Customer Segmentation Using Machine Learning Algorithms

[1]Akanksha Toppo, [2]Om Narayan Gupta, [3] Ragini Sinha, [4] Sana Danwani

[1,2,3]Student, [4]Assistant professor
Department of Computer Science and Engineering
Bhilai Institute of Technology
Raipur.

Abstract-
The emergence of numerous competitors and entrepreneurs has created a great deal of tension among competing business es as they seek new buyers while retaining existing ones. As a result of the preceding, the requirement for exceptional cust omer service becomes appropriate regardless of the size of the business. Furthermore, the ability of any business to unders tand the needs of each of its customers will provide greater customer support in the provision of targeted customer service s and the development of customised customer service plans. This comprehension is attainable through structured custom er service. Customers in each segment share similar market characteristics. Traditional market analytics always fail when the customer base is very large, so k-means clustering algorithm and DBSCAN clustering algorithm are used for this purpose. Additionally, t-SNE and PCA are applied for dimensionality reduction. Finally, a variance technique and feature importance are applied to select variables that specifically represent the customer base.

Key words: Customer Segmentation, Clustering, K-Means, DBSCAN, SNE, PCA.

INTRODUCTION

Customer segmentation is a method of categorizing customers who may be suitable for advertising based on criteria such as gender, age, interests, and various spending habits. The primary goal of any organization or business is to identify their target customers, how their customers perform and use their services.

Due to increased business competition and the availability of large-scale historical data, knowledge mining techniques have been widely used over the years to extract important and strategic information hidden in an organization's dataset. The process of extracting logical details from a dataset and presenting them in a human-readable format for decision support is known as data filtering. Statistics, artificial intelligence, machine learning, and data systems are all distinguished by data processing techniques. Bioinformatics, meteorology, fraud detection, financial analysis, and customer segmentation are just a few examples of data processing application.

The purpose of this paper is to use data mining to identify customer segments within a commercial business. Customer division entails dividing customers into groups called customer segments based on business characteristics, with each customer segment consisting of consumers who share similar market characteristics. These distinctions are auxiliary factors that will influence market or business-like product preferences or expectations, locations, behaviour, and so on. The importance of customer segmentation includes a company's ability to tailor market plans to each segment of its customers; support for business decisions made in a risky environment, such as debt relationships with their customers. Identification of products associated with individual components, as well as how to manage demand and supply power; reveal the interdependence and interchange between consumers, between products, or between customers and products, the ability to predict customer decline, and which customers are most likely to have problems, as well as consider other marketing research questions and provide clues to finding solutions.

RELATED WORK

A. Customer Classification-

The commercial world has become more competitive over time, as organisations like these have to meet their customers' needs and desires, attract new customers, and thus improve their businesses. [6] Identifying and meeting the needs and requirements of each customer in the business is a difficult task. This is due to the fact that customers differ in terms of their needs, desires, demographics, size, taste and taste, features, and so on. In business, treating all customers equally is a bad practise. This challenge has adopted the customer segmentation or market segmentation concept, in which consumers are divided into subgroups or segments, with members of each subcategory exhibiting similar market behaviours or characteristics. [9]. Customer segmentation is thus the process of dividing a market into indigenous groups.

B. Big Data –

Big Data research is gaining traction. Big data is defined as a large amount of formal and informal data that cannot be analysed using traditional methods and algorithms. Companies include billions of data about their customers, suppliers, and operations, and millions of internally connected sensors send sensing, manufacturing, and communications data to the real world on devices such as mobile phones and cars. [10] Capability to improve forecasting, save money, increase efficiency, and improve a variety of areas including traffic control, weather forecasting, disaster prevention, finance, fraud control, business transactions, national security, education, and healthcare. The three Vs of big data are volume, variability, and speed. Other 2Vs are available, such as authenticity and price, making it a 5V.

C. **Data repository –**
Data collection is the process of gathering and analysing information from an established system in order to answer pertinent questions and evaluate the results. [12] Data collection is an essential component of research in all disciplines, including physical and social sciences, humanities, and business. The goal of all data collection is to obtain high-quality evidence that will lead to concrete and misleading answers to the questions posed. We gathered information from the UCI machine learning repository.

D. Clustering data-
Clustering is the process of grouping data into datasets based on similarities. There are several algorithms that can be applied to datasets based on the condition provided. [7] However, because there is no universal clustering algorithm, it is critical to select the appropriate clustering techniques. We implemented three clustering algorithms in this paper using the Python scalar library.

E. k-Means-
The letter K indicates that an algorithm is one of the most widely used classification algorithms. This clustering algorithm is based on centro, which places each data point in one of the overlapping ones that are pre-sorted in the K-algorithm. Clusters are formed to correspond to hidden patterns in the data, which provide the necessary information to aid in decision-making process.

METHODOLOGY

Clustering Algorithms
There are numerous unsupervised clustering algorithms available, and while each has significant strengths in certain situations, I will focus on two that are commonly used.

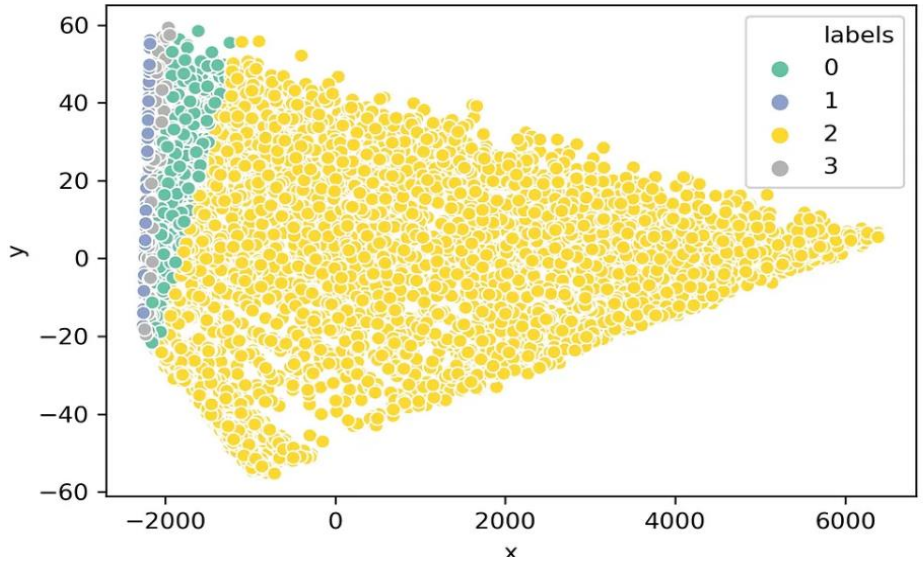
k-Means Clustering
This is by far the most commonly used algorithm for data clustering in my experience. k-Means begins by selecting k random centres that you can specify. Then, based on their Euclidean distance, all data points are assigned to the nearest centre. Following that, new centres are calculated and data points are updated. This process is repeated until no clusters change between iterations.

DBSCAN
Clustering can also be done based on data point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is one example, which clusters data points if they are sufficiently dense. DBSCAN detects and expands clusters by scanning neighbourhoods. If it cannot find any points to add, it simply moves on to the next point in the hope of discovering a new cluster. Any point that does not have enough neighbours to cluster is classified as noise.

Visualizing Clusters-
To visualise the clusters, use one of the most popular dimensionality reduction methods, such as PCA or t-SNE.

Principal Component Analysis (PCA)-
PCA is a statistical process that uses orthogonal transformation to convert observations of correlated features into a set of linearly uncorrelated features. The Principal Components are the new transformed features. It is one of the most widely used tools for exploratory data analysis and predictive modelling. It is a method for extracting strong patterns from a given dataset by reducing variances.

We can then visualize our data in 3d:



Although PCA was successful in reducing the dimensionality of the data, it does not appear to visualize the clusters in an intuitive manner. This is common with high-dimensional data; they tend to cluster around the same point, and PCA extracts that information.

Instead, we can employ an algorithm known as t-SNE, which was designed specifically to generate an intuitive representation/visualization of the data.

t-distributed Stochastic Neighbor Embedding (t-SNE)-

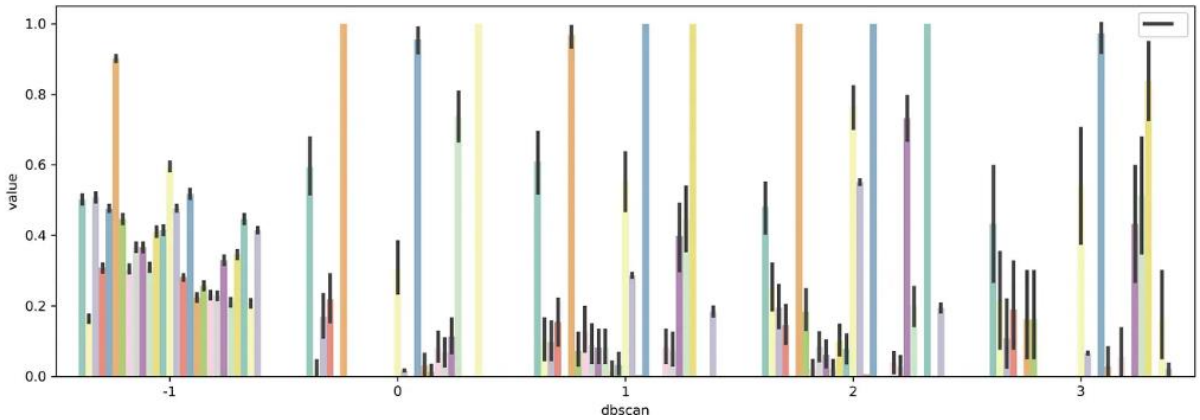
t-SNE is a high-dimensional data visualization algorithm. It captures non-linear structures by using local relationships between points to create a low-dimensional mapping.

It begins by generating a probability distribution (such as a Gaussian) that governs the relationships between neighboring points. Then, using the Student t-distribution, it creates a low-dimensional space that as closely as possible follows that distribution. You might be wondering why it uses a Student t-distribution at this point. A Gaussian distribution, on the other hand, has a short tail that squashes nearby points together. The tail of a Student t-distribution is longer, and points are more likely to be separated.

Interpreting Clusters

We'd like to know what distinguishes each cluster of customers now that we've segmented them. This will assist us in determining the types of customers we have.

One method is to simply plot all variables and look for differences between clusters. However, this approach fails when dealing with more than ten variables because it is difficult to visualize and interpret:



The solution would be to choose a subset of variables that are important when defining clusters. I'd like to demonstrate two methods here: variance between averaged groups and feature importance extraction via predictive modelling.

Variance within variables and between clusters

One variable importance assumption in cluster tasks is that if the average value of a variable ordered by clusters differs significantly from one another, that variable is likely to be important in the creation of the clusters.

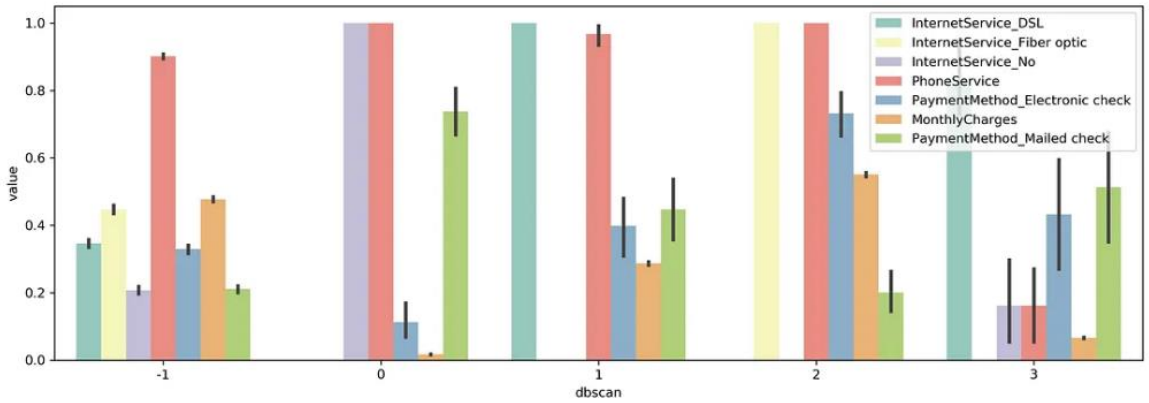
We begin by aggregating the data based on the generated clusters and obtaining the mean value for each variable:

dbscan	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	OnlineSecurity	OnlineBackup	...
0	0.593750	0.018750	0.168750	0.218750	0.0	1.000000	0.000000	0.000000	0.000000	...
1	0.609756	0.105691	0.097561	0.154472	0.0	0.967480	0.073171	0.130081	0.089431	...
2	0.480447	0.251397	0.195531	0.145251	0.0	1.000000	0.184358	0.022346	0.083799	...
3	0.432432	0.216216	0.108108	0.189189	0.0	0.162162	0.162162	0.000000	0.000000	...

Groupby of clusters generated by DBSCAN averaged per variable

DBSCAN defines the -1 cluster as noise, so I ignored it. To facilitate visualization, the data were scaled between 0 and 1.

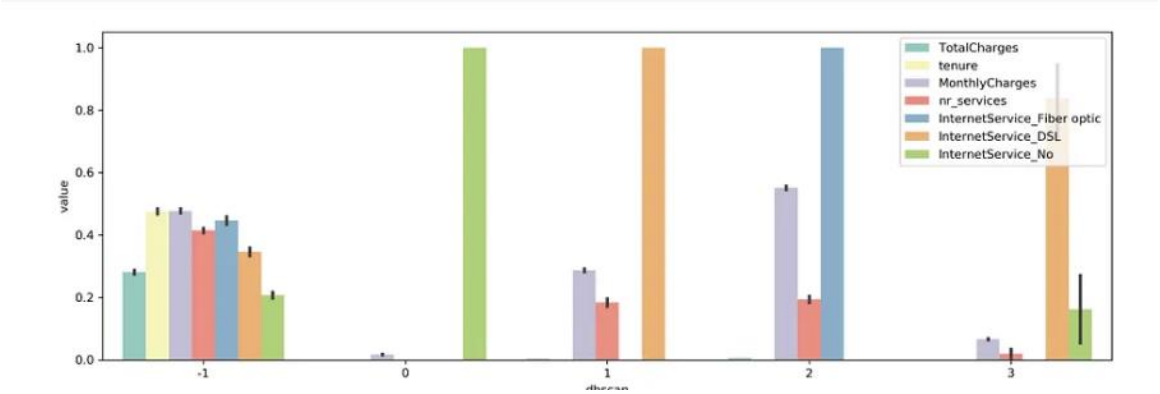
Following that, we simply compute the variance of means between clusters within each variable and choose the top seven variables with the highest variance:



Differences between clusters can now be seen more clearly. For example, in cluster 0, every single person has no Internet service, whereas most other clusters have Internet service. Furthermore, we can see that Cluster 2 only contains people who have both Fiber optic and Phone services, implying that they were purchased together or as part of the same package.

Random Forest Feature Selection

Finally, we can use the clusters as a target variable and then use Random Forest to determine which features are important in cluster generation. This method requires a little more effort because you will need to check the accuracy of your model in order to extract important features accurately.



When we compare the variance analysis to the previous one, we can see that similar features are chosen. Because this method requires more work in the form of validation, I would recommend using the variance method described previously.

CONCLUSION

Customer segments are created in this project using the k-means clustering and DBSCAN clustering algorithm models, and the dataset is analyzed in various ways. To gain a better understanding of all of the elements and their relationships to the data, the data set was visualized. We investigate cluster analysis methods, visualize clusters using dimensionality reduction, and interpret clusters by examining influential features. Although the use of supervised machine learning techniques in organizations has increased significantly, these methods typically have one major drawback: the requirement for labelled data. Customer segmentation is one of the most common applications of clustering in order to gain a better understanding of them, which can then be used to increase the company's revenue.

REFERENCES:

[1] J. N. Sari, L. E. Nugroho, R. Ferdiana, and P. I. Santosa, “Review on customer segmentation technique on ecommerce,” Advanced Science Letters, vol. 22, no. 10, pp. 3018–3022, 2016.

[2] M. Aryuni, E. D. Madyatmadja, and E. Miranda, “Customer segmentation in xyz bank using k-means and k-medoids clustering,” in 2018 International Conference on Information Management and Technology (ICIMTech), IEEE, 2018, pp. 412–416.

[3] M. T. Ballestar, P. Grau-Carles, and J. Sainz, “Customer segmentation in ecommerce: Applications to the cashback business model,” Journal of Business Research, vol. 88, pp. 407–414, 2018.

[4] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, “A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca,” Journal of Big Data, vol. 7, no. 1, pp. 1–23, 2020.

[5] F. Yoseph, N. H. Ahamed Hassain Malim, M. Heikkil¨a, A. Brezulianu, O. Geman, and N. A. Paskhal Rostam, “The impact of big data market segmentation using data mining and clustering techniques,” Journal of Intelligent & Fuzzy Systems, vol. 38, no. 5, pp. 6159–6173, 2020.

[6] A. Abdulhafedh, "Incorporating k-means, hierarchical clustering and pca in customer segmentation," Journal of City and Development, vol. 3, no. 1, pp. 12–30, 2021.
[7] E. Akar, "Customers' online purchase intentions and customer segmentation during the period of covid-19 pandemic," Journal of Internet Commerce, vol. 20, no. 3, pp. 371–401, 2021.
[8] S. Bandyopadhyay, S. Thakur, and J. Mandal, "Product recommendation for e-commerce business by applying principal component analysis (pca) and kmeans clustering: Benefit for the society," Innovations in Systems and Software Engineering, vol. 17, no. 1, pp. 45–52, 2021. 36
[9] J. J¨a¨askel¨ainen, "Segmentation of investor customers using machine learning in banking," 2021.
[10] C. Jamunadevi, S. T. Selvan, M. Govindarajan, C. Saravanan, and B. J. Raman, "Lrfm model for customer purchase behaviour using k-means algorithm," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol. 1055, 2021, p. 012 111.
[11] F. Khanizadeh, F. Khamesian, and A. Bahiraie, "Customer segmentation for life insurance in iran using k-means clustering," International Journal of Nonlinear Analysis and Applications, vol. 12, no. Special Issue, pp. 633–642, 2021.
[12] R. N. N. Naseri, F. Rahmiati, and M. M. Esa, "Consumer attitude and online purchase intention: A segmentation analysis in malaysian halal cosmetic industry," 2021.
[13] E. H. Sharaf Addin, N. Admodisastro, S. N. S. Mohd Ashri, A. Kamaruddin, and Y. C. Chong, "Customer mobile behavioral segmentation and analysis in telecom using machine learning," Applied Artificial Intelligence, pp. 1–21, 2022.