CUSTOMER SEGMENTATION

HADYA HUSSAIN

DATE:
EMAIL:hadyahussain7@gmail.com

CONTENT

ABSTRACT

Customer segmentation is the process of separating your customers into groups based on certain traits they share. To develop the business and marketing in an efficient and successful manner we need to analyze the customer data. For that process the grouping of customers into small segments of individuals who share the common interest &characteristics' are known as customer segmentation. To analyze more efficiently we need to segment the customers based upon various types of segmentation.
1. Demographic segmentation: Segmenting the market based upon Age, Gender, Income, Financial status, Education, Family status and so on.
2. Geographic segmentation: as the name itself suggested that this kind of segmentation is done based upon the physical location of person.
3. Behavioral segmentation, this kind of segmentation is based on the behavioral data of the customers like Purchasing habits, spending habits, Brand interaction are used in this type.

OBJECTIVE

The main objective of the project is to create a model for Customer segmentation.
In this project, We will be performing an unsupervised clustering of data on the customer's records from a groceries firm's database. Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster. We will divide customers into segments to optimize the significance of each customer to the business. To modify products according to distinct needs and behaviours of the customers. It also helps the business to cater to the concerns of different types of customers.

TECHNOLOGY USED

PYTHON

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small- and scale projects. Python is open-source, so it is free to use, modify and distribute the python source code . Python is a high-level programming language that has English-like syntax making it easier to read and understand the code also the standard library of python is very big, that any and all function needed for a project can be found minimizing the use of external libraries. Different python packages like numpy, pandas,matplotlib, will used in the project.

MACHINE LEARNING

Machine learning is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications where it is difficult to develop conventional algorithms to perform the needed tasks.

LIBRARIES USED

NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high -level mathematical functions to operate on these arrays. numpy is open-source software and has many contributors.

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. It is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Scikit-Learn

Scikit-Learn is a free machine learning library for Python. It supports both supervised and unsupervised machine learning, providing diverse algorithms for classification, regression, clustering, and dimensionality reduction. The library is built using many libraries such as NumPy and SciPy. It also plays well with other libraries, such as Pandas and Seaborn.

ALGORITHM

A **clustering algorithm** is a type of Machine learning algorithm that is useful for segregating the data set based upon individual groups and the business need. It is a popular category of Machine learning algorithm that is implemented in data science and artificial intelligence . There are two types of clustering algorithms based on the logical grouping pattern: hard clustering and soft clustering. Some of the popular clustering methods based upon the computation process are K-Means clustering, connectivity models, centroid models, distribution models, density models, hierarchical clustering. The use cases for clustering algorithms are image segmentation, market segmentation, and social network analysis.

1. Collect dataset.
2. Import required libraries.
3. Load dataset.
4. Data Cleaning

In this section we are using Data cleaning and Feature engineering.
In order to, get a full grasp of what steps should I be taking to clean the dataset. Let us have a look at the information in data.
After checking the dataset we can find that:
* There are missing values in income .

* Dt_Customer that indicates the date a customer joined the database is not parsed asDateTime .

* There are some categorical features in our dataset as there are some features in dtype: object. So we will need to encode them into numeric forms later.

For the missing values we can use different methods here I am simply going to drop the rows that have missing income values.

In the next step, I am going to create a feature out of **"Dt_Customer"** that indicates the number of days a customer is registered in the firm's database. However, in order to keep it simple, I am taking this value relative to the most recent customer in the record. Creating a feature **("Customer_For")** of the number of days the customers started to shop in the store relative to the last recorded date .

Then we will be exploring the unique values in the categorical features to get a clear idea of the data.

I will be performing the following steps to engineer some new features:

- Extract the **"Age"** of a customer by the **"Year_Birth"** indicating the birth year of the respective person.
- Create another feature **"Spent"** indicating the total amount spent by the customer in various categories over the span of two years.
- Create another feature **"Living_With"** out of **"Marital_Status"** to extract the living situation of couples.

- Create a feature **"Children"** to indicate total children in a household that is, kids and teenagers.
- To get further clarity of household, Creating feature indicating **"Family_Size"**
- Create a feature **"Is_Parent"** to indicate parenthood status
- Lastly, I will create three categories in the **"Education"** by simplifying its value counts.
- Dropping some of the redundant features

The data is quite clean and the new features have been included. I will proceed to the next step. That is, preprocessing the data.

5. Data preprocessing

In this section, I will be preprocessing the data to perform clustering operations.

The following steps are applied to preprocess the data:

- Label encoding the categorical features
- Scaling the features using the standard scaler
- Creating a subset dataframe for dimensionality reduction

6. Dimensionality Reduction

In this problem, there are many factors on the basis of which the final classification will be done. These factors are basically attributes or features. The higher the number of features, the harder it is to work with it. Many of these features are correlated, and hence redundant. This is why I will be performing dimensionality reduction on the selected features before putting them through a classifier.
Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

Steps in this section:

- Dimensionality reduction with PCA
- Plotting the reduced dataframe

7. Clustering

Now that I have reduced the attributes to three dimensions, I will be performing clustering via Agglomerative clustering.

Agglomerative clustering is a hierarchical clustering method. It involves merging examples until the desired number of clusters is achieved.

Steps involved in the Clustering

- Elbow Method to determine the number of clusters to be formed

- Clustering via Agglomerative Clustering
- Examining the clusters formed via scatter plot

8. Evaluating Models

Since this is an unsupervised clustering. We do not have a tagged feature to evaluate or score our model. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters' patterns.

9. Profiling

Now that we have formed the clusters and looked at their purchasing habits. Let us see who all are there in these clusters. For that, we will be profiling the clusters formed and come to a conclusion about who is our star customer and who needs more attention from the retail store's marketing team.

DATA COLLECTION

Data is mainly collect from internet. Relevant data can be obtained from websites like kaggle.
Data is also collected from google trends and google form.

RESULT

CONCLUSION