# Market Segmentation, Targeting, and Positioning Using Machine Learning

**Harish Paruchuri**

Senior AI Engineer, Department of Information Technology, Anthem, Inc., **USA**

Corresponding Email: harishparuchuri9999@gmail.com

## ABSTRACT

The occurrence of numerous rivals and industrialists has created a lot of pressure between rivalry companies to go in search of new customers and at the same time working to maintain the existing ones. Owing to this, the necessity for a new customer service policy or strategy becomes essential irrespective of the business size. Moreover, the capability of any company that wants to remain in existence and growth as well needs to understand its clients and also delivers necessary customer support to make available targeted client services and develop branded clients' service policy. This understanding is conceivable via organized client service. The respective segment devises consumers who share similar market potentials. Big data philosophies and machine learning devise appropriate means of promoting better recognition and approval of computerized client segmentation methods in good turn of outdated business analytics that frequently lead to any meaningful approach that can keep or source for a new customer while the client disreputable is very growing by the day. This study makes use of the k-means clustering algorithm for this determination. The Sklearn public library was established for the k-Means algorithm and the package is competent by means of a 100-model two-parameters dataset generated or collected from the retail trade. Features of an average figure of client buying and average figure of periodic consumers.

**Key Words:** Big data mining; machine learning; k-Mean algorithm; customer segmentation

**Source of Support:** None, **No Conflict of Interest:** Declared



## INTRODUCTION

For the past years, high rivalry among organizations and ease of use of huge weighbridge chronological data has given rise to the general use of data withdrawal methods to discover life-threatening and considered information, which is concealed in business' information (Blanch et al 2019). Data withdrawal is a procedure of isolating coherent data from a dataset and showing it in a human-availability method for resolution sustenance. Data withdrawal methods discriminate fields like artificial intelligence, statistics, data structures, and machine learning. Data withdrawal applications take account of, on the other hand, are not limited to

bioinformatics, meteorological conditions prediction, scam exposure, financial testing, and client dissection. Client dissection is a set of commercial client ignoble known as client fragment such that each client's fragment has clients who share the same market salient features (Puwanenthiren, 2012). These disparities are centered on the parameters that openly or unopenly impact the business or market such as product expectation or preferences, behavior, locations among other parameters. The significance of client segmentation comprises the capability of a firm to brand market models that would be appropriate for each dissection of its clients (Puwanenthiren, 2012). Backing for corporate resolution depends on risky atmosphere like credit associations with its clients; detecting products correlated to discrete constituents and in what way to bring about suppler and demand influence. Interaction and interdependence amidst products, consumers, or clients and products are shown that the business could not be alert of. The capability to estimate research questions and offer hints to discover keys (Griva et al., 2018).

Submerged in a database of combined data shown to be operative for identifying restrained relationships or patterns. This manner of learning is categorized under supervised learning. Incorporation algorithms consist of the K-nearest, K-Mean, Sorting map, and other algorithms (Vadlamudi, 2017). These algorithms do not require prior knowledge of the data but are capable to detect classes in them by recurrently equating response models, as long as stationary capacity in training instances is accomplished according to subject matter or procedure. Each group of data projects very close resemblances, on the other hand, vary immensely from the data opinions of other sets. Incorporation has immense uses in model detection, image testing, and bioinformatics, and the likes (Paruchuri, 2015; Vaishali and Rupa. 2011). The scalar archive of the K-Means algorithm was designed, and training was in progress based on standard silhouette-score with two property groups of 100 training models seen in the retail line of work. Subsequently, many indications, four stable intervals, or client distinctions were detected. Out of the 4, two parameters are well-thought-out in grouping with the number of things a client buys per month and the average number of clients per month. From the dataset, 4 clients or groups are classified and mark out as presented as cluster matrices 1 to 4.

## PROBLEM STATEMENT

The occurrence of numerous rivals and industrialists has created a lot of pressure between rivalry companies to go in search of new customers and at the same time working to maintain the existing ones. Owing to this, the necessity for a new customer service policy or strategy becomes essential irrespective of the business size (Griva et al., 2018). Also, due to the demographic data of an individual, how can a mail-order business proficiently secure new clients. Giving to this statement, a comparison between existing client data and the overall populace data in some way to infer a correlation. A manual approach of performing this comparison between statistics and the clients as well the overall populace (Paruchuri, 2017). However, this approach of testing or analysis will give rise to several outcomes which still require further analysis to arrive at the ultimate policy. This method is time-consuming and by the moment the analysis is completed, rivalry organizations will populate the market by winning more customers. It was to this advent that machine learning was introduced to address these challenges through machine learning algorithms.

### The objective of the Study

The fundamental aim of this study is to categorize client fragments in a commercial business utilizing the data withdrawal technique. In this study, the k-means clustering algorithm remained employed in the client fragment.

## LITERATURE REVIEW

### Client Sorting

Over an inordinate length of time, the profit-making organization has to turn out to be more viable, as firms try to meet the cravings and requests of their clients, draw new clients, and thus advance their businesses (Puwanenthiren, 2012). The job of detecting and attaining the desires and needs of every client in the organization is quite challenging. This is as a result of client's differences either according to wants, needs, size, demographics, taste, and potential, etc. it is a very outdated approach to treat all clients with the same methods or equally. This problem brought to light customer dissection or market fragment, where customers are grouped into segments, where participants of each subgroup display the same market performances or features (Neogy & Paruchuri, 2014). Consequently, client segmentation is a method of distributing the market into local clusters.

### Big Data

In recent times, big data investigation has gain popularity. Big data is defined as a term used in labeling a huge amount of prescribed and familiar data that is not possible to handle with traditional approaches and algorithms. Organizations take account of billions of data concerning their clients, operations, suppliers, and millions of on the inside linked sensors are directed to the actual domain on objects like smartphones, sensing, cars, manufacturing, and communication data (Vadlamudi, 2015). The capability of enhancing predictions, low-cost involvement, upturn productivity, and develop various sectors like weather forecast, traffic management, disaster avoidance, scam control, finance, national security, business relations, healthcare, and education. Big data is majorly come across in 3 categories (3Vs) namely; variability, volume, and speed. In addition to the 3Vs is another 2Vs which are accessible-validity and price, hence totaling 5Vs (Paruchuri, 2017).

### Data Depository

The gathering of data is a procedure of collecting and computing information in contradiction of targeted variations in a reputable organization, which allows one to attempt pertinent queries and estimate the outcome (Jain et al., 1999). Collection of data is a crucial aspect in all research irrespective of the field of study, be it social and physical sciences, business, and humanities. The goal of all data assemblage is to attain feature indication that leads the exploration to design substantial and ambiguous or confusing responses to the queries offered.

### Clustering data

Clustering is a procedure of assembling data into a dataset according to approximately cohesions. Many algorithms can be functional to the dataset according to the delivered situation (Sulekha, 2011). Nevertheless, there is no general clustering algorithm at the moment, thus it turns out to be essential to select the right clustering methods. Paruchuri, (2017) discussed the implementation of the 3 clustering algorithms utilizing the Python scalar archive.

### K-Means

This is an algorithm whereby one of the utmost prevalent ordering algorithms is involved. It depends on Centro, where each data opinion is positioned in one of the coinciding ones that are sorted before the K-algorithm. Clusters have twisted that match to unseen models in the data that deliver the essential data to assist in the implementation process (Hong and Kim, 2011). There are diverse methods to collect K-means, but the elbow technique is the approach we will be using in this study.

 

## METHODS

We collected data from the UCI machine learning repository. It is a group of geographical data, together with all businesses that take place between February 1st, 2010 and December 9th, 2011 in an unregistered United Kingdom broker. The organization majorly market single gifts to Tom, Dick, and Harry at the same time. Although several of the firm's clients are shopkeepers (McKinsey Global Institue, 2011). The database has eight features. These attributes include;

- *Invoice Number*: usually invoice figure by default should be six-digit and totaling number is allotted distinctly for each operation. For instance, if the invoice number starts with 'C', it means cancellation.
- *Stock code*: item name usually consists of a five-digit totaling number is given only to each item.
- *Definition*: Item name should be addressed only by name.
- *Price*: The value of each item number.
- *Invoice*: the time and date of the bidding. The time and date should be provided with digits for each operation.
- *Unit Price*: a price is a unit. Price, item price per piece of degree.
- *Customer*: the number assigned for each customer and the name, should be five-digit for each client.
- *Country*: the name of the counter where the consumer is located is very important.

This study deploys many steps to obtain a precise outcome. It consists of characteristics of Centro's initial stage, sorting stage, and update stage, which are the best common step in K-Means algorithms.

**Collection of Data**

This is the groundwork stage of the research. The characteristics generally support the upgrade of all data objects at a standard proportion to enhance the performance of clustering algorithms (Jain et al., 1999). Each data opinion differs commencing from rating 2 to +2 combination methods that consist minimum to maximum, z-point are customary of z-signing policy applied to make things unequal prior the dataset algorithm used the K-Means algorithm.

**Customer classification Techniques**

There are numerous methods to segment, and they differ in strictness, the requirement of data, and drive. Below are generally applied means, but is not limited to this (Vishish and Rupa, 2011). Some paper discusses on artificial neural connections, complex kind of ensemble and constituent part detection, nevertheless are not involved as a result of limited detection. Each succeeding segment of this paper will include a basic description of the techniques, all together with an encryption instance for the manner deployed. In case you are not endowed with the skill of coding samples, you can skip the stage and you need to get a decent switch on each of the four subgroups presented in this paper (Vadlamudi, 2015).

**Cluster analysis**

Cluster analysis is an incorporation or amalgamation, the method to customers according to their relationship. We have two core forms of definite cluster analysis in marketplace strategy: graded cluster analysis, and cataloging (Miller, 2015). For the time being, we will deliberate by what means to categorize clusters, known as k-methods.

**K-Means encounter**

The K-means grouping algorithm is a procedure frequently deployed in drawing perceptions into arrangements and variances in a database (Vishish and Rupa, 2011). In advertising, it is frequently deployed in building client clusters and recognize the performance of each single division. In trying to construct a muster model in Python's atmosphere.

**Centroids commencement**

Some chosen or beginners were shortlisted. The 4 shortlisted hubs for this study are presented in diverse sizes and they were chosen through the Forgy technique. Forgy approach explains data opinion or points in randomly chosen group centroids by k, (k=4 as in the present case). The methodological overview was the code that was developed in the Jupiter manual using Python 3x and Python packages for editing, analyzing, processing, and visualizing data (Vadlamudi, 2018). Although, some of the codes presented in this paper are generated using the Github application of a book known as Hands-on Data science for marketing. You can access the book on OilReilly or Amazon if you are a client. The exposed basis data cost employ in the code originates from Irwin's machine learning depository.

## RECOMMENDED MODEL

*a. Import Applications and data*

To commence, we introduce the needed application to handle our data analysis and then the Microsoft Excel Spreadsheet data file (Jain et al., 1999). In case you want to practice the process you can download the same data from the UCI database.

*b. Data Clean-Up*

The next thing to do after introducing the application and data is data cleaning. The data obtained was not that supportive, thus the data set was clean up and organize to suit our purpose and also produce further actionable insight.

*c. Data Normalization*

The k-means region is very sensitive to the information scale employed such that grouping or clustering algorithms can generalize the information (Vaishali and Rupa, 2011). A study also describes the reason for normalization or standardization is important for data employed in K-means clustering (Vadlamudi, 2015).

*d. Choose the Best number of Clusters*

It is true, we aim to run cluster testing, but it is crucial to determine how many clusters we want to employ in the analysis. We have many methods to choose the number of clusters to employed, but we are only covering two approaches in this paper namely; the silhouette coefficient and elbow techniques (Sulekha, 2011).

*e. Silhouette Clustering*

The silhouette clustering discusses the extent to which validation and interpretation consistency in the data system is maintained. This approach displays a picture of in what manner an individual item is systematized (Lakshmi Narayana et al., 2012). The score of a silhouette is a degree in what manner roughly is more like in its combination than

 

other clusters. The silhouette ranges from -1 to +1, anywhere a developed worth designates that a thing equals its group correctly and is likened to nearby clusters. If numerous items take a high rate, the additional shape is suitable. If the greatest points take a number or a negative number, the established scheme may take too numerous or too few groups. The silhouette may be designed with any detachment metric, such as the Euclidean detachment or the Manhattan detachment. Now that we distinguish an entire ratio of silhouettes, we employ code to discover the correct amount of collections.

*f. Elbow standard technique (with the number of sharpened errors)*

The main purpose of this approach is to run a k-mean relationship in the data specified for the k significance (num_clusters, e.g. k = 1 to 10), and for each k value, compute the number of sharpened errors or sum of squared error (SSE). Then, modify the sum of the squared error marks for individual k rates. If the line chart appears like a hand - a red sphere (in the procedure of an angle) underneath the line of the line, the "elbow" on the pointer is the precise value (group value) (Puwanenthiren, 2012). At this time, we aim to decrease the sum of squared error. The Sum of squared error generally cascades to 0 as we move up k (and the sum of squared error is 0 anywhere k is equal to the sum of data points, this is as a result where individual data point has its set, and we do not have error concerning it and its stem) The unbiased is consequently select a lesser value of k, which tranquil has a subordinate sum of squared error, and the conduit commonly characterizes anywhere it starts to reappearance negatively with accumulative.

*g. Explaining client segmentation*

We discover that combing the matrix of integration and check what can be derived from the standard data for the individual clusters. Tables 1 and 2 shows customers data and clusters, respectively (Jerry, 2015).

Table 1. Customers' Information

| Customer ID | | | | |
|---|---|---|---|---|
| 12346.0 | 1.72499 | -1.731446 | 1.731446 | 0 |
| 12347.0 | 1.457445 | 1.064173 | 1.401033 | 2 |
| 12348.0 | 0.967466 | 0.6573368 | 0.929590 | 2 |
| 12349.0 | 0.944096 | -1.730641 | 1.683093 | 0 |
| 12350.0 | -0.732148 | -1.729835 | 0.331622 | 0 |
| 12352.0 | 1.193114 | 1.309162 | 0.169639 | 2 |
| 12353.0 | -1.636352 | -1.729029 | -1.570269 | 3 |
| 12354.0 | 0.508917 | -1.727417 | 1.612961 | 0 |
| 12355.0 | -0.366472 | -1.727417 | 0.970690 | 0 |
| 12356.0 | 1.268868 | 0.158357 | 1.557375 | 2 |

Table 2. Cluster Data

| S/N | Total Sales | Order Count | Average Order Value |
|---|---|---|---|
| 0 | 0.244056 | 0.740339 | -0.640559 |
| 1 | -0.137710 | -0.851493 | 0.792034 |
| 2 | 1.203710 | 0.996813 | 0.879446 |
| 3. | -1.235415 | -0.784442 | -1.056848 |

### *h. Segmentation of Best-selling products*

We are working on four-segments and we can identify the total amount that will be spent on buying, their entire norm, and the number of their remits. The subsequent item we may ensure is to support purchaser segments improved comprehension which stuffs sell top in the respective segment.

Table 3. Stock code

| Description | Stock code |
|---|---|
| Jumbo Bag red retrospot | 1129 |
| Regency cake stand three-tier | 1080 |
| White hanging heart T-light holder | 1062 |
| Lunch bag red retrospot | 924 |
| Party bunting | 859 |

## RESULTS AND DISCUSSION

From the Silhouette clustering analysis which checks and validate as well interpreting the consistency within a system to suggest that cluster 4 show the furthermost and comprehensive silhouette proper, representing that cluster 4 might be the best figure of groups.

In the same vein, thriving with the precise appreciation of the elbow instrument at hand, comparing the elbow technique with the Silhouette, if it reaches an agreement with our preceding results suggestive of 4 sets. According to Puwanenthiren (2012), it looks approximating K = 4, or 4 clusters are the accurate figure of clusters in the study. Now interprets the client fragments providing by these workings.

From the data obtained on explaining clients' segmentation, we visualize clusters by tallying dissimilar columns in the x and y axes. And the following result was deduced from them.

Segmentation will help you to identify the set of clients or customers to invest more on, like giving them a discount on their next patronage at least within a month. Preferably, a delayed coupon can be provided at a checkpoint. Likewise, for consumers in the blue region or segment, you can improve on your sales and marketing policy, this might encourage them to patronage you better. Maybe the profligate offer according to market analysis.

In this design, there is an average amount and order likened to the total merchandising price. This design also strengthens the former 2 spots in classifying the oranges segment as the most promising customers, green segment shows the lowest valued consumers, and blue and red as the prospective customers that if work on might increase their price value. Basing on the growth perspective, more attention should be given to customers in the red and blue segments. One should work on better understanding each happenstance and their intellectual conduct on spot as to which crew to concentrate on and familiarizing some trial phase.

## CONCLUSION AND RECOMMENDATION

Because the dataset deployed for this study was uneven, we opted for internal grouping authentication instead of eternal grouping certification, which depends on some peripheral data like labels. Internal group authentication may be utilized to select the grouping algorithm that best ensembles the dataset and vice versa can suitably group the data in the

 

category. Client dissection can have an optimistic impression on the commercial if complete right.

However, customers found in the orange segment deserve special bunches of gifts or discount vouchers to keep them for a lengthy period and customers on the red and blue segment should be given discounts and promote vastly traded items to fascinate them. Also, customers in the green segment exhibit lesser value, which demands a kind of feedback supports to identify what can be put in place to draw them.

According to the above finding, it is clear that the Jumbo Bag Red Retrospot is the highest purchasing or best-selling product by the classiest group. With the available result obtained from this study, we recommend other impending clients in this segment.

## REFERENCES


Griva, A., Bardaki, C., Pramatari, K. and Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.

Hong, T. and Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. *Expert System Applications*, 39 (2), 2127-2131.

Jerry, W. T. (2007). Accessed at: www.decisionanalyst.com on May 12, 2021.

Lakshmi Narayana S., Suneetha Devi J., Bhargav Reddy I., Harish Paruchuri. (2012). Optimizing Voice Recognition using Various Techniques. *CiiT International Journal of Digital Signal Processing*, *4*(4), 135-141

Neogy, T. K., & Paruchuri, H. (2014). Machine Learning as a New Search Engine Interface: An Overview. *Engineering International*, *2*(2), 103-112. https://doi.org/10.18034/ei.v2i2.539

Paruchuri, H. (2015). Application of Artificial Neural Network to ANPR: An Overview. *ABC Journal of Advanced Research*, *4*(2), 143-152. https://doi.org/10.18034/abcjar.v4i2.549

Paruchuri, H. (2017). Credit Card Fraud Detection using Machine Learning: A Systematic Literature Review. *ABC Journal of Advanced Research*, *6*(2), 113-120. https://doi.org/10.18034/abcjar.v6i2.547

Puwanenthiren, P. (2012). Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). Volume 12 Issue 1.

Sulekha, G. (2011). The basis of market segmentation: a critical review of the literature. *European Journal of Business and Management*, *3*(9).

Vadlamudi, S. (2015). Enabling Trustworthiness in Artificial Intelligence - A Detailed Discussion. *Engineering International*, *3*(2), 105-114. https://doi.org/10.18034/ei.v3i2.519

Vadlamudi, S. (2017). Stock Market Prediction using Machine Learning: A Systematic Literature Review. *American Journal of Trade and Policy*, *4*(3), 123-128. https://doi.org/10.18034/ajtp.v4i3.521

Vadlamudi, S. (2018). Agri-Food System and Artificial Intelligence: Reconsidering Imperishability. *Asian Journal of Applied Science and Engineering*, *7*(1), 33-42. Retrieved from https://journals.abc.us.org/index.php/ajase/article/view/1192

Vaishali R. Patel and Rupa G. Mehta. (2011). Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. IJCSI.

Vishish R. P. and Rupa G. M. (2011). Impact for External Removal and Standard Procedures for JCSI. International Science Issues, 8(5,2): 1694-0814.


**--0--**