Vol. 05, No. 3 (2023) 539-544, doi: 10.24874/PES05.03.016

Proceedings on Engineering Sciences

www.pesjournal.net

CUSTOMER SEGMENTATION AND PROFILING FOR E-COMMERCE USING DBSCAN AND FUZZY C-MEANS

Snehalatha N[1]
Mohana Kumar S
Vitasta Kachroo

Received 21.03.2023.
Accepted 14.08.2023.
UDC – 005.6

Keywords:

Customer Segmentation, DBSCAN, Fuzzy C-Means, Clustering, RFM

A B S T R A C T

The Rapid Industrialization has led to a sudden overwhelming boost in business. With growing advancements in technology and the fast-paced growth of users that have access to said technology everything reaches customers faster. Businesses have also tried their hand at being better at technology andthus boosted growth by appealing to a much larger customer base. Market segmentation is an increasingly important part ofa strong marketing strategy and can make all the differencefor companies in competitive market landscapes, such as e- commerce. To be in business today means much more than quality of the product and a brand name. It involves the right kind of marketing to the right kind of people. This paper focuses on creating a tool to enable businesses to understand the various customers it has, by segmenting them into groups and providing a detailed statistical model for analyzing where the marketing is needed the most and what would be the best offers the company can afford.

1. INTRODUCTION

All businesses want to grow, develop and build a community of loyal customers who improve your business by advertising, promoting and come more often. Such a community of loyal customers can't be found naturally. Targeting an entire mass of audience is not an effective marketing strategy instead First we need to understand the customers we are selling to and therefore be able to target them effectively which drives greater loyalty. Understanding the customers will allow you to prevent the negative experience, provoke positive emotions and lets you identify the most effective marketing ways. To achieve this customer segmentation comes into picture. Customer segmentation is mainly about who you target, where you reach them and how you talk to them. This mainly helps to understand the target audience more clearly and create customization to them accordingly. It is more effective to target your customer by aligning your tactics and strategies which helps you market and sell more effectively. This will lead to develop a better understanding of your customers' needs and desires more efficiently.

Customer segmentation is a way of splitting customers into groups based on their similar and shared characteristics. Some basic customer segmentation categories mainly include, geography, products or service purchased, how customers found you, device used by customers including device type / brand /

[1] Corresponding author: Snehalatha N
Email: snehalathan@jssateb.ac.in

browser, payment method, etc. Companies may collect customer information during purchase or checkouts by which even more segmentation can be carried out by knowing the reason for purchase, what drove them to purchase, etc which will help you increase customer lifetime value. Most obviously well implemented segmentation will improve your marketing performance. We can perform customer segmentation by using any data available about them. Creating customer profiles or buyer persona helps you understand who your current or ideal customers are. Based on the actual data they segment customers who act and think similarly. By grouping them to different categories, we can focus on them according to their needs and desires and approach them differently.

As of today the general segmentation seems to be too narrow and a bit old fashioned. The general segmentation will be too basic to understand your customers deeply. That's when the customer profiling takes the stage. Customer profiling is mainly about the customers' experience and a better way for its improvement. In this approach we use segmentation characteristics and perform segmentation accordingly but we also pay more attention to the customer's previous experience, pain points, plus points, opinions, etc. The main aim is to understand customers more effectively, offer a better experience, services or products.

Hence data is collected from various resources, transferring them into understandable format. We perform data preprocessing to remove all the null missing values to make it a clean data. The preprocessed data is fed into the RFM model to carry out the recency, frequency and monetary which will segment our valuable customers. After extracting the valuable customers clustering methods are applied and characteristics of each customer groups are segmented and analyzed to which we are using DBSCAN and Fuzzy C-Means algorithm. The proposed approach is designed in python and the performance of the proposed approach is evaluated in the UK e-commerce dataset.

2. METHODOLOGY

All forms of data undergo 4 main stages in their life cycle. Data Collection followed by its preprocessing. Once said data is cleaned and essential data is grouped together, data is manipulated and is analyzed to produce a statistical record and observations are made on the usefulness of data and how it can impact the growth or sales of a certain product. To simplify the complicated algorithms and observations, the user is presented with a visualized model including various kinds of graphs and an attractive User Interface that helps them be in control. The project follows a similar trajectory where we utilize various algorithms and combine the power of these into a single product that aims to provide the maximum coverage of segments that can help a business owner maximize his growth and minimize his churn rate.

The dataset we will be using is from a UK e-commerce website. The workflow of the project includes the following steps:

A. Preprocessing

Data that is available to us in the raw format is corrupted and unclean, that is, data is usually very inconsistent when stored in databases. Since databases focus on minimizing the redundancy, they might eliminate data that is useful to conduct a thorough analysis, and thus effect the efficiency of the work done. Applying data mining algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing hence, is crucial for maintaining data quality. Data Preprocessing can be defined as the necessary steps that are taken to transform or encode data so that it may be easily parsed by the machine. This process plays a key role in the final results of the project and is the starting point of all data. Hence, we focus on removing the null values and rid the data of missing values and other inconsistencies and outliers to make it as clean as possible and ready for the next steps.

B. RFM analysis

The pre-processed data is now fed to the RFM system. RFM stands for 'Recency, Frequency and Monetary'. In orderto track this a quantitative analysis technique used rank and group customers based on the RFM approach for the net total of their recent transactions. Thus, identifying the best customers and performing targeted marketing campaigns to these segmented groups. The system assigns each customer numerical scores based on these factors to provide an ob jective analysis. RFM analysis is based on a belief that a few customers (about 20 percent) make up the maximum profits of your business (up to 80 percent). Explaining the three terms Recency, frequency and Monetary and what it means in business technology (figure 1):

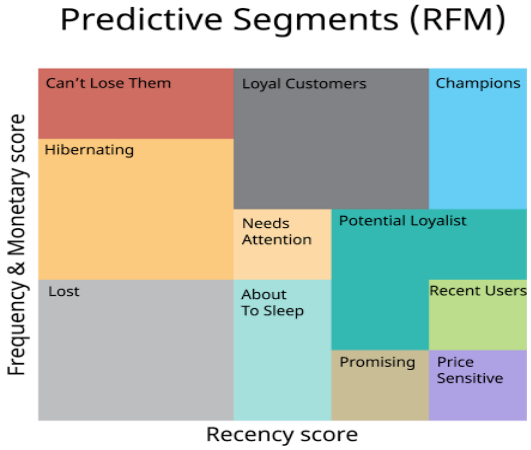

Figure 1. Predictive Segments

Recency defines how recent was the customer's last pur- chase. It targets the mentality of the customers who recently made purchases and are in the initial stages where the product and the company is still on their mind. These customers can be targeted to make similar purchases or suggested new things from the same brand. Recency is usually measured in days but depending on the depth of the analysis a more exhaustive or relaxed system can be designed, where they may measure it in years, weeks or even hours.

Frequency defines how often the said customer makes a purchase in a given time period. The logic being, customers who purchased once are often are more likely to purchase again. Additionally, targeting new customers or first timers for a follow-up advertising is universally exercised to convert them into more frequent customers.

Monetary stands for how much money a customer spend in a given period. Customers who spend a lot of money are more likely to spend money in the future and in turn be valuable assets to the business.

C. Customer Segmentation

The primal techniques of boosting trade and growing a business involve maintaining an excellent customer relationship. Communication is the bridge that is to be built to ensure a business reaches its customer base and the customers believe that they are being heard on a personal level. However, in case of big e-commerce websites, creating a space for a one-on-one connection with each and every customer is not feasible. This is where Customer Segmentation comes into effect. Segmentation according to Collica is a process to categorize or classify an item into a group that has a similarity in characteristic and in CRM (Customer Relationship Management) segmentation is used to classify customer based on some similarities by segmenting the records of customer database.

Clustering Algorithms can be applied for grouping these customer into broadly classified segments so that the tar- get market is created. In this project, we apply clustering methodologies on the data that is retrieved from RFM models and characteristics of each customer group is analyzed. The clustering mechanism commonly known as the DBSCAN clustering is applied with Fuzzy C-Means algorithms. A detailed description of these is mentioned further in the document.

Segmentation Model

Segmentation is the process of dividing your customer base into smaller segments. Segmentation can be of various types and there are innumerable ways to segment your customers. Some of those include the infamous:

1) Behavioral analysis:

This tests the way a customer purchases and tracks out details of what chronology or factors affect the purchases made by the user.

2) Demographic analysis:

This is the most popular kind of segmentation done. It uses the demographic description of customers to categorize the into gender, or age-based groups or segments and uses different approaches on said customer.

3) Psychographic analysis:

This categorizes a customer on the psyche or personality traits, it traces the beliefs and values and lifestyle of a custom and provides necessary recommendations. This type of segmentation is much more difficult as compared to others.

4) Geographic analysis:

This analysis categorizes customers into groups based on their geographical location on the map. If a business supplies world wide a much broader classification can be done. If it is internalized in cases of businesses limited to a particular country it is more detailed.

D. Evaluation of the Model

The simulation of the proposed model will be design using python programming language and with the help of the various libraries it provides to make visualizations easier in data mining and machine learning algorithms (Asuncion & Newman, 2013). The cluster performance can be measured by evaluating intra-cluster performance Sil- Houte Index and Calinski - Harabasz Index, Iterations, Time (in seconds). Various graphs and a statistical page is provided to the admin with administrator privileges that can be accessed only with a secret code.

2.1 Customer Segmentation

Customer Segmentation is carried out using the below mentioned algorithms.

A. DBSCAN Clustering Algorithms

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise, which is a typical representation of density algorithms. The data points are divided into multiple distinct bunches or groups using the unsupervised learning technique or clustering analysis, such that the data points within the same group have similar properties and the data points within different groups have somewhat dissimilar at- tributes. The reason we prefer DBSCAN clustering over the

various other clustering algorithms is because unlike K-means clustering algorithm, we need not specify the number of clusters in this algorithm. It also produces a more reasonable result over a variety of different distributions. Density-Based Clustering concept that is responsible of forming a cluster in data space that is a contiguous region of high point point density, separated from other similar clusters by contiguous regions of low point density, in an unsupervised learning approach that result in unique groups/clusters in the data.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers. The DBSCAN algorithm uses two parameters: minPt: The minimum number of points (a threshold) clustered together for a region to be considered dense. eps (): A distance measure that will be used to locate the points in the neighbor- hood of any point. The above parameters are directly related to Density Reachability and Density Connectivity. Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it. Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster.

There are three types of points after the DBSCAN clustering is complete:

Core — This is a point that has at least m points within distance n from itself.

Border — This is a point that has at least one Core point at a distance n.

Noise — This is a point that is neither a Core nor a Border. And it has less than m points distance n from itself.

Implementation

Input the minimum radius along with the mini- mum density threshold minPt. The data is then sequentially read into a text file and the results are plotted along a X and Y co=ordinates that hold the Point Structure. The final step involves determining if the point is a core point. We first read a point (data from the dataset) and analyze if the point is marked or not. If it is marked, which means it is part of a cluster, the distance between this point and all its immediate neighbors or the Density Connectivity is found. Based on how close the value is to that of the radius. The clusters are then created and a merged result list is generated.

B. Fuzzy Clustering

Fuzzy c-means (FCM) (Lee, 2018; Tsai et al., 2002) is a data clustering technique in which a data set is grouped into N clusters with every data point in the dataset belonging to every cluster to a certain degree (Babu et al., 2018). For example, a data point that lies close to the center of a cluster will have a high degree of membership in that cluster, and another data point that lies far away from the center of a cluster will have a low degree of membership to that cluster.

It starts with a random initial guess for the cluster centers; that is the mean location of each cluster. Next, FCM assigns every data point a random membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, FCM moves the cluster centers to the correct location within a data set and, for each data point, finds the degree of membership in each cluster. This iteration minimizes an objective function that represents the distance from any given data point to a cluster center weighted by the membership of that data point in the cluster.

Each cluster is defined as the largest set of dense connected points, and each point in a cluster has a minimum number of neighbors of MinPts or greater in a given radius Eps. The time complexity of DBSCAN is O(n*logn).

3. RESULT AND FUTURE SCOPE

A. Results

The primary results include various graphs and analytics that are represented on a website to the administrator with only admin privileges. In this paper DBSCAN algorithm identifies objects based on bivalent logic (Deng, 2020). The existing techniques have not focused on the hybridization of DBSCAN with fuzzy if then rules (figure 2).

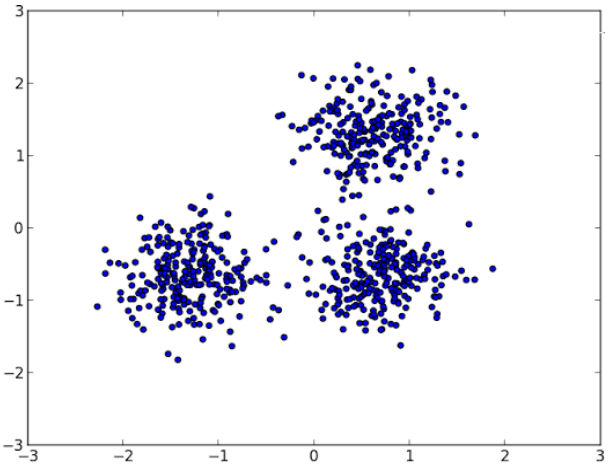

Figure 2. DBSCAN Clustering results

DBSCAN will be combined with fuzzy if then rules (Beri & Kamaljit, 2015; Lee, 2018). The hybridization will allow DBSCAN to decide the cluster in more efficient manner (figure 3). The simulation of the proposed approach will be designed in python. The cluster performance can be measured by evaluating intra- cluster performance Silhouette Index and Calinski-Harabasz Index, Iterations, Time Taken (in seconds).

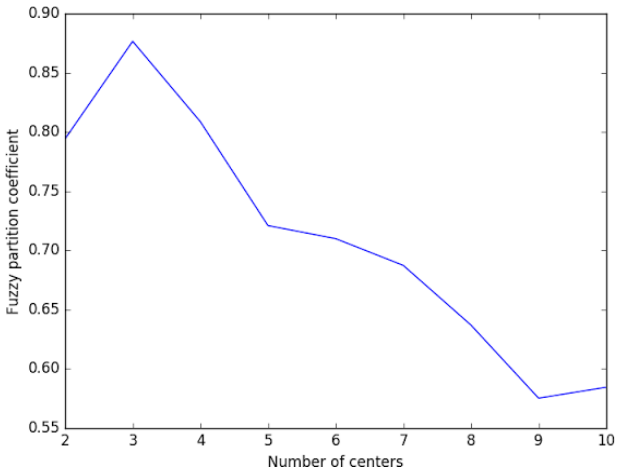

Figure 3. DBSCAN Clustering results

In this paper, we have tried to implement a model that will perform clustering techniques apart from the existing clustering algorithms. We also perform a comparison check against the accuracy score of all the conventional algorithms mentioned above with the accuracy score of the model that we have developed. This enables the users to understand the credibility of our enhanced model.

B. Future Scope

The future Scope involves creating a comprehensive system to enable these companies or E-Commerce to directly send targeted customers periodic advertisements and be a more evolved system to not only display the data in a static web application (Sari et al., 2016; He & Li, 2016). However, the results obtained from our model are purely visualizations which will visualize the consumers of an e-commerce into clusters according to the information provided in the dataset. The future improvements would be to generate insightful information from the already available visualizations. This would help generate reports which could be useful for further stages.

4. CONCLUSION

Customer segregation is a way to improve customer communication, customer awareness, customer service to build effective communication. Customer segregation is required for use by businesses to maximize profits. Potential customer data can be used to provide the customer with features that include e-commerce services such as online media buying and selling. Knowing the customer base is crucial to any business. The segments result in a better understanding of where the company is not meeting the customer demands or needs and these observations will help reduce the churn rate of the customers.

This paper discusses a few aspects of performing customer segregation, namely: Customer segregation is the task of separating customers or an object into groups with similar characteristics. Data needed to differentiate between internal and external data sources and consumers. External data comprises cookies and server logs, whereas internal data contains statistical data and data purchase history. External data can be obtained through a web server or other source, whereas internal data can be accessible on a website where a client registers or transacts.

Customer Separation Methods can be categorized into sim- ple method, RFM method, targeted method, and unchecked method. In Targets, the researcher focuses on one exception, either product or purchase. An unsupervised method was used when the integration process researcher had many variables

The Customer Classification process could be simplified to describe business purpose, data collection, data preparation, dynamic analysis, data processing, and performance evaluation.

References:

Asuncion, A., & Newman, D. J. (2013). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science.

Babu, B. S., Prasanna, P. L., & Vidyullatha, P. (2018). Customer Data Clustering using Density-based algorithm. *International Journal of Engineering & Technology*, *7*(2.32), 35-38.

Beri, S. K., & Kamaljit. (2015). *Hybrid framework for DBSCAN algorithm using fuzzy logic*. In 2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management.

Deng. (2020). *DBSCAN Clustering Algorithm Based on Density*. In 2020 7th International Forum on Electrical Engineering and Automation (IFEEA) (pp. 949-953). doi:10.1109/IFEEA51475.2020.00199.

He, X., & Li, C. (2016). The research and application of customer segmentation on e-commerce websites. *2016 6th International Conference on Digital Home (ICDH)*, 203–208. https://doi.org/10.1109/ICDH.2016.050

Lee, K. (2018). A Comparative Study between Fuzzy C-Means Algorithms and Density-based Spatial Clustering of Applications with Noise. *International Journal of Engineering Technology*, *7*(3.33), 131-133.

Lee, S. (2018). A Hybrid Framework using Fuzzy if-then rules for DBSCAN Algorithm. *International Journal of Computational Intelligence Research*, 403-412.

Sari, N., Nugroho, J., Ferdiana, L., & Santosa, R. P. (2016). Review on Customer Segmentation Technique on E-commerce. *Advanced Science Letters*, *22*, 3018-3022. doi:10.1166/asl.2016.7985.

Tsai, C.-F., Wu, H.-C., & Tsai, C.-W. (2002). *A new data clustering approach for data mining in large databases*. In Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks, 2002 (I-SPAN'02). IEEE.

Snehalatha N
Jss Academy of Technical Education,
Karnataka,
Bengaluru,
India
snehalathan@jssateb.ac.in

Mohana Kumar S
M S Ramaiah Institute of technology
Bangalore
Department of Computer Science and Engineering
Bangalore
India
mohanaks@msrit.edu
ORCID 0000-0003-4143-9450

Vitasta Kachroo
Jss Academy of Technical Education,
Karnataka,
Bengaluru,
India