Customer Segmentation

Aamir Sohail,

Almabetter, Bangalore

Introduction

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data companies can then out perform the competition by developing uniquely appealing products and services.

- **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**.

The most common ways in which businesses segment their customer base are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
2. **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence.
3. **Psychographics**, such as social class, lifestyle, and personality traits.
4. **Behavioral data**, such as spending and consumption habits, product/service usage, and desired benefits.

Advantages of Customer Segmentation

1. Determine appropriate product pricing.
2. Develop customized marketing campaigns.
3. Design an optimal distribution strategy.
4. Choose specific product features for deployment.
5. Prioritize new product development efforts.

Our Goal

Given a dataset related to a online retailer based out of the UK, we need to analyse and identify major customer segments using K Means algorithm and also using different verification method to confirm the result.

Dataset

now let me introduce our data set, a transnational data set with transactions occurring **between 1st December 2010 and 9th December 2011** for a UK-based online retailer. The company **mainly sells unique all-occasion gifts** and many customers of the company are **wholesalers.**

The dataset contains features like:

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 product.

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 name of the country where each customer resides.

- **Null values Treatment**

  Our dataset contains around 135080 null values in CustomerID which might tend to disturb our mean absolute score hence I removed the null values of CustomerID and have performed KNN imputer for numerical features and replaced categorical features d Description Column. After that our Dataset reduced to (406829,8).

Feature engineering

After that I did Feature engineering so it is the process of tansforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved accuracy on unseen data.

And i converted Invoice Date column into date time format and created new features like day, day_num, month_num, year, hours etc. From Invoice Date and again created Total Amount from Product of Quantity and Unit price columns. This can help improve machine learning accuracy since algorithms tend to have a hard time dealing with high cardinality columns.

Exploratory Data Analysis

After loading the dataset I looked for duplicate values in column. There were none. So I performed EDA by taking many features.This process helped me figuring out various aspects and relationships. It gaves a better idea of which feature behaves in which manner.

Create the RFM model (Recency, Frequency,Monetary 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: . Frequency: How often a customer makes a purchase. Monetary Value: How much money a customer spends on

Performing RFM Segmentation and RFM Analysis, Step by Step

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).

Calculating RFM scores

The number is typically 3 or 5. If we decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 333 to a low of 111. Generally speaking, the higher the RFM score, the more valuable the customer.

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. 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 $S=(b-a)/\max(a, b)$

The Elbow Method

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

The steps can be summarized in the below steps:

1. Computed K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).
3. I Plot the curve of WCSS vs the number of clusters K.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

The optimal K value is found to be 3 using the elbow method.

Finally I made a plot to visualize the spending score of the customers with their income. The data points are separated into classes which are represented in different colours as shown in the plot.

Dendrogram

Hierarchical clustering is 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, that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset. So what cluster that data point belonged to and the sequence of merges made for those clusters. So in summary, we see that the dendrogram is able to capture all the critical elements of the hierarchical clustering result.

DBSCAN

Density-based spatial clustering of applications with noise (**DBSCAN**) is a data clustering algorithm. It is a density-based clustering non-parametric algorithm-- given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbours), marking as outliers points that lie alone in low-density regions (whose nearest neighbours are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

The Challenge

so owing a supermarket mall and through membership cards, we have some basic data about our customers like Customer ID, age, gender, annual income and spending score. we wanted to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly and that was our challenge but I did it with k-means clustering algorithm perfectly.

K Means Clustering Algorithm

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

Conclusions

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

That's it for this presentation

Thank you