SYMBIOSIS
University of Applied Sciences, Indore
(Established by Govt. of M.P. vide Act No.23 of 2016 & Recognised by UGC u/s 2(f) of 1956 Act)

School of Computer Science and Information Technology

Project Title : Customer Segmentation using K-Means Clustering

Subject Name : Machine Learning

Subject Code : BTCS0505

Semester : V

Submitted by : Tejas Bhati

Submitted to : Ms. Shruti Jain

Date of Submission : 05/12/2022

Introduction:

These days, everything can be personalized. But, for business, this is actually a great thing. It creates a lot of space for healthy competition and opportunities for companies to get creative about how they acquire and retain customers.

One of the fundamental steps towards better personalization is customer segmentation. This is where personalization starts, and proper segmentation will help you make decisions regarding new features, new products, pricing, marketing strategies, even things like in-app recommendations.

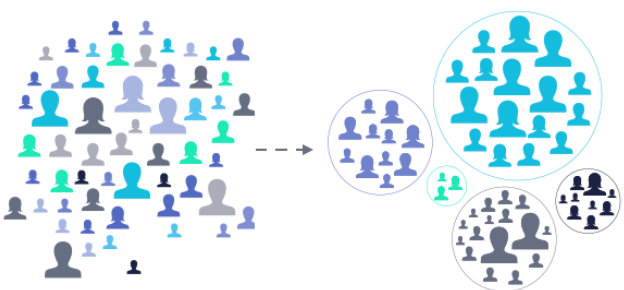
We actually try to find and group customers based on common characteristics such as age, gender, living area, spending behaviour, etc. So that we can market the customers effectively. But, doing segmentation manually can be exhausting and we use machine learning for this purpose.

Machine learning methodologies are a great tool for analyzing customer data and finding insights and patterns. Artificially intelligent models are powerful tools for decision-makers. They can precisely identify customer segments, which is much harder to do manually or with conventional analytical methods.

Customer segmentation is the practice of dividing customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

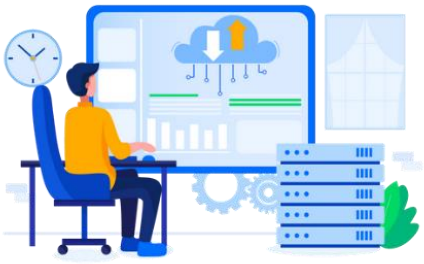
There are different methodologies for customer segmentation, and they depend on four types of parameters:

- geographic
- demographic
- behavioral
- psychological



Description of Dataset:

Before starting any project, We need to understand the dataset first.
The dataset named **customer_segmentation.csv** is used to create Machine Learning Model which is downloaded from Kaggle(An online community platform for data scientists and machine learning enthusiast.



Let's analyse the customer_segmentation dataset.
Our dataset has 1,999 data points and 08 features.
The features are:

1. ID: Unique identification of a customer.

2. Sex: Describes gender of a customer. In this dataset, there are only 2 different options.
- 0: male
- 1: female

3. Marital status: Marital status of a customer.
- 0: single
- 1: non-single (divorced / separated / married / widowed)

4. Age: The age of the customer in years, 18 the lowest age & 76 the highest age observed in the dataset.

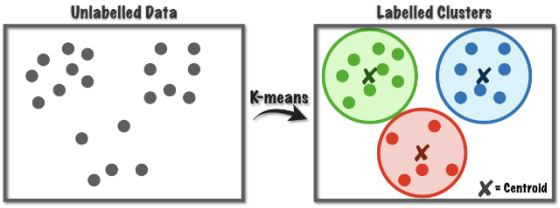
5. Education: Education Level of the customer.
- 0: other / unknown 1: high school
- 2: university 3: graduate school

6. Income: Annual income in US dollars of the customer.
- 35832 the lowest income & 309364 the highest income observed in the dataset.

7. Occupation: Category of occupation of the customer.
- 0: unemployed/unskilled
- 1: skilled employee / official
- 2: management / self-employed / highly qualified employee / officer

8. Settlement size: The size of the city that the customer lives in.
0: small city 1: mid-sized city 2: big city

Description of ML model implemented:



Here for this Problem Statement K-means clustering is used. It is a unsupervised machine learning algorithm used when we have unlabelled data. Unlabelled data means input data without categories or groups provided. Our customer segmentation data is like this for this problem.
The algorithm discovers groups (cluster) in the data, where the number of clusters is represented by the K value. The algorithm acts iteratively to assign each input data to one of K clusters, as per the features provided. All of this makes k-means quite suitable for the customer segmentation problem.
Given a set of data points are grouped as per feature similarity. The output of the K-means clustering algorithm is:

- The centroids values for K clusters,
- Labels for each input data point.

At the end of implementation, we're going to get output such as a group of clusters along with which customer belongs to which cluster.
We're going to use the elbow method. The K-means clustering algorithm clusters data by separating given data points in k groups of equal variances. This effectively minimizes a parameter named inertia. Inertia is nothing but within-cluster sum-of-squares distances in this case.

The elbow method finds the value of the optimal number of clusters using the total within-cluster sum of square values. This represents how spread-apart the generated clusters are from one another. In this case, the K-means algorithm is evaluated for several values of k, and the within-cluster sum of square values is calculated for each value of k. After this, we plot the K versus the sum of square values.
After analyzing this graph, the number of clusters is selected, so that adding a new cluster doesn't change the values of the sum of square values significantly.

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the **Principal Components**.

It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given

dataset by reducing the variances. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

Future Scope:

It’s not wise to serve all customers with the same product model, email, text message campaign, or ad. Customers have different needs. A one-size-for-all approach to business will generally result in less engagement, lower-click through rates, and ultimately fewer sales. Customer segmentation is the cure for this problem.

Customer segmentation simply means grouping your customers according to various characteristics.

It’s a way for organizations to understand their customers. Knowing the differences between customer groups, it’s easier to make strategic decisions regarding product growth and marketing.

The opportunities to segment are endless and depend mainly on how much customer data you have at your use.

Conclusion:

We segmented our customers into 4 groups. We are ready to start to choose our groups based on our goals. Segmentation helps marketers to be more efficient in terms of time, money and other resources.

They gain a better understanding of customer's needs and wants and therefore can tailor campaigns to customer segments most likely to purchase products. We can see the green segment well off is clearly separated as it is highest in both age and income. But the other three are grouped together.

We can conclude that K-Means did a decent job! However, it’s hard to separate segments from each other.

When we plotted the K means clustering solution without PCA, we were only able to distinguish the green segment, but the division based on the components is much more pronounced.