Review on Customer Segmentation Methods Using Machine Learning

Rishi Gupta, Tarun Jain, Aditya Sinha, and Vishwas Tanwar

1 Introduction

There are several ways a business tries to attract new customers all the while trying to retain its current customers. A few decades back, these businesses practiced what was known as mass marketing—in which companies tried to sell the most popular product to all their customers or they practiced product differentiation—in this, companies offered a variety of products to a large market.

But, as technology rapidly evolved, companies moved to a newer approach—personalizing the products and targeting them to a specific market segment. Customer segmentation (or market segmentation as it is widely known) is the most approachable method for obtaining the above result. Understanding their customers and the market has allowed businesses to service each of their customers with care and personalization and proved the value of focusing on them. The oldest and most traditional form of segmentation is demographic segmentation, although with recent developments in big data, newer forms have emerged [1].

These approaches have also taken into consideration buyer attitudes, motivations, patterns of usage, and preferences [2]. The platform that utilizes segmentation the most is the e-commerce industry. Since the birth of e-commerce, sellers are constantly looking for ways to expand their reach while also maintaining a stable relationship with their frequent customers. And, in the age of booming social media, consumer data is as readily available as daily bread. Companies like Facebook, Google sell billions of dollars worth of consumer data to corporations which are then used as a basis of market segmentation. These e-commerce giants like Amazon, Flipkart target each segment with personalized promotional offers as well as personalized advertisements driving in clicks and eventually large amounts of profits. In this

R. Gupta (✉) · T. Jain · A. Sinha · V. Tanwar
Manipal University Jaipur, Rajasthan 303007, India
e-mail: genieousrishi@gmail.com

R. Agrawal et al. (eds.), *International Conference on IoT, Intelligent Computing and Security*, Lecture Notes in Electrical Engineering 982,
https://doi.org/10.1007/978-981-19-8136-4_33

Fig. 1 Market segmentation

article, we shall look at the currently available approaches to customer segmentation and their merits and demerits [14].

2 Background

The applications of customer segmentation are irreplaceable. It has now become the heart of product marketing and strategy in any industry. It is an indispensable tool for organizations to understand the market, whom to target with what product, and how to optimize the marketing strategy [13].

With the increasing popularity of social media, consumer data is readily available for businesses to use and profile the customers, to understand them better, and to act accordingly.

Let us say a website has a new user. The user would be asked to register an account on the website, for which they would be offered a discount coupon. Instead of offering every new user the same coupon, the website could—with the help of customer segmentation—give out targeted discount coupons. This will increase the likelihood of that user buying a product rather than window shop. It is also useful for planning customer communications as it allows for personalized recommendations for each group (Fig. 1).

3 Customer Segmentation

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters) [7] (Fig. 2).

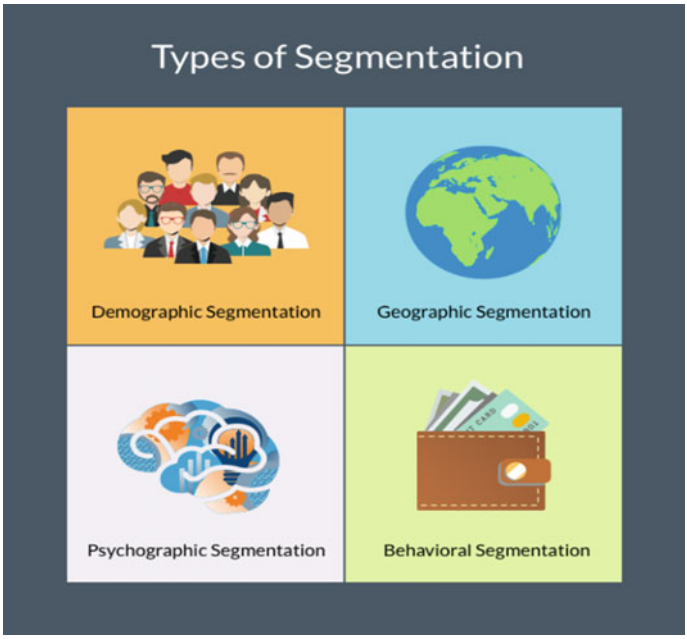

Fig. 2 Types of customer segmentation

Domains: Customer segmentation can be applied to many domains. These domains are defined by the type of data they are using for segmentation. The most prominent of these are listed below.

3.1 Demographic Segmentation

The oldest and the most traditional form of market segmentation, demographic customer segmentation, focuses on the structure of the population based on everyone's current living status. Age, gender, income, education, marital status, etc. are generally considered as demographic data. Segmentation performed on such data often yields basic groups which are the go-to for target marketing—if any other type of data is not available.

Most frequent segments formed after demographic segmentation include:

Students—this group involves unmarried, young adults/adolescents who have little or no income.
Parents—this group usually has married, middle-aged adults who have kids (dependents) and have a stable income.
Women restarting their careers—these middle-aged women tend to have a low academic background and low income.

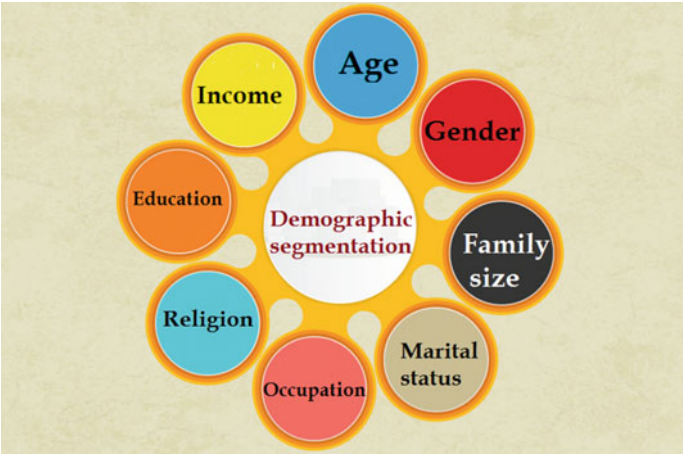

Fig. 3 Demographic customer segmentation

Senior citizens—this group includes old people who either live on a monthly pension or rely on their children, and they tend to live alone and so on.

These segments have different spending habits which can be directly related to their living scenario (Fig. 3).

3.2 Psychographic Segmentation

These are subjective attributes about a person that is not as easily available as other forms of data.

In this, data cannot be directly fed into an algorithm and is expected to yield results. The data and results must be interpreted by a psychological professional. Just as demographic segmentation emphasizes a person's current living situation, psychographic segmentation utilizes a person's real-life behavior which pertains to things like personality traits, values, attitudes, interests, lifestyles, subconscious, and conscious beliefs, motivations [2], etc. (Fig. 4).

3.3 Behavioral Segmentation

While demographic and psychographic segmentation focuses on who a customer is, behavioral segmentation focuses on how the customer acts [13]. The most widely used and most efficient form of segmentation—behavioral—relies on how the customer acts toward the company's services or services related to it. This involves

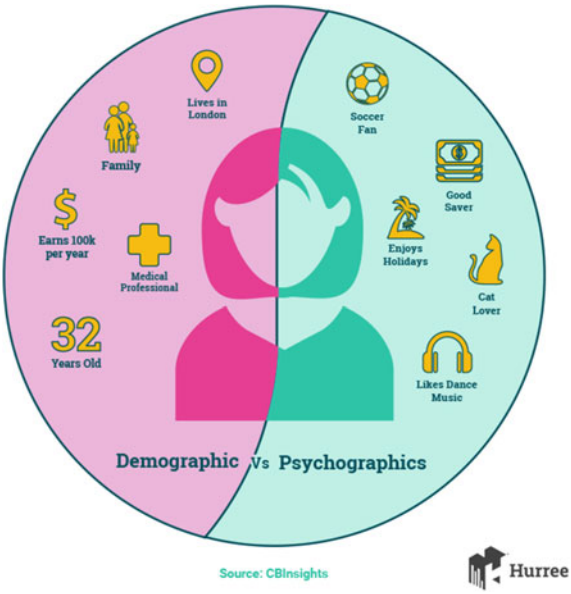

Fig. 4 Psychographic customer segmentation

gathering data, performing analysis, and targeting the customers in real time. Therefore, this type requires higher computational power than others. Purchasing habits, spending habits, user status, and brand interactions are the main attributes companies look toward for behavioral analysis (Fig. 5).

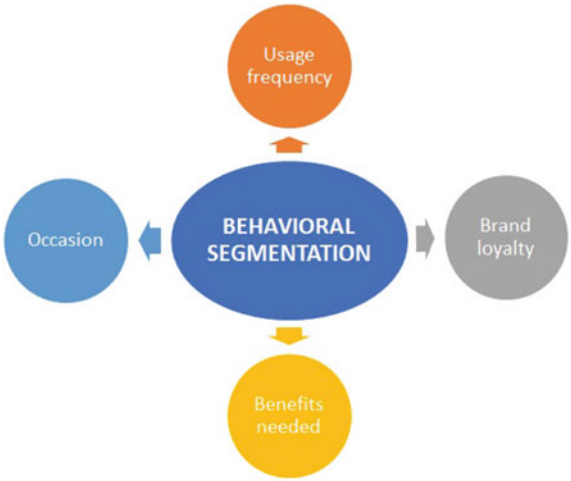

Fig. 5 Behavioral customer segmentation

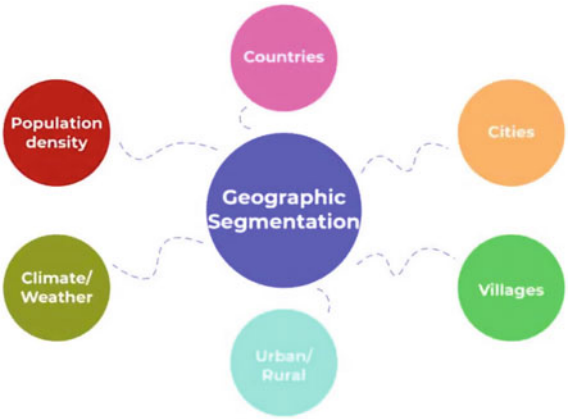

Fig. 6 Geographic customer segmentation

3.4 Geographic Segmentation

The simplest form of segmentation involves gathering the location of a user and segmenting them based on it. Geographic segmentation utilizes a customer's ZIP code, city, country, a radius around a certain location, climate, urban, or rural [14] attributes for segmentation.

Geographic segmentation is an effective methodology used by organizations with large national or international markets to better understand the location-based attributes that comprise a specific target market [11] (Fig. 6).

4 Pros and Cons

4.1 Demographic Segmentation

Pros:

- Demographic variables are quite easier to collect and measure when compared with other segmentation techniques.
- Targeting is usually more straightforward when using demographics as a metric—for example, you can target consumer groups, such as women or men between the ages of 40 and 50.
- Consumer profiles are more comprehensible across the board, making it easier to develop strategy among various departments (sales, customer service).

Cons:

- The model offers limited insight for marketers. Similar demographics among customers do not always mean that they all have the same needs.

- Because customers can have varying needs, a 'one-size-fits-all' approach to consumers based solely on broad demographics can make your marketing message ineffective.
- Skewed or problematic demographic data within a given region can produce unreliable assumptions, which can reduce the accuracy of your marketing methods.

4.2 Behavioral Segmentation

Pros:

- Marketers can build targeted consumer segments based on their responsiveness to certain product categories, promotion types, or path-to-purchase preferences.
- Monitoring and understanding the behavior of consumers online has become easier due to advances in data collection and tracking technologies.

Cons:

- While consumer behavior can be tracked, it is not always easy to pinpoint the motivations behind those behaviors with segmentation models, because they can vary greatly from person to person.
- Behavioral segmentation is often based on complex data constructs that are not always easy to understand without the help of a large team of data scientists and marketers.

4.3 Psychographic Segmentation

Pros:

- Marketers can get some insight into customer motivations.
- Psychographic segmentation can help brands in executing more emotive marketing to highly responsive segments.

Cons:

- Psychographic surveys which are self-reported can be inaccurate.
- Although marketers can use predictive modeling to create statistical projections, the accuracy of all these predictions firmly depends on the quality of the data used.
- Marketers need clear rules about how to interpret psychographic data to ensure a consistent approach among the individuals or departments that engage in customer segmentation analysis.

4.4 Geographic Segmentation

Pros:

- People in different communities have different needs. Something useful for someone in a more rural area, like gardening supplies, may not have any appeal to city dwellers. Separating consumers by where they live can ensure that marketers are only targeting those that may want or need your products or services.
- Marketers can adjust advertised pricing based on the cost of living of the customers that they are targeting.
- You can tailor products based on the local preferences of your customers. For example, certain clothes are more popular in some Canadian regions than others.
- You can use this segmentation to target customers in new areas where you want to grow your business.

Cons:

- This type of audience segmentation makes the assumption that people living in the vicinity of each other have got similar needs, which is not always true. This is true if you are targeting a wider geographic area.

5 Methodology

Although customer segmentation can be done through several methods, the structure is similar for all.

Prerequisites: Segmentation involves processing data; though any system can run the algorithms, larger data requires higher computational power.

5.1 Data Collection

There are many ways to collect consumer data, but since each domain requires a different type of data, data collection methods vary between them.

Demographic data can either be directly asked from the customer through surveys or be bought from data resources such as social media, census. Psychographic data is the least readily available data and contains sensitive personal data; therefore, it must be directly obtained from the users. Behavioral data is obtained through analysis of users' actions—be it clicks on advertisements or purchase habits, etc. Lastly, obtaining geographic data is a one-step process since it includes only the location.

5.2 Data Preprocessing

The raw data collected needs to be processed before analysis and segmentation. This involves removing outliers, null values, duplicates, and corrupted data. Principal component analysis also needs to be performed to compute the principal component and change the dimensionality of the data by selecting the first few principal components and ignoring the rest.

This ensures that machine learning algorithms are not overburdened and makes data visualization easier.

5.3 Data Analysis

In statistics, exploratory data analysis is an approach to analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods [6]. EDA is a valuable tool that is used to analyze the data at hand and develop a model based on which the machine learning algorithm is performed.

Mainly performed in two ways: univariate analysis—for data consisting of a single variable—and multivariate analysis (aka relationship analysis)—for data consisting of more than one variable.

Data visualization is a major step involved in EDA, different forms of visualization of data include—box plots and histograms (for univariate analysis)—scatterplot, and bar charts (for multivariate analysis) (Fig. 7).

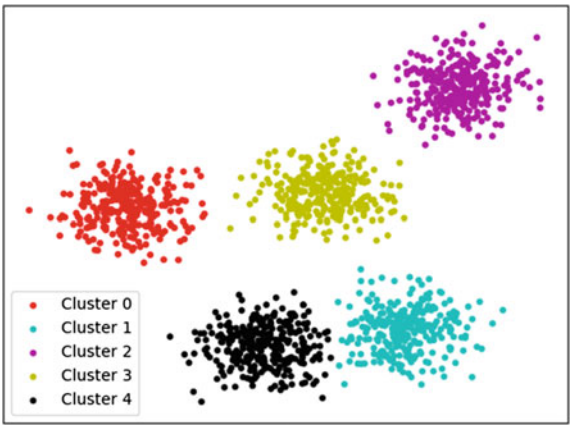

Fig. 7 Exploratory data analysis

5.4 Segmentation

This is the main step in the whole process. This involves performing machine learning algorithms to obtain the segments and various fields attributed to it. There are several ways to perform segmentation (some do not involve machine learning), but to achieve the highest efficiency, some are more suitable to a type of segmentation than others.

But, before segmentation is performed, we need to find the optimal number of clusters or segments. This can be done using hyperparameter tuning and elbow method.

Parameters that define the model architecture are referred to as hyperparameters, and thus, this process of searching for the ideal model architecture is referred to as hyperparameter tuning [8]. When it comes to clustering algorithms, the hyperparameter in question is the number of clusters that need to be made. This is performed with the help of the elbow method.

These methods are discussed in the next section (Fig. 8).

6 Methods Available

There are many ways to perform customer segmentation; in this article, we will mainly focus on how machine learning algorithms compare in terms of usability and efficiency for segmentation.

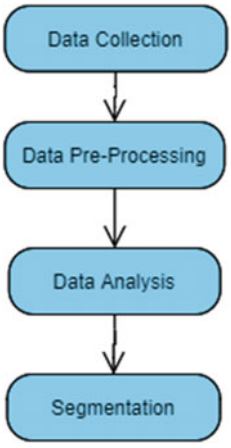

Fig. 8 Customer segmentation process

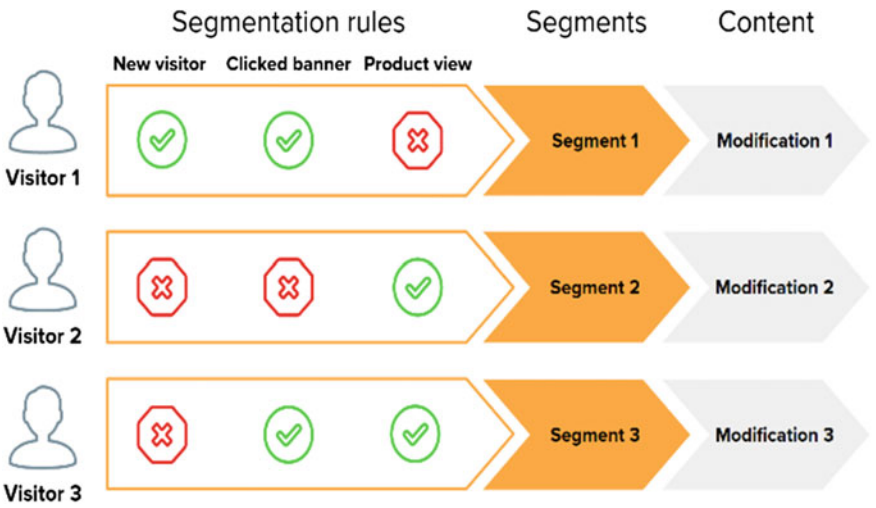

Fig. 9 Rule-based segmentation

6.1 Rule-Based

Rule-based segmentation is based on criteria such as Boolean logic or thresholds and is often two-dimensional; this is traditionally the easiest type of targeting for many professionals, as one can do it by filtering in Excel, e.g., marketers have traditionally segmented customers based on heuristics such as the industry, company size (in B2B) or age, income, etc. in B2C [3]. This follows an 'if A then B' type of rules and forms an algorithm (Fig. 9).

6.2 Supervised Clustering with Decision Tree

This method uses a specific target or dependent variable, and the target would predict differences in independent variables (input). Data utilized in this method is previous purchase patterns and customer demographic. The algorithm used is the decision tree with the target on their nodes. According to Baer, although this method connects the target with the other customer attributes, it shows only one aspect of customer behavior [4]. This is used for demographic and behavioral segmentation [13, 14].

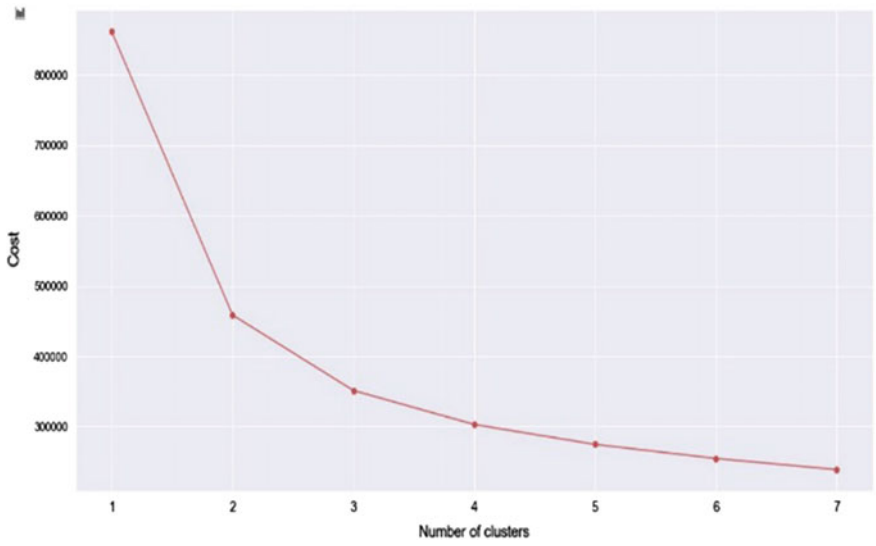

Fig. 10 Elbow method

6.3 k-means

First, for each centroid, the algorithm finds the nearest points (in terms of distance that is usually computed as Euclidean distance) to that centroid and assigns them to its category. Second, for each category (represented by one centroid), the algorithm computes the average of all the points which has been attributed to that class. The output of this computation will be the new centroid for that class [5]. To find the optimal number of clusters, we use the elbow method (hyperparameter tuning). The elbow method is a heuristic used in determining the number of clusters in a dataset. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use [9]. This algorithm is the best algorithm for behavioral segmentation (Figs. 10 and 11).

6.4 k-prototype

K-prototype is a clustering method based on partitioning. Its algorithm is an improved form of the k-means and k-mode clustering algorithm to handle clustering with mixed data types [10]. It is used for datasets that have both numerical and categorical data types. For numerical data types, it uses the k-means algorithm and calculates the average after each iteration. For categorical data types, it uses the k-mode algorithm which calculates Euclidean distance from each cluster center and calculates modes after each iteration. To find the optimal number of clusters, we use the elbow method.

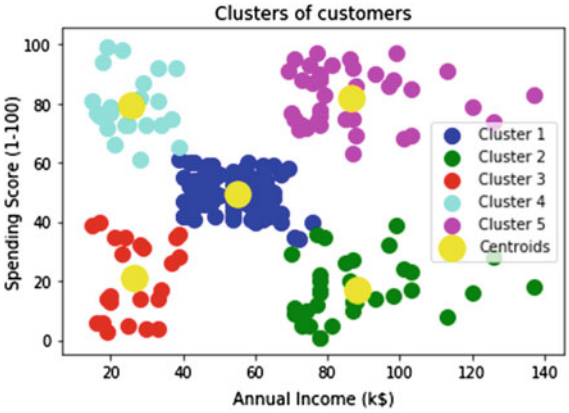

Fig. 11 k-means clustering

6.5 k-medoid (PAM)

A problem with the k-means and k-means++ clustering is that the final centroids are not interpretable. The idea of k-medoids clustering is to make the final centroids of the actual data points. This result makes the centroids interpretable [12].

The working is the same as the k-means, and the only difference comes in the updating centroids step in between iterations. Instead of computing the mean of points in a cluster (like in k-means), we swap the previous centroid with all other (m−1) (if there are m-point in a cluster) points from the cluster and finalize the point as new centroid that has a minimum loss [12] (Table 1).

7 Conclusion

We discussed several ways a company profits from customer segmentation, its methodology, and various ways it can be performed. Even though there are many types of and any ways to perform customer segmentation, the purpose of each remains the same to personalize a business service and experience for its customers. For the last two decades, market segmentation has taken over every big and small company's marketing scheme. Every corporate has a data analyst in its marketing department. It is left to be seen what the future holds for artificial intelligence incorporated with business.

Table 1 Methods for customer segmentation

Method	Working	Advantage	Disadvantage
Business rule-based	Traditional targeting by filtering in Excel	Easy to apply, use database query	Tends to rely on heuristics developed over time and is slow to adapt to changes [3]
Supervised clustering with decision tree	Uses a dependent variable to predict differences in independent variables	Classify customers according to target	Uses one variable to cluster
k-means clustering	Uses unlabeled data to find a significant number of clusters	Uses' any number of customers' attributes	Speed of computation depends on k-values
Hierarchical clustering	Initially treats each observation as a cluster, then repeatedly merges two similar clusters	Relatively straightforward to program, no need to specify the number of clusters required	Very high time complexity compared to k-means
PAM clustering (k-medoid)	Finds a sequence of objects called medoids that are centrally located in clusters, and clusters are constructed by assigning each observation to the nearest medoid	Effectively deals with the noise and outliers present in data	Since the first k-medoids are chosen randomly, different results may be obtained on the same dataset
k-prototype	Combines working of k-means (for numerical values) and k-modes (for categorical values) algorithms	Can be applied to datasets with mixed data types, whereas k-means can be applied to only numerical data types	Unclear what weights have to be given to categorical variables

References

1. Marcus C (1998) A practical yet meaningful approach to customer segmentation, 1st ed. J Consumer Mark 15
2. Yesbeck J (2021) Types of market segmentation, 18th May 2021. https://blog.alexa.com/types-of-market-segmentation/
3. Elizabeth (2019) Customer segmentation: rules-based vs. K-means clustering, 1st ed. d3mlabs, Denmark
4. Sari J, Nugroho L, Ferdiana R, Santosa P (2011) Review on customer segmentation technique on E-commerce, 1st ed. American Scientific Publishers, Indonesia
5. Alto V (2021) Unsupervised learning: K-means vs hierarchical clustering, 18th May 2021. https://towardsdatascience.com/unsupervised-learning-k-means-vs-hierarchical-clustering-5fe2da7c9554
6. "Exploratory data analysis" Wikipedia. Wikimedia Foundation, 7 May 2021. en.wikipedia.org/wiki/Exploratory_data_analysis

7. “Cluster analysis” Wikipedia. Wikimedia Foundation, 2 June 2021. en.wikipedia.org/wiki/Cluster_analysis
8. Jordan J (2021) Hyperparameter tuning for machine learning models, 4th June 2021. https://www.jeremyjordan.me/hyperparameter-tuning/
9. “Elbow method (clustering)” Wikipedia. Wikimedia Foundation, 11 December 2020. en.wikipedia.org/wiki/Elbow_method_(clustering)
10. Aprilliant A The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical), 4th June 2021. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb
11. The Benefits of Geographic Segmentation. Alchemer, 7th June 2021. https://www.alchemer.com/resources/blog/benefits-of-geographic-segmentation/
12. Kumar S (2021) Understanding K-means, K-means++ and, K-medoids clustering algorithms, 7th June 2021. https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms-ad9c9fbf47ca
13. Kaminskyi A (2021) Information technology model for customer relationship management of nonbank lenders: coupling profitability and risk, pp 234–237
14. Monil P (2020) Customer segmentation using machine learning. Int J Res Appl Sci Eng Technol 8(6):2104–2108. https://doi.org/10.22214/ijraset.2020.6344