CUSTOMER SEGMENTATION USING MACHINE LEARNING

AMAN BANDUNI[*1], Prof ILAVENDHAN A.[2]
[*1,2]School of Computing Science & Engineering,
Galgotias University, Greater Noida, U.P.
[*1]abanduni.5@gmail.com,[2]ilavendhan@galgotiasuniversity.edu.in

Abstract-The emergence of many competitors and entrepreneurs has caused a lot of tension among competing businesses to find new buyers and keep the old ones. As a result of the predecessor, the need for exceptional customer service becomes appropriate regardless of the size of the business.[2] Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service. Each segment has customers who share the same market features.[5] Big data ideas and machine learning have promoted greater acceptance of automated customer segmentation approaches in favor of traditional market analytics that often do not work when the customer base is very large. In this paper, the k-means clustering algorithm is used for this purpose.[8] The Sklearn liabrary was developed for the k-Means algorithm (found in the Appendix) and the program is trained using a 100-pattern two-factor dataset derived from the retail trade. Characteristics of average number of customer purchases and average number of monthly customers.

Keywords- data mining; machine learning; big data; customer segment; k-Mean algorithm; sklearn; extrapolation;

I. Introduction

Over the years, increased competition among businesses and the availability of large-scale historical data has resulted in widespread use of data mining techniques to find critical and strategic information that is hidden in organizations' information.[1] Data mining is the process of extracting logical information from a dataset and presenting it in a human-accessible manner for decision support. Data mining techniques distinguish fields such as statistics, artificial intelligence, machine learning, and data systems. Data mining applications include, but are not limited to bioinformatics, weather forecasting, fraud detection, financial analysis and customer segmentation. The key to this paper is to identify customer segments in a commercial business using the data mining method. Customer segmentation is a group of business customer base called customer segment such that each customer segment has customers who share the same market characteristics.[5] These differences are based on factors that directly or indirectly affect the market or business such as product preferences or expectations, location, behavior and so on. The importance of customer segmentation includes, inter alia, the ability of a business to customize market plans that would be appropriate for each segment of its customers;[6] Support for business decisions based on risky environments such as credit relationships with its customers; Identify products related to individual components and how to manage demand and supply power; Interdependence and interaction between consumers, between products, or between customers and products are revealed, which the business may not be aware of; The ability to predict customer declines, and which customers are likely to have problems and raise other market research questions and provide clues to find solutions.

Buried in a database of integrated data proved to be effective for detecting subtle but subtle patterns or relationships. This mode of learning is classified under supervised learning. Integration algorithms include the K-Means algorithm, K-nearest algorithm, sorting map (SOM), and more.[4] These algorithms, without prior knowledge of the data, are able to identify groups in them by repeatedly comparing input patterns, as long as static aptitude in training examples is achieved based on subject matter or process. Each set has data points that have very close similarities but differ greatly from the data points of other groups. Integration has great applications in pattern recognition, image analysis, and bioinformatics and so on.[15] In this paper the k-means clustering algorithm was implemented in the customer segment. The scalar library (Appendix) of the K-Means algorithm was developed, and training was started using a standard silhouette -score with two feature sets of 100 training patterns found in the retail trade. After several indications, four stable intervals or customer segments were identified. Two factors are considered in combination with the number of items a customer purchases per month and the average number of customers per month. From the dataset, four customers or categories are classified and labeled as follows: cluster_metrics_1, cluster_metrics_2, cluster_metrics_3, cluster_metrics_4.

II. Literature Survey

A. Customer Classification

Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and desires of their customers, attract new customers, and thus improve their businesses.[6] The task of identifying and meeting the needs and requirements of every customer in the business is very difficult. This is because customers can vary according to their needs, wants, demographics, size, taste and taste, features etc. As it is, it is a bad practice to treat all customers equally in business. This challenge has adopted the concept of customer segmentation or market segmentation, where consumers are divided into subgroups or segments, where members of each subcategory exhibit similar market behaviors or characteristics.[9] Accordingly, customer segmentation is the process of dividing the market into indigenous groups.

B Big Data

Recently, Big Data research has gained momentum. Defines big data - a term that describes a large number of formal and informal data, which cannot be analyzed using traditional methods and algorithms. Companies include billions of data about their customers, suppliers, and operations, and millions of internally connected sensors are sent to the real world on devices such as mobile phones and cars, sensing, manufacturing and communications data.[10] Ability to improve forecasting, save money, increase efficiency and improve various areas such as traffic control, weather forecasting, disaster prevention, finance, fraud control, business transactions, national security, education and healthcare. Big data is mainly seen in three Vs: volume, variability, and speed. Other 2Vs are available - authenticity and price, thus making it 5V.

C. data repository

Data collection is the process of collecting and measuring information against targeted changes in an established system, which enables one to answer relevant questions and evaluate the results.[12] Data collection is part of research in all fields of study including physical and social sciences, humanities and business. The purpose of all data collection is to obtain quality evidence that leads the analysis to construct concrete and misleading answers to the questions presented. We collected data from the UCI machine learning repository.

D. Clustering data

Clustering is the process of grouping information into a dataset based on some commonalities. There are several algorithms, which can be applied to datasets based on the provided condition.[7] However, no universal clustering algorithm exists, hence it becomes important to choose the appropriate clustering techniques. In this paper, we have implemented three clustering algorithms using the Python scalar library.

E. K-mein

K-means that an algorithm is one of the most popular classification algorithms. This clustering algorithm relies on centro, where each data point is placed in one of the overlapping ones, which is pre-sorted in the K-algorithm. Clusters are created that correspond to hidden patterns in the data that provide the necessary information to help decide execution. process. There are many ways to make assembling K-means, we will use the elbow method.

III. Methodology

The data used in this paper were collected from the UCI Machine Learning Repository. It is a set of geographic data, including all transactions that occur between 1/1/2/10 and 9/12/2011 in an unregistered and unregistered UK broker. The company mainly sells unique gifts to everyone at once. Many of the company's customers are shopkeepers.[10] The database has 8 attributes. These features include:

"Invoice: invoice number. By default, a 6-digit total number is assigned separately for each transaction. If this code starts with the letter 'c', it indicates a cancellation. "

Stockcode Code: Product (Item). Name, a 5-digit number assigned only to each unique product. "

"Definition: Product Name (Item). By Name."

“Price: The price of each product (item). Number. "

"Invoice: The date and time of the invitation. In terms of numbers, the date and time of each transaction. "

"UnitPrice: Price is one unit. Price, product price per unit of measure. "

"Customer: Customer Number. Name, 5-digit number to each customer. "

Country: Country name. Name, the name of the country where each customer resides. "

In this paper several steps were taken to obtain an accurate result. It includes a feature with Centro's first stage, allocation phase and update phase, which are the most common phase k-means algorithms.

A. Collect data

This is a data preparation phase. The feature usually helps to refine all data items at a standard rate to improve the performance of clustering algorithms.[12] Each data point varies from grade 2 to +2. Integration techniques that include min-max, decimal, and z-point are the standard z-signing strategy used to make things uneven before the dataset algorithm applies the k-Means algorithm.

B. Methods of customer classification

There are many ways to partition, which vary in severity, data requirements, and purpose. The following are some of the most commonly used methods, but this is not an incomplete list.[13] There are papers that discuss artificial neural networks, particle determination and complex types of ensemble, but are not included due to limited exposure. In future articles, I may go into some of these options, but for now, these general methods should suffice.

Each subsequent section of this article will include a basic description of the method, as well as a code example for the method used. If you do not have the expertise, well, just skip the code and you have to get a good handle on each of the 4 sub-sections included in this article.[14]

C. Group analysis

Group analysis is an integration or unification, approach to consumers based on their similarity.

There are 2 main types of categorical group analysis in market policy: hierarchical group analysis, and classification (Miller, 2015). In the meantime, we will discuss how to classify groups, called k-methods.

D. K. Means encounter

The K-means clustering algorithm is an algorithm often used to draw insights into formats and differences within a database.[13] In marketing, it is often used to build customer segments and understand the behavior of these unique segments. Let's try to build an assembly model in Python's environment.

E. Centroids initiation

Selected cents or initials were selected. Figure 1 introduces the beginning of graduate centers. The four selected centers, shown in different sizes, were selected using the Forgi method. In Forgy's method, data points are randomly selected as cluster centroids using k (k = 4 in this case).

Technical introduction: -

The code below was created in the Jupiter manual using Python 3.x and some Python packages for editing, processing, analyzing, and visualizing information.[11]

Most of the codes below come from the Github package of a book called Hands-on Data Science for Marketing. The book is available on Amazon or OilReilly if you are a customer.

The open source data cost used in the following code comes from Irwin's machine learning repository.

IV. Proposed Model

A) Import packages and data:

To begin, we import the necessary packages to do our analysis and then the xlsx (Excel spreadsheet) data file.[12] If you want to follow up with the same data, you have to download it from UCI. For this example, I place the xlsx file in the folder (directory) where I present Jupiter's notebook.

B) Data cleaning:

After importing the package and data, we will see that the data is not as helpful as that, so we need to clean and organize this data in a way that we can create more actionable insights.

C) Normalize the data:

The K-means area unit is sensitive to the scale of the information used, such as clustering algorithms, so we would like to generalize the information.[15]

A screenshot of the StackExchange answer below discusses why standardization or normalization is necessary for data used in K-means clustering. The screenshot is linked to the StackExchange question, so you can click on it and read the entirety of the discussion if you want more information.[10]

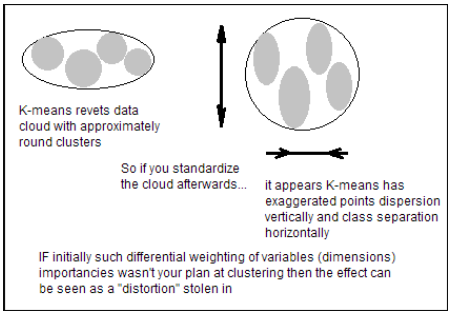

Fig. 1 Standardistion and Normalisation

D) Select the optimal number of groups:

Okay, we are ready to run cluster analysis. But first, we need to find out how many groups we want to use. There are several approaches to selecting the number of groups to use, but I am going to cover two in this article: (1) the silhouette coefficient, and (2) the elbow method.[7]

E) Silhouette (clustering):

The silhouette refers to how to interpret and validate consistency within data structures. This method shows a diagram of how well each item is organized. [1]

The value of a silhouette is a measure of how something is more similar in its collection (combination) than other groups (partitions). The silhouette goes from -1 to +1, where a higher value indicates that an object matches its collection properly and is compared to neighboring groups. If several objects have a high value, the integration configuration is appropriate. If most points have a value or a negative value, the coordinate system may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

Now that we know a whole lot of silhouettes, we use code to find the right number of groups.

```
Silhouette Score for 4 Clusters: 0.4114
Silhouette Score for 5 Clusters: 0.3773
Silhouette Score for 6 Clusters: 0.3785
Silhouette Score for 7 Clusters: 0.3913
Silhouette Score for 8 Clusters: 0.3810
```

Fig. 2 Silhoette Score

Cluster 4 had the most complete silhouette fit, indicating that 4 may be the best number of clusters. But we'll see twice the way to the elbow.

F) Elbow criterion method (with the sum of squared errors) (SSE):

The idea behind the elbow method is to run a k-mean correlation in the data given for the k value (num_clusters, e.g. k = 1 to 10), and for each k value, calculate the sum of the squared errors (SSE). is.

Then, adjust the SSE line for each k value. If the line graph looks like a hand - a red circle (in the form of an angle) below the line of the line, the "elbow" on the hand is the correct value (collection value).[6] Here, we want to reduce SSE. SSE usually falls to 0 as we go up k (and SSE is 0 where k is equal to the number of data points, because where each data point has its own set, and there is no error between it and its trunk) .

The objective is therefore to select a smaller value of k, which still has a lower SSE, and the cone usually represents where it begins to return negatively with increasing.

Well, with the correct understanding of the elbow mechanism at hand, let's use the elbow method to see if it agrees with our previous results suggesting 4 sets.

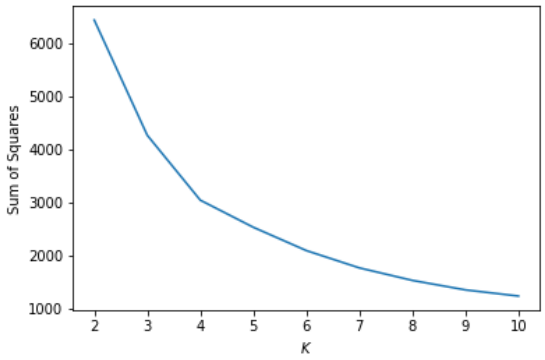

Fig. 3 Elbow Graph exported from my working Jupyter notebook

Based on the graph above, it looks like K = 4, or 4 clusters is the correct number of clusters in this analysis. Now translates the customer segments provided by these components.

G) Explaining customer segment

CustomerID	TotalSales	OrderCount	AvgOrderValue	Cluster
12346.0	1.724999	-1.731446	1.731446	0
12347.0	1.457445	1.064173	1.401033	2
12348.0	0.967466	0.573388	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	-1.728223	1.612981	0
12355.0	-0.386422	-1.727417	0.970690	0
12356.0	1.268868	0.158357	1.557375	2

Fig. 4 Customer table

Now we have to combine the matrix of integration and see what we can gather from the standard data for each cluster.

	TotalSales	OrderCount	AvgOrderValue
0	0.244056	0.740339	-0.640559
1	-0.137750	-0.851493	0.792034
2	1.203710	0.996813	0.879446
3	-1.235415	-0.784442	-1.056848

Fig. 5 Clusters

In the following section, we need to visualize clustering by adding different columns in the x and y axes. Let's see what we say.

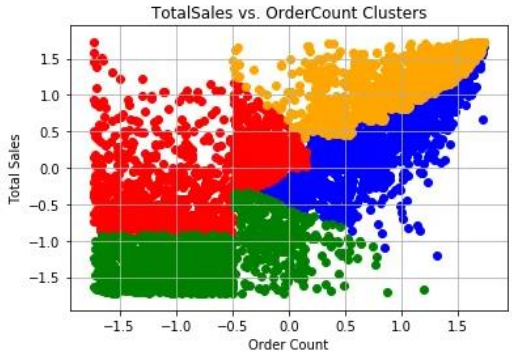

Fig. 6 TotalSales vs OrderCount Clusters

Green customers have the lowest price and lowest order count, meaning they are the lowest bidder. On the other hand, orange customers have the highest total sales and highest order count, indicating that they are the highest priced customers.

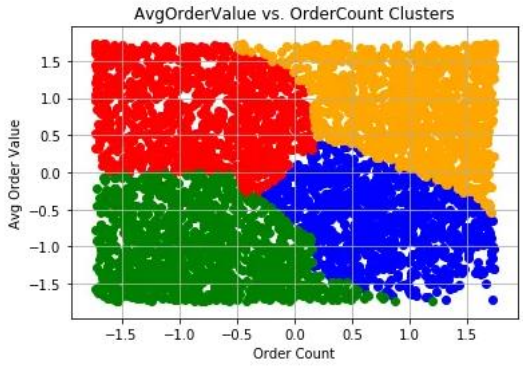

Fig. 7 AvgOrderValue vs OrderCount Clusters

In this structure, we consider the average order value versus the order value. Once again, green buyers have the lowest prices and orange has the highest customer prices.

You can see it this way. You can target customers in red graphics and try to find ways to increase your order count via email reminders or SMS notifications directed to other identification features. Maybe you can give them a discount when they come back within 30 days. Ideally, you can provide a delayed coupon (which will be used at some point) at checkout.

Similarly, with customers who are in the blue segment, you may want to try other sales and marketing strategies for the cart. Possibly the fastest offer based on market basket analysis (see section on market basket analysis below).

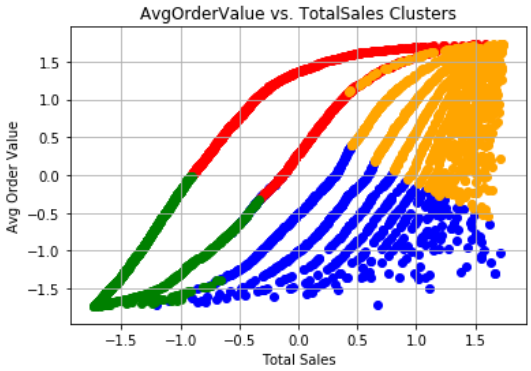

Fig. 8 AvgOrderValue vs TotalSales Clusters

In this building, it has an average price and order compared to the total retail price. This structure also reinforces the previous 2 sites in identifying the orange group as the highest value customer, green as the lowest priced customer, and blue and red as the high potential customers.

From a development perspective, I focus my attention on the blue and red collections. I try to better understand each encounter and their intelligent behavior on site as to which team to focus on first and introduce some test cycles.

H) Best-selling item by segment

We know that we have 4 categories and we know how much they spend on each purchase, their total usage and the number of their orders. The next thing we can do is to help customer segments better understand which items sell best in each segment.

Description	StockCode
JUMBO BAG RED RETROSPOT	1129
REGENCY CAKESTAND 3 TIER	1080
WHITE HANGING HEART T-LIGHT HOLDER	1062
LUNCH BAG RED RETROSPOT	924
PARTY BUNTING	859

Fig. 9 StockCode

V. Result

Here, the result suggests that the orange cluster as the highest value customers, green as the lowest value customers, and blue and red as the high opportunity customers.

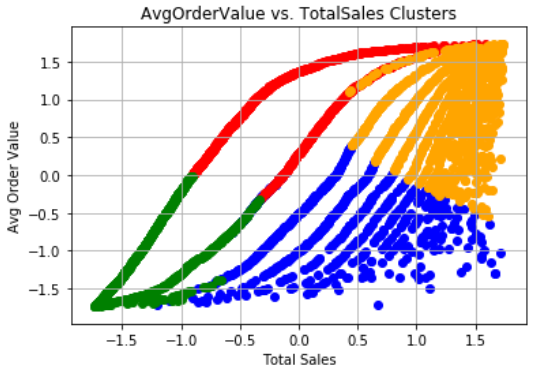

Fig. 8 AvgOrderValue vs ToatalSales Clusters

Result also concludes that the Jumbo Bag Red Retrospot is the best-selling item.

Description	StockCode
JUMBO BAG RED RETROSPOT	1129
REGENCY CAKESTAND 3 TIER	1080
WHITE HANGING HEART T-LIGHT HOLDER	1062
LUNCH BAG RED RETROSPOT	924
PARTY BUNTING	859

Fig. 9 StockCode

VI. Conclusion

As our dataset was unbalanced, in this paper we opted for internal clustering validation rather than external clustering verification, which relies on some external data such as labels. Internal cluster validation can be used to choose the clustering algorithm that best suits the dataset and vice versa can correctly cluster the data in the cluster.

Customer segmentation can have a positive impact on business if done properly.

So we can give people of orange bunches special discounts or gift vouchers to keep them for a long time and we can give discounts to people in blue and red clusters and advertise highly sold items to attract them , And for those of lower value who are in green clusters, we can organize feedback columns to find out what we can change to attract them.

Based on the above information, we now know that the Jumbo Bag Red Retrospot is the best-selling item by our most expensive team. With that information available, we can make recommendations for other potential customers in this section.

VII. References

[1] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited

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

[3] Hong, T., 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.

[4] Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited

[5] Puwanenthiren Premkanth, - 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). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[6] Puwanenthiren Premkanth, - 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). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

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

[8] By Jerry W Thomas. 2007. Accessed at: www.decisionanalyst.com on July 12, 2015.

[9] T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. AnuManufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.

[10] McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015.

[11] Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf July 14, 2015.

[12] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.

[13] Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814

[14] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar "Telecom Customer Classification Based on Group Analysis of K-methods", JIRCCE, Year: 2015.
[15] Vaishali R. Patel and Rupa G. Mehta “Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm”, IJCSI,Year: 2011.