform clustering using K-means and Hierarchical. By knowing clusters based on K Means and Hierarchical clustering, we mapped each customer id to a cluster group. K Means Clustering generated the optimal K value = 2. Now that cluster has been set, the mean value of each cluster.

Cluster	Recency	Frequency	Monetary	R	F	M	RFMScore
0	34.501024	137.489765	2362.510876	1.763562	1.575742	1.582395	4.921699
1	143.607298	23.169852	410.934853	3.103388	3.308862	3.278888	9.691138

- **Cluster 0:-**
  Comprises of customers who have very high frequency and monetary value but it's RFMScore is very less as compared to Cluster 1 and also contribute largely to the sales.

- **Cluster1:-**
  Comprises of customers who have high recency and monetary value which results in high RFMScore value.

Cluster 0 who have high mean monetary value and also have high mean frequency value. Apart from the numbers, the visualization of clusters in Silhouette Analysis show that both customer segments are quite distinct with very little overlap between them.

In Hierarchical clustering, the dendrogram also depicted the number of clusters as 2. We saw in Hierarchical clustering the customers were well grouped using 2 clusters. Looking at the below summary we can observe the same as K-means clustering.

Here, mean recency for both clusters have decreased whereas mean frequency and monetary have increased as compared to K-means clustering.

Wholesale Customers - 'Cluster 1' is the high value customer segment as the customers in this group place the highest value orders with a very high relative frequency than other members. They are also the ones who have transacted the most recently. These are the wholesale customers of the retail store.

Average Customers - 'Cluster 0' is the average customer segment. These customers order less frequently than the wholesale customers and their orders are pretty low valued.

7. Conclusion:-

- Dataset consists of 541909 observations and 8 features which includes various dtypes such as object d-type , integer d-type , floating number d-type , and feature 'InvoiceDate' is of datetime d-type.
- There were null values as well as duplicate values in our dataset.

- There were more 'United Kingdom' r esidents in 'Country' feature in our da taset.
- There were 89% United Kingdom re sidents in 'Country' feature because o f highest value counts and the least w as Spain in our dataset
- Using the boxplot, we detect outliers for Recency , Frequency and Moneta ry and we get to know that there so much outliers in Monetary so we app lied Inter-
  Quartile Range Method to eliminate outliers but that didn't help much to r emove it.
- We used Standard Scaler for Normal ization purpose which helps us to eva luate values between the range of 0 a nd 1.
- For building the model we firstly app ly K-
  Means Clustering and it we used Elb ow method to get the right numbers of clusters as well as Silhouette Anal ysis had been performed for each clu ster values with their respective Silh ouette score.
- We have also used Hierarchical Clus tering with various linkages methods such as Single Linkage , Complete L inkage , Average Linkage , Centroid Linkage as well as Ward's
  Linkage and plot their respective den drograms.

K-Means Clustering with 3 Cluster Id's:-

- Customers with Cluster Id 1 are the c ustomers with high amount of transa ctions as compared to other customer s.
- Customers with Cluster Id 1 are freq uent buyers.
- Customers with Cluster Id 2 are rece nt buyers and hence most of importa nce from business point of view.

Hierarchical Clustering with 3 Cluster Labels:-

- Customers with Cluster_Labels 2 are the customers with high amount of t ransactions as compared to other cust omers.
- Customers with Cluster_Labels 2 are frequent buyers.
- Customers with Cluster_Labels 0 are recent buyers and hence most of imp ortance from business point of view.