Capstone Project

Online Retail Customer Segmentation

(ML - Unsupervised)

(MANISH GUPTA)
Data science
Cohort- Warshaw Alma Better

ABSTRACT

The customer base is usually quite small and individually targetable. But, as a business grows in size, it will not be possible for the business to have an intuition about each and every customer. At such a stage, human judgments about which customers to pursue will not work and the business will have to use a data-driven approach to build a proper strategy.

For a medium to large size retail store, it is also 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.

PROBLEM STATEMENT

Customer segmentation has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers. For example, a company can make a high-end product, a budget product, or a cheap alternative product, depending upon whether the product is intended for its most high yield customers, frequent purchasers or for the low-value customer segment. It may also fine-tune the features of the product for fulfilling the specific needs of its customers.

Customer segmentation can also help a company to understand how its customers are alike, what is important to them, and what is not. Often such information can be used to develop personalized relevant content for different customer bases. Many studies have found that customers appreciate such individual attention and are more likely to respond and buy the product. They also come to respect the brand and feel connected with it. This is likely to give the company a big advantage over its competitors. In a world where everyone has hundreds of emails, push notifications, messages, and ads dropping into their content stream, no one has time for irrelevant content.

Finally, this technique can also be used by companies to test the pricing of their different products, improve customer service, and upsell and cross-sell other products or services.

ATTRIBUTE INFORMATION

- **InvoiceNo**
  Invoice number. Nominal, digit integral number uniquely assigned to each transact ion. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**
  Product (item) code. Nominal 5digit integral number uniquely assigned to each distinct product.
- **Description**
  Product (item) name

- **Quantity**
  The quantities of each product (item) per t ransaction.
- **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 5digit integral number uniquely assigned to each customer.
- **Country**
  Country name. Nominal, the name of the c ountry 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. Insights from customer segmentation are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning. A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the company philosophy and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy.

- **Zero segments**
  This means that the company is treating all of its customers in a similar manner. In other words, there is no differentiated strategy and all of the customer base is being reached out by a single mass marketing campaign.

- **One segment**
  This means that the company is targeting a particular group or niche of customers in a tightly defined target market.

- **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.

- **Thousands of segments**
  This means that the company is treating each customer as unique and is making a customized offer for each one of them.

Once the company has identified its customer base and the number of segments it aims to focus upon, it needs to decide the factors on whose basis it will decide to segment its customers.*F*actors for segmentation for a business to consumer marketing company:

- **Demographic**
  Age,Gender,Education, Ethnicity, Income, Employment, hobbies, etc.

- **Recency, Frequency and Monetary**
  Time period of the last transaction, the frequency with which the customer transacts, and the total monetary value of trade.

- **Behavioral**
  Previous purchasing behavior, brand preferences, life events, etc.

- **Psychographic**

Beliefs,personality,lifestyle,personal interest, motivation, priorities, etc.

- **Geographical**
  Country, zip code, climatic conditions, urban/rural areal differentiation, accessibility to markets, etc.

STEPS INVOLVED

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.
4. Removing future warnings in seaborne plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Checking the Unique values in the column (if any)
8. Converting the data types to similar objects as the Analysis Demands.
9. Formatting the "size" column into a single column in the dataset.
10. Eradicating special characters from the dataset columns.

- **EXPLORATORY DATA ANALYSIS**

  Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it..

- **EXAMINING NULL VALUES**

  The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analysing the first thing which we will do is to examine the null or missing values on the Dataset. It is the first step that will make the results "more" accurate &should be handled before it affects the performance of the models that predict the outcome.

- **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).

  RFM analysis helps to answer the following questions: Who was our most recent customer? How many times has he purchased items from our shop? And what is the total value of his trade? All this information can be critical to understanding how good or bad a customer is to the company.

  After getting the RFM values, a common practice is to create 'quartiles' on each of the metrics and assigning the required order. For example, suppose that we divide each metric into 4 cuts. For the recency metric, the highest value, 4, will be assigned to the customers with the least

recency value (since they are the most recent customers). For the frequency and monetary metric, the highest value, 4, will be assigned to the customers with the Top 25% frequency and monetary values, respectively. After dividing the metrics into quartiles, we can collate the metrics into a single column (like a string of characters {*like '213'*}) to create classes of RFM values for our customers. We can divide the RFM metrics into lesser or more cuts depending on our requirements.

- **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.

ALGORITHM

K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

What is K-Means Algorithm?

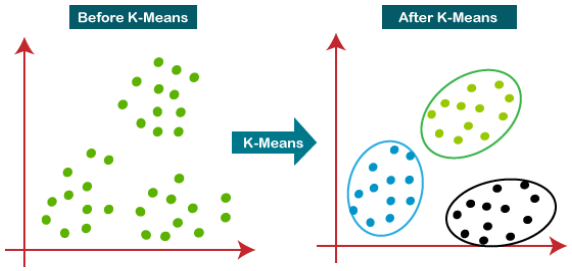
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled 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 algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

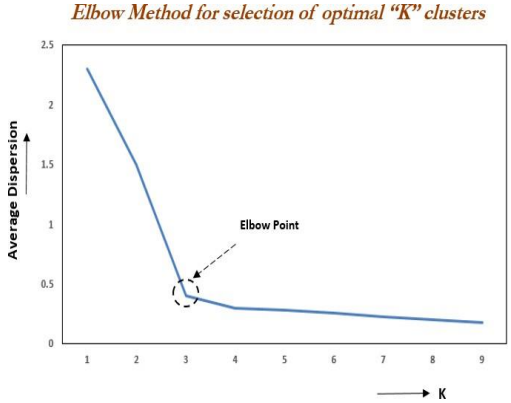
The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to the particular k-center, create a cluster.Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



Elbow method to find optimum k value:



Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

CONCLUSION

Customer segmentation is highly effective strategy for organizations because it allows them to know which customers care about them and understand their needs enough to send a message that ensures brand success.

We used RFM Modeling to see the relation betwee n Recency,Frequncy and Monetary.After RFM mo del we used this data to perform clustering with the help of k mean clustering Algorithm.

At the end we make 4 clusters of customer's name d as:

Cluster 0

- New Customer = Customer who transacte d recently and have lower purchase freque ncy, with low amount of monetary spendi ng.

Cluster 1

- Lost Customers = Customers with the leas t Monetary spending and the least number of transaction.

Cluster 2

- Best Customwers = Most frequent spender s with the highest monetory spending amo unt and had transacted recently.

Cluster 3

- At Risk Customers = Customers who made their last transaction a while ago and made less frequent and low monetary purchases.

References:

- **Python Pandas Documentation:** https://pandas.pydata.org/pandas-docs/stable
- **Python Numpy Documentation:** https://numpy.org/doc/
- **Python MatPlotLib Documentation:** https://matplotlib.org/stable/index.html
- **SciKit Documentation:** https://scikit-learn.org/stable/
- **Towards Data Science:** https://towardsdatascience.com