Customer Segmentation

Vridhi Parmar, Hrithik Chourasia
Data science trainees,
AlmaBetter, Bangalore

Instructions:

i) Please fill all the required information.
ii) Avoid grammatical errors.

Abstract:

For a medium to large size retail store, it is imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their 'best' or high-valued customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, companies use the technique of customer segmentation.

It is said that 20% of customers contribute 80% share of the total revenue of a company. That's why finding this set of people is important.

Problem Description

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.

Data Description:

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

Introduction:

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.

Types of segmentation:

1. **Zero segments:** This means there is no differentiated strategy and all of the customer base is being reached out by a single mass marketing campaign.
2. **One segment:** This means that the company is targeting a particular group or niche of customers in a tightly defined market.
3. **Two or more segments:** This means that the company is targeting 2 or more groups within its customer base and is making specific marketing strategies for each segment.
4. **Thousands of segments:** This means that the company is treating each customer as unique.

Factors for segmentation for a business:

1. Demographic: Age, Gender, Education, Ethnicity, Income, Employment, hobbies, etc.
2. Recency, Frequency, and Monetary: Time period of the last transaction, the frequency with which the customer transacts, and the total monetary value of trade.
3. Behavioral: Previous purchasing behavior, brand preferences, life events, etc.
4. Psychographic: Beliefs, personality, lifestyle, personal interest, motivation, priorities, etc.
5. Geographical: Country, zip code, climatic conditions, urban/rural areal differentiation, accessibility to markets, etc.

Steps Involved:

Exploratory Data Analysis:

It refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Null Value Treatments:

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

Standardization of features:

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

RFM Segmentation:

RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).

Fitting the Model:

Here we are using K- Means Clustering Our dataset is large so Hierarchical clustering is not well suited for analysis. So, we will use the K-Means clustering algorithm, which is easy to apply fast, and accurate for clustering problems.

K-means clustering requires parameters as per to group the parameters.

Methods to find the number of groups here are:

1. Elbow method:

   In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.

2. Silhouette Analysis:

   Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Conclusion:

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

Starting with loading the data so far we have done EDA , null values treatment, feature selection and then model building.

In our model our optimum value of k was 4 because at 4 we saw the error is decreasing as K increases. For values of k at 4 or 5 slope of the curve is decreasing very fast this means errors do not decrease much faster as the increase in number of clusters.

Cluster	RFM Interpretation	Type of Customer
0	Least purchase long ago Least no of transaction Least Monetary Spending	Churned
1	Recent Transaction Most Frequent Transaction Highest Monetary Spending	Targeted Customer
2	Recent Transaction Low Frequent Transaction Low Monetary Spending	New
3	Last Purchase While ago Low Frequent Transaction Low Monetary Spending	At Risk

Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.
Github_link:- https://github.com/Vridhip/Online_Retail_Customer_Segmentation Google_drive_link:- https://drive.google.com/drive/folders/1obS643tMnRLLSaRy4OfOmWORdCotWd4Y?usp=sharing