Online Retail Customer Segmentation

Sonica Sinha, Mohd Taufique
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Customer segmentation is the method of distributing a customer base into collections of people based on mutual characteristics so organizations can market to group efficiently and competently individually. And the present dataset is such kind of dataset which helps us in analyzing the reason why customers cancelled the product and what is the general behavior or tendency of a customer while placing an order through transactions. For that, dividing the categories of customers into groups will be very important. The present dataset of UK based registered online retail non-store, that really helps us in understanding the way via which the customers are doing transactions and on which product more and also why they canceled the product. And hence, understanding will help the company in expanding its business further without facing any loss or by minimizing it.

Our experiment can help in creating the segments of the different groups of customers into different clusters from the given transnational data of the products. And this has been performed by development of Machine Learning algorithms using Unsupervised Classification.

Keywords: machine learning, segmentation, unsupervised classification, clustering.

1. Introduction

In a world where businesses are growing tremendously, and cater to a large number of customers on a regular basis. It becomes very essential for businesses to categorize their customers, this would not only lead to better customer service but also would help businesses understand how different customers can impact their business. And hence, doing the customer segmentation can help in finding similarities in behavior, characteristics, needs and habits of different customers. This further helps businesses to customize its strategies and marketing way to lessen their product cancellation and also improve their user experience with wide varieties of product to choose from.

Our goal here is to build a cluster of segments using different models by classification which can categorize the customers according to their transactional behavior and the impact which can occur over the business.

2.Problem Statement

The challenge of the problem 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.

This is an unsupervised ML classification problem.

Following columns has been provided with the dataset:

- **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.
- **Unit Price:** 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.

The CustomerID column is one of the important features which includes unique id for each customer, helps in identifying particular during analysis.

The InvoiceNo column is another feature which includes the information on the cancellation.

3. Steps involved:

- **<u>Data Cleaning & Null Value Treatment</u>**

  After loading the dataset, head, tail, data type, shape were explored. Further , the dataset is checked for its null value and **CustomerID** columns contain the null value of 24.93%, where we have a large percentage of cancelled orders of 35%. Studying these cancelled orders may help in preventing future cancellation. This made it compulsory to do treatment of both the features by dropping them. As we dont have much idea about the missing customer IDs person, so it's better to drop them. Also few duplicate values were also removed from the dataset. Hence, after dropping the null values our dataset is ready for the beginning of the project inorder to get a better result.

- **<u>Exploratory Data Analysis</u>**

  EDA was performed in the following steps:

  - **CustomerID:** here we checked the top 5 customer IDs present in the dataset, who shopped more and showed maximum transaction.
  - **Country:** All together there were 37 different unique countries. Showing the distribution of transactions from around the world. As per order, UK showed the highest position among the top 10 countries.
  - **Descriptiont:** here we tried to find out the top 5 and

bottom 5 product names on which customers showed maximum and least interest. This also helped us to analyze the interest of customers towards considering costly products for gifting purposes.

- **StockCode:** Topl 5 and bottom 5 stock codes were checked which were directly associated with the products availability.

And henceforth, different kinds of plots were plotted like count plot, bar plot, heatmap, etc.

- **<u>Feature Engineering</u>**

  In feature engineering we created some new features from the already present features. We added features with InvoiceDate for e.g.,Year, Month, Day, Hour, Month_Num and Day_Num. Total Amount also calculated using as product between unit price and quantity. And further proceed in the following manner.

  - Firstly we dropped all the canceled items from the Invoice No. feature which had letter **"C"** in them.
  - Checked the distribution of all the numerical features from the dataset including the newly added features.
  - We do applied normalization on the features like, quantity, unit price and total amount for its proper distribution.
  - Further we explored newly added features- **Year:** year in which maximum purchases were made.
  - **Month:** Top 5 month names were extracted. In which month of November showed the maximum trend of transactions for purchase.
  - **Day:** Top 5 days' names were extracted.
  - **Hour:** Most purchased hours were extracted.
  - Duration of maximum purchase for a particular day was also explored.

4. Regency, Frequency & Monetary (RFM) value:

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:** Customers who made purchases recently.
- **Frequency:** How often a customer makes a purchase.
- **Monetary Value:** How much money a customer spends on.

For performing the RFM segmentation we used following steps:

- The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer.

- The second step is to divide the customer list into tiered groups for each of the three dimensions (R, F and M).

5. Quantile Split:

In the next step after segmenting the dataset into the RFM model, we proceeded for the quantile split by taking the ratio as 0.25, 0.5, 0.75, respectively.

This process helped in generating the RFM group and RFM score values, where RFM group: it is the combination of R,F,M values and RFM score: it is the sum of R,F,M values assigned during the quantile split.

Before proceeding towards the model development or the cluster development via models, we performed the pre-processing on the values of R, F and M using the math log for its normalization.

6. Model Training:

After quantile splitting the variables and normalizing it using math log, different unsupervised classification machine learning models were implemented for doing segmentation of customers on the basis of transactions made by them for a period of one year.

The used model for clustering are:

1. K-Means Clustering
2. DBSCAN
3. Hierarchical Clustering

- **<u>Scaling of the features</u>**

  Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

  Here we have used Standard Scaling which is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

7. Algorithms Used:

A. K-Means Clustering:

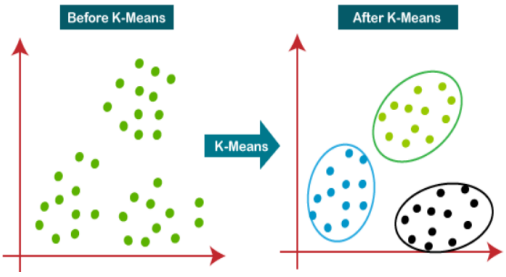
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here, K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.

- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.



There are many ways to select the best number of clusters for the K-mean algorithm. In this project we applied two methods called the Elbow method and Silhouette Analysis method.

B. DBSCAN :

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise) is the most productive kind of unsupervised machine learning algorithms. These algorithms are based on the 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.

DBSCAN requires two parameters to work upon, i.e.,

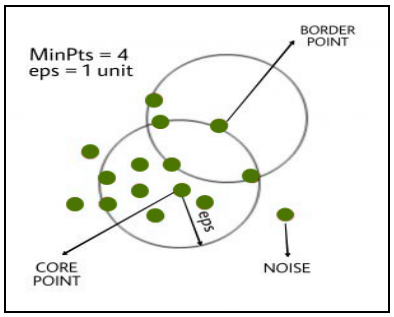
- Eps: It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered neighbors.
- MinPts: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. The minimum value of MinPts must be chosen at least 3.

In this algorithm, we have 3 types of data points:

➔ **Core Point:** A point is a core point if it has more than MinPts points within eps.

➔ **Border Point:** A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

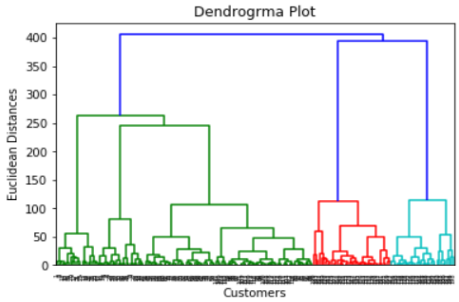
➔ **Noise or outlier:** A point which is not a core point or border point.



C. Hierarchical Clustering :

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA. 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**.

It works on two approaches: **agglomerative** (bottom-up approach) and **divisive** (top-down approach).



To measure the similarity & dissimilarity in clustering using hierarchical clustering we have used Ward's minimum variance method: it reduces the overall within-cluster variation to the lowest level possible. At each phase, the clusters with the shortest distance between them are merged.

D. Model performance:

Clustering algorithm can be evaluated by two approaches as:

1. **Silhouette Analysis Method-**

   This method calculates the average silhouette value for each data point in the cluster, this value represents how similar a data point is to its own cluster. The range of this measure from -1 to 1. A value of 1 means the sample is far away from the neighboring clusters. The negative value refers to samples that might have been assigned to the wrong cluster.

2. **Elbow Method-**

   Elbow is the common method used to determine the best value of K. This method calculates the variance between data points within a cluster using the Sum of Squared Error. The best value of k to select is the point of inflection on the curve.

   For K-means, DBSCAN and Hierarchical clustering all these scoring methods were used for analyzing the result of models. Dendrogram is also used for hierarchical clustering.

8. Conclusion:

That's it! We reached the end of our exercise.

Starting from importing of libraries, dataset, treating of null values, EDA of important features to get the insight hidden inside the dataset, reaching towards the quantile split to preparation of Regency, Frequency & Monetary model which is an important part for the customer segmentation analysis. For the algorithm development K-means, DBSCAN, Hierarchical clustering method were used. The best silhouette score came up with 2 clusters and DBSCAN resulted in 3, for RFM. And hence, the business can focus on these different clusters and provide to customers of each sector in a different

way, which would not only benefit the customer but also the business at large.

9. References:

1. AnalyticalVidhya
2. Geeksforgeeks
3. Medium Blog
4. Javatpoint
5. AlmaBetter Resources