Technical Document

Online Retail Customer Segmentation

Prepared By

Nitin Solanki
(solankinitin1210@gmail.com)

Contents

1. Objective

Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviours, so you can market to those customers more effectively.
These customer segmentation groups can also be used to begin discussions about building a marketing persona. This is because customer segmentation is typically used to inform brands' messaging, and positioning and to improve how a business sells – so marketing personas need to be closely aligned to those customer segments to be effective.
In this Analysis Our problem is a dataset is given which content transaction recorder of UK based company and from data set we need to obtain the optimum number of cluster of customers

2. Introduction

Customers are the most important people in any business. Many businesses get most of their revenue from their 'best' or high-valued customers. it is crucial to find these customers and target them. For that customer segmentation is required, Customer segmentation is the process of separating customers into groups based on their shared behaviour or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other.

Customer segmentation helps a company to develop an effective strategy for targeting 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.

3. Dataset

We have been given a 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.

Attribute Information

- invoice no: Invoice number. Nominal, is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, is a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- 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, is a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

4. Data Modification and Exploration

In this section, Data modification, exploration and pre-processing will be explained, which is required before we input our data set to any machine learning algorithm to improve prediction for machine learning models.

Data exploration and pre-processing part will be done using pandas and NumPy packages. All the graphs and figures will use Matplotlib and seaborn package.

RFM Analysis

RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns. The system assigns each customer a numerical score based on these factors to provide an objective analysis. RFM analysis is based on the marketing adage that "80% of your business comes from 20% of your customers."
RFM analysis ranks each customer on the following factors:

Recency: How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their minds and are more likely to purchase or use the product again. Businesses often measure recency in days. But, depending on the product, they may measure it in years, weeks or even hours.

Frequency: How often did this customer purchase in a given period? Customers who purchased once are often more likely to purchase again. Additionally, first-time customers may be good targets for follow-up advertising to convert them into more frequent customers.

Monetary: How much money did the customer spend in a given period? Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business.

	CustomerID	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFMGroup	RFMScore
0	12346	327	1	77183.60	4	4	1	441	9
1	12747	4	103	4196.01	1	1	1	111	3
2	12748	2	4412	33053.19	1	1	1	111	3
3	12749	5	199	4090.88	1	1	1	111	3
4	12820	5	59	942.34	1	2	2	122	5
5	12821	216	6	92.72	4	4	4	444	12
6	12822	72	46	948.88	3	2	2	322	7
7	12823	76	5	1759.50	3	4	1	341	8
8	12824	61	25	397.12	3	3	3	333	9
9	12826	4	91	1474.72	1	2	2	122	5

5. Cluster Evaluation Metrics

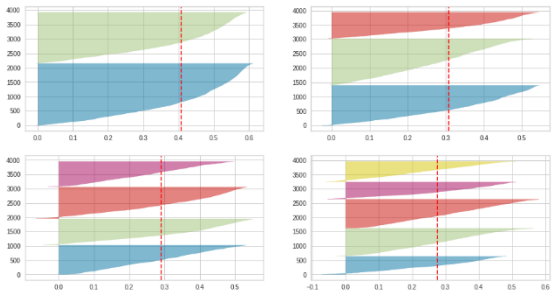
The evaluation metrics which do not require any ground truth labels to calculate the efficiency of the clustering algorithm could be used for the computation of the performance evaluation. There are three commonly used evaluation metrics: Silhouette score, Calinski Harabaz index, and Davies-Bouldin Index.

5.1 Silhouette score

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 for each observation belonging to all the clusters:

- 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
- 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

Silhouette score, S, for each sample is calculated using the following formula:
S = frac{ (b - a)} {max(a, b) }



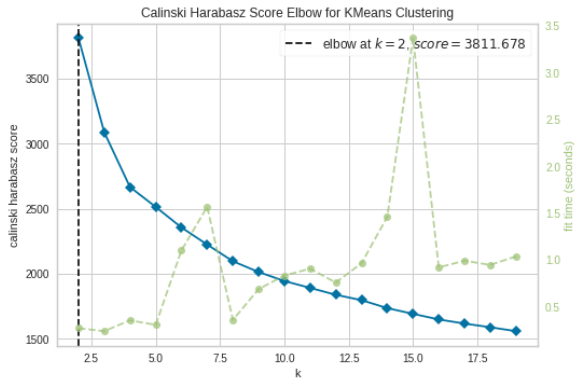
The value of the Silhouette score varies from -1 to 1. If the score is 1, the cluster is denser and more well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighbouring clusters. A negative score [-1, 0] indicates that the samples might have got assigned to the wrong clusters.

5.2 Calinski Harabaz index

The Calinski Harabaz index is based on the principle of variance ratio. This ratio is calculated between two parameters within-cluster diffusion and between cluster dispersion. The higher the index the better is clustering. The formula used is

CH(k)=[B(k)W(k)][(n−k)(k−1)]

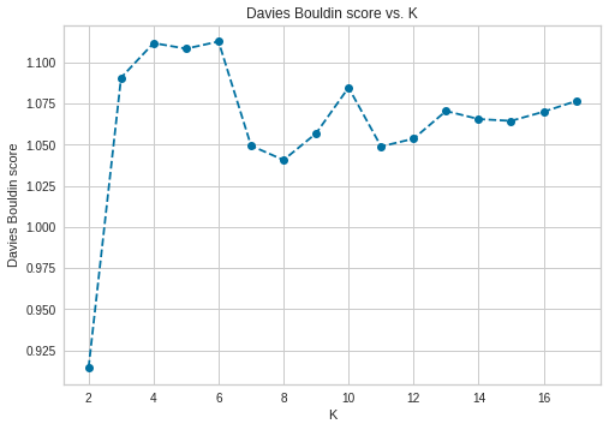
where n = data points , k = clusters, W(k) = within cluster variation, B(k) = between cluster variation.



5.3 Davies Bouldin index

Davies Bouldin index is based on the principle of with-cluster and between cluster distances. It is commonly used for deciding the number of clusters in which the data points should be labelled. It is different from the other two as the value of this index should be small. So the main motive is to decrease the DB index.
The formula is used to calculate the DB index.

DB(C)=1Ci=1kmaxjk,jiDij
Dij=di + dj dij
where Dij= within-to-between cluster distance ratio for the ith and jth clusters. C = no of clusters
i,j = numbers of clusters which come from the same partitioning

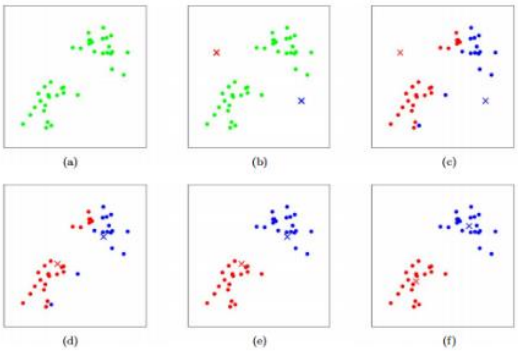


6. Different Clustering Model

6.1 K-Means Clustering

K-Means Clustering is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties. Here K defines the number of pre-defined clusters that need to be created in the process, if K=2, there will be two clusters, and for K=3, there will be three clusters,
The following stages will help us understand how the K-Means clustering technique works-

- Step 1: First, we need to provide the number of clusters, K, that need to be generated by this algorithm.
- Step 2: Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
- Step 3: The cluster centroids will now be computed.
- Step 4: Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.



6.2 Hierarchical clustering

Hierarchical clustering is one of the popular and easy-to-understand clustering techniques. This clustering technique is divided into two types:

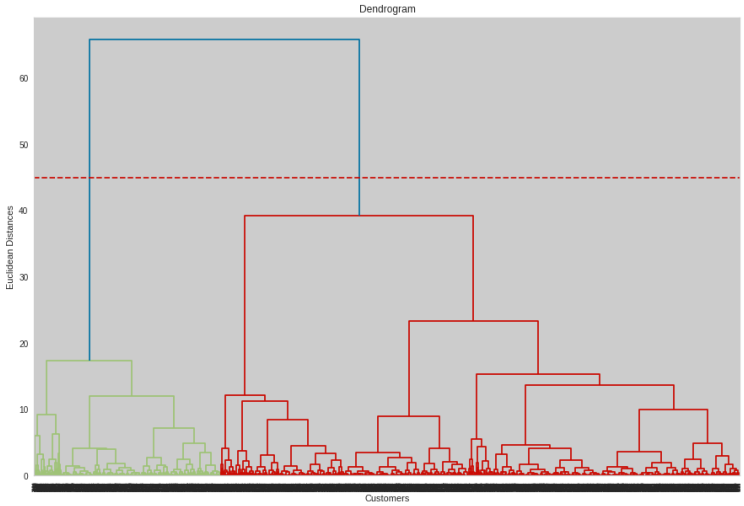
- Agglomerative
- Divisive

Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

The basic algorithm of Agglomerative is straightforward.

- Compute the proximity matrix
- Let each data point be a cluster
- Repeat: Merge the two closest clusters and update the proximity matrix
- Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters

A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data.

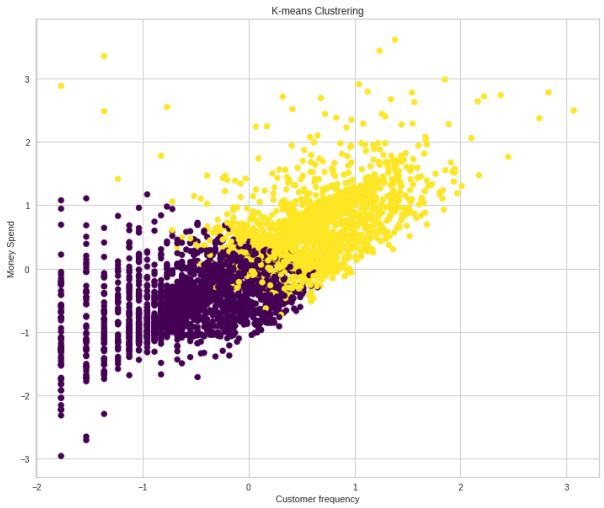


we can say that the Divisive Hierarchical clustering is exactly the opposite of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we consider all the data points as a single cluster and in each iteration, we separate the data points from the cluster which are not similar. Each data point which is separated is considered an individual cluster. In the end, we'll be left with n clusters.

7. Score Board

Model Name	Optimal Number of Cluster
K-Means with Silhouette Score	2
K-Means with CH Index	2
K-Means with DB-Score	2
Hierarchical Clustering with Silhouette Score	2
Hierarchical Clustering with Dendogram	2
Hierarchical Clustering with DB-Score	2

Here we got our final results in the form of clusters segmented into different clusters



8. Conclusion

- Built a clustering model using K-means and hierarchical clustering to identify major customer segments on transaction data to optimize the impact of marketing.
- Engineered features to obtain new features such as RFM-Score for getting more details about the customers' purchasing behaviour.
- Evaluated the optimal clusters using the silhouette score, CH index, DB-score and dendrogram. The optimal number of clusters was 2 with a silhouette score of 0.4.