# ONLINE RETAIL CUSTOMER SEGMENTATION

**KOTA LAKSHMANA RAO**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

### Practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits

### Allows us to better understand our customers helping us target these customers in a more efficient manner and improve the customer experience

***Keywords :unsupervised machine learning, customer segmentation***

**1.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.**

.

**2.Data Description:**(**Attribute Information):**

* **Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.**
* **Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct**
* **Produce Description: Product (item) name. Nominal.**
* **Quantity: The quantities of each product (item) per transaction. Numeric.**
* **Invoice Date: 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.**
* **Customer ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.**
* **Country: Country name. Nominal, the nameofthecountry where each customer resides.**

**3. Introduction**

### **The online retail customer segmentation ,the main task identify major customer segmentation on transnational data This algorithm automatically segregate the customers**

### **The machine learning algorithm general****ly segregate different types of customers based on the transactional data**

## **4. Reasons for less amount of selling materials**

**The reasons for less selling of material:**

* **Less amount of quantity**
* **Less amount of its price**

# **5. Reasons for more amount of selling**

**There are times when so many people are requesting ordering bulk amount of material**

**This causes shows more amount selling**

**6. Steps involved:**

* **Exploratory Data Analysis**

**After loading the dataset we performed . This process helped us figuring out various aspects and relationships among independent/ feature variables. It gave us a better idea**

* **Null values Treatment**

**Our dataset contains some null values like customer\_id, AND Description which may tend to our project so will come good and at the beginning of our project in order to get a better result**.

* **Encoding of categorical columns**

**We are not used Label Encoding because no need to unsupervised learning**

* **Feature Selection**

**In these steps which more important to produce clustering data , that will taken**

* **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 silhouette\_score**
2. **K-means with Elbow method**
3. **DBSCAN**
4. **Hierarchical \_clustering**

**7.1. Algorithams:**

1. **K-means with silhouette\_score:**

### **The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:**

### **The "cluster center" is the arithmetic mean of all the points belonging to the cluster.**

### **Each point is closer to its own cluster center than to other cluster centers.**

### **Those two assumptions are the basis of the k-means model.**

### **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. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:**

### **Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.**

### **Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.**

### **The Silhouette Coefficient for a sample is S=(b−a)/max(a,b)**

.

**2.K-means with Elbow method:**

### **The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:**

### **The "cluster center" is the arithmetic mean of all the points belonging to the cluster.**

### **Each point is closer to its own cluster center than to other cluster centers.**

### **Those two assumptions are the basis of the k-means model.**

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.

**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.

**4.Hierarchical \_clustering:**

### **Hierarchical clustering takes away the problem of having to pre-define the number of clusters. So, let’s see what hierarchical clustering is and how it improves on K-means.**

### **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

**7.2. Model performance:**

Model can be evaluated by various metrics such as:

1**. K-means clustering with silhoette :**

**In k-means clustering with silhoette clusters of**

**RM DATA MODEL IS -2**

**FM DATA MODEL IS -2**

**RFM DATA MODEL IS -2**

**2. k-means clustering with Elbow method**:

**In K-means clustering with Elbow method clusters**

**RM DATA MODEL IS -2**

**FM DATA MODEL IS -2**

**RFM DATA MODEL IS -2**

. **3. DBSCAN**:

**In DBSCAN clustering the number of clusters**

**RM DATA MODEL IS -2**

**FM DATA MODEL IS -2**

**RFM DATA MODEL IS -3**

**4.Hierarchical clustering:**

**In Hierarchical clustering the number of clusters**

**RM DATA MODEL IS -2**

**FM DATA MODEL IS -2**

**RFM DATA MODEL IS-2**

**8. Conclusion:**

**Thats it! We have come to an end of this long exercise. Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.**

**Next, we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented KMeans clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster**.

**However, there can be more modifications on this analysis. One may choose to cluster into more no. 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, finding out customer lifetime value (clv), 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.**

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