Study on Customer Segmentation using K-Means Clustering

Shewatha Arul[1] Mithun DJ[2] Shinu Abhi[3]

[1, 2, 3] REVA Academy for Corporate Excellence, Bengaluru, India
shewatha.ba05@reva.edu.in
mithun.dj@reva.edu.in
shinuabhi@reva.edu.in

Abstract - Customer segmentation is the classification of customers based on their attributes that are specific and similar. By using the clustering method, the decision of which customer segment to target is determined, which has been many retailers challenge, it provides an excellent opportunity for internal customer analysis. The dataset used for this analysis has over 3 Lakhs transaction of customers, it gets challenging to derive at the decision as to which group of customers to target to Upsell or Cross sell any product. Hence that is where a company loses most of its marketing investment as there is a high marketing expenditure that goes into the campaigning of any sales, but not many sales are successful and there is a revenue to the company from it. Hence low Return on Investment (ROI). This is the problem that can be resolved by Customer Segmentation. The research focuses on segmenting clients utilising customer-specific data such as Customer IDs and over all consumption levels and it is achieved by using machine learning methods - Specifically used in this research are, the Density Based Clustering Algorithm (DBSCAN) and the Partitioning Algorithm (KMeans Clustering Algorithm). The findings indicate that there are 3 distinct groups of customers or the campaign target, each with unique purchasing characteristics on whom the company's campaigning marketing design can be based on.

Keywords – Customer Segmentation, Cluster, DBSCAN, Kmeans

Introduction

We live in a world where massive amounts of data are collected regularly. This information must be analyzed. A company's business plan should adapt to the present era of innovation when everyone is trying to outdo one another. Because many potential clients are unclear about what to buy, businesses nowadays rely on fresh concepts. Furthermore, the businesses involved are unable to identify the target potential clients. Machine learning is used in this instance to uncover hidden patterns in the data, allowing

for better decision-making. This section will go deeper into the business problems that retailers confront when it comes to campaign marketing. A. The Business Problem

Every retail company, regardless of its industry, collects, creates, and manipulates data during its lifetime. For better or worse, each of these pieces of information offers information about how the firm is evolving as a brand. The more data a person possesses, the more comprehensive a picture the data will generate. Companies that use competent data science and data mining methodologies may dive deep into their operational strategies, allowing them to improve their business processes. As a result, there is an increasing interest in looking into events and data that are difficult to explain. Who is most likely to return to our store? Can we figure out who is most likely to respond to our marketing? Who are the customers we can market and concentrate on? This research will focus on the last part of the final question. A particular cluster of customers we can cross-sell and up-sell with is the desired output.

When we have large data set of over 3 Lakhs transactions, it gets challenging to derive the decision to whom and which customer to target to perform any kind of campaigning to sell any product. Hence that is where Customer Segmentation comes into the picture to resolve this challenge

By combining ML practices with conventional business approaches to answering these questions, the paths to answering a similar question were increasingly intertwined: What segments or groups of customers does our company have? Having discussed it in numerous other contexts, examining customer clustering became a useful tool in understanding the purchasing habits and behaviours of its customers.

Value-based customer extraction and recommending the company to target them for any mode of campaigning, hence there will be a change in the marketing model and higher ROI is the objective of this analysis.

The challenges faced by every retail company with customer segmenting and our solution of clustering the customers to campaign the right target for higher profit. In the following section, the existing methodology and techniques for clustering and our solutions shall be reviewed to lay the foundation for this paper.

Literature Review

The entire basis of marketing methods is mutual consumer-retailer connections. One way to increase income is to determine consumer needs through conversation with them. It's almost impossible to communicate with customers on a personal level, yet without it, marketing disasters are unavoidable. To address this issue, merchants might use data provided by customers to communicate. Retailers can segment their customers based on their habits and then design business strategies based on that information.

Customer segmentation is a method of improving communication with customers by identifying their interests and desires so that appropriate communication can be produced. But why segment customers? Well, every firm is built around the production of specific items, and each product has its own set of clients. Targeting those designated clients will ensure that the organizations' sales are ensured. Customer segmentation,

like all the other critical procedures in company strategy, is a crucial phase. This integration enables businesses to communicate with a specific segment of their consumer base depending on their current interests and requirements. This strategy not only assists in identifying customers, but it also aids in channeling communication to these specific audiences.

Clustering tasks can be performed using a variety of approaches and algorithms, which can be classified into three sub-categories:

- Partition-based clustering: E.g. k-means, k-median
- Hierarchical clustering: E.g. Agglomerative, Divisive
- Density-based clustering: E.g. DBSCAN

Organizations can improve the quality of their goods, services, and relationships with their customers. Businesses may focus their attention on the most profitable customers through Customer Segmentation.

So far, the proposed customer segmentation initiatives are only valid until they are integrated into their appropriate categories, after which the gathered data is used for further analysis. This helps validate our data obtained after segmentation of customers into various required groups [1].

Through clustering, K-means is utilized to identify a company's most important customer. The major aim is to find relevant, valuable customers and utilize their information to generate fresh digital marketing campaigns [2].

And today's business is based on new ideas because there are many potential customers who are unsure what to purchase and what not to purchase. The businesses themselves are unable to diagnose the target potential clients. This is where Machine Learning comes in [3].

It is necessary for enterprises to identify the potential customers in the market by mining the customer data to gain profitable insight. One of the efficient ways to identify the different customer characteristics is by applying clustering analysis [4].

Today's business run based on such innovation having ability to attract the customers with the products, but with such a large raft of products leave the customers confuse. Three different clustering algorithms (k-Means, Agglomerative, and Meanshift) are being used to segment customers and then compare the results of the algorithms' clusters [3].

Clustering techniques consider data tuples as objects. They organize the data items into groups or clusters so that things inside a cluster are similar but not identical to objects in other clusters. According to [5], segmentation is based on commonalities in

many aspects significant to marketing, such as gender, age, hobbies, and various purchasing behaviors.

According to [6] , the DBSCAN method allows clustering large datasets with similar results to the outcome of DBSCAN on the entire dataset. Experiments results suggest that the proposed technique presents good results and consistency compared to other algorithms with similar approach.

Customer Relationship Management (CRM) can support the customer segmentation process by implementing appropriate marketing strategies so that companies can identify the quality and behavior of customers. Customer segmentation is the process of dividing customers into groups based on past data with the demands, characteristics, and the same functioning. DBSCAN method works better than the K-Means [7].

Objective

This research aims to identify customer segments using a data mining approach, specifically the Density-Based Clustering Algorithm (DBSCAN) and the Partitioning Algorithm (K-means Clustering Algorithm). The elbow technique is implemented to find the optimal number of clusters.

This paper research demonstrates 2 kinds of clustering methodology - DBSCAN and K-Means as explained above and decides the best method that is suitable for this data and recommends the same.

Project Methodology

This project was structured using the CRISP-DM approach, which is a prominent framework for data mining project planning. This is a framework that has been demonstrated to be useful in a range of industrial applications. This is an unlimited procedure in which you can go back and forth between phases. The arrows expressing the requirement of the phases are equally important; the outside circle depicts the framework's cyclical qualities. CRISP-DM is not a one-time process, as the outer circle graphic illustrates. Every step is a new learning experience from which we may get new insights and potentially solve other business issues. In the below figure, the CRISP-DM methodology is pictured. And each of these steps are explained regarding our analysis in this study.

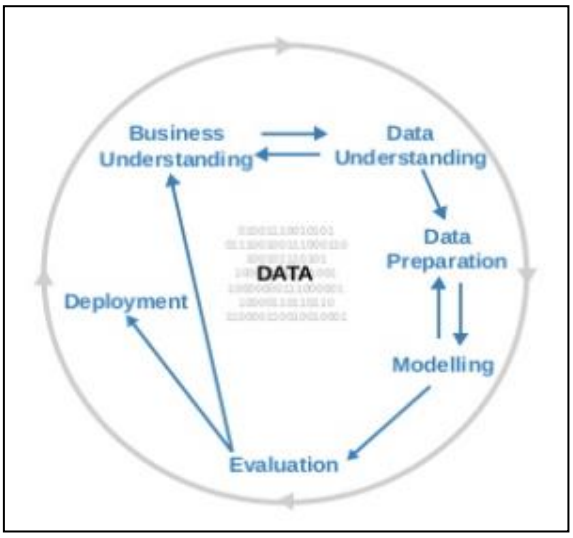


Fig 1 - CRISP-DM Methodology

Here all the project activities are grouped into the following five predefined steps –

Business Understanding:

Over the years, the vast amount of historical data accessible has resulted in the widespread usage of data mining techniques for extracting relevant information and strategic information from an organization's database as commercial competitiveness has intensified. Methods are used to extract data patterns and display them in a human-readable way that may be utilised for decision assistance in data mining. The main challenge for businesses today is having access to customer profile data but not being able to integrate insights into their business strategies.

Businesses today find it difficult to reach the ideal prospects since they have many products to sell and multiple channels to market. Our value-based customer segmentation solutions enable businesses to better understand their customers, enabling them to target the right customers through the right channels. This assists organizations to improve customer engagement through smart marketing schemes that significantly improve spend and ROI.

Data Understanding:

This phase begins with data collection and progresses to steps to familiarise yourself with the data, identify data quality concerns, and get early insights into the data. Also discovered are several intriguing subsets that can be utilised to develop assumptions regarding hidden data.

Data Preparation:

This stage takes up most of the data mining' time. Most data used in data mining is collected and kept for other purposes, and it must be modified before it can be utilised in modelling. All processes that go into developing the final dataset from raw data are included in the data preparation step: choosing, cleaning, building, integrating, and formatting.

Modelling:

This step is about choosing and employing the most effective modelling methods. There could be a loop back to data preparation here.

Approaches, such as DBSCAN have more stability to outliers in modelling steps but result in a higher number of clusters. And some approaches, such as K-Means, do great in large datasets and are more efficient.

Evaluation:

The result of this step is to select the best model based on the cluster and segmentation that is efficient for business decisions and further actions. The model is chosen in such a way that it can be generalised by performing finals test against unseen data and ensuring the model adequately address all critical business challenges.

Deployment:

In most cases, this entails deploying a model's code representation into an operating system. test against unseen data and ensuring the model adequately address all critical business challenges. It also includes a scoring or ranking mechanism when new invisible data appears. All data preparation actions must be incorporated in the code representation before modelling; this assures that the model will treat the new unseen data the same way it did when it was produced.

Business Understanding

Understanding the components of a retailer’s customer base is critical for maximizing their market potential; the retailer with the most consumers gain the greatest market share. The hefty expenses of getting a new consumer or recovering an existing one push businesses to take shortcuts.

Furthermore, in the retail business, it is commonly known that the Pareto Principle typically applies to the company: 80% of earnings originate from 20% of clients. Retail companies depend on recurring purchases, which is one of the main reasons why this

theory stays true. As a result, the net shift of a single client can have a large long-term influence on a business's earnings. As a result, it is generally in the best interests of the retailer to devote efforts to customer retention by understanding them as thoroughly as possible.

Customer segmentation analysis is the process of categorizing consumers so that those in one group are similar yet distinct from those in other groups. There are two types of segmentation methods: a priori and post hoc. A priori analysis requires generating the segments in advance and then placing each consumer into them after analysing the data. Rather than having customer data influence the sorts of segments generated, the ideal segmentations would be dictated by outside knowledge or structure. As a result, the formed segments, rather than the consumers themselves, are the primary unit of study in this scenario.

While customer segmentation research has been a goal for retailers for a long time, past methodologies relied on far less sophisticated analytical tools than those available today. It's nonsensical to blame firms in the past for not effectively utilizing their data; technology and data infrastructure were just not available or affordable enough to allow them to collect vast volumes of data in the way they do today. Despite this, many firms continue to rely on archaic approaches to gain a better understanding of their clients, the most frequent of which is merely demographic analysis.

The technique of segmenting clients purely based on demographic variables like age, sex, gender, race, and ethnicity is known as demographic analysis. It is assumed that the demographics of a store's customer region define retail behaviour. The reduction of consumers to only a few well-understood and categorized demographic variables made it easier for merchants to acquire and use data from their customers since it was very trivial to take a limited number of characteristics and establish suitable pre-set groups. Furthermore, demographic research grew in popularity as a rapid, low-cost, and straightforward method for forecasting how new consumers would engage. As a result, demographic segmentation enables merchants to gather just relevant data, requiring little labor and hence expense, while keeping analysis and communication of the study at the same level.

Despite the success of several well-known marketing organizations, the growing availability of retail technology demonstrated that demographic segmentation was unable to provide insight into customer purchase histories. When merchants and marketers started experimenting with alternative segmentation strategies, it became evident that the deeper segmentation process would rapidly replace just demographic segmentation.

A conventional RFM implementation is affordable and simple: if each of the components is described in a way that makes them easy to gather, a retailer's work of visualizing the findings is very simple, making interpretation simple. The inferred segments and their defining properties are often displayed in three plots, one for each combination of two variables (e.g., recency and frequency). Because of its ease of use and low cost

of implementation, as well as its ability to communicate effectively, RFM analysis has become a modern marketing mainstay. In some ways, the visualization aspect alone provided utility to the RFM model, allowing managers to effectively manage resources [7]. However, as the retail business expanded in lockstep with the technological revolution, it became substantially simpler for merchants to gather data on a greater scale, making it easier to mine data on a larger scale as well.

Retailers may employ this information to establish targeted marketing efforts, provide particularly customized discounts to specific consumer categories, and even welcome previous customers back into the store. Retailers may use this data to create ultra-targeted marketing campaigns, which have changed the way they compete in the Big-data era. To do high-level consumer segmentation research, retailers have begun to include components of machine learning in their customer analysis.

Rather than concentrating on just a few features or customers at a time, it is possible to write programs and implement algorithms that can consider several more features or several more instances than traditional spreadsheets can hold or process. Retailers from many sectors are striving to apply clustering algorithms, as a result of this vast potential, like DBSCAN (Density-Based Clustering) and K-Means (hierarchical clustering) to segment their customers more accurately and quickly. The more quickly and effectively retailers can cluster their customers, the more quickly they can market to them and thus gain market share.

In data understanding, business questions are framed, and we will explore the data and how the data helps to answer each question. In this paper test data is used which mimics actual production data.

Data is collected from Alshaya and it consists of 354,335 customers' data and it consists of customer demographic, for easy viewing, importing, and analysis.

Data Preparation

Following business and data understanding, the next key purpose of the CRISP-DM process is to prepare the data for modelling and analysis. This includes selecting, cleaning, and transforming the data that will be used in the paper. Although processing raw data for analysis often demands a great deal of effort, the phase is crucial, as the old saying "garbage in, garbage out".

At the end of the data extraction, there were over 5 lakh rows of transactions. The no. of transactions for clustering was very high. There was a dire need to clean the data further by removing the Null Values, NaN, and junk values from the data, the same was achieved by data cleansing and data extraction, yielding 3,54,335 rows final for data modelling.

The below table describes the distribution of several rows for 1 customer id as per the transaction.

CustomerID	InvoiceNo	StockCode	Quantity	InvoiceDate	Time	UnitPrice	Country
12747	537215	85124C	12	05-Dec-18	15:38:00	2.55	USA
12747	537215	85124B	6	05-Dec-18	15:38:00	2.55	USA
12747	537215	84879	16	05-Dec-18	15:38:00	1.69	USA
12747	537215	85062	24	05-Dec-18	15:38:00	1.65	USA
12747	537215	85064	6	05-Dec-18	15:38:00	5.45	USA
12747	537215	82484	36	05-Dec-18	15:38:00	5.55	USA
12747	537215	21136	8	05-Dec-18	15:38:00	1.69	USA
12747	538537	22795	16	13-Dec-18	10:41:00	5.95	USA
12747	538537	48138	2	13-Dec-18	10:41:00	7.95	USA
12747	538537	82494L	24	13-Dec-18	10:41:00	2.55	USA
12747	538537	84879	24	13-Dec-18	10:41:00	1.69	USA
12747	538537	85062	12	13-Dec-18	10:41:00	1.65	USA
12747	538537	21754	3	13-Dec-18	10:41:00	5.95	USA
12747	538537	82484	12	13-Dec-18	10:41:00	5.55	USA
12747	538537	82482	12	13-Dec-18	10:41:00	2.55	USA
12747	541677	21136	16	20-Jan-19	14:01:00	1.69	USA
12747	541677	82484	36	20-Jan-19	14:01:00	5.55	USA
12747	541677	82494L	12	20-Jan-19	14:01:00	2.95	USA
12747	541677	82482	12	20-Jan-19	14:01:00	2.55	USA
12747	541677	71459	12	20-Jan-19	14:01:00	0.85	USA

Table 1 - Rows of transaction across table.

All the given data variables are explained below as what they denote and represent.

CustomerID – CustomerID carries a unique identifier for each customer.
InvoiceNo - The invoice number is a unique identifier for each transaction.
StockCode – identifies the one-of-a-kind item purchased by the buyer.
Quantity – is a number that represents the amount of items purchased.
InvoiceDate – The date on which the transaction took place is known as the invoice date.
Time – The time when the transaction was completed.
UnitPrice – Each unique item's unit price as indicated by StockCode.
Country – This transaction took place in the following country.

After the data file is received, we remove the Null Values, NaN and junk values from the data and labelling of data is done as per the understanding with respective names. Once the data is cleansed, we need to create features for the analysis. Features are the fundamental elements of datasets. The quality of variables in a dataset has a significant impact on the quality of output expected from any ML algorithm. As the saying goes, if garbage in, garbage out. Below is the figure - 2 and 3, are the snapshot of the features generates and after data preparation for modelling.

UNIQUE_CU	Num_of_	Autu	Sprin	Sumr	Winte	noon	even	Morn	Total it	Items	DollarsPT
12747	103	22	12	31	38	40	11	52	1275	12.379	40.73796
12748	4586	765	391	756	2674	2855	741	990	25700	5.604	7.332983
12749	199	85	0	43	71	160	0	39	1471	7.392	20.55719
12820	59	36	0	0	23	45	0	14	722	12.237	15.97186
12821	6	0	0	6	0	6	0	0	70	11.667	15.45333
12822	46	46	0	0	0	37	0	9	550	11.957	20.62783
12823	5	2	3	0	0	2	1	2	230	46	351.9
12824	25	25	0	0	0	25	0	0	232	9.28	15.8848
12826	91	13	0	7	71	64	0	27	1058	11.626	16.20571
12827	25	15	0	0	10	20	0	5	197	7.88	17.206
12828	56	38	0	0	18	0	16	40	494	8.8214	18.19125
12829	11	0	0	0	11	5	0	6	376	34.182	26.63636
12830	38	9	0	27	2	9	11	18	9848	259.16	179.3326
12831	9	0	9	0	0	9	0	0	135	15	23.89444

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average.T	Sum of Cost
5	0	12	0	23	8	0	7	0	15	11	22	36.7	4196.01
25	100	169	102	268	262	226	135	410	220	1681	968	3.3303571	33629.06
0	0	0	0	43	0	0	85	0	0	32	39	70	4090.88
11	0	0	0	0	0	0	0	14	22	0	12	107.66667	942.34
0	0	0	0	6	0	0	0	0	0	0	0	1	92.72
0	0	0	0	0	0	0	0	46	0	0	0	17	948.88
0	2	1	0	0	0	0	1	1	0	0	0	55.5	1759.5
0	0	0	0	0	0	0	0	0	25	0	0	1	397.12
40	0	0	0	0	7	0	0	13	0	15	16	60.5	1474.72
0	0	0	0	0	0	0	0	0	15	5	5	19.5	430.15
0	0	0	0	0	0	0	21	4	13	0	18	25.6	1018.71
6	0	0	0	0	0	0	0	0	0	0	5	24	293
0	0	0	0	0	13	14	0	9	0	2	0	33.5	6814.64
0	0	9	0	0	0	0	0	0	0	0	0	1	215.05

Fig 2 & 3 – Data generated

As the result of the feature engineering, we have generated a total 27 variables of data that is derived from the given 6 variables.

The goal of feature engineering is to accomplish two primary goals:

- Create input data that is compatible with and best fits the k-mean function.
- Improving the performance of machine learning models

Created a separate variable from the transaction as the “Invoice date” and “Time” are together, and both are different data types. Hence it would not be possible to access them while used in the modelling. It is best practice to create as many variables as possible as it helps a model in better accuracy and results.

Computed total number of transactions and derived unique customer id from the data. We have created 12 new variables called with Months name that were derived from transaction date and have the transactions listed below each month made by each of the customers namely **Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.**

Computed the seasonality based on variables created previously called Autumn, Spring, Summer, and Winter.

If the transaction date falls in between Dec to March, then it is called “Winter”.

If the transaction date is in between March to June, then it is called “Spring”.

If the transaction date falls in between June to Sept, then it is called “Summer”.

If the transaction date is in between Sept to Dec, then it is called “Autumn”.

Derived Average transaction from all the monthly transactions made by the user (**Average. Transaction**).

Extracted Average of dollars spent per transaction. (**DollarsPT**) by computing the average sum of all items by total number of transactions for each customer. Also computed the Value of each customer by multiplying Average Dollar Per Transaction and total number of transactions made by the customer as “**Sum of Cost**”.

Found Average of Items purchased per transaction. (**ItemsPT**) by computing the average sum of all items by total number of transactions for each customer.

FEATURE DESCRIPTION

Variable name	**Feature Engineering variables**		
	No of values	***Type***	***Datatype***
UNIQUE_CUST_ID	3920	Non-null	Int64
Num_of_trans	3920	non-null	int64
Autumn	3920	non-null	int64
Spring	3920	non-null	int64
Summer	3920	non-null	int64
Winter	3920	non-null	int64
Noon	3920	non-null	int64
evening	3920	non-null	int64
Morning	3920	non-null	int64
Total Items	3920	Non-null	int64
ItemsPT	3920	non-null	float64
DollarsPT	3920	non-null	float64
Jan	3920	non-null	int64
Feb	3920	non-null	int64
Mar	3920	non-null	int64
Apr	3920	non-null	int64
May	3920	non-null	int64

Variable name	**Feature Engineering variables**		
	No of values	***Type***	***Datatype***
Jun	3920	non-null	int64
Jul	3920	non-null	int64
Aug	3920	non-null	int64
Sep	3920	non-null	int64
Oct	3920	non-null	int64
Nov	3920	non-null	int64
Dec	3920	non-null	int64
Average.Transaction	3920	non-null	float64
Sum of Cost	3920	Non-null	float64

Table 2 - Features generated

Data Modelling

This step is about choosing and employing the most effective modelling methods. There could be a loop back to data preparation here. Because some approaches, such as DBSCAN have easier data modelling steps but result in a higher number of clusters. Because some approaches, such as K-Means, do great in large datasets and are more efficient.

The Fig. 3, shows the flow for this research, such as Input of data, followed by Data selection and cleaning, Data Transformation such as like Categorization and Standard-Scalar and PCA generation. Data modelling that is developed here is DBSCAN and K-Means. Followed by optimal number of cluster validation by elbow method. And deciding the best method that is suitable for the business.

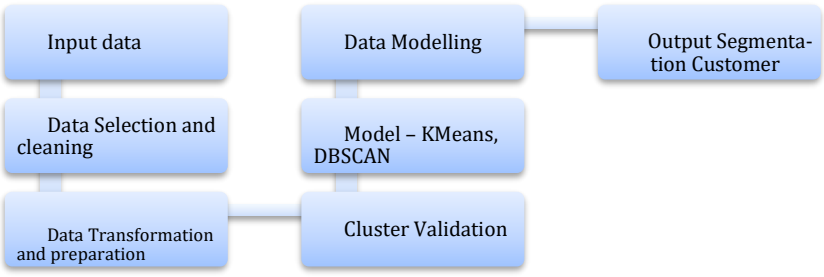


Fig. 4 – Project Flow of proposed model

Stepwise approach of the research:

- After the data file is received, we remove the Null Values, NaN and junk

values from the data and labelling of data is done as per the understanding with respective names.

- Created a separate variable from the “transaction” as the date and time are together and both are different data types.
- Created as many variables as possible as it helps a model in better accuracy and results.
- Computed total number of transactions and derive unique customer id from the data.
- Created 12 new variables called with Months name from transaction date and have the transactions listed below each month made by each of the customers namely Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.
- Computed the seasonality based on variables created previously called Autumn, Spring, Summer, and Winter.
- Computed Average transaction from all the monthly transactions made by the user (Average. Transaction)
- Found Average of Items purchased per transaction. (ItemsPT) by computing the average sum of all items by total number of transactions for each customer.
- Found Average of dollars spent per transaction. (DollarsPT) by computing the average sum of all items by total number of transactions for each customer.
- Computed the value of each customer by Average Dollar Per Transaction and total number of transactions made by the customer.
- Finally integrated all the above variables to create an input data for demonstration, wherein we categorize if the user as the high, medium, or low spender and make recommendations based on the clusters for campaigning.

Clustering with DBSCAN Algorithm

Clustering by density is an unsupervised method for identifying distinct groups in data, based on the notion that a cluster is a contiguous area in the data.

Density-based clustering is built on the **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** technique. It can detect clusters of diverse forms and sizes in enormous volumes of data with noise and outliers. DBSCAN clustering is exceptionally resilient in the face of outliers, which makes it very intriguing.

The DBSCAN algorithm uses two parameters:

- **minPts**: The number of points that must be packed together to qualify as dense (a threshold).
- **EPS:** An EPS is a distance metric that is used to find the locations of points in the neighborhood of a point.
- **Core point:** The point is a core point if there are at least minPts number of points with radius eps in its immediate surrounds (including the point itself).
- **Border point:** A border point is one where the number of points within a point's immediate surrounds is fewer than minPts and it can be accessed from a core point.
- **Outlier:** If a point isn't a core point and can't be reached from any core point, it's

an outlier.

To determine whether two points are connected, transitivity-based chaining is used. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a [8].

We gave the epsilon value as 3, and minimum number of points within each epsilon to be 4. Metric chosen as Euclidean.

The DBSCAN method resulted in 16 clusters of users' categories which is obviously is higher number of clusters. Below is the cluster that was created.

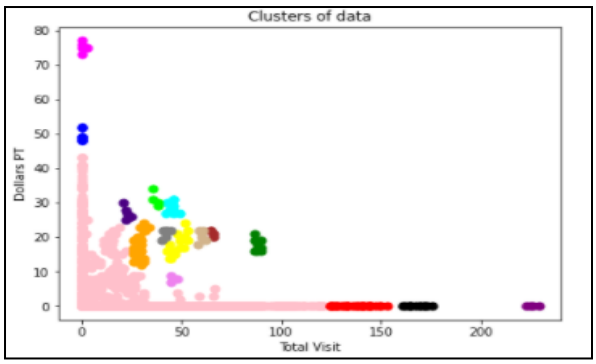


Fig 5 - Snippet of DBSCAN clusters that was generated.

The DBSCAN method resulted in 16 clusters of users' categories which is obviously is higher number of clusters. Below is the cluster that was created. The whole idea of customer segmentation is to reduce the target audience on whom we can market for any sales. It is not very efficient method when it gives us a wide spectrum of users.

Hence the DBSCAN method was not very efficient in our customer segmentation analysis as it had large number of datasets.

Clustering with K-Means Algorithm

Clustering method produces comparable clusters within clusters based on parameters. The distance between two objects is used to determine similarity. One of the most often used centroid-based algorithms is K-means. A K-means algorithm for partitioning, where the mean value of the objects in a cluster is used as the cluster's center.

Input: k: the number of clusters, D: a data set containing n objects.

Output: A set of k clusters.

Method: (1) randomly select k objects from D as the initial cluster centres; (2) repeat (3) (re)assigns each object to the cluster to which it is most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; and (5) repeat until no change occurs.

Once data is loaded, All the data variables of our analysis are **categorized**, hence it will be easier for data analysis with all the variables being categorical value. After creating the categorical variables, we fit the data to transformed data by using Standard-Scaler.

StandardScaler is a dataset standardisation methodology that is required by many ML models. If an individual feature does not resemble standard normally distributed data, the model may behave unexpectedly. It will change the data so that its distribution has a mean of 0 and a standard deviation of 1.

We do PCA, or Principal Component Analysis, after standardising the variables, which is a typical way for speeding up a machine learning process. Because scale affects PCA, we must first scale the features in your data before using PCA. Because the variables were already standardised in the previous phase. As a result, PCA is now complete, and PCA variables will be created as a result. If the model is excessively sluggish due to the input dimensions being too huge, PCA may be a good way to speed it up.

The generated PCA would be able to explain 75% of the actual data behavior. PCA generated variables are 6 variables with names P1, P2, P3, P4, P5 and P6. Running the K-means algorithm with the generated PCA variables, here we suggest the range of clusters to be created given from 2 to 10.

Here the output displays the number of clusters that’s created based on the range specified and the error_term with each clusters created. We can also see that the more clusters there are, the lower the error. We also can't choose the number of clusters that are suitable for the study based on our personal preferences.

As a result, we use the elbow approach to choose the best number of clusters to analyse. One of the most prominent approaches for determining the appropriate number of clusters with the lowest error factor is the Elbow Method.

The below image shows the output graph for running the elbow method to decide the number of clusters.

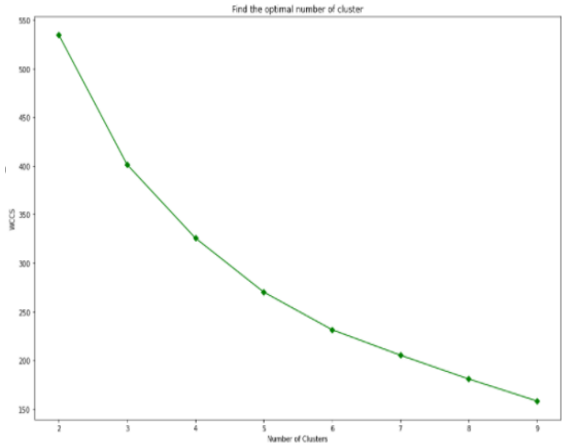


Fig 6 - Elbow graph

To determine the appropriate number of clusters, we must determine the value of k at the "elbow," or the point when distortion/inertia starts to decline linearly. As a consequence, we discover that 3 clusters are the optimal number for the provided data.

Once the building the clusters based on the optimal number of clusters, we have the output based on the segmenting of users into their purchase traits.

The below image (Fig No. 9.3.5) shows the code for running the clusters and shows the output.

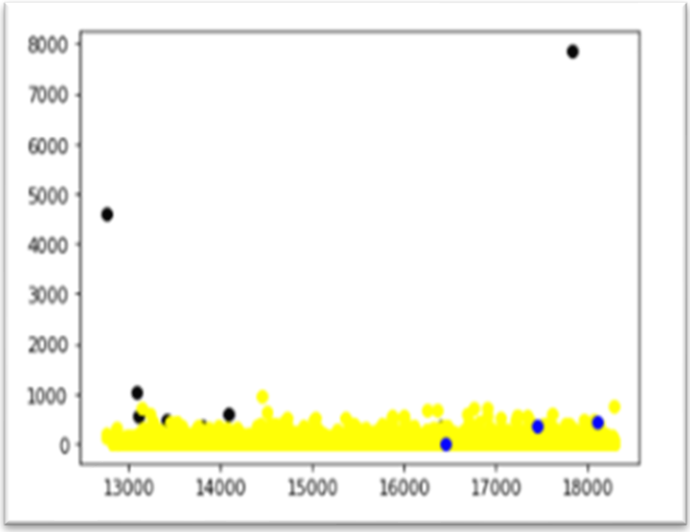
Cluster	x	y
0	2223.114032	-2546.100913
1	39081.729950	-5122.323423
0	2207.857748	-2598.943503
0	-970.656469	-2670.947192
0	-2011.798697	-2608.014565

Fig 7 - Snippet of output of clusters

From the above output we have 3 clusters of customer groups that is created for the given list of users.

If we consider the elbow method of cluster numbers and run the kmean program, we are resulted with lesser number of clusters and groups the high spenders, medium spenders, and low spenders as clusters as cluster 2, cluster 1 and cluster 0 respectively.

Dollars PT



Number of visits

Fig 8 : Kmeans clusters output

As a result, KMeans model resulted in 3 groups of clusters of customers and they are as follows,

- Yellow set of clusters being Low spenders of customers.
- Blue set of clusters being Medium spenders of customers.
- Black set of clusters being High Spenders of customers.

Result Analysis

After studying data and uncovering relevant patterns that solve the business purpose in the first four phases of the CRISP-DM process, it's time to raise the question: Are the outcomes good? Not only are models evaluated, but also the method used to create

them, as well as their potential for practical application. In general, data evaluation consists of two key tasks: task assessment and process review.

The whole idea of customer segmentation is to reduce the target audience on whom we can market for any sales. It is not very efficient method when it gives us a wide spectrum of users as the output as it will not be very successful in reducing the cost involved in sales and have much profit.

Comparison between DBSCAN Clusters Versus KMeans Cluster

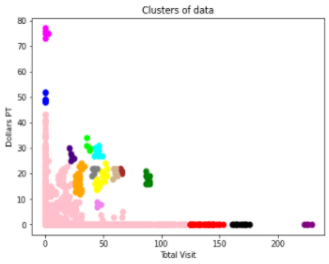


Fig 9.1 DBSCAN clusters

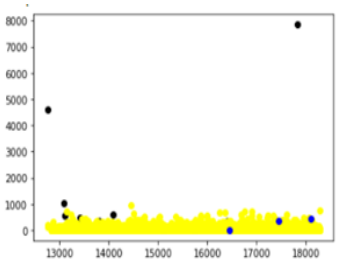


Fig 9.2 Kmeans clusters

The results of DBSCAN method was not very efficient in our customer segmentation analysis as it resulted in large number of clusters - 16 clusters were created as the result. But where as in Kmeans it resulted in lesser number of clusters from the elbow methos and groups the high spenders, medium spenders, and low spenders as clusters as cluster 2, cluster 1 and cluster 0 respectively.

To compare the outcomes of the two approaches, the Silhouette score is employed. Based on how well samples are grouped with comparable data, it is used to test the validity of clusters created using clustering methods like K-Means and DBSCAN.

The Silhouette score is calculated for each sample of distinct clusters. To establish the Silhouette score for each observation, the following distances must be computed for every observation belonging to all clusters.

The average distance between the observation and the cluster's remaining data points. The mean intra-cluster distance is also known as this ***a***. A Mean distance between the observation and all other data points in the next closest cluster is calculated. This is also known as the mean nearest-cluster distance. The mean distance is denoted by ***b*** [1].

Silhouette score, S, is calculated using the below Formula (1):

$$S = \{(b - a)\}/\{\max(a, b)\} \quad (1)$$

With a score of 1, the cluster is compact and well-separated from its neighbors. A score around 0 suggests overlapping clusters with samples close to the decision border of neighboring clusters. A negative score suggests that the samples may have been grouped incorrectly as per [1].

The Silhouette score for DBSCAN model is **"-0.1374"** for which it signifies that this method is not a very efficient for this dataset.

Whilst the Silhouette score for K-Means model is **"0.7343"** which signifies that this modelling method is a very efficient model for this dataset. Silhouette analysis for K-Means model with suggested number of clusters as **3** which signifies that modelling method is a very efficient model for this dataset. We see better distribution of customers in their clusters 0, 1 and 2.

- Cluster 0 - Medium Spenders Customer
- Cluster 1 - High Spending Customer
- Cluster 2 - Low Spending Customer

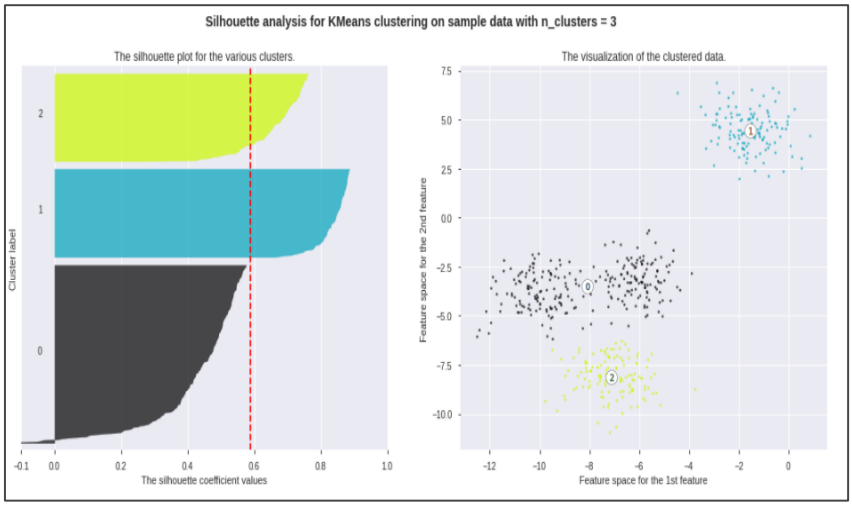


Fig 10 - Silhouette analysis for KMeans

Hence, we consider K-Means algorithm to be best methodology for this dataset of customer segmentation to win more sales with the existing customer.

Conclusion

As discussed in the previous section, the overall goal of this paper is creating a customer segmentation for a retail company market strategy. Since the algorithm DBSCAN gave us 16 clusters of groups to analyse with. We are going with K-means of analysis which gives us 3 clusters of spenders based on the Silhouette score.

References

[1] S. (. Patil, "CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL DATA: A MACHINE LEARNING APPROACH.," *International Journal for Research*, 2021.

[2] H. J. S. G. B. &. S. A. Stefanovic, "Science and Higher Education in Function of Sustainable Development.," *Mećavnik-Drvengrad.*, 2020.

[3] T. Kansal, "Customer Segmentation using K-means Clustering.," *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 135–139.*, 2018.

[4] M. K, "Customer Segmentation using Machine Learning.," *International Journal for Research in Applied Science and Engineering Technology, 8(6), 2104–2108.* , 2020.

[5] M. Kamber, Data Mining Concepts and Techniques, 2014.

[6] I. L. D. R. A. L. &. V. F. M. de Moura Ventorim, "BIRCHSCAN: A sampling method for applying DBSCAN to large datasets. Expert Systems with Application," 2021.

[7] R. W. M. F. A. &. C. K. Sembiring Brahmana, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 2020.

[8] P. Sharma, "The Most Comprehensive Guide to K-Means Clustering You'll Ever Need," 2019.

[9] S. Ajitesh Kumar. (2020, "K-Means Silhouette Score Explained With Python Example.".