Customer Segmentation Using Unsupervised Machine Learning Techniques

1st Tanishq Verma, MSc Computer Science,
Queen Mary University of London,
vermatanishq573@gmail.com

***Abstract*—Customer Segmentation is the process through which large organization and business can segment customer into different groups based on their different factors. These factors may include age, demographic, income, relationship status and others. Customer Segmentation gives an insight and better understanding of their customers. Using Customer Segmentation, the organization and business get a better understanding of the needs of the customer and can create different business strategies for their customer. This study focuses on applying Customer Segmentation using the widespread and most popular strategy called RFM (Recency, Frequency, Monetary) by choosing an effective and reliable open-source dataset. This study first uses a dataset and explore and visualize the dataset. Further, it applies data-cleaning and data-preprocessing techniques. After getting the final pre-processes data, it creates a RFM data frame. This data frame will be used to cluster data using unsupervised machine learning techniques. In this study, we have applied some unsupervised machine learning techniques which includes Hierarchical Clustering, K-means, DB-SCAN, Gaussian Mixture Model and BIRCH Clustering including the visualizing the data into different clusters to further understand the data and how each technique is different from other. This study is concluded by comparing the cluster similarities for each of the techniques.**

Keywords - Customer Segmentation, Data Cleaning, Data pre-processing, Clustering Techniques, Agglomerate Clustering, K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Model, BIRCH (Balance Iterative Reducing and Clustering using Hierarchy)

I. INTRODUCTION

Customers behavior and shopping ways have changed over the past decade. Online shopping and E-commerce have seen a significant growth over the year compared to traditional shopping ways. Comparing the E-commerce sales from 2010 to 2020, the sales increased from 572 billion to 4.2 trillion according to red stag fulfillment [1]. This indicates that customers have changed the way to shop products and customers behavior have changed over the years. One of the key benefits of E-commers or online shopping is that all the sales purchases receipt is stored safely. Business is using these data to explore and understand the customers behavior to increase their sales. These detailed data contain the item, price, date, and geographical location. Based on these data, the business are able to explore more and get a better insight of their customers. The business can further group these customers into similar groups to better target and satisfy their needs. There are various sort of techniques and bases that can be used segment customers into similar segments.

The primary focus of this study is to provide a meaningful clustering of customers for business to be able to better target their customers. This study also targets the different types of clustering techniques that can be used, giving an in-dept analysis of which is the efficient and desirable. For this study, we have used one of the most popular and widely used the RFM (Recency, Frequency and Monetary) model. The RFM model is basically the basis on which the customers will be clustered into similar clusters. The dataset which we will be using is an Online Retail Dataset. Further, for applying the different clustering techniques efficiently, we are scaling the data using standard scaler. For modelling, we are using Hierarchical Clustering, K-means, DBSCAN, Gaussian Mixture Model and BIRCH Clustering. The visual representation for each of the techniques has been explored. Finally, we have compared the similarity of the clusters generated by each of these techniques. In conclusion, some recommendations are provided for businesses.

A. Customer Segmentation

This section contains the key and relative ideas related to customer segmentation, CRM and customer usability.

With the growing market and rapid changes into the market, the competition between different leading organization is increasing. To keep up with other, the originations are increasing their marketing budget and investing heavily into modern techniques. One of the vital investments is investing into Information Technology (IT) which can further develop their business strategies. Customer Segmentation is typically splitting the customers into similar categories based on their behavior. Segmenting customers into groups can lead to creating better strategies for different clusters of customers to better target and increase sales of the organization. One of the key benefits of customer segmentation is that clustering can be done according to specific factor such as age, demographic, spending behaviors, and others. Marketing for various organization can be enhance using customer segmentation. Unique marketing strategies and techniques can be made for unique cluster. Instead of researching each customer individually, which will be time consuming and burdensome, we can aggregate similar customer into similar groups and then create marketing strategies to benefit the company.

B. Customer Segmentation Methodology

With the growing market and rapid changes into the market, the competition between different leading organization is increasing. To keep up with other contenders in the market, many organizations are increasing their marketing budget. Customer Segmentation is used by Industrial business to target specific customer groups by grouping them into similar groups. Client Segmentation or Customer Segmentation is a method through which clients or customers are groups into similar clusters on basis of different factors so that specific marketing strategies can be made for each of the groups.

C. Use of Customer Segmentation

Customer Segmentation are applied to various markets. One such place is when the organization are working towards “target marketing” according to research study [2]. Target Marketing is when an organization is grouping customers based on some properties that company want to sever. For any organization to achieve target marketing, there are three fundamental steps. First, customers are first divided into groups based on their preferences. Second, the features and behaviors of the group are explored so that numerous marketing strategies can be applied. Finally, comparison of brands and customer behaviors can be made explored.

D. Why Unsupervised Machine Learning

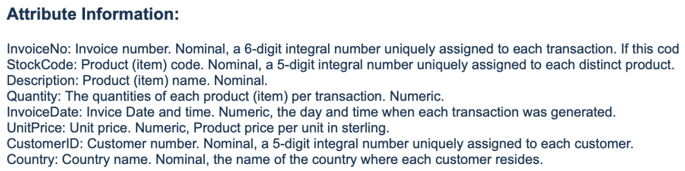
As the name suggests, Supervised Machine Learning requires to train the data using label. Supervised Machine Learning uses ‘Labelled’ data. This means that some of the data samples are already tagged with correct answers. The machine or model uses this ‘labelled’ to learn about the data and further can apply predictions to the unseen data. Using the learning of labelled data, the machine then uses its knowledge to predict on the unseen data.

Unsupervised Machine Learning does not require model supervision. The model uses on its own to understand the data and work on it with the information provided. The data set is an unlabeled data. Unsupervised machine learning can perform more complex processing.

For this study, we have used Unsupervised Machine Learning Techniques instead of Supervised Machine Learning. The main reason for using Unsupervised machine learning is that Supervised machine learning requires labelled data which is difficult to find. Further, our dataset was big, so even for training sample to apply labeling, we needed labelling a lot of data samples. In addition, wrong labelling or incorrect labelling can make the model predict incorrectly. However, this may not be the case for Unsupervised Learning. Apart from this, Unsupervised can discover unknown patterns and help to discover features which may be useful.

E. Dataset

The dataset creating model is an Online Retail Dataset. The dataset contains transaction occurring from 1/12/10 to 09/12/11 for UK based non-store online retail. The dataset is available at UCI Machine Learning Repository [6]. This dataset contains approximately 541909 data samples. Each sample contains 8 attributes which includes InvoiceNo, StockCode, Description, Quantity, Invoice date, UnitPrice, CustomerID and Country. Each attribute detailed information is given below in the picture.



Attribute Information:

InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this cod
StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice: Unit price. Numeric, Product price per unit in sterling.
CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

Fig. 1. Online Retail Dataset Attributes

II. LITERATURE REVIEW

Over a few decades, the market is expanding and changing to meet the needs of the customers. Business also needs to keep on expanding and changing its strategies to meet the needs of the customer and to increase its sales and profit. The method of identifying the needs of the customer satisfying their needs and create new business strategies to attract similar customer can be done using Customer Segmentation. Achieving customer segmentation is a tedious task since it involves customers to segment according to different factors which include age, customer behavior, demographic, monetary, recency etc. One fit for all customers can be a bad practice. According to [3], which is segmenting customers into different categories which include High Buyer Regular Customer, High Buyer Irregular Customers, Low Buyer Regular Customer and Low Buyer Irregular Customer. This study uses a MATLAB program of K-means used to train on z-score. The data set contains 100 sample of two features which includes the average amount of money spent by the customer and the average number of times customer visits the store. Further, using K-means, it has successfully achieved an accuracy of 95 percent.

According to other research [4], this study applied supervised machine learning technique to companies segregated data. As it has applied supervised machine learning, they need to label the data. The study labeled the data into ‘standard’ and ‘premier’. For the dataset, a real customer data from Google AdWords Premier Agency was taken which included two features the number of payments and the total number of payments by each customer. Further, linear regression and logistic regression is applied to it. The conclusion of the study is that logistic regression had much better method efficiency of 89.43 percent compared to linear regression which has a method efficiency of about 37.18 percent. This study clearly concludes that logistic regression is a good technique for segmentation problem using supervised machine learning.

According to another study [5] which includes customer segmentation. This study is segmenting customer by determining the travelers/customers choice towards green hotel to normal hotels. This study works with existing online travel reviews by TripAdvisor. [5] applies k-means and Techniques for Order of Preference by Similarity to Ideal Solution (TOPSIS) techniques to segment customer and traveler based on the reviews provided. Further, it states that sleep quality was number one priority when customers are looking for a hotel.

Understanding and using this study, the green hotels can understand the behavior of the customer and create strategies or marketing techniques to attract customers to their hotels. This study first pre-processes the data and applies k-means clustering to it. It further uses TOPSIS to analyze which feature is important and gives ranking to each one of the segments whenever customer is looking for a hotel.

According to different research [7], which takes from same repository UCI machine learning naming Whole Customer Dataset. This study uses unsupervised machine learning and applying centroid based and density-based clustering on the dataset. For centroid based, the study applies K-means and for density based it applies DBSCAN and OPTICS (Ordering Point to Identify the Clustering Structure). Instead of using Elbow method, they have used different values of k using different methods which includes Euclidean distance and Manhattan Distance. They have compared the sum of squared distance using different values of k and computed the time taken by each. Using DBSCAN they have used different parameter and identified which number of clusters generated and number of data point un-clustered. The conclusion made when compared K-means and DBSCAN states that both can be used for clustering however, DBSCAN provides an extra benefit to give details of the customer with unusual behavior.

According to another study [8], the study uses a dataset to apply customer segmentation using K-means. The study uses Elbow method to determine the value of k and plot a graph for silhouette score to get the optimal cluster value or K. The study visualizes the dataset on different plot to get the better understanding of the dataset. Further, all of the essential data cleaning and pre-processing techniques have been applied.

From [9], it uses customer segmentation using Unsupervised hierarchical and K-means. It uses silhouette to compare the techniques, Dunn Index and Davis-Bouldin Index to compare techniques. Clustering customer into different types of similar kind. All pre-processing n data set done with visualization on Elbow for K-means and dendrogram for Hierarchical Clustering.

Reading [10], shows that k-means is one of the effective ways of customer segmentation. It uses pre-process and using elbow method determines the clusters for the dataset. Similar to [10], [11] also uses k-means and elbow method to cluster customers. They have created a group of based on their finances and spending habits. However, for this study, the data sample is less with only 201 data samples. [12] follows k-means with elbow method as Hyper Parameter Turning to find optimal clusters. [13] does similar to the above using k-means and visualizing the clusters generated by it.

While studying [14] which uses supervised machine learning, it uses linear and logistic regression, random forest method to apply customer segmentation. It must label the data based on RFM to apply techniques. Further, it uses accuracy and precision to determine which techniques was the most effective.

According to another study [15], which uses Gaussian Mixture, Agglomerative Clustering, and k-means to cluster data. Further, it has also used SVC support vector classification which is part of supervised machine learning. Using the precision, recall and accuracy, it can determine the best technique. Analysis is done based on RFM. Finding of [16], have similar methodology of clustering using k mean. Hyper parameter for k is determined using elbow method and silhouette analysis. The study also applies winsorization which is a process of removing the outliners. [17] and [18] has also implemented k-means to achieve customer segmentation. Parameter taken are spending vs annual income for [17] while [18] uses RFM.

III. Methodology

This section includes the methodology, the tools required along with the data set used. In addition, it also comprises of the explaining of the unsupervised techniques applied in the study.

A. Tools Required

For implementing the machine learning model, we had to use some tools. To begin with, we have used Jupyter Notebook on the local machine to run the python code. Further, we had to use different types of python libraries to implement unsupervised machine learning on the selected dataset.

- Pandas - Pandas is a python library used for data manipulation and visualization. Data set created by Pandas is a multi-dimension table or data frame. Further, it has read csv and json functions which is used to read csv or json to load and create a pandas data frame. Pandas is an excellent library for making changes or updates according to features in the data set.

- NumPy – NumPy or Numerical Python is a python library which is used for working with arrays. NumPy main advantage is that it is written in C/C++ which makes it a fast library to work with large arrays particularly large dataset. NumPy have lots of function for arrays which include indexing, slicing, reshape, and so on. For our study, we have used NumPy to create an array to visualize it on the plot taking it from pandas.

- Matplotlib – Matplotlib is a low-level python library which is used for visualization by plotting different types of graphs. It is open source and free. It is written in C/C++ and JavaScript. Most of the work is done by a sub-module of the Matplotlib called pyplot. Pyplot is used for plotting Lines, Labels, scatter plot, Bars, Histograms, and others. For our study, we have used pyplot to plot scatter plot, plot figures and others.

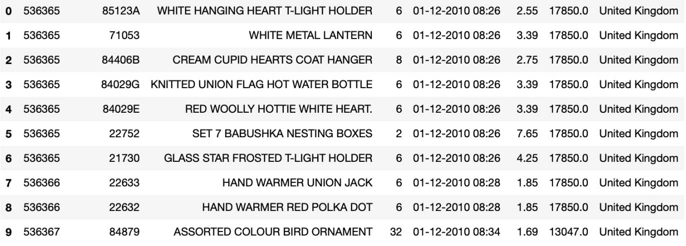
- Seaborn – Seaborn is a Python library used for plotting statistical graphs. It is used above matplotlib and gives a better plot with different colors for visualization. For our study, we have used relplot and boxplot to visualize the data set based on different parameters.

• Sklearn – Sklearn is an open-source machine learning library for Python. It supports NumPy and SciPy. It supports different sort of algorithm which includes Support Vector Machine (SVM), K-means, Random Forest, DBSCAN. For our study, we have used Sklearn almost everywhere. We have also used sub-module of Sklearn which includes preprocessing, metrics, and clusters. Pre-processing is used of standard scaler. Metric is used for calculating Silhouette Score, Adjusted Rand Score while clusters include K-means, DBSCAN, Agglomerative and Birch. For four out of five unsupervised machine learning techniques applied in this study we have used Sklearn. For the gaussian mixture model, we have used another sub-module of Sklearn called mixture which has Gaussian Mixture used for implementing Gaussian Mixture Model.

• SciPy – SciPy or Scientific Python is an open-source library used for computing mathematical and scientific calculations. It is an extended version of NumPy which comprises to solve even more complex calculation when compared to NumPy. SciPy is used for manipulating data and visualizing data. For our study, we have used sub-module of SciPy for 'cluster.hierarchy' to import Dendrogram for Agglomerative Hierarchical Clustering visualization. We have also used another sub-module of SciPy called 'spatial.distance' to import cdist which is used for distort elbow method to find the value of K to implement K-means.

1) Data Preprocessing: Data Preprocessing is the most important part while creating a model. Almost all the data needs to be pre-processed before applying any of the techniques. The open-source data usually have a lot of redundant value, null values, duplicate samples, and vague values for the data samples. These needs to process before applying any technique. In this model, we have applied various pre-processing techniques. These include checking the null values and removing it if any, checking for duplicate values and further checking the unique values. After applying all the essential pre-processing techniques, we further visualized the dataset. Although the dataset is UK based, still there are data sample from across the world but most of the data samples are UK based. We further decided to use the data sample of UK based only.

After Complete Pre-processing, the data frame generated is given below.



	Invoice	ProductCode	Desc	Qnt	InvoiceDate	Price	CustID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047.0	United Kingdom

Fig. 2. Dataset Data Frame

2) Unsupervised Machine Learning Techniques: As part of Feature Selection, we selected all of the features values except for product description. All of the features were important however we used the existing features to create new features as well. To exemplify, we used the invoice data to calculate the recency for the model attribute.

3) RFM (RECENCY FREQUENCY MONETARY) : There are various types of methods on which we can analyze the customer and apply clustering based on these methods. One such famous and most widely used method is RFM model. RFM stands for Recency, Frequency and Monetary is a model which is used for market research to study the behaviors of the customers. We can apply clustering based on the RFM model. As the name suggests, it is used to analyze each customer based on three essential factors which includes Recency (How recent a customer bought a product), Frequency (How frequently a person is buying a product) and Monetary (How much money a customer is spending). Using these three factors, we can categories the customer into different clusters. If all these factors are high, then customer is a loyal customer or a big spender and if they are low, then customer is not loyal. According to these clustered customers groups, business can further create business strategies accordingly (one fit for all) to better target that section of the customers and increase business sales.

To apply RFM to the model, we first need to compute all the factors i.e. Recency, Frequency and Monetary. For computing the Recency, we first founded the most recent date for the transaction. Further, subtracted the most recent date from the transaction date for each of the customers. For Frequency, we counted the number of invoice number for each unique customer. Finally, for Monetary, we founded how much money the customer spends during the period. After computing all the values, we added it to the data frame.

After creating the RFM data frame which will be used for applying different clustering techniques, we first used the seaborn python library to visualize data on RFM. For Recency vs Frequency, we discovered that recent customer is more frequent as well. For Frequency vs Monetary, it depicts that frequent buyer spend less compared to less frequent buyers which spends usually more. For Recency vs Monetary, the recent customers spend much more compared to less recent customers.

The figures below describes customer based on RFM with Recency vs Frequency, Frequency vs Monetary and Recency vs Monetary.

Finally, after complete visualization of the RFM on the created data frame, we need to apply standard scaler before applying any of the techniques.

4) Hierarchical Clustering: As the name suggest, Hierarchical Clustering is a clustering technique which creates a hierarchy of clusters. Hierarchical Clustering are of two types Agglomerative and Divisive. Agglomerative Hierarchical Clustering is the process which starts from the bottom and keeps creating clusters until a single cluster is created. In the beginning, each of the data points are considered a separate individual clusters. Divisive is opposite of Agglomerative Clustering a top-down approach. The result of Hierarchical Clustering is represented in dendrogram.

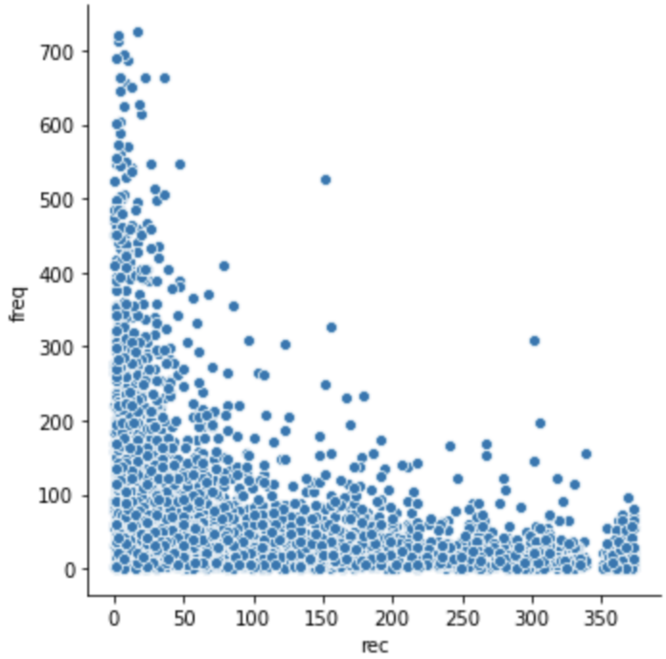


Fig. 3. Recency VS Frequency

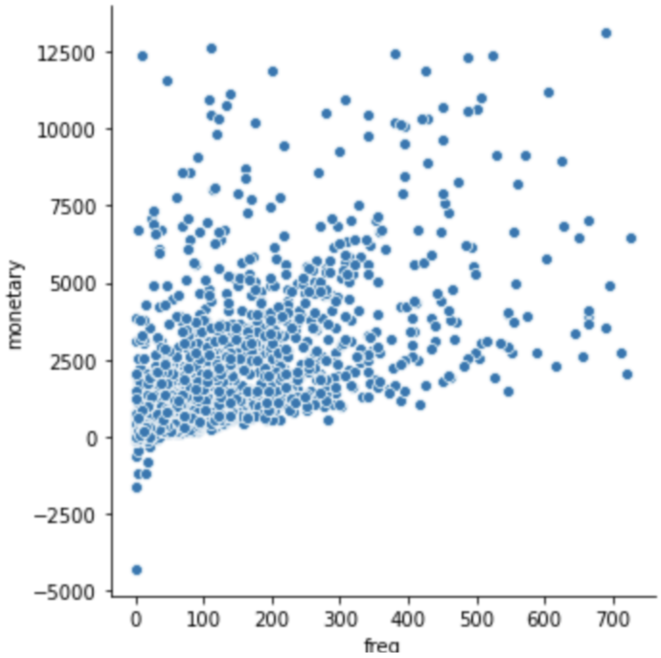


Fig. 4. Frequency VS Monetary

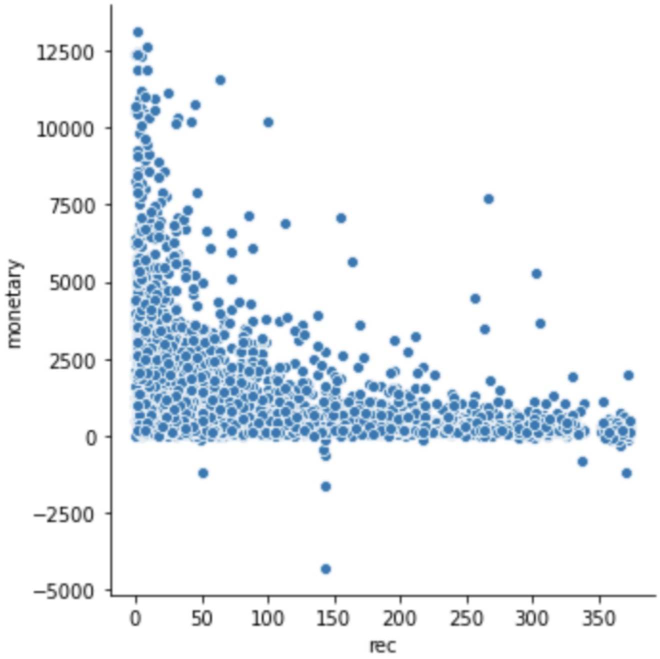


Fig. 5. Recency VS Monetary

Agglomerative Hierarchical Clustering further uses a variety of possible matrices. These include Single Linkage, Complete Linkage, Average Linkage, ward method. Single Linkage or shortest distance works on between the pair of the observation. Complete Linkage work on the farthest pair of data points in the cluster. Average Linkage work on calculating the distance of each pair of the observation in the cluster and dividing it by the number of pair to get the average inter-cluster distance. Finally, ward method uses the sum of squared distance between the clusters data point to apply clustering at each step of the Agglomerative Hierarchical Clustering.

For this study, we have applied Ward method, Complete Linkage and Average Linkage and created dendrogram for each of the following methods. After complete visualizing and understand all the three generated dendrogram by each of the following method, we decided that ward method is the best from all the above.

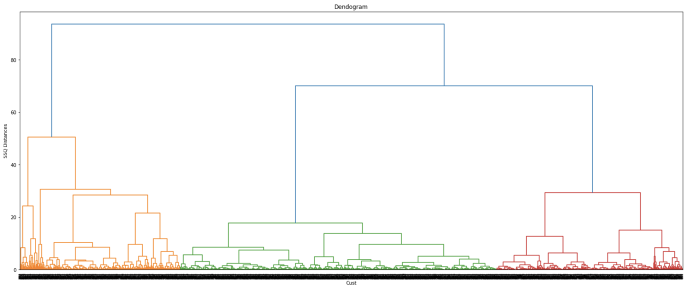


Fig. 6. Agglomerative Hierarchical Clustering using Ward Method Dendrogram

The dendrogram generated by wards method is much better compared to Complete Linkage and Average Linkage. Finally, we decide the threshold distance between the data point to get the number of clusters to be generated. We used the different threshold and created cluster three, four and five using the ward method. After generating for all different number of clusters and visualizing it, we finally decided to go with three number of clusters for the data set and label the data point with it and add it to the data frame so we can compare the cluster generated by Hierarchical Agglomerative to the other cluster as well. We have added a new data frame which includes generating three clusters and labeled each data point to their specific cluster number. Finally, we have created a 3d plot based on RFM to get a visualization of how the clusters are formed.

After generating the clusters using agglomerative Hierarchical Clustering, we further, used the Silhouette Score to generate Average, Minimum, and Maximum Silhouette Score. Silhouette Score gives an estimation or validation of goodness of clustering techniques. Silhouette Score varies from -1 to 1. 1 is the ideal score that clustering is excellent and -1 is an example of bad clustering. Value close to 0 can be due to overlapping of clusters.

For this study, for Hierarchical Clustering, the avg value is 0.43244, the maximum silhouette score is 0.7674 and minimum silhouette score is -0.61550.

5) K-Means Clustering: K-Means clustering is a very popular unsupervised machine learning clustering algorithm. It is one of the most famous clustering algorithms where n observation is clustered in k clusters where each observation belongs to the cluster with nearest mean. In k-means clustering, any arbitrary value for creating clusters is given. Once defined, it chooses random centroids and then calculate the sum of square distance to position the data points in the cluster. Further, it keeps iterating and repeating the steps until one of the following is achieved. First, the centroids are stabilized, or a definite number of iterations have been achieved.

For calculating the value of k or how many clusters should be used to clusters, we use Elbow Method. Elbow method is a heuristic way which helps in determining the optimum number of clusters to be created for a given data set. It is a graph which contains number of clusters on the X-axis and sum of the squared distance on the Y-axis. The graph is a downwards decreasing curve, meaning as the number of clusters increases, the sum of the square distance keeps reducing. As the name suggests, we need to find the elbow point where there is no significant decrease in the sum of square distance after a specific number of clusters. We choose that cluster number to be the value of K. Elbow method is an effective way of determining the optimal number of clusters to be selected for a given data point.

For this study, we first apply Elbow method and plot the graph for it. We can clearly see that the elbow point is near cluster three after which the sum of the squared distance keeps decreasing gradually. Finally, we select the value for K to be 3.

We have used distortion as well which is used to get the error difference, but we get similar result in that as well. We have also generated the average silhouette score value for each of the cluster to get an estimation of the goodness of the clusters generated by K-means. Finally, after getting the value of K, we applied K-means clustering to it. We have also computed the time taken by it and added the predicted values generated by k-means to the data frame assigning cluster to each data point in the data set. This will be used further for visualizing and comparing the similarity of the cluster to other techniques. Finally, updating the cluster id to each data point in the cluster, we visualized it by creating a 3d plot on RFM and showing clusters using different colors. In the end, we have also computed the silhouette score for k=3, where average score is 0.5799, with maximum score of 0.729 and minimum score of -0.694. With the average silhouette score turning out to be 0.5799, we can say that the good clusters are generated.

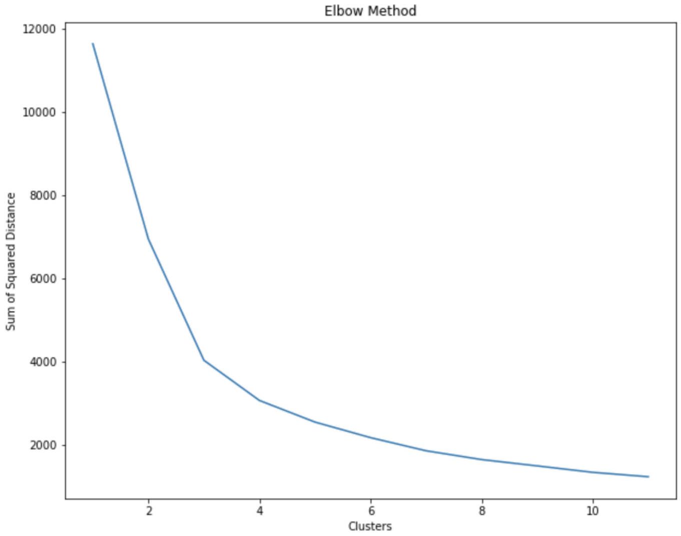


Fig. 7. Elbow Method for finding K for K-Means

6) DBSCAN (Density based spatial clustering of application): DBSCAN or Density based spatial clustering of application with noise is one of the famous data clustering algorithms. It is an unsupervised machine learning algorithm which separates clusters of high density from low density. DBSCAN is best when we need to cluster data based on high density to low density for a large data sample. It can cluster data based on different shapes and sizes. DBSCAN works better when the data is large. DBSCAN most important feature is that it can create a robust outline. DBSCAN does not require you to give the number of clusters to be created unlike K-Means. Instead, it uses some parameters to be defined and using those parameters, it can calculate efficiently the ideal number of clusters that can exist. DBSCAN as the name suggest, works on the principal that clustering can be done based on the density where high density defines a cluster and low-density data points can either be classified into the closest point or can be considered and ignored as an arbitrary data point in the data sample.

For DBSCAN, since we do not have to give cluster numbers (k) to cluster the data, we do however have to define some parameters before applying so that the algorithm can use these parameters to calculate the number of clusters. These parameters include first Epsilon and second midpoints. Epsilon can be defined as distance or the radius around the core data point to consider the data points lying in the radius be declared as a cluster. midpoints or min Sample can be defined as the number of data points that needs to be to consider it a separate cluster. Using DBSCAN, will give an estimation of how correct the other techniques are since the number of clusters that it generates can be used to match it to K-means as well. If similar cluster are generated, meaning that both are correct.

For this study, we have defined the value of epsilon to be 0.5 and min sample to be at least 6. This means it will create the cluster based on the radius of 0.5 and will be considered a cluster only if at least 6 data points lies inside the cluster. We have also calculated the time it took for the machine to apply DBSCAN so we could compare it to other techniques. Further, when we check how many clusters DBSCAN created, we get k=3 which is same as that of the K-means. We also calculated the silhouette score for DBSCAN turning out to be 0.5656 which is a good score stating that the cluster are not overlapping and overall, the clusters generated a good. Finally, we assigned the individual cluster id to each of the individual data sample and added it to the data frame. The figure below gives a visual representation on 2d plot of the clusters generated by DBSCAN based on RFM. Since it is 2d, only two parameter can be added so we visualized it on Monetary and Frequency.

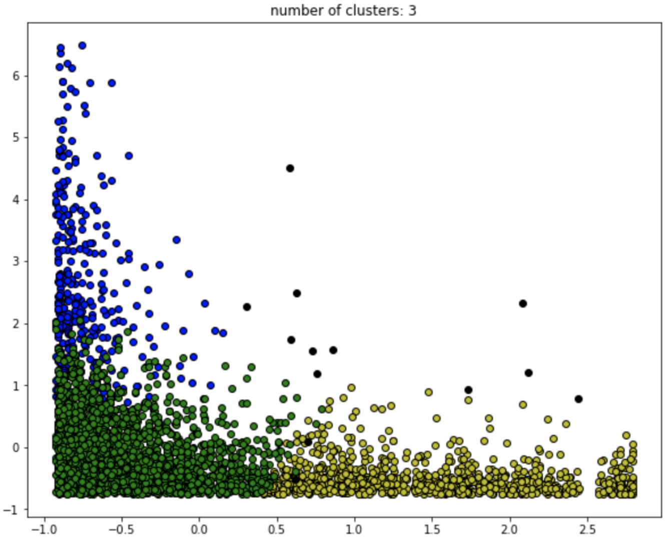


Fig. 8. DBSCAN 2D Plot (Monetary VS Frequency)

Similar for the other techniques, we used it to visualize it on 3d plot based on RFM with different cluster assigning different color to each cluster. Further, we will also use this data frame to check the similarity between other clusters.

7) Gaussian Mixture Model : Gaussian Mixture Model is an unsupervised machine learning clustering algorithm. Gaussian Mixture Model works on the principle of Gaussian Distribution or Normal Distribution. Gaussian Mixture is usually used when the data points are vague or overlapping and sometimes the clusters may also overlap one over the other. It gives a detailed overview of how the cluster are formed and how each cluster belongs to specific cluster. One of the key benefits of using the Gaussian Mixture Model is that it can be used for both soft clustering and hard clustering.

When the data points are unevenly scattered after visualizing the plot and are denser in one region than the other, then it is best to use Gaussian Mixture Model. In addition, it also uses the weight of the data points and create a better cluster accordingly. As stated above, GMM can be used for both soft clustering and hard clustering. Soft clustering works on the principle that if a data point overlaps between the two clusters, then it can be classified into both the cluster depending on which cluster the data point lies more towards or plotting it using the Gaussian Distribution. Hard clustering is when the data point overlap but are either classified into one cluster or the other and does not lie on both the cluster depending on the weightage.

Gaussian Mixture Model first creates a random gaussian. It then adds the closest data point into the gaussian and then re calculate the center of the gaussian. Again, it adds more data point to the gaussian and re calculate the center. It keeps on iterating until the centroid of the gaussian does not change in the next step.

For this study, we have defined the number of gaussian or the cluster to be created three. This is because it will be easy to compare it using the other techniques as well. We fit the model and add the prediction to the data frame. We have visualized the cluster using the Gaussian Mixture Model on 3D plot based on RFM. The average silhouette score generated by Gaussian Mixture Model is 0.315 indicating that it was not able to create better clusters.The visualization is given below in the figure.

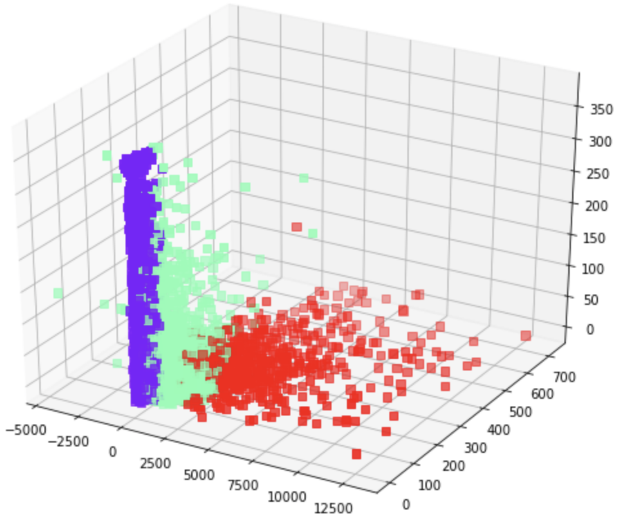


Fig. 9. Gaussian Mixture Model 3D Plot based on RFM (X axis=Monetary,Y axis=Frequency,Z axis=Recency)

We assign the individual cluster id to each of the data point and add it to the data frame. Similar to all of the above techniques, we can compare and visualize later using the data frame we created.

8) BIRCH (Balance Iterative Reducing and Clustering using Hierarchy) : BIRCH stands for Balance Iterative Reducing and Clustering using Hierarchy is an unsupervised machine learning algorithm. It is used to perform hierarchical clustering on large dataset. BIRCH is sometimes also used to accelerate K-means and Gaussian Mixture Model. One of the key benefits of BIRCH is to cluster the incoming data point to get the best clustering possible. BIRCH requires only single run through the data. Being a hierarchical clustering algorithm, it usually performs better on large dataset. Since it does not require to go through the whole data before applying, it can read some data and apply clustering to it that is why it is called incremental learning.

BIRCH is one of the most efficient unsupervised algorithms since it gives a lot of key benefits like incremental learning, does not need to go through the whole data at once, considering the individual weights of the data point and not all the data points are equally important. It can work with vague and scattered data points as well. Being Hierarchical clustering, BIRCH works by creating small clusters and further groups these clusters into bigger clusters. BIRCH requires only one scan of the data making it efficient and time saving to work on large data-sets. Just like hierarchical clustering which uses a dendrogram to visualize the clustering starting from bottom to the top, BIRCH creates something called as clustering feature tree which shows how the clustering is done and each stage of it.

For this study, we implemented BIRCH. While implementing it on RFM, we first need to define some parameters. These include threshold and the number of clusters. As it is easy to understand, still explain that threshold means the minimum number of samples to be able to consider it a cluster in CF tree while no of cluster defines how many clusters you want to create for the data. We used the cluster equal to three and put the default value for the threshold. We applied the technique and added the predictions for each of the cluster to the data frame so we could visualize it and compare it with other clustering algorithm. We created a separate data frame with each data point containing it separate cluster id the cluster it belongs to. The average silhouette score generated by BIRCH was the highest among all of 0.710 creating a good quality clusters.

After applying and adding predication to the data frame, we have further visualized each on Recency vs Frequency, Frequency vs Monetary and Recency vs Monetary with each plot also showing the different clusters using a different color. Like all of the above we have also visualized it on 3d plot based on RFM on each scale.

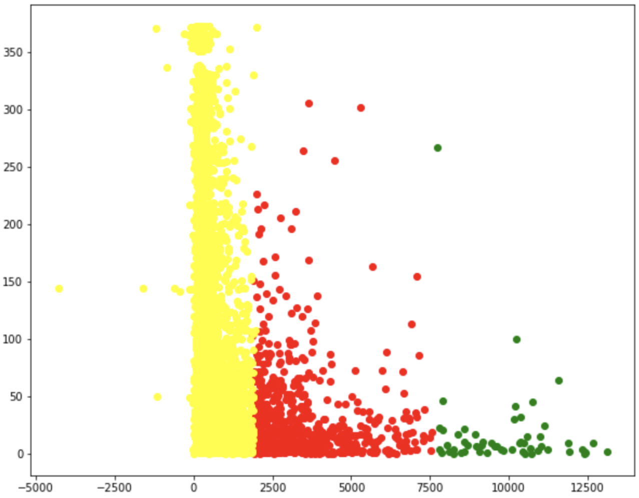


Fig. 10. BIRCH (Monetary VS Recency) respectively on the axis

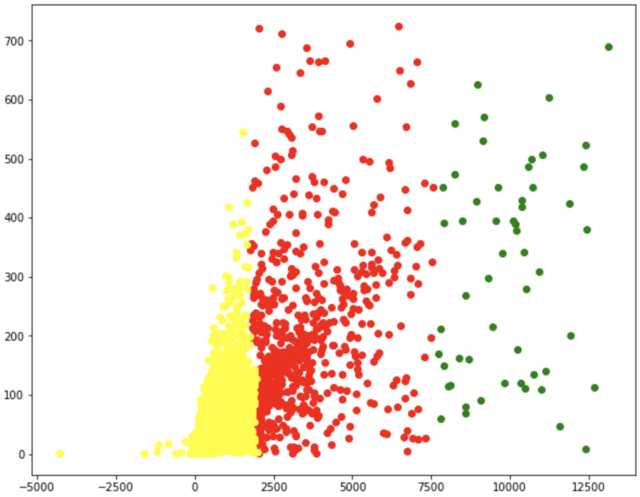


Fig. 11. BIRCH (Monetary VS Frequency) respectively on the axis

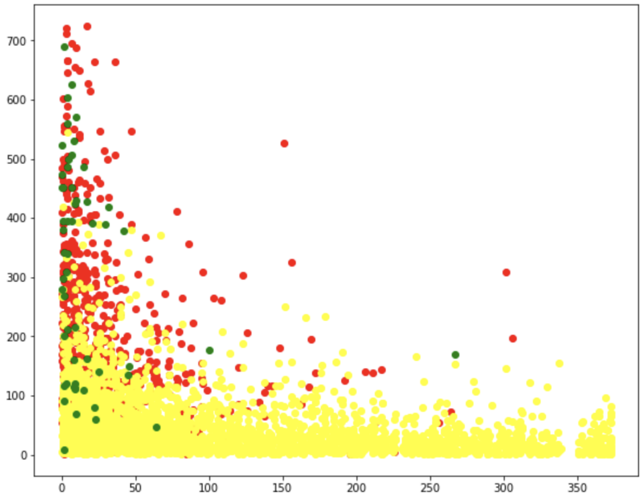


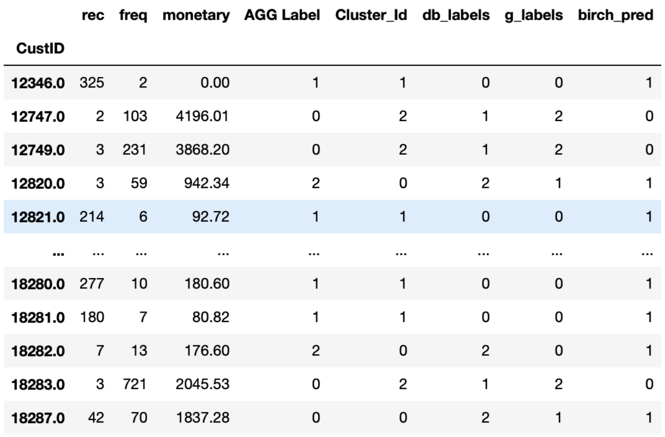
Fig. 12. BIRCH (Recency VS Frequency) respectively on the axis

IV. Result

In this study, we have implemented five different unsupervised machine learning technique on the given data set. The techniques were implements on RFM (Recency, Frequency and Monetary) which is the deciding factor to which cluster each of the individual data points should be put into. Based, on RFM we created the data frame and added each of the RFM factors to it by calculating.

We further applied unsupervised machine learning techniques starting with the Agglomerative Hierarchical Clustering. We applied it using different method including single linkage, complete linkage, and wards method. We calculated the time take and calculated the silhouette score. In conclusion for the Agglomerative Hierarchical Clustering, it takes a significant time to run and create the dendrogram especially on large data set when compared to other techniques. Further, when calculated the silhouette score, it gives an average value of 0.4324 which defines that the clustering is good. We have also visualized the cluster on 3d plot.

Further, the figure given below lables each individual customer based on RFM and each has attribute with regard the cluster id provided by these five techniques.



CustID	rec	freq	monetary	AGG Label	Cluster_Id	db_labels	g_labels	birch_pred
12346.0	325	2	0.00	1	1	0	0	1
12747.0	2	103	4196.01	0	2	1	2	0
12749.0	3	231	3868.20	0	2	1	2	0
12820.0	3	59	942.34	2	0	2	1	1
12821.0	214	6	92.72	1	1	0	0	1
...	...	...	...	...	...	...	...	...
18280.0	277	10	180.60	1	1	0	0	1
18281.0	180	7	80.82	1	1	0	0	1
18282.0	7	13	176.60	2	0	2	0	1
18283.0	3	721	2045.53	0	2	1	2	0
18287.0	42	70	1837.28	0	0	2	1	1

Fig. 13. Dataframe having cluster assigned for each techniques for every customer (AGG Label for Agglomerative Hierarchical Clustering, Cluster id for K-Means, dblabels assigned by DBSCAN, g label by Gaussian Mixture Model and birch pred by BIRCH)

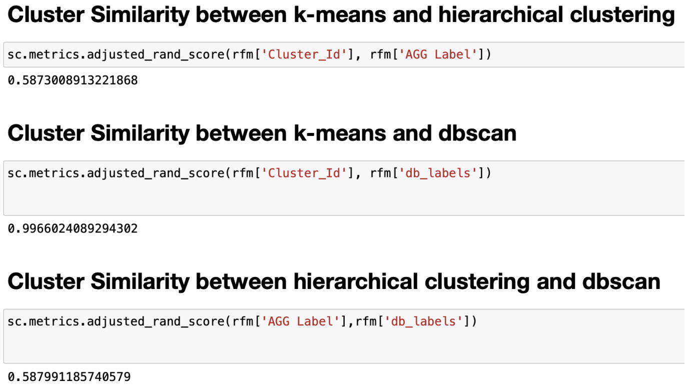
For K-means clustering, it was easy to implement. First calculated the value of k using Elbow method. We got the value of k to be 3. Further, we implemented k means which was fast which took less than a second compared to Hierarchical Clustering. We calculate the average silhouette score which was 0.5799. We have also visualized the cluster 3d plot based on RFM.

Next unsupervised machine learning technique that we implemented was DBSCAN. Implemented it based on RFM, we calculated the time it took for it to implement on the taken dataset which was approximately 1 second. We have given all the required parameter too implement it and it calculates cluster base on it. The number of clusters it calculated was three. For comparing it to the rest of the above technique and see how good the clustering is, we calculated the average silhouette score which was 0.5656. Like the above, we also visualized the clustering on 3d plot.

For the gaussian mixture model, we defined the number of clusters to be three since it will be easy to compare it with the rest of the techniques. Further, implementing gaussian mixture model on the selected data was quick. We calculated the average silhouette score which was 0.33 which is much less compared to other techniques. Finally, we plotted the cluster it created on 3d plot.

Finally, for BIRCH, based on RFM, we passed the required parameters while defining the model which includes the threshold and the number of clusters. Further, implementing it was easy and quick just like the above techniques if Hierarchical clustering is excluded. Like the above, we calculated the average silhouette score which was much higher than the above techniques about 0.710 which explains that it creates very good clusters and showed the visualization based on different RFM factors. Finally, plotted a 3d plot on RFM depicting the three clusters it created.

Instead of explaining each of the techniques individually, we have also compared the cluster each technique created and checked the similarity of it. For this we have used sklearn library which contains a function called adjusted rand score. It is a function which takes two parameters. Each of these parameters will be the clusters generated by the unsupervised machine learning techniques. The score generated will lie between 0 and 1 where 0 meaning that clusters and data point in the cluster for the techniques are very different from each other while 1 will show that each cluster and data points in the cluster are quite like each other.



Cluster Similarity between k-means and hierarchical clustering

```
sc.metrics.adjusted_rand_score(rfm['Cluster_Id'], rfm['AGG Label'])
```

```
0.5873008913221868
```

Cluster Similarity between k-means and dbscan

```
sc.metrics.adjusted_rand_score(rfm['Cluster_Id'], rfm['db_labels'])
```

```
0.9966024089294302
```

Cluster Similarity between hierarchical clustering and dbscan

```
sc.metrics.adjusted_rand_score(rfm['AGG Label'],rfm['db_labels'])
```

```
0.587991185740579
```

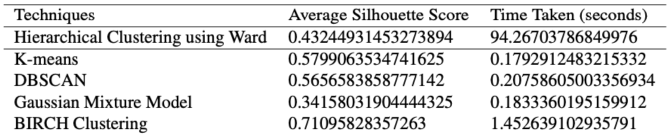
Fig. 14. Clustering Similarity between K-means, Hierarchical and DBSCAN

When we used the adjusted rand score function and compared the similarity of each of the techniques with each other, we got that DBSCAN and k-means clustering generated the most similar cluster and got a score of 0.99 while k-means and hierarchical and hierarchical and DBSCAN also gave a similar result with was almost same about 0.587. When compared the rest of the techniques with each other, we got that all of them yield almost a same result. Comparing Hierarchical with BIRCH got 0.277, BIRCH and Gaussian Mixture Model got score of 0.267, while Hierarchical and Gaussian Mixture gave a score of 0.227 and result of them gave similar score of about 0.212.

From the table below, we can estimate which unsupervised machine learning technique is the most efficient. When we compare time taken by technique, we can see that Hierarchical Clustering takes up a significant amount of time. The reason is because it takes time on large data set to generate Dendrogram. Second, BIRCH takes more time when compared to K-means, DBSCAN and Gaussian Mixture Model. When comparing K-means, DBSCAN and Gaussian with each other, we can see that time taken by each is competitive.

However, BIRCH clustering generated the best Average Silhouette Score for the data set with value equal to 0.71 making it one of the best technique. K-means and DBSCAN also yield good average silhouette with 0.57 and 0.56 respectively. Hierarchical gives bad silhouette score of 0.43 while Gaussian Mixture performs the worst generating the average silhouette score of 0.34.

From the above observation we can say that BIRCH clustering is the best performing in terms of the goodness of the clusters generated and time taken while K-means and



Techniques	Average Silhouette Score	Time Taken (seconds)
Hierarchical Clustering using Ward	0.43244931453273894	94.26703786849976
K-means	0.5799063534741625	0.1792912483215332
DBSCAN	0.5656583858777142	0.20758605003356934
Gaussian Mixture Model	0.34158031904444325	0.1833360195159912
BIRCH Clustering	0.71095828357263	1.452639102935791

Fig. 15. Comparison Table for Techniques

DBSCAN also give good results for the data set. However, Hierarchical and Gaussian Mixture Model should be avoided.

V. Conclusion

In conclusion for this study, we have successfully achieved clustering of a data set using different types of unsupervised machine learning techniques. We have achieved clustering based on the famous RFM model that is used to cluster the customers into different groups. We also discussed the goodness, similarity and time taken for clustering using each of the techniques. Further, we can distinguish which technique to be used based on the type of dataset and how an individual wants to perform clustering. Each of the clustering technique that is used in this study, have been described with in-dept analysis the result it obtained. Each techniques have its own key advantages and disadvantages.

The conclusion that can be made by this study is that BIRCH achieve the highest average silhouette score indicating the best clusters formed. K-means and DBSCAN generated similar silhouette score with good cluster but not better than BIRCH. Apart from this, Hierarchical generated a good silhouette score as well however, it took a lot of time for computation which is something to take into account. Gaussian Mixture Model generated the worst silhouette score. The clustering similarity was highest for k-means and DBSCAN and second most similar k-means and hierarchical and hierarchical and DBSCAN. We can conclude that BIRCH is the best technique for this particular dataset following k-means and DBSCAN.

As explained at the beginning of the study, Customer Segmentation or Market Segmentation is very important for business to grow their sales and achieve target. In addition, it also gives an insight of want the customer wants and how can new business strategies can be created to target specific groups of customers. The machine learning model generated for this study, can be used by a varied variety of business to segment their customers into different groups, analyze and get an in-dept knowledge of their customer.

After reading the study and using the model, business can get an understanding of how machine learning can be used to cluster customers. Further, after applying customer segmentation, business can develop new business strategies to target their customer. Business can group customers and give a loyalty badge of Bronze, Silver and Gold based on different factor (RFM according to this study). Business can further, target specific groups of the customer using a single tailored message and strategies of giving huge discount to customer who buy products at specific period of the year. In contrast, they can give discount to customers who are frequent buyer based on giving a reward coupons or reward system based on how much money they spend and help business achieve increase its sales.

In the market which is rapidly expanding and changing at the same time, Customer Segmentation can help the businesses to keep up with its competitors by constantly applying Customer Segmentation based on new strategies to get insight of the customer and further use to increase its business. Finally, Customer Segmentation can also be used to identify which kind of products are in demand and can create new of such product can be in demand to increase its business.

In general, Customer Segmentation can be applied by a varied variety of Business to increase its business, get knowledge of their customer, help understand which products are in demand, creating new business strategies like advertisement and marketing to better reach that section of the customer and also develop product which will be in demand in the future based on their current sales.

VI. Future Work

For this study as discussed above, we applied customer segmentation using famous RFM model and clustered customers based on different types of unsupervised machine learning techniques. However, customer segmentation is vast and very large domain to explore.

Many different domains of the customer segmentation can be explored if one wants to do research. To illustrate, this study focuses on exploring using the RFM model and applied unsupervised clustering techniques only using selected techniques. Customer Segmentation can further be explored using a different data set and using an RFM analysis as well where RFM create each individual score based on RFM factor for each of the data point. Further, customer segmentation can be explored using demography as well which gives an even better understanding of the customer based on their geography. Customer Segmentation can also be explore further using the customer behaviors as well. In general, there are various ways to apply clustering to customer based on the market.

In addition to using the strategies to apply customer segmentation, there are various techniques which can be used to apply customer segmentation. Using unsupervised machine learning techniques as used in this study, there are still many techniques which can be explored further and so how good is the clustering achieved by each of these techniques. These include fuzzy k-means, mean shift, mini batch k-means and others. Apart from unsupervised machine learning techniques, there are supervised machine learning techniques too which could be used to explore and get an analysis of comparison between supervised and unsupervised machine learning. Supervised machine learning techniques include classification and regression. Since supervised machine learning requires labelled data, the data needs to be labeled before applying any of the techniques.

References

[1] E-commerce growth from 2010 to 2020 e-commerce trend,Jake Rheude,Publisher, Red Stag Fullfillment

[2] Alam, Mohd Faraz, Raushan Singh, and Sandhya Katiyar. ”Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning.” 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE, 2021.

[3] Ezenkwu, Chinedu Pascal, Simeon Ozuomba, and Constance Kalu. ”Application of K-Means algorithm for efficient customer segmentation: a strategy for targeted customer services.” (2015).

[4] Ozan, Şükrü. ”A case study on customer segmentation by using machine learning methods.” 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE, 2018.

[5] Yadegaridehkordi, Elaheh, et al. ”Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques.” Technology in Society 65 (2021): 101528.

[6] Online Retail Dataset — UC Irvine Machine Learning Repository Dr Daqing Chen, Director: Public Analytics group. chend ’@’ lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

[7] A. S. M. S. Hossain, ”Customer segmentation using centroid based and density based clustering algorithms,” 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), 2017, pp. 1-6, doi: 10.1109/EICT.2017.8275249.

[8] E. Y. L. Nandapala and K. P. N. Jayasena, ”The practical approach in Customers segmentation by using the K-Means Algorithm,” 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 344-349, doi: 10.1109/ICIIS51140.2020.9342639.

[9] Abdulhafedh, Azad. ”Incorporating k-means, hierarchical clustering and pca in customer segmentation.” Journal of City and Development 3.1 (2021): 12-30.

[10] V. R. Maddumala, H. Chaikam, J. S. Velanati, R. Ponnaganti and B. Enuguri, ”Customer Segmentation using Machine Learning in Python,” 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022, pp. 1268-1273, doi: 10.1109/ICCES54183.2022.9836018.

[11] Keerthi, K., et al. ”Customer Segmentation Analysis and Visualization.” International Journal of Scientific Research in Computer Science, Engineering and Information Technology(2021): 280-284.

[12] Patankar, Nikhil, Soham Dixit, Akshay Bhamare, Ashutosh Darpel, and Ritik Raina. ”Customer Segmentation Using Machine Learning.” (2021).

[13] S. R. Regmi, J. Meena, U. Kanojia and V. Kant, ”Customer Market Segmentation using Machine Learning Algorithm,” 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1348-1354, doi: 10.1109/ICOEI53556.2022.9777146.

[14] Venkatakrishna, Ms Ramamani, Mr Pradeepta Mishra, and Ms Sneha P. Tiwari. ”Customer Lifetime Value Prediction and Segmentation using Machine Learning.” Int. J. Res. Eng. Sci 9 (2021): 36-48.

[15] A. A. Aktaş, O. Tunalı and A. T. Bayrak, ”Comparative Unsupervised Clustering Approaches for Customer Segmentation,” 2021 2nd International Conference on Computing and Data Science (CDS), 2021, pp. 530-535, doi: 10.1109/CDS52072.2021.00097.

[16] R. H. Khan, D. F. Dofadar and M. G. Rabiul Alam, ”Explainable Customer Segmentation Using K-means Clustering,” 2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2021, pp. 0639-0643, doi: 10.1109 UEMCON53757.2021.9666609.

[17] Dileep, Pyla Srinivas, and M. Seshashayee. ”CUSTOMER SEGMENTATION USING MACHINE LEARNING.”

[18] Dogan, Onur, Ejder Ayçin, and Zeki Atıl Bulut. ”Customer segmentation by using RFM model and clustering methods: a case study in retail industry.” International Journal of Contemporary Economics and Administrative Sciences 8.1 (2018): 1-19.