Societal Transformation: AI and Big Data Journal
Vol: 3(1): 27-49, 2025
DOI: 10.20547/aibd.253103

A comparison of k-means and mini-batch k-means algorithm for customer segregation analysis

Hania Marfani[1], Hira Kamal[1], Abdul Samad Hussain[1], Darakhshan Syed[2]

Abstract: This study applies k-means and mini-batch clustering dataset for customer segmentation. Based on the analysis, various clusters were formed and validated. It was found that the results of both clustering techniques are same with an enhancement in processing speed via the use of mini-batch k-means. Customer segmentation can be used for market intelligence to identify interested clients by giving corporate entities in the retail sector pertinent and relevant facts periodically. It can be used as methods for examining customer purchase patterns and sales trends. The clustering has been applied over a dataset extracted from Kaggle. After performing the exploratory data analysis, the customers were divided in to platinum, gold, silver and bronze categories. The data were clustered using both the clustering methods (elbow methods were used to cluster the data). The results were found to be same for both clustering techniques. The customers were segmented into platinum, gold, silver and bronze. Results highlighted that mini-batch is more efficient computationally along with providing similar segmentation results as k-means. The validation of the results was performed using elbow method, Silhouette coefficients score was used for further analysis.

Keywords: *K-Means, mini-batch, clustering, customers relationship management, business intelligence, segmentation*

1. Introduction

Customer segmentation is the technique of assembling groups of individuals according to shared traits. These client groupings are helpful for promotional activities, for finding clients who might be productive, and for fostering customer trust. Following are examples of typical consumer segmentation: client experience segmentation, geographical segmentation, behavioral segmentation and psycho-graphic segmentation.

When we have customer segmentations, we can develop appropriate results i.e., decide on the ideal dissemination and placement, and tailor the sales process to each client. This can be done all while gradually improving the segments. If used correctly, the strategy can instantly help everyone at the firm and provide understanding of the clients. It is however noted that segmentation is not capable of accounting for all strategic factors or alterations in products,

[1]*NED University of Engineering and Technology, Pakistan*

[2]*Bahria University, Pakistan*

prices, and packaging over the period. Instead, it is a continuous activity to identify important variations among your clients.

For understanding the consumers, the targeted market and to communicate this information across collaborative individuals, segmentation is essential for a developing organization. Beyond a certain organizational structure, one can't live without it. There are various segmentation aspects that should be considered before its application and one should carefully consider what is best and adaptable for the company.

Demographic trends are often the main emphasis of the consumer understanding process. This will take into account factors like age, location, where is urbanization occurring—urban or rural?, revenue, the status of relationships, the family and the nature of employment. By extending the valuable database of customers, segmentation strengthen customer loyalty and also strengthen enterprise-level for long-term relationships in the face of escalating competitiveness and a more complex situation of business [1].

Based on the literature, the qualitative perspectives and the quantitative views are the two most common segmentation types utilized in k-means. In the context of the present study, quantitative analysis is employed for segmentation clusters. Well-defined customer segmentation helps marketers to target a particular segment of clients and fosters the development of positive, long-lasting relationships with those customers. The retail business [2] is one of the primary sectors where behavioral segmentation and information extraction can be implemented because it necessitates a large number of data on purchases, shipping, intake ratio, reinstatement service, and many other areas.

Clustering is a crucial tool for data mining studies and applications. It can be used for customer segmentation. Many academic disciplines, including computer science, predictive analysis, statistical data, pattern classification, cognitive computing, and machine learning, are actively working on this topic[3]. Numerous clustering methods have been studied in literature, and the majority of them have been able to provide good results in the aforementioned fields. According to the commonalities of the input data features, it is define as a way of categorizing data into numerous groups called clusters. Various data clustering strategies have been developed and used in recent years to address various data clustering issues[4]. The two primary categories of clustering analysis approaches are hierarchical and partitioned. Despite the fact that the approaches in these two groups have shown to be quite effective and efficient, they often rely on the user having knowledge of the precise quantity of clusters for each sample that needs to be grouped and analyzed beforehand[5].

In this research, the problem of customer segmentation is solved using mini-batch k-means clustering algorithm. k-mean is one of the most straightforward approaches to clustering. However, mini batch k-means approach is a variation that employs mini-batches to speed up computations. It reduces the computing cost of clustering. The k-means technique can be generally chosen, but when dealing with a massive dataset, one should choose the mini-batch technique[6, 7].

The proposed solution has been compared with k-means clustering algorithm. Rest of the sections of the paper has been organized as follows: The next section provides a review of the literature. Then, the methodology applied is discussed. Exploratory data analysis is performed over the dataset which are briefly discussed. The results are then presented. The paper

concludes with discussions on future work. Before moving forward, we discuss the major objectives of this research. The objectives of this research are as follows:

- To investigate the current state-of-the-art research on customer segmentation analysis
- To perform customer segmentation analysis to identify various types of customer
- To apply k-means and mini-batch k-means algorithms over the customer segmentation and evaluate the results

While there are a lot of studies already available on analysis of k-means and min-batch k-means, this study is differentiated from those studies based on employing this for a specific domain of CRM. The evaluation is not only limited to accuracy but other parameters as well.

2. Literature Review

We will investigate the results from following two dimensions. First the literature review on customers' segmentation is performed. Then, a brief overview of clustering techniques proposed in literature for customer's segmentation is provided.

2.1 Literature review on customers segmentation

Marketing to particular groups is prevalent and crucial from the standpoint of conventional mass marketing. The Customer Relationship Management value chain includes a number of operations that include customer segmentation. The author uses multidimensional segmentation to present an unique modeling approach for the postal region. In postal service organizations, this CRM design is helpful [1, 8, 9].

Khalili et al. [1] proposed a hybrid strategy which is intended to forecast types of organizations for customer-centric businesses. The classification tree methodology, clustering, rule collection, and other techniques are the foundation of this work. Based on past consumer's behavior in distinct segmentations, the authors' used k-means to forecast upcoming activities. The hybrid feature extraction for filtering producing and decisions-making technique is employed to help with this. To forecast overall sales and profit targets for a company, the current study can also be used in the insurance and telecommunications sectors. This approach is particularly good at forecasting profitable clients as well as spotting new customer behavior.

According to Khajvand et al. [9], the idea of client's segmentation can be used by businesses in the aesthetics and healthcare industries, which consequently contributes to the management of customer relationships. The researchers employed two segmentation methods, the first of which is based on the Recency, Frequency, and Monetary (RFM) framework and the second of which extends RFM by include the count the parameter of the entity. By applying weighted RFM to calculate the value customer long term engagement, sales and marketing tactics are described.

According to Song et al. [10], numerous analytics based on the combination of the customer relationship management (CRM) and RFM models is crucial for researching CRM in big data. You et al. [11] while presenting a methodology for marketing forecasts, proposed that the RFM model be used to forecast the monthly supply amount by cluster analysis the clients

using the K-Means technique. Utilizing CHAID decision trees centered on attribute readings, each group is identified.

The connection between customer behavior and order fulfillment is discussed by Nguyen et al.[12]in the context of marketing, and activities are discovered using a variety of marketing techniques, which improves customer satisfaction standards. The traders who use the RFM model which is based on Electronic Funds Transfer at Point of Purchase in commercial settings are grouped by Singh et al. [13] using a clustering algorithm.

Researchers [14-16] claim that a business intelligence technique for consumer segmentation is offered depending on client visitation to the site, acquired from the entire sales reports. An attribute selection strategy is also suggested, using customer category outputs as applied sources and product taxonomies as outputs. The transactional datasets used in the current work is used to perform RFM analytics, which analyses the very same dataset utilizing unsupervised methods like K-Means and Fuzzy C-Means. Additionally, the author has developed a novel concept by choosing centroids in the K-Means technique and contrasting the outcomes.

Brito et al.[17] proposes that the data mining method of clustering is able to overcome a number of challenges in the production and advertising of clothing. Segmentation is crucial for identifying trends in consumer requirements, as is obvious.

Qualitative characteristics are necessary for audience segmentation in order to recognize behavioral traits. Nevertheless, descriptive parameters are inadequate in some fields. In light of this, Murray et al.[18] introduced the segmentation technique as a solution to the issue, resulting in enhanced efficiency. Customer segmentation is the process through which a market is divided into different buyers based on the specific behaviors and characteristics they exhibit. Customer segmentation helps business organizations to develop high-quality and lasting relationships with their clients and identify the characteristics of various consumer groups. K-means clustering is applied in customer segregation to divide the customer population into a number of groups such that the customers in one group are similar to customers in in the same group and different to the customers in other groups in particular ways and attributes. K-means clustering algorithm is the most widely used algorithm for customer segregation and has provided high computation speed and less clustering time in customer segregation analysis.

k-means and self-organizing SOM methods are utilized by Qadadeh et al. [13] to classify client attributes using an RFM framework for insurance datasets. It assists in determining the demands, expertise, and qualities of the customers. According to the study, SOM- and fuzzy-based clustering techniques are more effective than conventional ones. Arunachalam and colleagues suggest [15] It employs an sequential exploratory research strategy that blends qualitative and quantitative techniques. Customer segmentation is one of the prerequisites for the crucial CRM in the field of customer loyalty, and it may also be implemented to clients of e-pharmacies to boost e-customer loyalty [14]. Customer categories shift throughout time as a result of the market's significant unpredictability.

The sensitivity of k-means in customer segregation to initialize the cluster center is very high because it selects the average value to be the cluster center. It is an iterative process in which the cluster center keeps being recalculated and the process stops when there are finally no changes to the cluster. K-means provided the highest level of validity in customer segregation

analysis with a greater possibility of the best customer class to emerge [19]. K-means clustering was applied to data set without dimensionality reduction to observe the impact of clustering besides the process of reduction and different number of dimensions was chosen each time. Internal and external cluster validations were performed and clustering performance was evaluated through Silhouette index and DB index[15].

K-means algorithm is employed to apply business intelligence approach to identify the potential customers of a retail industry. The data set encompasses the buying patterns as well as behavior of customers from various regions. It is an RFM-based research and an optimal solution was found through a variety of dataset clusters which were evaluated using Silhouette analysis. K-means was applied on a customer segregation analysis which was based on PFM model (Presence, Frequency, Monetary value). The output of k-means was compared with that of other grouping algorithms, accompanied by the implementation of Apriori algorithm which was used to assess the associations among the products for respective customers. Each cluster number was assigned a specific level which described the loyalty of respective customers which served as a decision attribute to generate classification rules [20]. One major problem with k-means clustering algorithm is that it has a high computational cost and consumes much time to compete the analysis. Furthermore, the efficiency of k-means significantly decreases while working on large data sets because it does not iterate over the entire data sets. In contrast, working with mini-batch k-means is considered to be a more useful and efficient approach. This approach works by dividing large data into small batches randomly which are saved in memory and then each batch is collected on each iteration for updating the clusters[21].

Mini-batch k-means approach is applied in a study in which the idea of 'divide and conquer' is applied to logically divide the large data into multiple small data sets called batches. This not only reduced the computation time of the algorithm but also optimized the objective function of clustering. Mini-batch k-means algorithm proved to be simple as well as easy-to-understand and -implement which speeded up the clustering speed and reduced the operation time while maintaining the accuracy of data [22].

Researchers explain about how the important performance factors for CRM are assessed in relation to the implications of big data. The findings include three main features, five assertions, and prior contributions. To improve decision-making for repeat purchase, a data mining CRM methodology is designed using naive Bayes and artificial neural networks categorization technique. To create an integrated platform, research sources from several disciplines—such as advertising, business, managing the information systems, and other aspects are connected [23-25].

Despite the earlier studies employing RFM, hybrid clustering and traditional k-means, there are some gaps in the literature. Most of them focused on smaller datasets or static environments, limiting applications for larger environments. In addition, rarely any study has addressed the issue of high computational cost and scalability issues. Therefore, studies need to address scalable and accurate segmentation techniques in modern CRM applications.

2.2 Algorithms for cluster analysis

Despite the existence of a number of cluster analysis approaches, hierarchical-agglomerative and iterative partitioned clustering is among the most popular and widely employed. In

specifically, the k-means method for partitioned clustering and single, average, complete, center, and Ward linkages for hierarchical are described.

Recursively locating nested groups can be done using either agglomerative or split methods through hierarchical clustering algorithms. Each data sample begins in its own group in an aggregative clustering, which then combines gradually similar pairs of clusters to produce a hierarchical structure. As an alternate, divisive clustering breaks each group into smaller ones after starting with all the sample points in one group. Readjustment is not possible with cluster analysis since once partitions or combinations are made, they cannot be undone [26]. To calculate how identical two items are, you need to know their distance or closeness. The Euclidean distance is used most commonly. A dendrogram, that is a two-dimensional architecture similar to tree showing the series of nested groups, is the product of a hierarchical clustering technique[27].

In hierarchical methods, several proximity metrics are employed to combine clusters. Single, complete, average, center, and Ward linkage are typical examples.

Non-hierarchical or partitional clustering algorithms divide the data into all the clusters at once. Scholars[28] first developed the k-means clustering technique more than 50 years ago, and later the researchers[29, 30] refined it. The updated technique has resulted in its use in a number of industries, including psychology, marketing research, healthcare, and biology. Although numerous additional clustering techniques have been created since then, k-means continues to be one of the most popular techniques because of its convenience, flexibility to use, and effectiveness[3].

K-means does discover a local minimum for a provided initial set of centroids, however it does not always locate the global minimum. K-means is repeatedly run to look for variations in clustering caused by various beginning centroids. Additionally, the k-means technique is a member of the family of clustering methods that call for an a prior selection of the required quantity of clusters.

In hybrid clustering, the results from one clustering approach is utilized as the inputs for the other, combining the benefits of both. To be more precise, the first phase is hierarchical clustering, and the next is k-means. The purpose of combining these two approaches in this manner is to utilize the hierarchical clustering to determine how many clusters should be used as "seeds" in the k-means approach. With no a priori cluster count determination necessary in the first phase and the speed of the second phase, this technique incorporates the best aspects of both.[31]

The k-medoids, hidden Markov modeling approach, mixture modeling, and fuzzy c-means (FcM) are a few other well-known clustering techniques in addition to those already covered[32-34]. The benefits and drawbacks of the various clustering techniques are compared in Table 1.

Table 1: Comparative summary of the benefits and drawbacks of various clustering techniques

Clustering technique	**Reference**	**Benefit(s)**	**Drawback(s)**
k-means	Jain et al. [3]	Minimum complexity Faster computation Able to handle huge data volumes Adjustable cluster membership is possible	Mandatory to predetermine the no. of clusters. Not consider outliers Inability to handle quasi clusters of different densities and sizes. Unable to deal with large scale the data sets The effectiveness totally depends on the initial centroids.
k-medoids	Madhulatha and T. Soni[32]	Adjustable cluster membership is possible Better consideration of outliers in contrast to k-means	Mandatory to predetermine the no. of clusters. Results vary depending on the starting centroids.
Hierarchical Clustering	Liao and T. Warren[26]	No mandatory requirement to predetermine the no. of clusters. A dendrogram offers a visual illustration. Capable of recognizing groups of diverse attributes.	Higher complexity Slower processing Clusters cannot be altered once they have created. It could be challenging to choose the dendrogram's extent of cutting. Clusters premised on the chosen distance measure.
Hidden Markov model		Adaptability in managing different data kinds.	Need a lot of parameters. Work with huge data sets.
Mixture models	Gómez-Losada et al. [35] Lozano-Garca author Environmental, Antonio[36]	A few variables can be used to create clusters.	Cost - effective if there are many distributions or if there are few observed data sources in the data collection. Mandatory requirement of large data sets. Estimating the number of clusters is challenging.
FcM	Chu et al. [37]	Flexible cluster placement is possible. Particularly accurate in its capacity to estimate the likelihood of becoming a member of a cluster.	Extremely complex. Mandatory to predetermine the number of clusters. Possibly leading to a local optimal solution

This study employed two popular clustering algorithms, k-mean and mini-batch k-means. These algorithms serve as a benchmark because of their simplicity and effectiveness. Mini-batch is popular for its reduction in computational cost using small random batch despite maintaining clustering quality. By comparing these two methods, this paper checks if mini-batch k-means can deliver similar segmentation accuracy as k-means.

3. Methodology

This section demonstrates the proposed study of the marketing dataset. The study performs the following steps: data preprocessing and preparation, feature engineering, exploratory data analysis, cluster modeling and then analysis based on that modeling. The proposed methodology has been briefly described in Figure. 1. We started with the data acquisition phase; the data is cleaned, preprocessed, customer segmentation is performed and then the final results can be used for precision marketing strategy. These steps are now described in detail in next few sections.

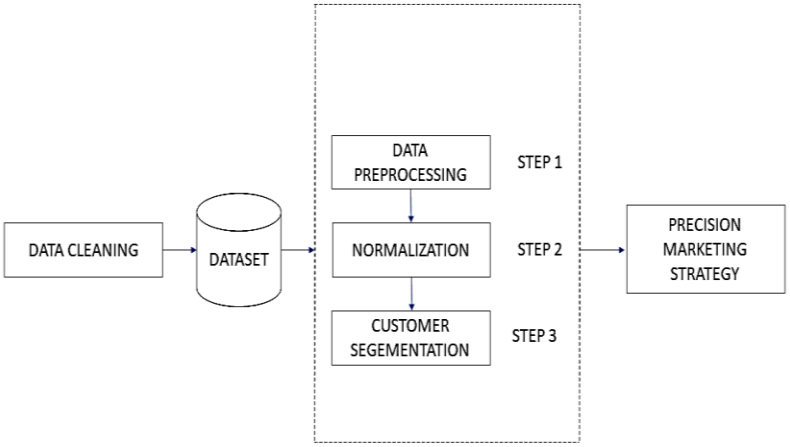


Figure 1:Major steps for proposed methodology

3.1. Data Preprocessing and Preparation

A dataset 'Marketing campaign' is retrieved from the Kaggle(Figure 2) which consists of four main heads: people, product, promotion, and place. A response variable titled 'target' is labeled in the dataset in which clustering is needed to be performed in order to determine the segments. The dataset contains some noise and missing values, therefore, is needed to be cleaned for the clustering to be performed. Firstly, all the missing values are evaluated and then those attributes having more that 10% missing values are separated. In this case, customer column has more than 10% NA values and has little impact on the dataset, so it is dropped from the analysis. The rest of the data is checked for any more outliers, and they are also removed.

1	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases
2	5524	1957	Graduation	Single	58138	0	0	2012-09-04	58	635	88	546	172	88	88	3
3	2174	1954	Graduation	Single	46344	1	1	2014-03-08	38	11	1	6	2	1	6	2
4	4141	1965	Graduation	Together	71613	0	0	2013-08-21	26	426	49	127	111	21	42	1
5	6182	1984	Graduation	Together	26646	1	0	2014-02-10	26	11	4	20	10	3	5	2
6	5324	1981	PhD	Married	58293	1	0	2014-01-19	94	173	43	118	46	27	15	5
7	7446	1967	Master	Together	62513	0	1	2013-09-09	16	520	42	98	0	42	14	2
8	965	1971	Graduation	Divorced	55635	0	1	2012-11-13	34	235	65	164	50	49	27	4
9	6177	1985	PhD	Married	33454	1	0	2013-05-08	32	76	10	56	3	1	23	2
10	4855	1974	PhD	Together	30351	1	0	2013-06-06	19	14	0	24	3	3	2	1
11	5899	1950	PhD	Together	5648	1	1	2014-03-13	68	28	0	6	1	1	13	1
12	1994	1983	Graduation	Married		1	0	2013-11-15	11	5	5	6	0	2	1	1
13	387	1976	Basic	Married	7500	0	0	2012-11-13	59	6	16	11	11	1	16	1
14	2125	1959	Graduation	Divorced	63033	0	0	2013-11-15	82	194	61	480	225	112	30	1
15	8180	1952	Master	Divorced	59354	1	1	2013-11-15	53	233	2	53	3	5	14	3
16	2569	1987	Graduation	Married	17323	0	0	2012-10-10	38	3	14	17	6	1	5	1
17	2114	1946	PhD	Single	82800	0	0	2012-11-24	23	1006	22	115	59	68	45	1
18	9736	1980	Graduation	Married	41850	1	1	2012-12-24	51	53	5	19	2	13	4	3
19	4939	1946	Graduation	Together	37760	0	0	2012-08-31	20	84	5	38	150	12	28	2
20	6565	1949	Master	Married	76995	0	1	2013-03-28	91	1012	80	498	0	16	176	2
21	2278	1985	2n Cycle	Single	33812	1	0	2012-11-03	86	4	17	19	30	24	39	2
22	9360	1982	Graduation	Married	37040	0	0	2012-08-08	41	86	2	73	69	38	48	1
23	5376	1979	Graduation	Married	2447	1	0	2013-01-06	42	1	1	1725	1	1	1	15
24	1993	1949	PhD	Married	58607	0	1	2012-12-23	63	867	0	86	0	0	19	3
25	4047	1954	PhD	Married	65324	0	1	2014-01-11	0	384	0	102	21	32	5	3
26	1409	1951	Graduation	Together	40689	0	1	2013-03-18	69	270	3	27	39	6	99	7
27	7892	1969	Graduation	Single	18589	0	0	2013-01-02	89	6	4	25	15	12	13	2

Figure 2: Screenshot of the dataset 'Marketing campaign' which is retrieved from www.kaggle.com

3.2 Dataset description

Before proceeding further, we will first discuss the dataset used for analaysis.

Features of the dataset

Listing 1 shows the features of the dataset.

Listing 1: Features of the dataset

- AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise
- Complain - 1 if customer complained in the last 2 years
- DtCustomer - date of customer's enrolment with the company
- Education - customer's level of education
- Marital - customer's marital status
- Kidhome - number of small children in customer's household
- Teenhome - number of teenagers in customer's household
- Income - customer's yearly household income
- MntFishProducts - amount spent on fish products in the last 2 years
- MntMeatProducts - amount spent on meat products in the last 2 years
- MntFruits - amount spent on fruits products in the last 2 years
- MntSweetProducts - amount spent on sweet products in the last 2 years
- MntWines - amount spent on wine products in the last 2 years
- MntGoldProds - amount spent on gold products in the last 2 years
- NumDealsPurchases - number of purchases made with discount
- NumCatalogPurchases - number of purchases made using catalogue
- NumStorePurchases - number of purchases made directly in stores
- NumWebPurchases - number of purchases made through company's web site
- NumWebVisitsMonth - number of visits to company's web site in the last month
- Recency - number of days since the last purchase

The missing values distribution of the data is shown in Figure 3.

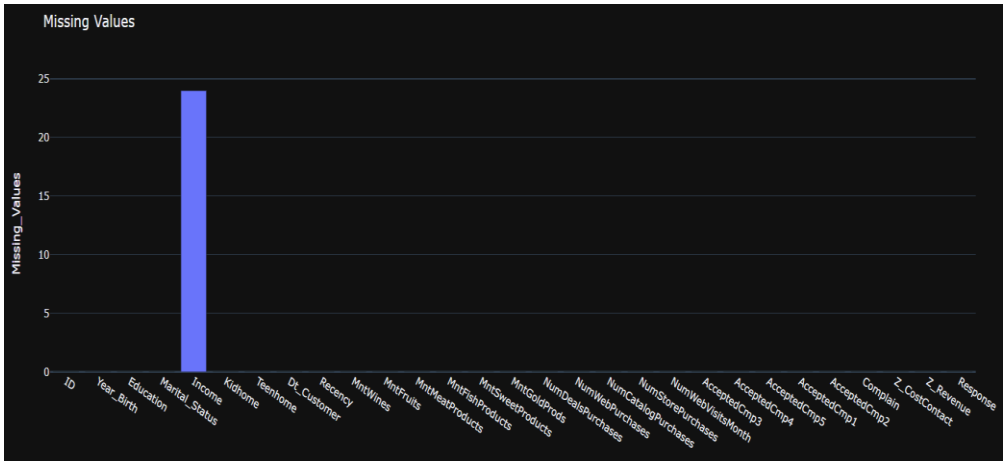


Figure 3: Missing values distribution of the data

3.3 Feature Engineering

During this process, different features are evaluated. The features with high correlation are combined, while others are further derived to form more clarity such as age instead of date of birth (DOB). Age range is determined in order to analyze impact of children in a home on purchasing. Marital status is also further defined. Total spending is calculated to evaluate the spending habits of people. Then outliers are further deducted.

4. Application of mini-batch k-means for cluster analysis

By minimizing the noise in the dataset during the pre-processing phase of data collection as well as by using the k-means method and mini batch k-means method, the effectiveness of clustering can be further enhanced. Although k-means is effective for clustering, it has limitations including slow processing, sensitivity to outliers, inefficiency for large datasets, and a lack of reliability[38]. Due of this, a number of techniques to lower the algorithm's cost have been suggested. The Mini batch k-means algorithm is another method. The primary innovation behind the mini batch k-means method is using short, fixed-size randomized groups of data so they can be retained in memory. At each step, a random subset of the dataset is taken. These random dataset is used to construct the clusters. This whole process is repeated till convergence is achieved. During each micro batch, learning rate is decreased with the increase in number of iterations. The clusters are updated using a converging assortment of the parameters of the samples and the input. It is observed that the size of data employed for a cluster is inversely proportional to the learning rate. The overall convergence can be achieved when there are no further adjustments in the clusters during a set of successive iterations. This is because the impact of fresh data is diminished as the iteration count increases. For every successive round, the method employs a small, random subset of the data. Based on the previous cluster position, the data in the batch is allocated to a cluster further. Based on the current data, modifies is made for the centroids of the cluster's positions. Contrary to a typical batch k-means change, the change we made is based on the gradient descent.

Listing 2 shows the pseudo-code for mini-batch k-means algorithm.

Listing 2: Pseudo-code for mini-batch k-means

```
Input dataset: D = {d1, d2, d3, .....,dn},
Set t = total iterations
Set b = size of batch
Set k= total clusters formed.

Set Clusters_formed C= {c1, c2, c3, ......ck}

initialize k cluster centers P = {p1, p2, .......pk}
# _initialize each cluster
Ci = 0 (1=< i =< k)
# _initialize no. of data in each cluster
Nci = 0 (1=< i =< k)

for j=1 to t do:
    # M is the batch dataset and xm
    # is the sample randomly chosen from D
    M = {xm | 1 =< m =< b}

    # catch cluster center for each
    # sample in the batch data set
for m=1 to b do:
pi(xm) = sum(xm)/|c|i (xm ε M and xm ε ci)
end for
    # update the cluster center with each batch set

     for m=1 to b do:
        # get the cluster center for xm
        pi = pi(xm)
        # update number of data for each cluster center
        Nci = Nci + 1
        #calculate learning rate for each cluster center
        lr=1/Nci
        # take gradient step to update cluster center
        pi = (1-lr)pi + lr*xm
     end for
end for
```

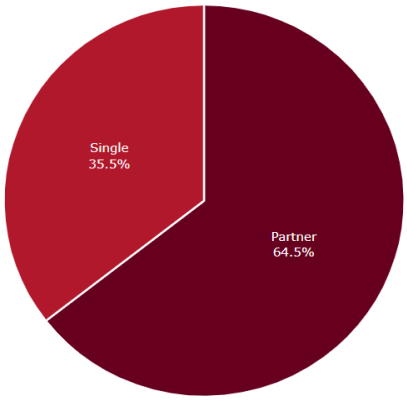
For both clustering algorithms, we determine convergence based on the stability of centroids across iterations. The convergence is achieved when successive iterations don't lead to any change in clusters. The same criteria were employed for both algorithms. In addition, for mini-batch k-means, reduction in inertia was also observed. The employment of a consistent policy for each clustering algorithm ensures unbiased evaluation of both approaches.

5. Results and Discussion

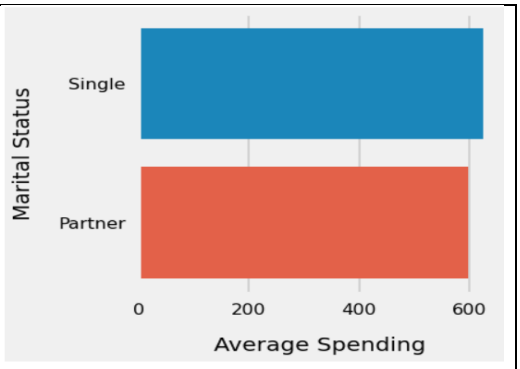
5.1 Exploratory data analysis

The attributes extracted were analyzed and then visualized via exploratory data analysis. The marital status of the customers is 35.5% single while 64.5 people having partners. Figure 4 shows that results of the exploratory data analysis. It is found that the average spending of single people is higher than people with partners. Almost half the customers are graduates while 21.6% having Ph.D. and third highest degree holders with 16.6% are Masters Students. Customers having a single child are half the total while with two children are 18.8% people

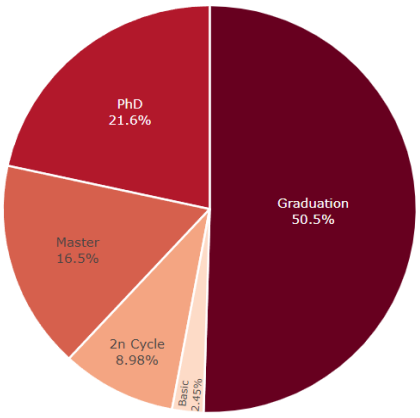
and without any child are 28.5%. It is found through the exploratory data analysis that the customers with no kids tend to have highest spending. Finally, the age of customers were found to follow normal distribution as shown in Figure 3. Most of the customers belong to middle-aged group (53.8%), 31.7% belong to adult , and 14.5% belong to senior Adult group. Income distribution of customer is also normally distributed. The most favorite products are the wine and meat, fruits and sweets are less bought so need more proper marketing.



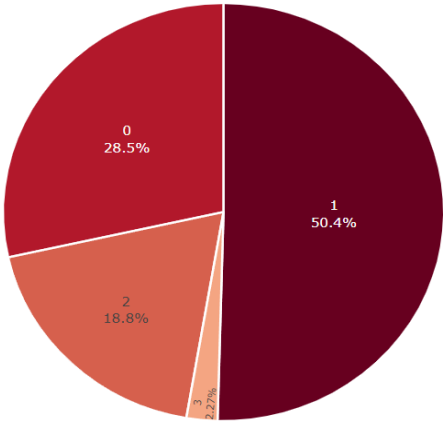
a. Marital status of customers



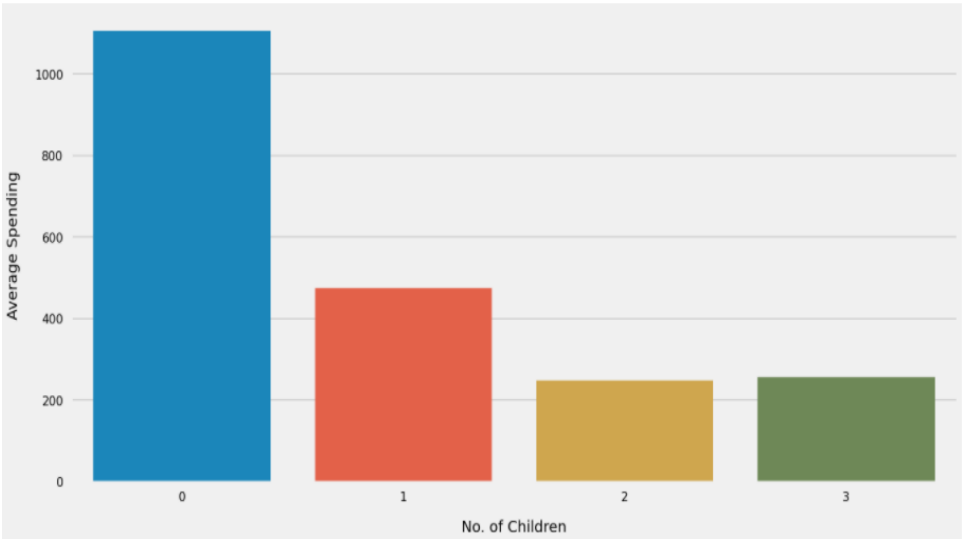
b. Status of spending based on marital status



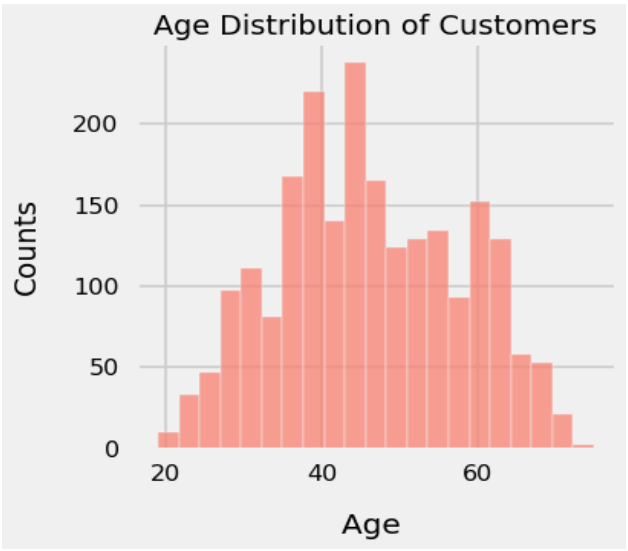
c. Education status of customers



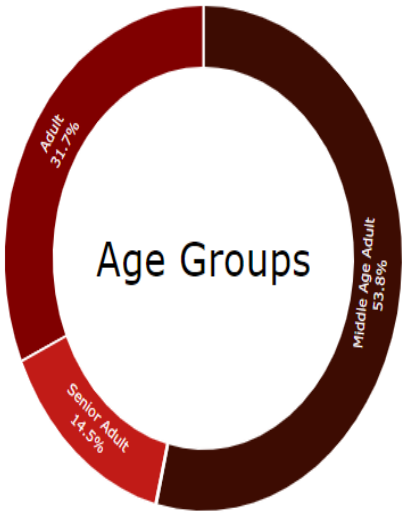
d. Customers status children-wise



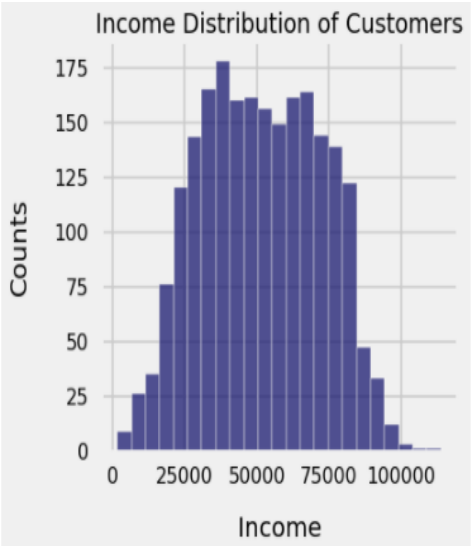
e. Spending of customers based on children



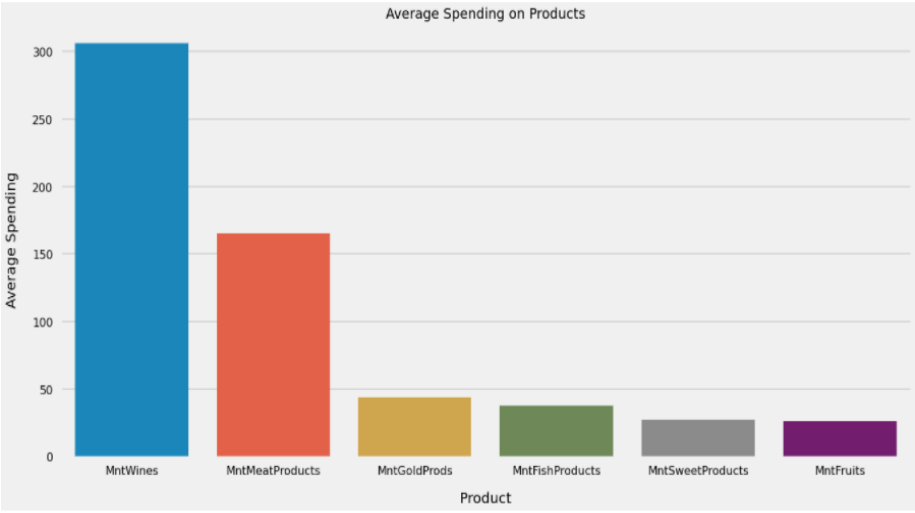
f. Age distribution of customers



g. Age group of customers



h. Income distribution of customers



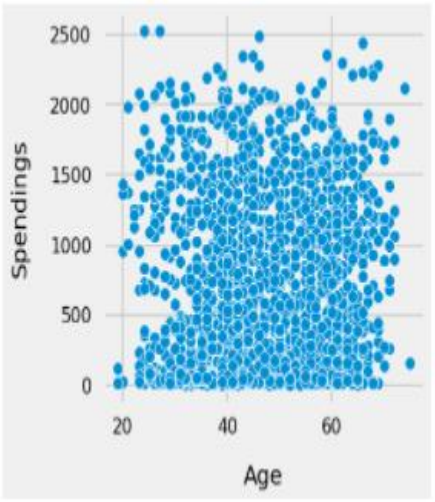
i. Average spending of various products

Figure 4: Exploratory data analysis

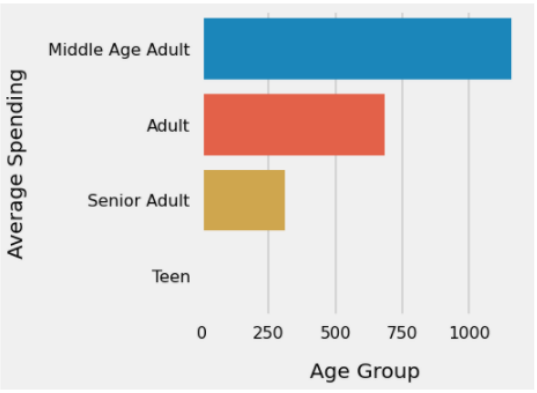
5.2 Analysis of the relationship between data items

In the next step, the data were analyzed to determine the relationship between data elements (Figure 5). It is found that age and spending have somewhat positive relation. So, a younger person can spend lesser amount as compared to a person of old age. This can be seen from the correlation analysis shown in Figure 5a. In fact, middle- aged adults tend to spend the most. There is a positive relationship between income and spending. The people with more income tend to spend more.

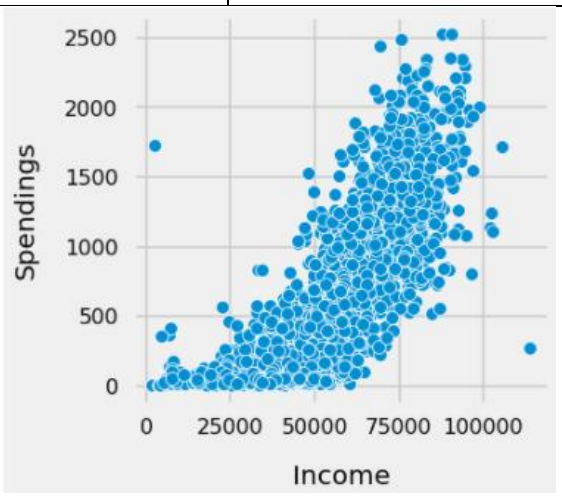
Despite a positive relation between income and spending, the causation is not necessarily implied. Other latent factors such as lifestyle and family size may have an impact on this trend. Further, the analysis doesn't take care of the outliers which may influence the trends. Further validation can be performed to strengthen the findings.



a. Relationship between age and spending



b. Average spending age-wise



c. Relationship between income and spending

Figure 5: Analysis of relationship between features

5.3 Analysis of k-means and mini-batch k-means

In this research, k-means and mini-batch model is selected as a model to form clusters for the analysis of customer. Most models were selected in order to measure the best for the current dataset.

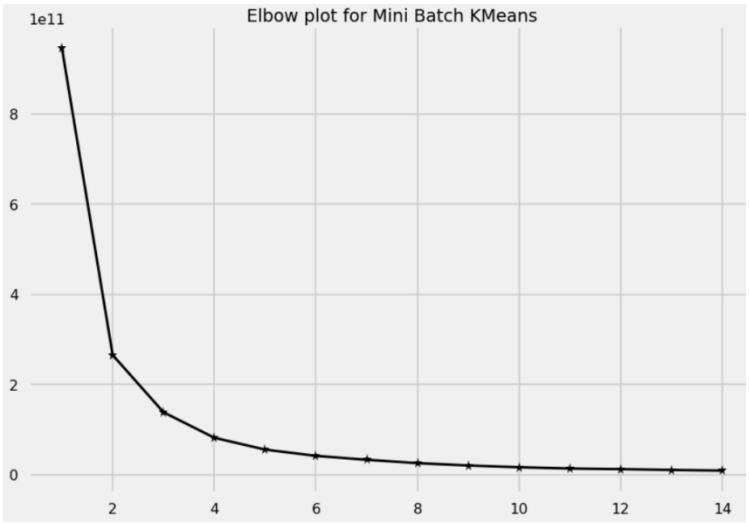


Figure 6: Elbow method applied to min-batch k-means

Elbow method is used to determine the number of clusters to be initialized instead of randomly selecting and choosing the clusters.

The performance of the algorithms is dependent heavily on the hyperparamters such as batch size. The batch size was set to 500 and number of iterations to 500. The elbow method is used to select the number of clusters.

Customer segmentation analysis

Figure 7 shows the scatter plot of mini batch k-means on the given dataset.From the above analysis, we have segmented the customers into 4 groups based on their income and total spending:

- Platinum: The ones with highest earnings and highest spending
- Gold: The ones with high earnings and high spending
- Silver: The ones having low salary and less spending
- Bronze: The ones having lowest salary and least spending

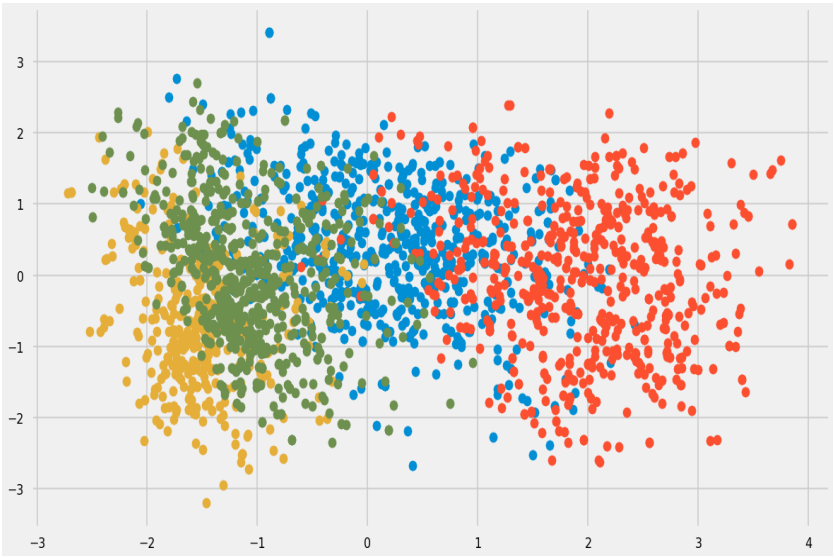
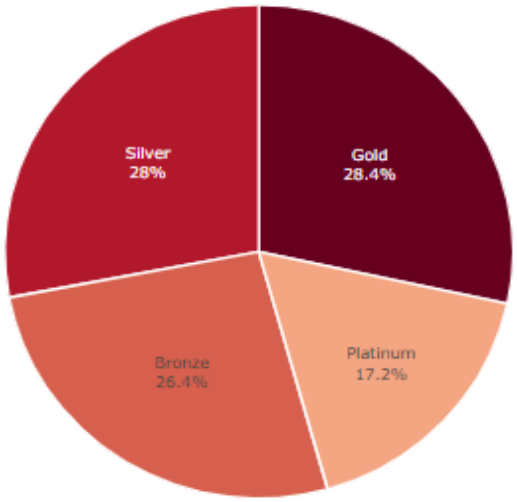


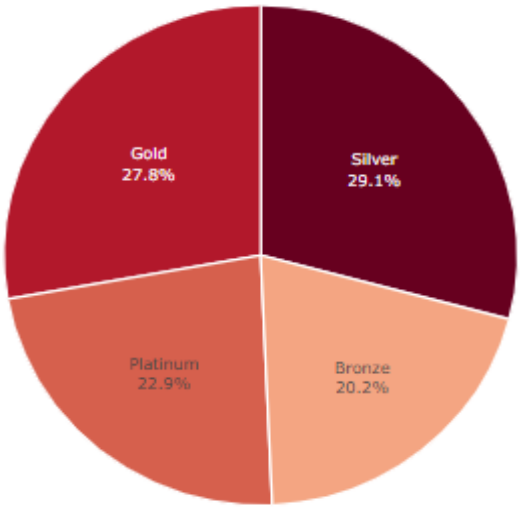
Figure 7: Scatter plot for mini-batch k-means

Comparing mini-batch k-means and k-means

Figure 8 shows the result of k-mean and mini-batch k-means overall clustering. Both the methods have just a slight difference in terms of clusters.



5.4 Distribution of the data using mini-batch k-means clustering



5.5 Distribution of data using k-means clustering

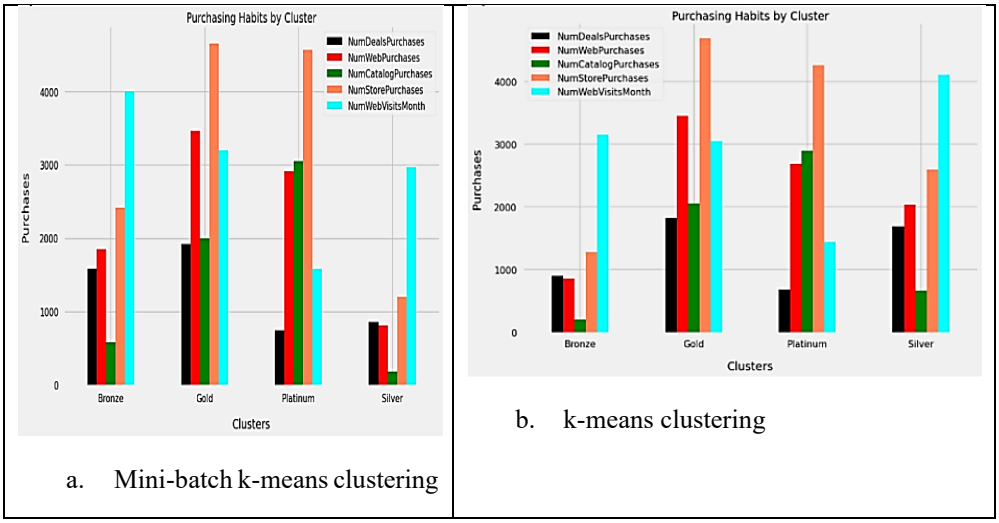
Figure 8: Results using k-means and mini-batch k-means clustering

We will now discuss the results obtained by applying the mini-batch clustering and compare it with the k-means clustering algorithm. Figure 9 shows the results obtained via applying the algorithms for finding the purchasing habits of the customer. It can be seen that the results of both the algorithms are almost same.

Consumer habits concentrate on the analysis of consumer behaviors and the methods individuals employ to select and reject goods and commodities. This element also considers psychological, intellectual, and emotional reactions.

Figure 10 shows the results obtained when clustering the customers based on promotions acceptance. The results are similar. Thus, it has been substantiated that mini-batch k-means gives similar results to k-means but with a faster speed. The mini-batch k-means have been tested for the segmentation analysis.

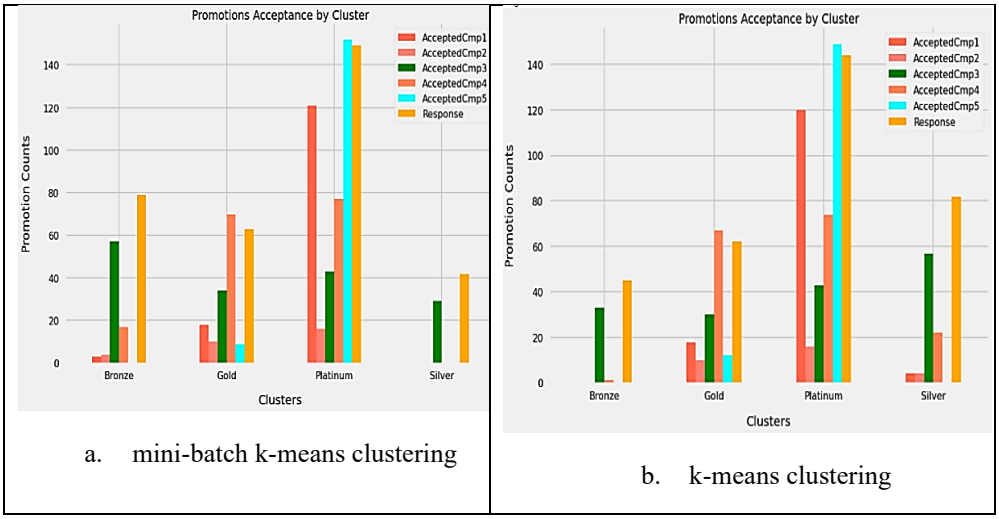
Advertising promotions have an impact on specific customer behavior. When creating effective strategies for marketing promotions and other related components of the communications mix, it's essential to comprehend how consumers react to promotions.



a. Mini-batch k-means clustering

b. k-means clustering

Figure 9: Purchasing habits results using both the clustering algorithm



a. mini-batch k-means clustering

b. k-means clustering

Figure 10: Promotions acceptance results using both the clustering algorithms

Summarizing, both the methods provide comparable groupings that suggest mini-batch as a reliable alternative to k-means. The results provide insights for targeted marketing as the clustering suggests likely customer segmentation to respond to different types of product promotion. The formed clusters are interpretable and align with demographics and spending trends.

However, the results are based on visual interpretation. In future, more robust solutions can be applied such as statistical evidence and cluster validity indices to quantitively verify the similarity of both clustering methods. The absence of these statistical tests such as adjusted

rand index (ARI) and normalized mutual information (NMI) may have limitations for the findings of this study.

6 Conclusion

In this paper, customer purchasing behaviors were analyzed based on the transactions done online and in store of an organization by using k-means and mini-batch clustering algorithm. Following interesting insights were observed from our analysis:

- Customers were classified into four groups: Platinum, Gold, Silver and Bronze, based on their income and total spending.
- The study concludes that the majorities of the customer prefers to shop from the actual store and are hesitant towards the online shopping.
- Most of the purchasing performed on website was from Gold and Bronze customers.
- Majority of the deals were taken by the gold and silver users. However, silver made the most website visits and bronze did the least number of visits to the official website.
- Bronze accepted the most promotions, while platinum showed the least interest in the promotion campaigns.
- Campaign 3 and final campaign seems to be the most successful ones out of all the campaigns.
- In the results, fruits and other sweets were the favorite in every category while wine and meat products were the top among bronze and gold category.

The paper applied k-means and mini-batch k-means algorithm over the dataset and the results are reported. Both of the clustering algorithms gave similar results. Hence for scenarios where fast processing is required, mini-batch k-means clustering can be used. In the future, there are two major ways to further the current study. One is to understand better embedding algorithms into the customer relationship management (CRM) system to support decision making of an organization to help better performance and higher revenue; where else the second approach is to search more suitable algorithms to match new datasets for theoretical analysis.

For more evaluation, a base line clustering or supervised clustering can be used. DBSCAN or hierarchical clustering can be used to help with benchmarking results. In addition, if labelled data can become available, techniques such as logistic regression, decision trees or neural networks can be used to provide an upper bound.

Both k-means and mini-batch k-means offer simplicity and scalability but with some limitations. They are not good for datasets with irregular cluster shapes, varying densities and significant noise. To address these limitations, advanced techniques such as DBSCAN and GMMs can be used.

Summarizing the discussion, embedding the proposed clustering algorithms in to CRM systems can enhance decision making by focusing on specific customer segmentation for customer retention, upselling and personalized engagement. We can provide customized offers for platinum and bronze customers. This can lead to improved customer satisfaction and maximized revenue potential. Future research can focus on integrating the proposed

clustering algorithms into CRM platforms. The organizations would be able to adopt marketing campaigns based on dynamics of customer behavior.

In addition, future studies can be done to incorporate quantitative comparisons to have better evaluation of clustering algorithms. For instance, the current analysis primarily rely on internal validation and external validation metrics such as ARI or NMI can be applied to have a better understanding. However, it is to be noted that this study primarily emphasize on feasibility and efficiency of mini-batch comparatively. The purpose is not optimization, but to analyze if mini-batch can provide comparable clustering capability. Therefore, similarity in customer segmentation was analyzed as these parameters are most important for real-time and large scale marketing applications. The incorporation of other metrics was beyond the scope of the study.

References

[1] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," *Appl. Soft Comput.*, vol. 73, pp. 816–828, 2018.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques* [Online]. Waltham, MA, USA: Elsevier, 2011.

[3] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[4] L. M. Q. Abualigah, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*. Cham, Switzerland: Springer, 2019.

[5] Y. Liu, X. Wu, and Y. Shen, "Automatic clustering using genetic algorithms," *Appl. Math. Comput.*, vol. 218, no. 4, pp. 1267–1279, 2011.

[6] M. K. Yadav and M. J. Baria, "Mini-batch K-means clustering using Map-Reduce in Hadoop," *Int. J. Comput. Sci. Inf. Technol. Res.*, vol. 2, no. 2, pp. 336–342, 2014.

[7] T. Kanungo *et al.*, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[8] P. Kolarovszki, J. Tengler, and M. Majerčáková, "The new model of customer segmentation in postal enterprises," *Procedia-Soc. Behav. Sci.*, vol. 230, pp. 121–127, 2016.

[9] M. Khajvand *et al.*, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Comput. Sci.*, vol. 3, pp. 57–63, 2011.

[10] M. Song *et al.*, "Statistic-based CRM approach via time series segmenting RFM on large scale data," in *Proc. 9th Int. Conf. Utility Cloud Comput.*, 2016.

[11] Z. You *et al.*, "A decision-making framework for precision marketing," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3357–3367, 2015.

[12] D. H. Nguyen, S. de Leeuw, and W. E. Dullaert, "Consumer behaviour and order fulfilment in online retailing: A systematic review," *Int. J. Manag. Rev.*, vol. 20, no. 2, pp. 255–276, 2018.

[13] A. Singh, G. Rumantir, and A. South, "Market segmentation of EFTPOS retailers," in *Proc. AusDM*, 2014.

[14] A. Griva *et al.*, "Retail business analytics: Customer visit segmentation using market basket data," *Expert Syst. Appl.*, vol. 100, pp. 1–16, 2018.

[15] A. J. Christy *et al.*, "RFM ranking–An effective approach to customer segmentation," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 10, pp. 1251–1257, 2021.

[16] Y.-H. Hu and T.-W. Yeh, "Discovering valuable frequent patterns based on RFM analysis without customer identification information," *Knowl.-Based Syst.*, vol. 61, pp. 76–88, 2014.

[17] P. Q. Brito *et al.*, "Customer segmentation in a large database of an online customized fashion business," *Robotics Comput.-Integr. Manuf.*, vol. 36, pp. 93–100, 2015.

[18] P. W. Murray, B. Agard, and M. A. Barajas, "Market segmentation through data mining: A method to extract behaviors from a noisy data set," *Comput. Ind. Eng.*, vol. 109, pp. 233–252, 2017.

[19] R. W. S. Brahmana, F. A. Mohammed, and K. Chairuang, "Customer segmentation based on RFM model using K-means, K-medoids, and DBSCAN methods," *Lontar Komput.: J. Ilm. Teknol. Inf.*, vol. 11, no. 1, pp. 32–43, 2020.

[20] J. Silva *et al.*, "Association rules extraction for customer segmentation in the SMEs sector using the apriori algorithm," *Procedia Comput. Sci.*, vol. 151, pp. 1207–1212, 2019.

[21] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Mark. Customer Strategy Manage.*, vol. 19, no. 3, pp. 197–208, 2012.

[22] B. Xiao *et al.*, "SMK-means: An improved mini batch k-means algorithm based on MapReduce with big data," *Comput., Mater. Contin.*, vol. 56, no. 3, 2018.

[23] P. Zerbino *et al.*, "Big Data-enabled customer relationship management: A holistic approach," *Inf. Process. Manage.*, vol. 54, no. 5, pp. 818–846, 2018.

[24] T. F. Bahari and M. S. Elayidom, "An efficient CRM-data mining framework for the prediction of customer behaviour," *Procedia Comput. Sci.*, vol. 46, pp. 725–731, 2015.

[25] J. Sheng, J. Amankwah-Amoah, and X. Wang, "A multidisciplinary perspective of big data in management research," *Int. J. Prod. Econ.*, vol. 191, pp. 97–112, 2017.

[26] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[27] R. Dubes and A. K. Jain, "Clustering techniques: The user's dilemma," *Pattern Recognit.*, vol. 8, no. 4, pp. 247–260, 1976.

[28] H. Steinhaus, “Sur la division des corps matériels en parties,” *Bull. Acad. Polon. Sci.*, vol. 1, no. 804, p. 801, 1956.

[29] G. H. Ball and D. J. Hall, *ISODATA: A Novel Method of Data Analysis and Pattern Classification*. Menlo Park, CA, USA: Stanford Res. Inst., 1965.

[30] J. MacQueen, “Classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, 1967.

[31] S. Tufféry, *Data Mining and Statistics for Decision Making*. Hoboken, NJ, USA: Wiley, 2011.

[32] T. S. Madhulatha, “An overview on clustering methods,” *arXiv preprint* arXiv:1205.1117, 2012.

[33] M. G. Omran, A. P. Engelbrecht, and A. Salman, “An overview of clustering methods,” *Intell. Data Anal.*, vol. 11, no. 6, pp. 583–605, 2007.

[34] G. Yu *et al.*, “clusterProfiler: An R package for comparing biological themes among gene clusters,” *OMICS: J. Integr. Biol.*, vol. 16, no. 5, pp. 284–287, 2012.

[35] Á. Gómez-Losada *et al.*, “Finite mixture models to characterize and refine air quality monitoring networks,” *Sci. Total Environ.*, vol. 485, pp. 292–299, 2014.

[36] A. Lozano-García, “Finite mixture models to characterize and refine air quality monitoring networks,” 2015.

[37] H.-J. Chu *et al.*, “Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9451–9457, 2012.

[38] H. Cho and M. K. An, “Co-clustering algorithm: Batch, mini-batch, and online,” *Int. J. Inf. Electron. Eng.*, vol. 4, no. 5, pp. 340–344, 2014.