ONLINE RETAIL CUSTOMER SEGMENTATION

Shadab Shaikh,
Data Science Trainee,
AlmaBetter, Bangalore

Abstract

Businesses, especially retail, heavily depend upon customer behavior and historical data to scale and manage business strategies. Keeping a record of customers, their demographics and interaction with the business is essential for product and service based industries.

Customer segmentation is a strategy which proves to be traditionally helpful for the organizations to group their customers, based on common characteristics. This enables them to plan new and unique strategies to target the customers based on their own needs and behavior.

Introduction

The process of categorizing consumers based on their frequent behaviors or other characteristics is known as customer segmentation. Both within and in relation to one another, the groups should be homogeneous.

The major objective is to identify the most profitable, devoted, and churned-out consumers in order to redefine company rules to stop additional customer loss. It might be challenging to identify which of a big number of clients is most crucial to the success of the company and to target them with an effective plan. Each customer has distinct wants.

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.

Problem Statement

In this project, your task is to identify major customer segments on 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.

With the increase in customer base and transactions, it is not easy to understand the

requirement of each customer. Segmentation can play a better role in grouping those customers into various segments.

Data Description

Our dataset contains product and order description, order date, customer and location information.

Attribute Information:

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: 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: 1 : Country name. Nominal, the name of the country where each customer resides.

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
For the scope of this project we deal with cleaning missing observations, duplicate rows and unwanted data.

Missing values:

Our dataset consists of about 25% of total data which has missing values. Out of CustomerID and Description, the former constitutes most of these values. Here, with the objective being segmentation of customers, we cannot do that without a unique identifier for customers, here, CustomerID. Hence, we drop these rows.

Duplicate Observations:

There were about 5268 duplicate observations present in the dataset. These were also simply dropped out of the data.

Unwanted Data:

The data dictionary suggests the existence of canceled orders recorded in the dataset. These were represented by the letter ‘c’ in the InvoiceNo feature. To proceed building the models, we do not require these observations. We cleaned these observations.

Feature Engineering

The approach of leveraging domain expertise to extract features (characteristics, qualities, and attributes) from unprocessed data is known as feature engineering or feature extraction. The goal is to employ these additional features to enhance the quality of machine learning process results as opposed to only giving the process raw data.
We carried out the following:

- Extracted new features from the ‘InvoiceDate’ column to get ‘Year’, ‘Month’, ‘MonthNum’, ‘Day’, ‘DayNum’, ‘Hour’, ‘Minute’
- construct a new feature, ‘TotalAmount’ from the ‘Quantity’ and ‘UnitPrice’ columns.
- 'DayPart' gives us the part of the day(Morning, Afternoon, Evening) based on the hour of that day. We extracted this feature from the newly engineered ‘Hour’ column.

Exploratory Data Analysis

The exploratory data analysis that we performed helped us derive important business insights. We drew the following hypothesis:

- WHITE HANGING HEART T-LIGHT HOLDER is the most frequently bought product
- PAPER CRAFT LITTLE BIRDIE is the product most bought in bulk
- Manual is the costliest product and PADS TO MATCH ALL CUSHIONS is the cheapest product
- PAPER CRAFT LITTLE BIRDIE makes the most revenue.
- November is the busiest month, followed by October and December.
- Thursday is the busiest day of the week.
- 12 pm is the peak hour of transactions.
- Most orders are placed in the
- afternoon.
- Most Transactions are from the UK.
- Most customers are based in the UK.
- Most bulk orders are placed from the Netherlands.

RFM Analysis

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 Value: How much money a customer spends on

We calculate the above attributes as follows:

- Recency = Latest Date - Last Invoice Data
- Frequency = Count of invoice no. of transaction(s)
- Monetary = Sum of total amount for each customer

We then get the R, F and M scores for each customer by dividing each attribute into four quantiles. This is interpreted as follows:

We will group the R, F and M values to get the groups for each customer. Explanation:

- Best Customer - If a customer belongs to group 444, they have made a purchase very recently and has high frequency and monetary value.
- Worst Customer - If a customer belongs to group 111, it means made a purchase a long time ago and has low frequency and monetary value.

We will give every customer an RFM score based on their individual R, F and M values. Explanation:

- Best Customer - High RFM Score
- Worst Customer - Low RFM Score

After inspection we have found out the Recency, Frequency and Monetary features are skewed towards the right. For a smooth model building process, we carry out log transformation on these features to make them close to normal

Model Building

Prerequisites:

Feature Matrix

To build the model for prediction and formulating clusters, we formulate the feature matrix X, by taking the CustomerID feature and the log transformed Recency, Frequency and Monetary columns..

Feature Scaling

Scaling data is the process of increasing or decreasing the magnitude according to a fixed ratio, in simpler words you change the size but not the shape of the data.

Various methods of feature scaling:

1. Standardization

It calculates the z-score of each value and replaces the value with the calculated Z-score. The Z-score can be calculated by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where σ is the variance and $\bar{x}$ is the mean.The features are then rescaled with $\bar{x}$ =0 and σ=1.

Library used: StandardScalar

2. Min-Max Scaling:

It is also referred to as Normalization. The features are scaled between 0 and 1. Here, the mean value remains the same as in Standardization, that is, 0. The formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Library used: StandardScalar

3. Normalizing

It is used to rescale each sample. Each sample (i.e. each row of the data matrix) with at least one non zero component is rescaled independently of other samples so that its norm (l1 or l2) equals one.

Library used: Normalizer

For this project, we utilized Standard Scaling.

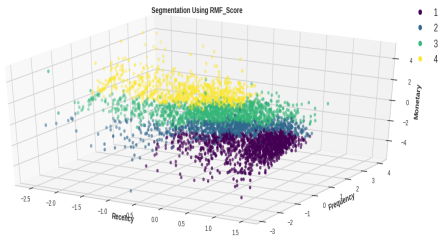
Model Selection:

To come up with customer segments, we have utilized two approaches; Heuristic approach and Machine Learning Approach.

Heuristic approach:

From the RFM analysis we carried out earlier, we have derived RFM scoring for each customer, we can simply bin these scores into major clusters based on their quantiles.

Segmentation Using RMF_Score



Here, we use the RFM_Score for each customer and divide it into 4 major segments by cutting the distribution based on its quantiles.

This method does not involve any ML computation and is easy and faster to implement.

Machine Learning approach:

We have implemented K-Means clustering, Agglomerative Hierarchical Clustering and DBSCAN for the scope of this project.

1. K-Means Clustering

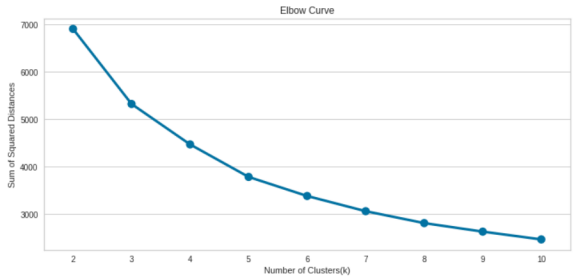
K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

The dataset is divided into "k" pre-defined, non-overlapping subgroups (clusters) by an iterative method, with each data point belonging to a single group.

To find the number of clusters, we use the Elbow Curve Method, Silhouette score Method.

Elbow Curve

It entails repeatedly iterating the method over a loop with increasing cluster options before plotting a score as a function of the number of clusters. The centroids are nearer cluster centroids as "k" rises. The entire reason this strategy is called the elbow method is because at some point the progress will quickly fall, forming an elbow-like curve in the graph. We count the cluster and calculate the k-value at the bend in the elbow.

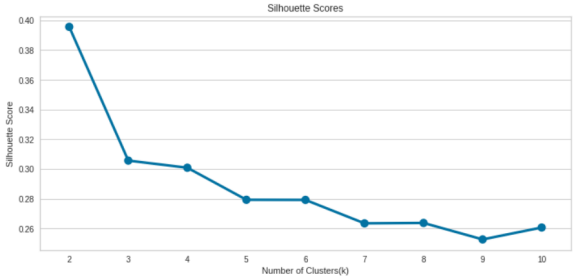


We have achieved the above plot from the analysis.

Silhouette scores

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.

We picked up a range of the k values and drew the silhouette graph by calculating the silhouette coefficient of every point.

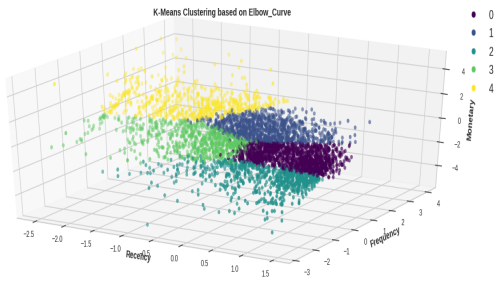


We have achieved the above plot from the analysis.

Optimal Number of clusters:

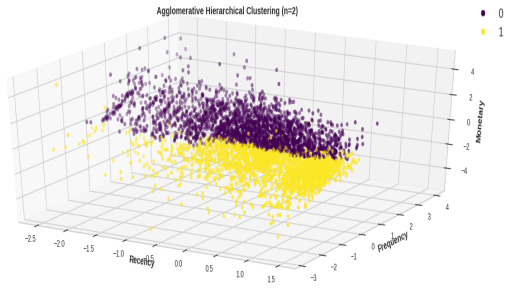
- We have considered the elbow curve and to start with 2 as the minimum number of clusters and calculated their respective silhouette scores.
- 5 appears to be the elbow, looking at the elbow curve.
- Silhouette score for 2 clusters is the highest
- Considering the tradeoff between the two plots we can also consider 4 to be the optimal number of clusters.

K-Means clustering based on Elbow Curve Results



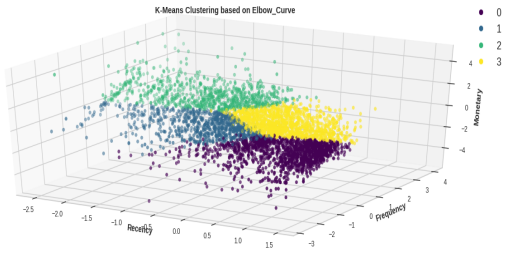
The result of the elbow curve gave us the optimal number of clusters as 5. We initiate the K-Means algorithm giving it the number of clusters n=5 and get the results as seen in the figure above.

K-Means clustering based on Silhouette Score Results



The Silhouette score is highest for n=2. We ran K-Means with giving it the number of clusters as 2 and get the above shown results.

K-Means clustering based on Tradeoff between Elbow Curve and Silhouette Score Results



From observing the results obtained from the elbow and silhouette analysis, we then proceed to consider the value of n to be something in between both the analysis and conclude it to be 4 as the optimal number of clusters. We can see the results as shown in the figure above.

There are certain difficulties with the K-means clustering algorithm, which are:

- It always attempts to produce clusters of the same size
- It has a predetermined number of clusters.

Because we don't need to be aware of the specified number of clusters while using the hierarchical clustering technique, we can choose this algorithm to address these two problems.

3. Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning algorithm, which is used

to group the unlabeled datasets into a cluster. 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. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

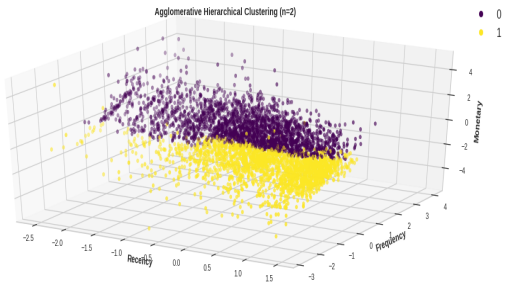
The hierarchical clustering technique has two approaches:

- 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.
- Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

For the scope of this assignment, we chose the agglomerative approach.

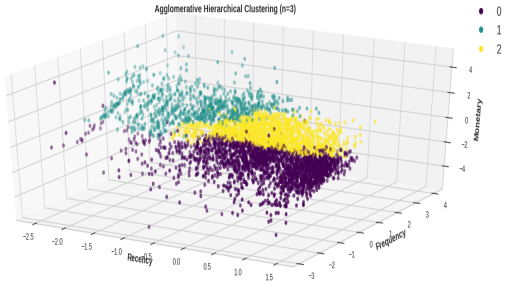
After plotting the dendrogram we have a task in hand to decide the threshold value, such that it cuts the tallest vertical line in the plot.

Agglomerative Hierarchical Clustering for Dendrogram threshold y=70



We first set the threshold value to be 70. This cuts the tallest line and yields us with two major clusters. We then perform Agglomerative Hierarchical Clustering with n=2 on the data. We can see the result in the figure above.

Agglomerative Hierarchical Clustering for Dendrogram threshold y=45



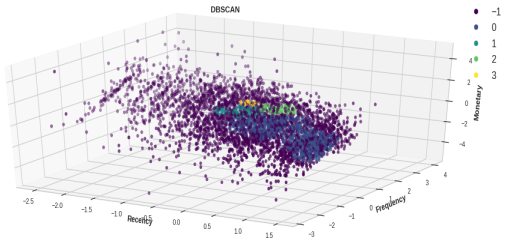
Aiming for a lower threshold value, we went with 45 and obtained three major segments. Again, performing Agglomerative Hierarchical Clustering with n=2 on the data, we get the results as seen in the figure.

K-means and Hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words, they are suitable only for compact and well-separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data. To overcome these difficulties, we proceed to move on with the DBSCAN algorithm to obtain the optimal number of clusters.

3. DBSCAN

Density-Based Spatial Clustering Of Applications with Noise (DBSCAN) does not require an initial specification of the number of clusters to be obtained. 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.



We ran DBSCAN on our data and got 5 major segments from the model.

Conclusion

Initiating the task, we found that our raw data contains about 25% missing values and we also found some duplicate values. We performed data cleaning to get rid of these values and also canceled orders. We derived important business insights based on products, time and location.

We also performed feature engineering. This included deriving new date and time features and getting a new feature giving the total amount of transaction using price and quantity of order.

We then performed RFM analysis of the transaction data. This helped us gain important metrics to build models for customer segmentation.

Model building included Segmentation Using RFM Scores(Heuristic Model) which gave us 4 major segments.

Further we used the Machine Learning Models giving us the results as follows:

- K-Means Clustering
    - Elbow Method - 5 Clusters
    - Silhouette Score - 2 Clusters
    - Elbow Curve & Silhouette Score - 4 Clusters
- Agglomerative Hierarchical Clustering
    - Dendrogram Threshold=70 - 2 Clusters
    - Dendrogram Threshold=45 - 3 Clusters
- DBSCAN - 5 Clusters

Based on the company's goals, we can use these models and their obtained clusters to build business strategies, marketing campaigns, etc. to target, incentivise and attract customer base.