Online retail Customer Segment

Ankit Sharma
Pankaj Kumar
Data science trainees,
AlmaBetter, Bangalore

company mainly sells unique all-occasion gifts. Many customers are whole sellers

Introduction

We have dataset of a company which is quite large. As we understand that when a small company is started it has a small customer base of which keeping a customer base or record is quite easy but when the company grows with time then it becomes difficult for the Human's inside the company to keep record of shopping habits and customers information of each customer by themselves at that time Machine learning and data driven approach is required and is very helpful at that stage to build a business approach and strategy.

Understanding 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.

Problem Statement

In this project, your task is to identify major customer segments on a transnational data set which contain all the transactions occurring 01/12/2010 and 09/12/2011 for a UK- based and registered non -store online retail. The

Data Description

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, 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: Invice 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 country where each customer resides.

Steps to perform following analysis

Step 1-Mounting the google drive

Step 2 -Loading the data.

The dataset we use is customer data we'll load the CSV file with the help of a function read csv

Step 3 -Checking the null values

We will check null values and we have null values in Description and Customer Id as null values is in huge amount, we can drop it as customer id and Description are not much important

Step 4 -Exploring the data.

We will look into the data for better understanding of data. we will use function like head and tail to print five rows of data from head and bottom. With the help of function like shape and info we will dig into data more and get to know about all the columns.

- We will find top 5 product with maximum sale and bottom 5 product with minimum purchase with the help of count function and plotted a graph for better understanding.
- We will find out top 5 countries with maximum number of purchase and bottom 5 countries with maximum number of purchases
- For all this graph we will use bar plot and use count function
- We have two integer type values price and Quantity feature we will see distribution of values. Both the values are highly skewed.
- we will use log function on both the data and applied displot as a result we will get gaussian distribution

Step 5- Feature engineering

We will split date time into different columns to do in-depth EDA on date time columns

- Plotted bar graph on month, year and time column.

Conclusion Of EDA

- ✓ We can conclude that most of the customers have purchase the items in Thursday, Wednesday and Tuesday
- ✓ The most numbers of customers have purchased the gifts in the month of November, October and December September
- ✓ From this graph we can see that in After Noon Time most of the customers have purchase the item.
- ✓ Most of the customers have purchase the items in Afternoon, moderate numbers of customers have purchased the items in Morning and least numbers of customers have purchase the items in Evening
- ✓ Top countries with respect to purchase are **UK Germany France.**
- ✓ Top Items are WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER JUMBO BAG RED RETROSPOT ASSORTED COLOUR BIRD ORNAMENT

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 Score

Recency = Latest Date - Last Invoice Data,

Frequency = count of invoice no. of transaction(s),

Monetary = Sum of Total

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.

Quartiles

After getting the RFM values, a common practice is to create 'quartiles' on each of the metrics and assigning 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.

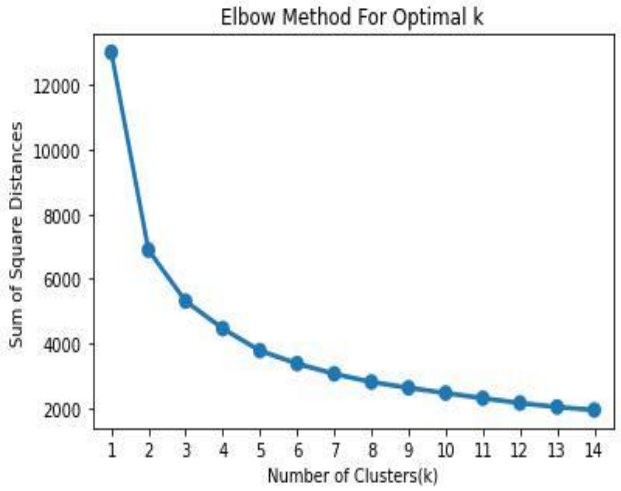
Model Implementation

K Means- Clustering

K-means is a well-known clustering algorithm that is frequently used for unsupervised learning tasks. For our purpose, we need to understand that the algorithm makes certain assumptions about the data. Therefore, we need to preprocess the data so that it can meet the key assumptions of the algorithm, which are:

1. The variables should be distributed symmetrically
2. Variables should have similar average values
3. Variables should have similar standard deviation values
4. We will form clusters from 2 to 15 and check how many clusters should be needed to Separate data
5. After k means clustering, we got this result and graph for elbow method on Recency Frequency and Monetary

Cluster	Silhouette score
2	0.39
3	0.30
4	0.30
6	0.27
7	0.27
8	0.26
9	0.25
10	0.26
11	0.25
12	0.26
13	0.26
14	0.26
15	0.27



6. From the above result we can infer that might be at cluster =2 will be optimal k for the given data

Elbow method

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e., the point after which the distortion/inertia start decreasing in a linear fashion.

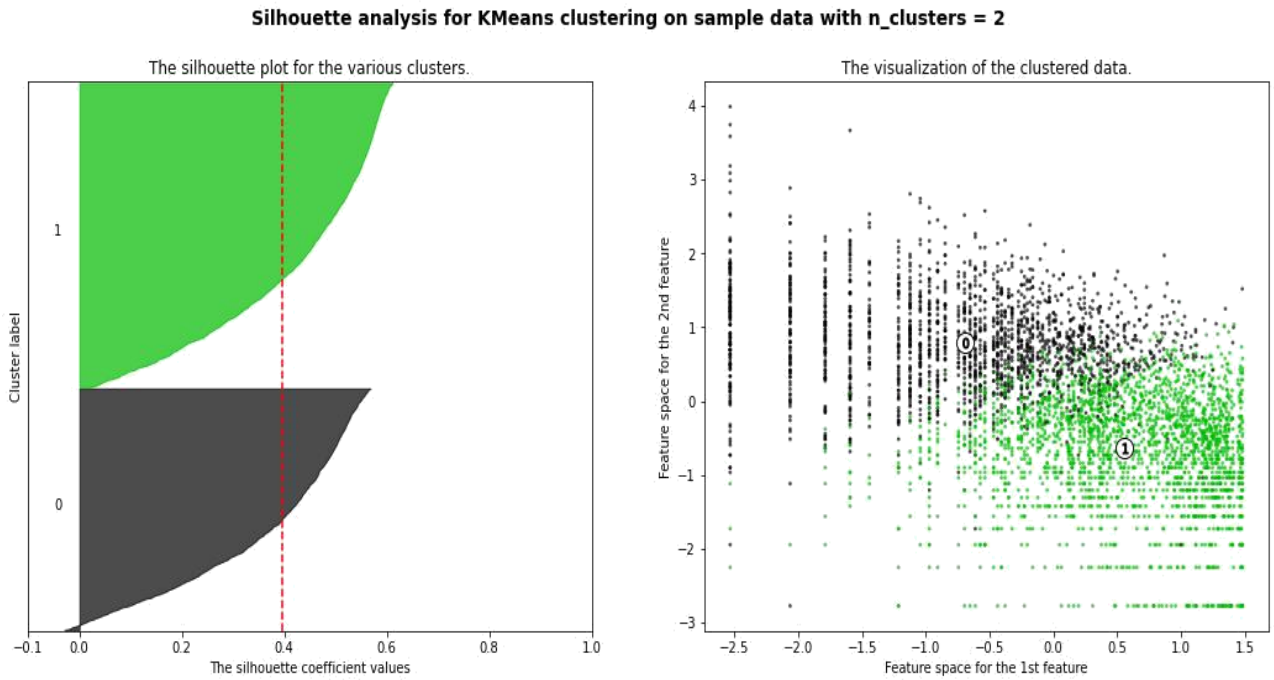
Silhouette (clustering)

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

cluster =2 can be optimal k for the given data from the above elbow method and silhouette score

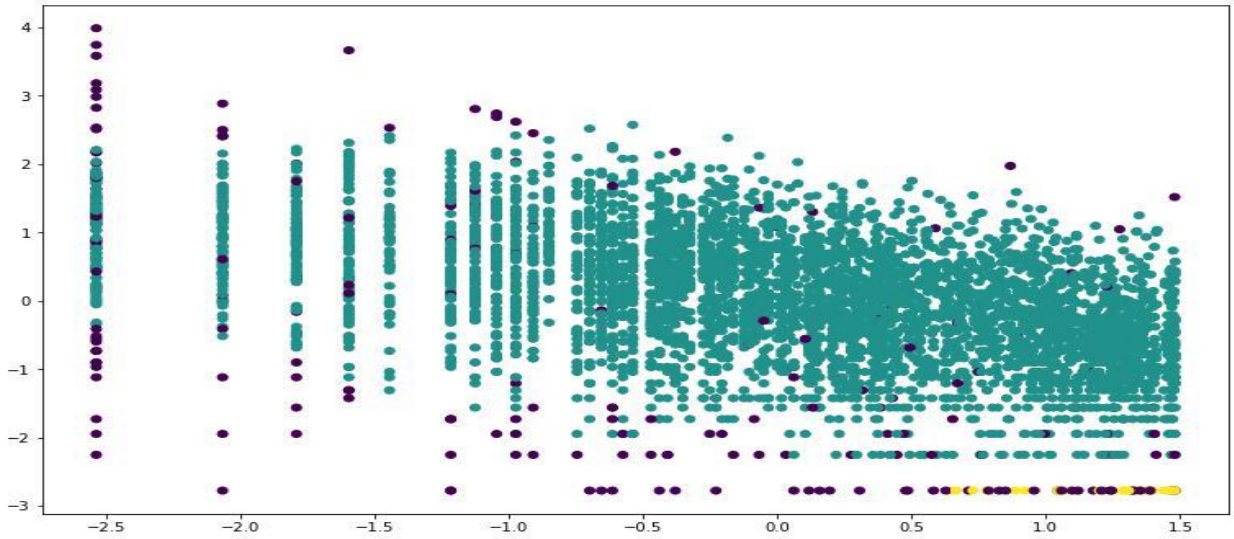


From the above we can say that cluster =2 is the optimal k values

DBSCAN Method

DBSCAN stands for **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise.

It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers**.** It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.



We see that, Customers are well separate when we cluster them by Recency, Frequency and Monetary and optimal number of clusters is equal to 3

Conclusion

	Model_Name	Data	Optimal_Number_of_cluster
0	K-Means with silhouette_score	RFM	2
1	K-Means with Elbow methos	RFM	2
2	DBSCAN	RFM	3

- ✓ K-Means with silhouette_score of RFM Optimal_Number_of_cluster are 2
- ✓ K-Means with Elbow methos of RFM Optimal_Number_of_cluster are 2
- ✓ DBSCAN of RFM Optimal_Number_of_cluster are 3
- ✓ Through the model we didn't get clearly separated clusters but we were able to build a model that can classify customers into 2 categories that are: low value" and "high value" customer groups. A customer could only be considered as a "high value customer" if he/she comes in at least top 50^{th} percentile in monetary spendings despite of the fact that how many times he/she have shopped with

us. The cluster assignments found to be little muddled which could be due to outliers that were not removed.