Online Retail Customer Segmentation

Ajit Kumar Toppo
Data Science trainee,Almabetter

Abstract

Customer Segments (or Market Segmentation) allow the companies to be able to utilize their resources (time, finance) to serve their goals: increasing sales, increasing profits, retaining important customers as well as implementing marketing campaigns more effectively, which is based on the understanding of the customer's behavior, habits, and preferences.

Keywords: Pandas, Matplotlib, Seaborn, Numpy, Scikit-learn, Unsupervised machine learning, segmentation, RFM

Introduction

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. For this project we have transaction records of online purchases done by customers from a UK based online retail. Our objective will be to segment the customers based on RFM analysis and K-means clustering. Companies use the clustering process to foresee or map customer segments with similar behavior to identify and target potential user base. Building a model to predict the optimal number of customers is essential for a business to understand customer behavior, plan business strategies, marketing campaigns, etc. to target, incentivise and attract customer base.

About Dataset

This is 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. Features of the dataset are explained below:

InvoiceNo:
A unique identifier for the invoice. An invoice number shared across rows means that those transactions were performed in a single invoice (multiple purchases).

StockCode:
Identifier for items contained in an invoice.
Description:
Textual description of each of the stock items.
Quantity:
The quantity of the item purchased.
InvoiceDate:
Date of purchase.
UnitPrice:
Value of each item.
CustomerID:
Identifier for customer making the purchase.
Country:
Country of customers.

Data Cleaning

There are 5,41,909 rows and 8 columns present in the dataset. There are 1,454 missing values in the 'Description' column and 1,35,080 missing values in 'CustomerID'. In order to do segmentation we need a unique id of customers, all the rows where customerID are missing will not be useful so I dropped all the rows with missing values present.

EDA and data preparation

After removing all the missing values the dataset has 4372 unique customer records,3684 unique products. The total number of unique transactions present is 22,190. Also it was found that 3,61,878 CustomerID are from the United Kingdom. Customer clusters vary by geography, so here we'll restrict the data to the United Kingdom only. For segmentation we will consider only those customers who are from the United Kingdom. In feature Quantity there were some negative values which represent the cancelled orders which were dropped for further analysis. For the segmentation we will consider only the transactions that are done between 9/12/2010 to 9/12/2011 because it's better to use a metric per Months or Years in RFM.

RFM

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 numerical scores based on these factors to provide an objective analysis. RFM analysis is based on the marketing adage that "80% of business comes from 20% of

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 mind 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. Here we have measured recency in days.

Frequency: How often did this customer make a 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. In our data the amount spent is given in terms of sterling(£).

Next we categorized RFM data obtained into four quartiles based on whether they lie between 0-25%,25-50% ,50-75%,75-100% as 1, 2, 3 and 4.

Customers who have an RFM score of 444 are our best customers.

Customers who have a Frequency score of 4 are our loyal customers.

Customers who have a Monetary score of 4 are big spenders.

Customers who have a RFM score of 244 are almost lost.

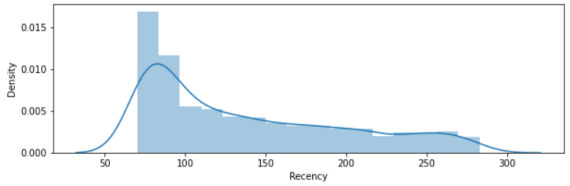
Customers who have a RFM score of 144 are lost customers.

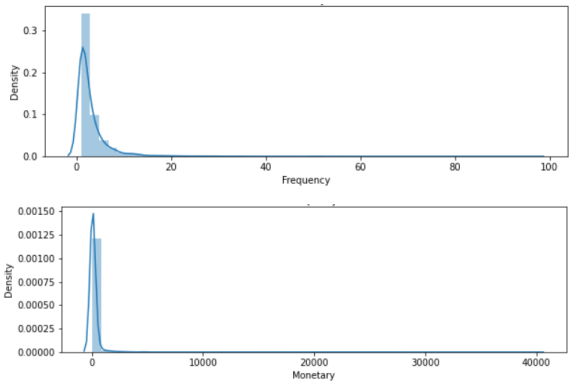
Customers who have a RFM score of 111 are lost cheap customers.

CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore
12747.0	109	5	191.85	3	4	4	344
12748.0	70	96	1054.43	4	4	4	444
12749.0	130	3	67.00	2	3	3	233
12820.0	74	1	15.00	4	1	1	411
12821.0	214	1	19.92	1	1	2	112

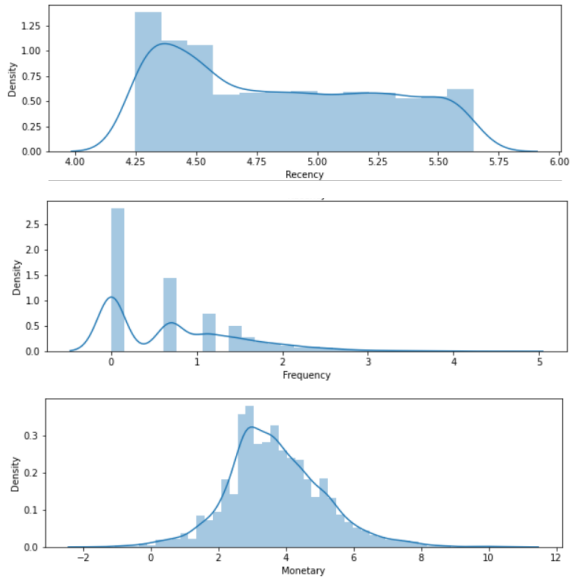
Preprocessing data for clustering

Checked the distribution of data in our variables Recency, Frequency and Monetary and found that the data is skewed towards right which needs to be treated before actually applying the model.





To deal with the skewed data distribution applied log transformation on the data and the results are shown below.

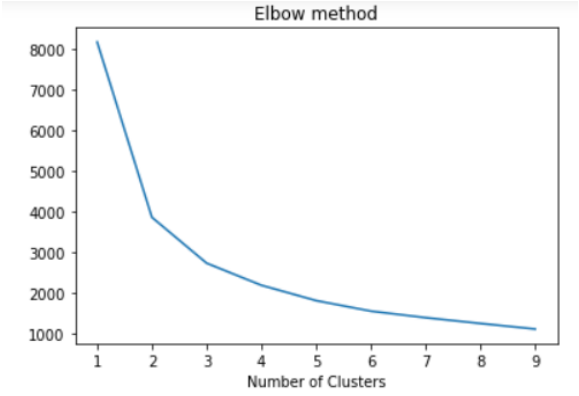


After applying log transformation we see Recency and Monetary have better distribution, more normalized. But this is not the case for Frequency which has improved little only.

K-Means Clustering Algorithm

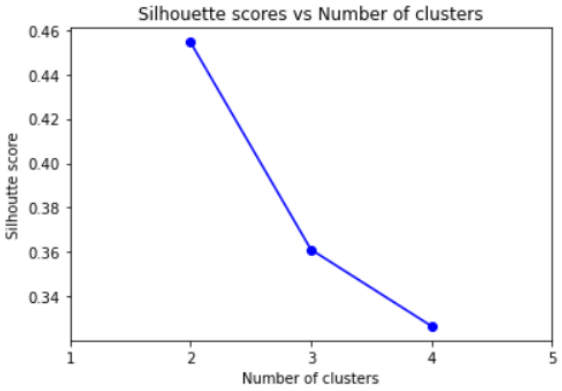
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

Elbow method is used to find the elbow in the elbow plot. The elbow is found when the dataset becomes flat or linear after applying the cluster analysis algorithm. The elbow plot shows the elbow at the point where the number of clusters starts increasing. To find the suitable number of clusters I plotted a elbow curve which is shown

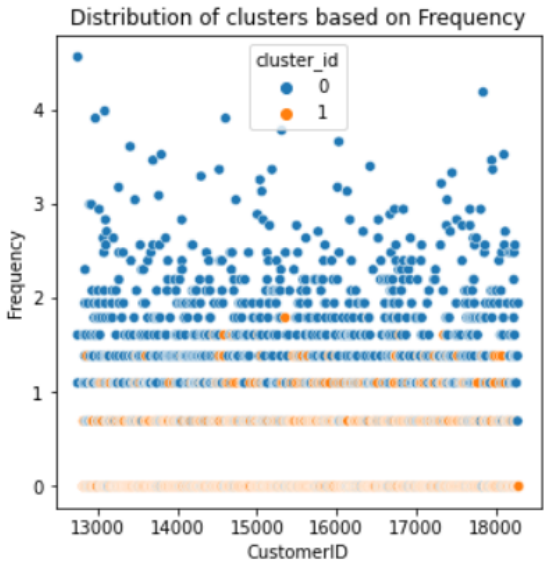


From the above elbow method we see that the optimal number of clusters is 2 or 3.

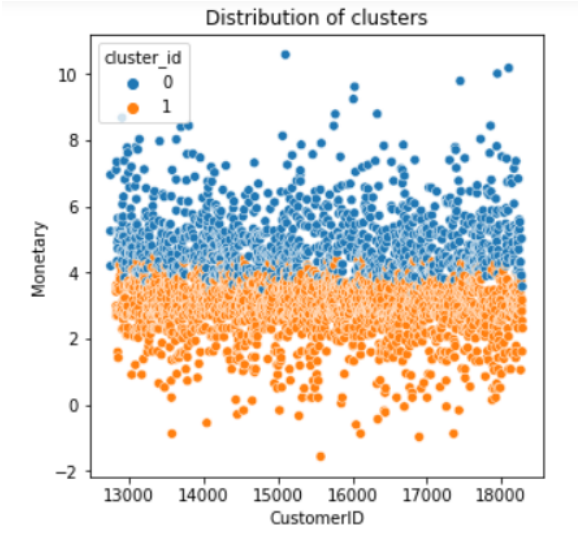
Silhouette score of a point measures how close that point lies to its nearest neighbor points, across all clusters. It provides information about clustering quality which can be used to determine whether further refinement by clustering should be performed on the current clustering



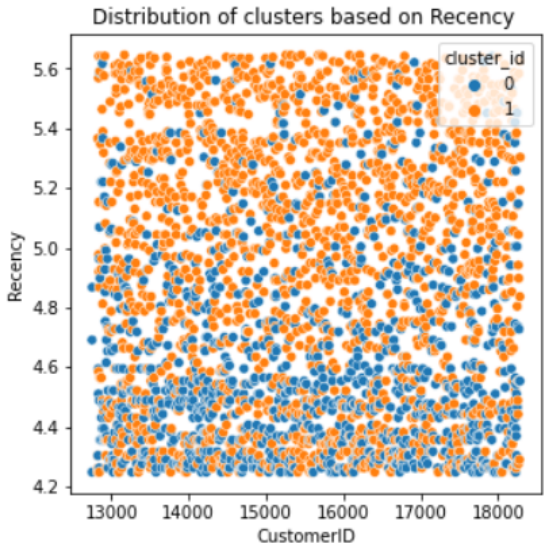
From the above graph we found the silhouette score is highest for the value of k=2, after that applied the K-means model on the data with k=2. The model classified CustomerID into two clusters: 1100 are in the cluster_id 0 and 1764 are in cluster _ id 1. Graphical representation of clusters are shown below based on Frequency, Monetary and Recency respectively



We see from the above graph customerID with low buying frequency are placed in cluster_id 1 and customerID with high frequency belong to cluster 0.



customerID with low monetary value are placed in cluster_id 1 and customerID with high monetary value are placed in cluster_id 0.



There is no clear distinction between clusters based on recency but still customers with low recency are grouped towards both cluster_id 0 & 1 whereas customers with high recency are more grouped in cluster_id 1.

Conclusion

From all the graphs we see the k-means has created two clusters for CustomerID 0 and 1. Customers with relatively high frequency of buying, high monetary value who purchased recently are grouped in cluster 0. Customers with relatively low frequency of buying, low monetary value some of them purchased recently while for some its relatively long time are placed in cluster 1. We observed clusters from RFM analysis also by k-means which can be used by business to understand customer behavior, plan business strategies, marketing campaigns, etc. to target, incentivise and attract customer base.

References:

1.Towards Data Science
2.Rpubs
3.kaggle