Customer Segmentation and Classification

Abstract

Most of the service providers and product based companies while launching brand new products, services or releasing new versions of existent products need to campaign to reach at the potential customers. While doing so they target their already existing customers who are the ambassadors of their company. To address the existing customers, they maintain the detailed customer data at all levels as customer master data.

Overloading information occurs when customers get too much information about a product then feel confused. Personalization will become a solution to overloading problem. In marketing, personalization technique can be used to get potential customers in a case to boost sales. The potential customer is obtained from customer segmentation. The process of segmenting the customers with similar behaviours into the same segment and with different patterns into different segments is called customer segmentation.

In this paper hierarchical clustering is been implemented to segment the customers. A python program has been developed and the program is been trained by applying standard scaler onto a dataset training sample taken from local retail shop. By applying clustering, 5 segments of cluster have been formed labelled as Careless, Careful, Standard, Target and Sensible customers. After segmenting customer on their shopping trend, classification has been added to classify new customer.

I. INTRODUCTION

As many businesses and startups being coming up every day, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition increased a lot. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer.

Clustering has proven efficient in discovering subtle but tactical patterns or relationships buried within a repository of unlabelled datasets. This form of learning is classified under unsupervised learning. Clustering algorithms include k-Means algorithm, k-Nearest Neighbour algorithm, hierarchical clustering algorithm, Self-Organising Map (SOM) and so on. These algorithms, without any knowledge of the dataset beforehand, are capable of identifying clusters therein by repeated comparisons of the input patterns until the stable clusters in the training examples are achieved based on the clustering criterion or criteria. Each cluster contains data points that have very close similarities but differ considerably from data points of other clusters. Clustering has got immense applications in pattern recognition, image analysis, bioinformatics and so on. In this paper, the hierarchical clustering algorithm has been applied in customer segmentation

Customer segmentation helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to each segment, providing discounts for a specific segment

II. LITERATURE REVIEW

A. Customer segmentation

Over the years, the commercial world is becoming more competitive, as such organizations have to satisfy the needs and wants of their customers, attract new customers, and hence enhance their businesses[1]

The task of identifying and satisfying the needs and wants of each customer in a business is a very complex task. This is because customers may be different in their needs, wants, demography, geography, tastes and preferences, behaviours and so on. As such, it is a wrong practice to treat all the customers equally in business. This challenge has motivated the adoption of the idea of customer segmentation or market segmentation, in which the customers are subdivided into smaller groups or segments wherein members of each segment show similar market behaviours or characteristics. According to [2], customer segmentation is a strategy of dividing the market into homogenous groups. [3] posits that ―the purpose of segmentation is the concentration of marketing energy and force on subdivision (or market segment) to gain a competitive advantage within the segment.

It's analogous to the military principle of concentration of force to overwhelm energy.

The main objective of segmentation is to separate the objects that are homogenous and heterogenous with external market/consumers. The outcome of the segmentation mainly depends on the knowledge variables, which can be gathered from market, psychographic, regional, life-style, etc., Steenkamp, Jan-beneddict EM and Frenkel [4].

Jayant et al. [5] states that in customer segmentation, the customers are divided into different groups where customers of same group are similar to each other in terms of marketing. Customers are divided into different clusters based on various attributes such as age ,interests, age, spending habits etc. Customer segmentation includes geographic segmentation, demographic segmentation, media segmentation, price segmentation, psychographic or lifestyle segmentation, distribution segmentation and time segmentation [3].

B. Agglomerative Clustering

This clustering comes under hierarchical clusters which are form based on some hierarchy. It is based on the concept that objects that are closer are more related to each other in comparison of the objects that are far from each other., T.Nelson et al. [5]. The main challenge of Hierarchical method is that once it undergoes split or merge operation it can never be undone. This challenge is profitable as it leads to smaller computation costs by not worrying about a combinatorial number of different choices. Yogita et al. [6].

There are two strategy in hierarchical clustering, first is top-down strategy also known as divisive clustering and second is bottom – up strategy also known as agglomerative clustering. Agglomerative clustering process is allows more flexibility because it permits the user to supply any arbitrary similarity function defining

III. METHODOLOGY

The data used in this paper was collected from a local retail business .The data set used to implement clustering and Hierarchical algorithm was collected from a store of shopping mall. The data set contains 8 attributes and has 1000 tuples, representing the data of 1000 customers. The attributes in the data set has CustomerId, Gender, Age, Maritial, Address, retired, Annual income(k$), Spending score on the scale of 1-100.

A. Visualization of data

i. Visualize gender of customers

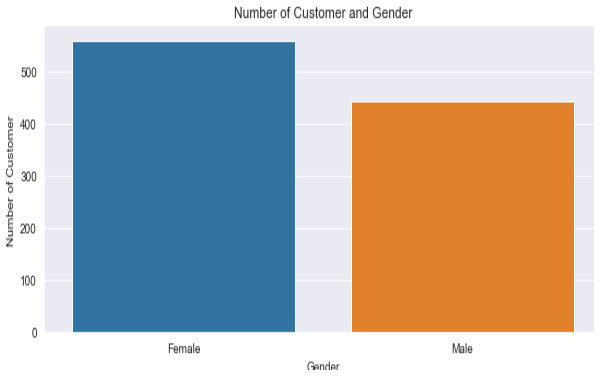

Fig 1: Bar graph of customer's gender

ii. Visualize age of customers vs frequency

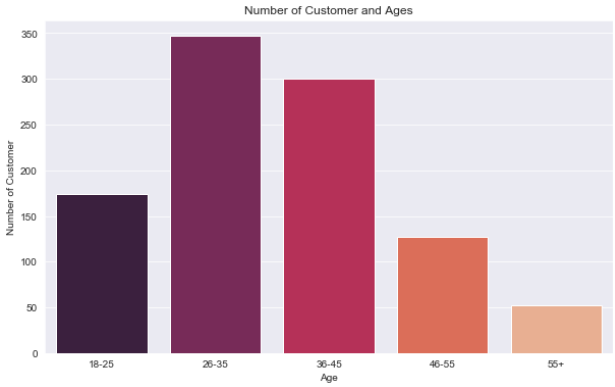

Fig 2: Bar graph of customers and their age

iii. Visualize Annual income

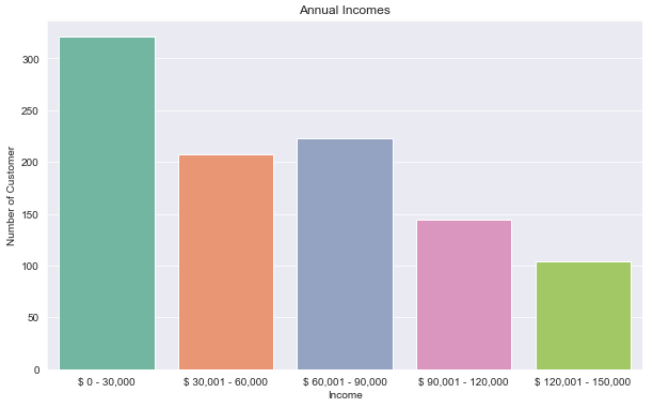

Fig 3: Graph of customers and their income

iv. Visualize Spending Score

Fig 4: Graph of customers and their score

B. **Hierarchical Clustering**

Hierarchical clustering is a method of cluster analysis which builds a hierarchy of data points as they move into a cluster or out of it. Strategies for this algorithm generally fall into two categories [7] agglomerative and Divisive .

Agglomerative - This is a bottom-up approach where each observation begins as an initial cluster and then merges into clusters as they move up the hierarchy. Divisive technique is a top-down approach where there is only one cluster initially and is then split into finer cluster groups as they move down the hierarchy. This merging and splitting of clusters takes place in a greedy manner and the hierarchical algorithm yields a dendrogram which represents the nested grouping of patterns and the levels at which groupings change

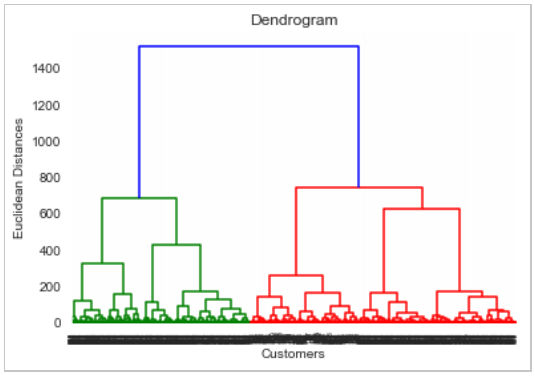

Fig. 5: Visualization of the formation of clusters in the studied dataset with the help of a dendrogram

We have the option of choosing the required number of clusters from the dendrogram itself by selecting the range of maximum distance and then placing a cut-off line at that position. It simply indicates that the distance between the formed clusters is maximum and distinction can be made among them. Hence, according to Figure 3, for satisfactory results, we can choose five clusters (K=5).

]*Algorithm:*

i. Each data point is taken as to be a cluster.

ii. Merge the two closest cluster.

Step ii needs to be repeated until all the data points are merged together to form a single cluster. However, as we have defined the value of K as 5, the algorithm will stop when all the data points are part of any of the 5 clusters.
The clusters are depicted as follows in Figure 6

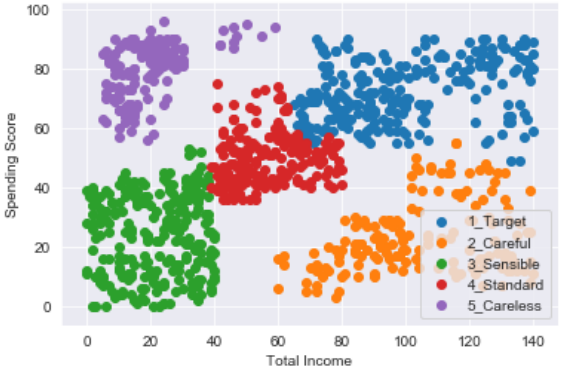

Fig. 6: Clusters formed as a result of applying Hierarchical Clustering on the dataset taken for study

Each color represents a different cluster and data points are plotted against Annual Income on the X-axis and the Spending Score on Y-axis.

Hierarchical clustering has been extensively used for segmentation purposes due to its ability to produce results in a visual way [7]. It helps in determining the number of clusters for any analysis. It can be used for varied datasets like categorical, spatial and time series with numerical data set being the most common as it consists of data as just real numbers [8]. It was also used and compared with other clustering algorithms in bank customer segmentation [9]. [10] and [11] have used Chameleon in their customer segmentation process, which is a hierarchical clustering algorithm as well. [12] have made use of a hierarchical pattern-based clustering. The main advantage of Hierarchical clusteringis that the output is in the form of a hierarchy(dendrogram) which tells us exactly at which point the clusters merged or split. Hence it is easy to choose and decide on the number of clusters that we wish to take by looking at the dendrogram. However, for a large number of observations its computational speed is very low as compared to the nonhierarchical methods of clustering [13]. Hence, the size and order of the data have an impact on the

final results obtained. Nevertheless, one does not need to plug the number of clusters as the input to the algorithm and hence we can have different partitioning groups, the choice of which completely depends on the end user.

C. Classification

Customer classification refers to the automated process of recognizing individuals based on their spending score patterns. After clustering the dataset of customers we train the model using Gradient Boosting Feature Selection.we selected Gender, Age, Maritial status, Address, retirement , Annual income(k$) and Spending score for feature selection and got the cluster of customer where they belong to.

We train 80% of data and set 20 % of data to test using xgboost [14] . We classify them according to the clusters we have made using hierarchial clustering. This helps the market when new customer add to their database.

IV. RESULT

The Hierarchial clustering algorithm was able to cluster almost the entire data points correctly, clusteres were formed on the basis of annual income and spending score of customer. After clustering and labelling all 5 clusteres we train our dataset and fitted the Train_Data using xgboost for a given dataset. When new data of customers being given to this model it will predict their clusters where they belong with the accuracy of 99.0 %.

V. CONCLUSION

The aim of the paper is to identify similar customers depending on their income and spending score and after clustering classify new customers. From the above visualization it can be observed that Cluster 1 (Target) denotes the customer who has high annual income as well as high spending score. Cluster 2 (Careful) represents the cluster having high annual income and low annual spend. Cluster 3 (Sensible) represents customer with low annual income and low annual spend. Cluster 4 (standard) denotes the customer with medium income and medium spending score. Cluster 5 (Careless) denotes the low annual income but high yearly spend.

Forming clusters and classifying them gives a good percentage accuracy which businessmen can use. Classification help company to target their customer accordingly.

VI. REFERENCES

[1] Puwanenthiren Premkanth, ―Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[2] Sulekha Goyat. ―The basis of market segmentation: a critical review of literature‖. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online).Vol 3, No.9, 2011

[3] By Jerry W Thomas.“Market Segmentation‖. 2007. Retrieved from www.decisionanalyst.com on 12-July, 2015.

[4] Steen Kamp, Jan-Benedict EM and Frenkel TerHofstede, “International Market segmentation: Issues and perspectives” International Journal of research in marketing, 19, no.3 (2002)

[5] Jayant Tikmani, Sudhanshu Tiwari and Sujata Khedkar, "Telecom customer segmentation based on cluster analysis An Approach to Customer Classification using k-means", IJIRCCE, 2015.

[6] Yogita Rani and Harish Rohil, "A Study of Hierarchical Clustering Algorithm", IJICT, 2013.

[7] O. Maimon and L. Rokach, “Clustering methods”, in Data Mining and Knowledge Discovery Handbook. Boston: Springer US, 2005, pp. 321-352.

[8] D. Gaur and S. Gaur, “Comprehensive analysis of data clustering algorithms”, in Future Information Communication Technology and Applications. Dordrecht: Springer Netherlands, 2013, pp. 753-762.

[9] D. Zakrzewska and J. Murlewski, “Clustering algorithms for bank customer segmentation”, in 5th International Conference on Intelligent Systems Design and Applications (ISDA‟05), Warsaw, 2005, pp. 197-202.

[10] S. Yoon, et al., “A data partitioning approach for hierarchical clustering”, in 7th International Conference on Ubiquitous Information Management and Communication, Kota Kinabalu, 2013, p. 4.

[11] J. Li, K. Wang and L. Xu, “Chameleon based on clustering feature tree and its application in customer segmentation”, Annals of Operations Research, vol. 168, no. 1, pp. 225-245, 2008.

[12] D. Suib and M. Deris, “An efficient hierarchical clustering model for grouping web transactions”, International Journal of Business Intelligence and Data Mining, vol. 3, no. 2, pp. 147-157, 2008.

[13] T. Sajana, C. Sheela Rani and K. Narayana, “A Survey on Clustering Techniques for Big Data Mining”, Indian Journal of Science and Technology, vol. 9, no. 3, 2016.

[14] F. O. Catak, "Classification with boosting of extreme learning machine over arbitrarily partitioned data," *Soft Computing*, vol. 21, May 2017 pp. 2269-2281,