International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217654

Customer Segmentation Using Machine Learning

Varad R Thalkar
Masters in Computer Science, Somaiya University, Mumbai, Maharashtra, India

Article Info
Volume 7, Issue 6
Page Number: 28-37

Publication Issue :
November-December-2021

Article History
Accepted : 02 Dec 2021
Published : 10 Dec 2021

ABSTRACT

Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. Segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decisions; identification of products associated with each customer segment and to mange the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.

Keywords : Customer, Segmentation, Data, Product, K-means.

I. INTRODUCTION

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, preferences and so on. Instead of "one-size-fits-all" approach, customer segmentation clusters the customers into groups sharing the same properties or behavioural characteristics.

Customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, data geographical conditions, economic conditions, demographical conditions as well as behavioural patterns.

Demographic information, such as gender, age, familial and marital status, income, education, and occupation.
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.
Psychographics, such as social class, lifestyle, and personality traits.
Behavioural data such as spending and consumption habits, product/service usage, and desired benefits.

The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, determining new market opportunities, making better brand strategy, identifying customers retention.

II. METHOD

Customer Segmentation :

Step 1: Collecting Customer Data (Transactional data): This step involves the collection of transactional customer data comprises of their static (Eg: Age, Gender etc.) and dynamic data (Eg: Purchase frequency etc.) [1] from shopping vendors.

Step 2: Preprocessing of Data: Pre processing of the data is one of the important step for the accuracy of predictive model. In this step, the collected data will be cleaned and relevant features will be extracted. Feature selection is a data reduction technique which is responsible for extracting relevant features required for input vector of predictive model. This acts as pre-processing steps for creating subset of original features by excluding those features which are redundant. Correlation measures the relationship between two features. Basically it simply filters those features which are not redundant to form subset of original features. To measure the association between features correlation coefficient is calculated between two features and based on its value.

Clustering :

During clustering, unlike during needs/value segmentation, attributes are not grouped with regard to their effect on a specific target variable. Instead, it is performed without a fixed objective as it attempts to recognize patterns in the existing data sets. There are various algorithms that can be used to perform clustering. Once segments have been formed via the clustering method they can be further analyzed and characterized to make them usable for campaigns. Hypotheses can help to better understand the clusters found and define their relevance for further campaign development.

K Means Clustering Algorithm :

K-Means is one of the most widely used clustering algorithms and is simple and efficient. The K-Means clustering beams at partitioning the 'n' number of observations into a mentioned number of 'k' clusters. The K-Means is an unsupervised learning algorithm and one of the simplest algorithms used for clustering tasks. The K-Means divides the data into non-overlapping subsets without any cluster-internal structure. The values which are within a cluster are very similar to each other but, the values across different clusters vary extremely. K-Means clustering works really well with medium and large-sized data. Despite the algorithm's simplicity, K-Means is still powerful for clustering cases in data science.

K-means technique for customer segmentation due to its following advantages:

This technique suits for the data with numeric features and often terminates at local optimum.

It is highly scalable and efficient for large data sets.

It is fast in modeling and its result is more understandable.

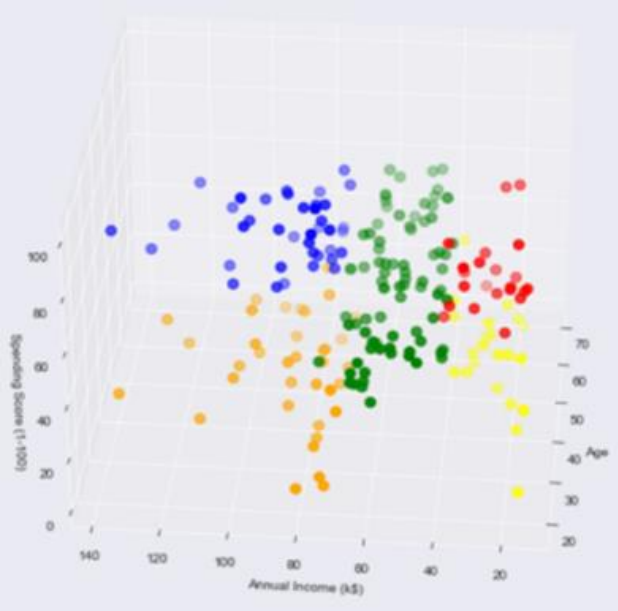
The aim of the K-Means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori. These centroids should be placed in a wise fashion so that the results are optimal which otherwise can differ if locations of the centroids change. So, they should be placed as far as possible from each other. Each data point is then taken and associated with the nearest centroid until no data points are pending. This way an early grouping is done and at this point, k new centroids have to be recalculated as these will be the centers of

the clusters formed earlier. After having calculated these centroids, the data points are then allocated to the clusters to the nearest centroids. In this iteration, the centroids change their position stepwise until no further modifications have to be done and the location of the centroids remains intact.

The K-Means algorithm is relatively simple. The "K" cluster points, which will be the centroids, are placed in the space among the data points. Each data point is assigned to the centroid for which the distance is the least. After each data object has been assigned, centroids of the new groups are re-calculated. The above two steps are repeated until the movement of the centroid ceases. This means that the objective function of having the least squared error is completed and it cannot be improved further. Hence, we get K clusters as a result.

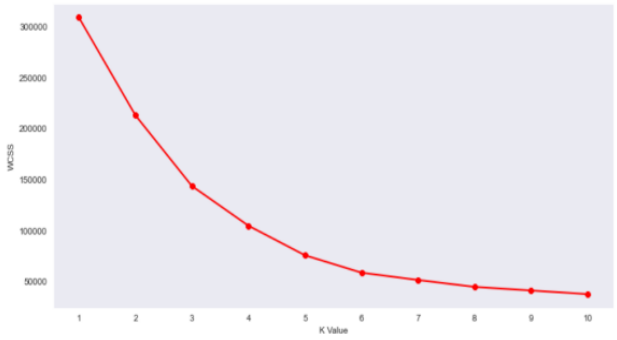
K-Means algorithm aims at minimizing an objective function, which here, is squared error. It is an indicator of the distance of the data points from their respective cluster centers. The process in this algorithm always terminates but the relevance or the optimal configuration cannot be guaranteed even when the condition on the objective function is met. The algorithm is also sensitive to the selection of the initial random cluster centers. That is why it runs multiple times to reduce this effect but for a large number of data points, it tends to perform very well even though it is iterative. One major advantage of K-Means clustering is that the computational speed of this algorithm is higher than other hierarchical methods of clustering and it is also easy to implement.

The algorithm works as follows:
Step-1 : Specifying the number of clusters – k value.
Step-2 : Centroids are initialized by shuffling the dataset and then randomly selecting k data points for the centroids without replacement.
Step-3: Repeat the iteration until there is no change to the centroids. i.e, assignment of data points to the clusters does not change.
Recency, Frequency and Monetary are brought to the same scale and the data is normalized before clustering process. It is important to determine the optimum number of clusters i.e, "k value". For this we used Elbow method.



Elbow Method:

Elbow method which uses the within cluster sums of squares by looking at the total within-cluster sum of square as a function of the number of clusters. The location of a knee or elbow in the plot is usually considered as an indicator of the appropriate number of clusters.



The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other.

Tool :

Jupyter Notebook : The Jupyter Notebook is an open source web application that can use to create and share documents that contain live code, equations, visualizations, and text.

Jupyter Notebook was created to make it easier to show one's programming work, and to let others join in. Jupyter Notebook allows you to combine code, comments, multimedia, and visualizations in an interactive document — called a notebook, naturally — that can be shared, re-used, and re-worked. And because Jupyter Notebook runs via a web browser, the notebook itself could be hosted on your local machine or on a remote server.

I.

Jupyter Notebooks can include several kinds of ingredients, each organized into discrete blocks:

Text and HTML : Plain text, or text annotated in the Markdown syntax to generate HTML, can be inserted into the document at any point. CSS styling can also be included inline or added to the template used to generate the notebook.

Code and output : The code in Jupyter Notebook notebooks is typically Python code, although you may add support in your Jupyter environment for other languages such as R or Julia. The results of executed code appear immediately after the code blocks, and the code blocks can be executed and re-executed in any order you like, as often as you like.

Visualizations : Graphics and charts can be generated from code, by way of modules like Matplotlib, Plotly, or Bokeh. Like output, these visualizations appear inline next to the code that generates them. However, code can also be configured to write them out to external files if needed.

Multimedia : Because Jupyter Notebook is built on web technology, it can display all the types of multimedia supported in a web page. You can include them in a notebook as HTML elements, or you can generate them programmatically by way of the "IPyhon.display" module.

Data : Data can be provided in a separate file alongside the ".ipynb" file that constitutes a Jupyter Notebook notebook, or it can be imported programmatically—for instance, by including code in the notebook to download the data from a public Internet repository or to access it via a database connection.

III. RESULTS AND DISCUSSION

The goal of customer segmentation is to identify the customer's behaviour and buying patterns which indirectly helps in boosting the sales of the company.

IV. CONCLUSION

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation needed to get potential customers used to increase profits.

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.

V. REFERENCES

[1]. Azad Abdulhafedh, “Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation”.

[2]. Yash Kushwaha, Deepak Prajapati, “Customer Segmentation using K-Means Algorithm”.

[3]. Kishana R. Kashwan, Member, IACSIT, and C. M. Velu, “Customer Segmentation Using Clustering and Data Mining Techniques”.

[4]. Shreya Tripathi, Aditya Bhardwaj and Poovammal E, “Approaches to Clustering Customer Segmentation”.

[5]. Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance

[6]. kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation”.

[7]. V.Vijilesh, A. Harini, M. Hari Dharshini, R. Priyadharshini, “Customer Segmentation using Machine Learning”.

[8]. Balmeet Kaur, Pankaj Kumar Sharma, “Implementation of Customer Segmentation using Integrated Approach”.

[9]. https://www.infoworld.com/article/3347406/what-is-jupyter-notebook-data-analysis-made-easier.h

Cite this article as :

Varad R Thalkar, "Customer Segmentation Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 207-211, November-December 2021. Available at doi : https://doi.org/10.32628/CSEIT217654
Journal URL : https://ijsrcseit.com/CSEIT217654