Online Retail Customer Segmentation

Sumit Kumar Dhibar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketingsuch as gender, age, interests, and miscellaneous spending habits.

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. In this machine learning project, we will make use of K-Means clustering which is the essential algorithm for clustering unlabelled dataset.

1. Problem Statement

In this project, our 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.

- ***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 eachtransaction 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.

2. Introduction

Companies that are under the notion that every customer has different requirements and require a specific marketing effort to address them

appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. So in our project our aim is to divide the customers in clusters so that wecan have a look for clear customer's spendingtrends.

- **Product Design**:- Clustering helps a company to understand what a users need. A company can identify the most active users/customers, and optimize their application/offers/products accordingto their needs.
- **Promotion**:- Properly implemented clustering helps a company to plan special offers and deals. Frequent deals have become a staple of e- commerce and commercial softwarein the past few years. If they reach a customer with just the right offer, atthe right time, there's a huge chancethey're going to buy. Customer segmentation will help them tailor tailor special offers perfectly for particular class of customers.
- **Budgeting**:- Nobody likes to invest in campaigns that don't generate anynew customers. Most companies don't have huge marketing budgets, so that money has to be spent right. Clustering helps them a lot.
- **Marketing**:- The marketing strategycan be directly improved with segmentation because company canplan personalize the marketing campaigns for different customer segments.

3. Steps involved:

- **Exploratory Data Analysis** :- Before applying our model it is must for us to have a look at our data. Because there is no point of applying a ML model which takes a plenty of time and effort before knowing what we are dealing with. So we explore our dataset to have a look that what are the different information and patterns in our dataset.
- **Null values Treatment** :- Our dataset contains some null values like customer_id, and Description. So we dropped the rows of null values at the beginning of our project in order to get a better result.
- **Feature Engineering** :- Created new features from the existing features for better understanding of the dataset.
- **Standardization of features:-** Create the RFM model (Recency, Frequency , Monetary value) for clustering made easy.

- **Fitting different models**

  For modelling we tried various clustering algorithms like:

1. **K-means with Elbow method**
2. **K-means with Silhouette score**
3. **Hierarchical _clustering**
4. **DBSCAN**

4. Algorithms:

4.1 K-Means Clustering

Segmentation is achieved through the acquisition of products in right quantity and in real time – for the right customer at optimum cost. To meet these requirements, the k-means clustering technique can be applied to segment the customers for efficient forecasting and planning decisions. K-means clustering is an iterative crisp clustering technique which aims at partitioning n data points into k disjoint clusters. The objective of k-means is to generate k cluster centers that minimize the sum of squared distances between each point and its nearest cluster center. The k-means algorithm is summarized as follows :

- select randomized k initial centroids (k denotes the initial number of clusters);
- assign every point in the data to the closest centroid (every group of data points linked to a centroid form a cluster).
- update the centroid of each cluster based on the new cluster, and repeat this process until no further change in cluster point is observed;
- repeat the steps above until a stopping criterion is met (no further point changes in cluster assignment is adopted as the stopping criterion in thispaper)

- **Elbow Method** :-The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.
- **Silhouette Score** :-Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plotdisplays a measure of how close eachpoint in one cluster is to points in theneighboring clusters and thus provides a way to assess

parameters like number of clusters visually.

4.2 Hierarchical clustering

The hierarchical clustering is a commonly used text clustering method, which can generate hierarchical nested classes. It clusters similar instances in a group by usingsimilarities of them . This requires the use of a similarity (distance) measure which is generally Euclidean measure in general, and cosine similarity for documents. Therefore, a similarity (distance) matrix of instances has to be created before running the method. Wehave used this method to form clusters of ourdataset. Finding the Number of Clusters:

For applying the KMeans algorithm to our dataset we have to decide the number of clusters before applying the ML models. And To know the optimum number of clusters we use elbow method and Silhouette method.

There are mainly two types of hierarchical clustering:

- **Agglomerative hierarchical clustering**
- **Divisive Hierarchical clustering**

The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as *AGNES* (*Agglomerative Nesting*). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named *dendrogram*.

The divisive clustering algorithm is **a top-down clustering approach**, initially, all the points in the dataset belong to one cluster and split is performed recursively as one moves down the hierarchy.

4.3 DBSCAN:

DBSCAN stands for **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise. Clustering is a way to group a set of data points in a way that similar data points are grouped together. Therefore, clustering algorithms look for similarities or dissimilarities among data points. Clustering is an unsupervised learning method so there is no label associated with data points. The algorithm tries to find the underlying structure of the data.

5. Conclusion:

In this project, the customer segments thus deduced can be very useful in targeted marketing, scouting for new customers and ultimately revenue

growth. After knowing the types of customers, it depends upon the retailer policy whether to chase the high value customers and offer them better service and discounts or try and encourage low/medium value customers to shop more frequently or of higher monetary values.

However, there can be more modifications on this analysis. One may choose to cluster into more number of depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefer to buy often, segmenting on the basis of time period they visit and much more.

As machine learning has become more of an ART, there is nothing such as right or wrong. We only try to get the best outcomes that can suit our final objectives. There is, and always will be, a need to improve, going forward.

References-

1. Towards datascience
2. GeeksforGeeks
3. Analytics Vidhya