**Online Retail Customer Segmentation**

**Kiran Ahire**

**Data Science Trainee,**

**AlmaBetter , Bangalore**

**Abstract:**

Segmentation is the process of dividing your customers up into different groups, with each group sharing similar characteristics, to improve engagement, sales and loyalty.

**Need of Customer Segmentation –**

* Helpsto detect and exploit new market opportunities
* Improves how to predict customer behaviour
* Increased customer retention and policy
* Helps to improve customer lifetime value
* Streamlines and improves workflow

### The provided dataset contains information about Invoiceno, Stockcode , Description, Quantity , InvoiceDate , Unitprice , CustomerId and Country . In this project our goal is to identify major customer segments on transnational dataset.

**1.Problem Statement**

### We have given a dataset with the about Invoiceno, Stockcode , Description, Qauntity , InvoiceDate , Unitprice , CustomerId and Country . In this project our goal is to identify major customer segments on transnational dataset. 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.

We have given the following information in our 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.

* **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**

We have given a dataset with the about Invoiceno, Stockcode , Description, Qauntity , InvoiceDate , Unitprice , CustomerId and Country . In this project our goal is to identify major customer segments on transnational dataset. 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.

### Our dataset has 541909 rows and 8 columns. The goal of the project is to identify major customer segments on transnational dataset beforehand which could us to get basic understanding of our dataset, finding outliers, finding correlation between the data, perform Data visualization, models implementation and finally getting conclusion.

## **3. Steps Involved**

* **Data Summary**

After loading the dataset we performed this step **to understand the data better** in the dataset and creating various graphical representations to get some conclusion from it.

This step is an important method for getting a grasp on the meaning of the content of a large collection of data. It basically simplify data for better understanding of our dataset.

* **Data Preprocessing**

First we check for missing values our dataset has large number of missing values so we dropped those rows then check for duplicated data , We drop some InvoiceNo which starts with C because it indicates cancellation.

* **Identify relationships in dataset**

**Correlation** **matrix** measures the statistical relationship between two different variables. Correlation evaluates the direction as well as strength of a relationship between continuous variables. Correlation coefficient can **range from -1 to +1**, which signifies strong negative to strong positive relation between the variables

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is default payment next month with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable. With the help of Exploratory data analysis we analyzed the categorical as well as numerical features in the dataset.

* **Feature Engineering**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

In our project we have created new feature ‘Day’ from InvoiceDate and ‘TotalAmount’ from product of Quantity and UnitPrice.

* **RFM Model**

**RFM model which stands for Recency, Frequency, and Monetary is one of such steps in which we determine the recency - days to last visit, frequency - how actively the customer repurchases and monetary - total expenditure of the customer, for each customer. RFM analysis allows marketers to target specific clusters of customers with communications that are much more relevant for their particular behaviour – and thus generate much higher rates of response, plus increased loyalty and customer lifetime value.**

* **Segmentation using K means Clustering**

K means is an unsupervised machine leaning algorithm that divides given data into the given number of clusters. Here, the “K” is the given number of predefined clusters, that need to be created. It is a centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid. The algorithm takes raw unlabelled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.

### **Calculation of Silhouette score**

**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.**

### **Applying elbow method**

Elbow is one of the most famous methods by which you can select the right value of k

and boost your model performance. We also perform the hyperparameter tuning to

chose the best value of k. Let us see how this elbow method works. It is an empirical

method to find out the best value of k. it picks up the range of values and takes the best

among them. It calculates the sum of the square of the points and calculates the

average distance*.*

* **Conclusion**

**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. Then, 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.**