Customer Segmentation

(IT Workshop Project Work)
(March-May 2021)
Swati Basu(06004092020) , Hritika Verma(01204092020)
Sangeetha Panicker(04804092020)

ABSTRACT

The world we live has a vast amount of data collected daily and analysing them is a important need.In this 21st century a new modern era where there is large competition to be better than everyone, so the business strategies need to be updated according to modern era. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. A RFM (recency, frequency, and monetary) model and K-means clustering algorithm are utilized to conduct customer segmentation and value analysis by using E-commerce dataset.

1 INTRODUCTION

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to, Clustering techniques consider data tuples as objects. They partition the data objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as miscellaneous spending habits.The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to mange the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.The thrust of this paper is to identify customer segments using the data mining approach, using the RFM modelling to find the RFM scores of customers by diving them into certain groups such as Bronze, Silver, Gold and Platinum.

2 METHODOLOGY

2.1 Dataset Description

The data set used to implement clustering and Kmeans algorithm on transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. It contains 541909 tuples, 8 attributes.The attributes in the data set has InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.

2.2 Data Preprocessing

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm. To Perform Data Preprocessing here, We've done Data Cleaning :

- We first checked for any null and missing values.
- We removed null values from the data.
- We dropped duplicate data from the dataset too.
- Then we observed that 80 percent of the data in this dataset is from United Kingdom which can,if used seperately, can give better results of Customer Segmentation. So we filtered data from United Kingdom only and then check for any negative values. If found, we removed those negative values too.
- Then, We converted Data of Date column into standard DATETIME format and added one more column of Total Amount by performing the operation of multiplication between Unit-Price and Quantity for each row.

2.3 Feature Engineering and Data Visualization

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself. Following Features are the outcome of which we performed:

Figure 1: Monthly Revenue

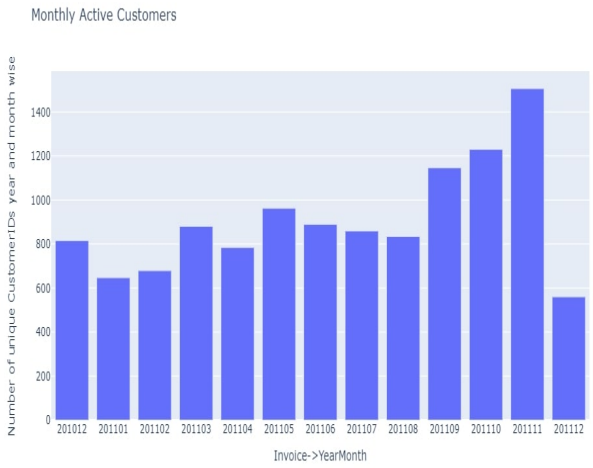
The figure clearly shows that Revenue is gradually increased in august 2011 to november 2011 and suddenly decreased in the month of november-december 2011 and up-down phase was there in monthly revenue since december 2010 to july 2011 and then a sudden increase in revenue is observed in August 2011 then it falls down in november-december at a big rate.

Figure 2: Monthly Revenue Growth Rate

	Invoice->YearMonth	Total Amount	Monthly-Growth
0	201012	498661.850	NaN
1	201101	442190.060	-0.113247
2	201102	355655.630	-0.195695
3	201103	467198.590	0.313626
4	201104	409559.141	-0.123372
5	201105	551568.820	0.346738
6	201106	524915.480	-0.048323
7	201107	485612.251	-0.074875
8	201108	498453.320	0.026443
9	201109	796780.272	0.598505

(1)

Figure 3: Monthly Active Customers



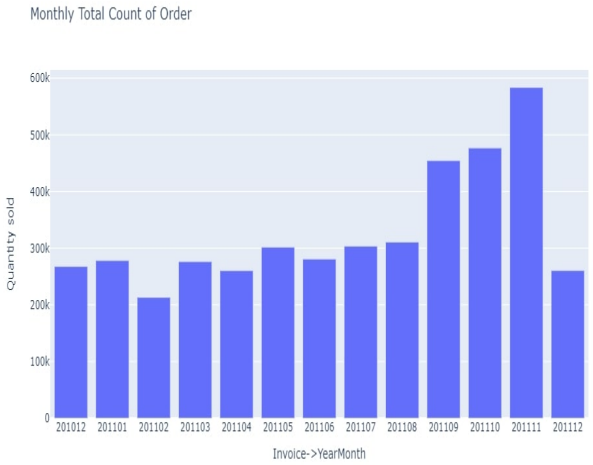
It also clearly shows sudden increase in active customers in the month of decemeber 2010 and then the normal up-down phase continued till november 2011 followed by sudden decrease in the rate of active customers in the month of decemeber 2011.

Figure 4: Monthly Order Count

	Invoice->YearMonth	Quantity
0	201012	267771
1	201101	278300
2	201102	213391
3	201103	276345
4	201104	260450
5	201105	301828
6	201106	280974
7	201107	303602
8	201108	310832
9	201109	454561
10	201110	476991
11	201111	583820
12	201112	260607

It also clearly shows sudden increase in no. of orders in the month of november 2011 and the normal up-down phase continued till november 2011 followed by sudden decrease in the orders count in the month of decemeber 2011.

Figure 5: Monthly Order Count



Monthly Order Average

Figure 6: Monthly Order Average

Monthly order average rised suddenly in december 2011. We observed wavy or up and down phase throughout the year including lowest order average in the month of november just before the highest order average.

New Customer Ratio

In our dataset, we can assume a new customer is whoever did his/her first purchase in the time window we defined. We will do it monthly for this example. We used .min() function to find our first purchase date for each customer and define new customers based on that.

Figure 7: New Customer Ratio

	Invoice->YearMonth	UserType	Total Amount
0	201012	New	498661.850
1	201101	Existing	199589.910
2	201101	New	242600.150
3	201102	Existing	219596.330
4	201102	New	136059.300
5	201103	Existing	296888.220
6	201103	New	170310.370
7	201104	Existing	299309.750
8	201104	New	110249.391
9	201105	Existing	456518.640

Created a column called User Type and assign New as default. Compared the person's invoice date with the minimum purchase date for each row for whichever row, invoice purchase date > min. purchase date, Assigned the person's user type to be existing in that row.

Figure 8: New Customer Ratio table

	Invoice->YearMonth	CustomerID
1	201101	1.238754
2	201102	1.002950
3	201103	0.908894
4	201104	0.546351
5	201105	0.362606
6	201106	0.317037
7	201107	0.244928
8	201108	0.203463
9	201109	0.317241
10	201110	0.357616
11	201111	0.246689
12	201112	0.064639

2.4 Proposed Approach

Our proposed model consists of the flowing six steps:

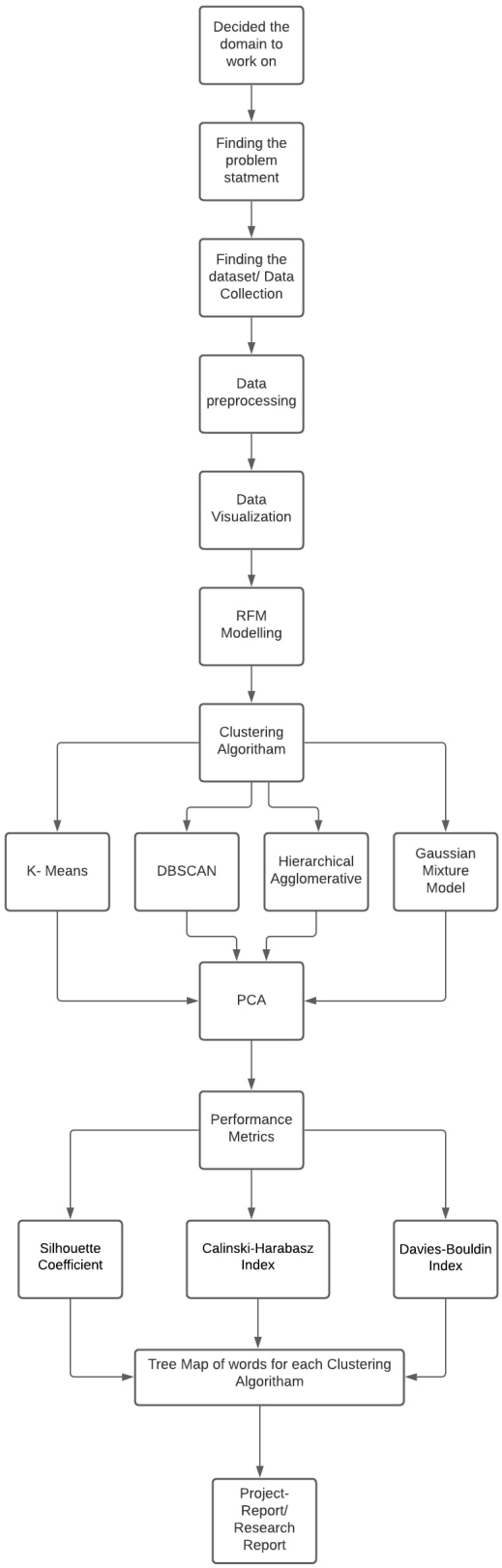
- **Data preparation and cleansing** : In this step,the cleaning and transforming of raw data prior to processing and analysis were done. It involves reformatting data, making corrections to data and the combining of data sets to enrich data.
- **Feature Engineering and Data Visualization** : In this step, the features are extracted from raw data via data mining technique and visualized them using the Bar charts, Scatter plots, and heat-maps.
- **RFM Modelling** : In this step, the RFM scores are calculated, and using this the customers are divided based on loyalty.
- **K-means Clustering and PCA Analysis** : In this step, the k-mean clustering is done based on RFM modeling. After this PCA analysis is used to reduce the dimensionality of data.
- **Different Clustering Algorithms and Accuracy Performance** : In this step, clustering algorithm such as DBSCAN Clustering Algorithm, Hierarchical Clustering, Agglomerative Hierarchical Clustering was done.

3 EXPERIMENT DESIGN

3.1 Workflow Design

Workflow Design is presented in figure 9

Figure 9: Workflow Diagram for whole project



3.2 RFM Modeling

Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:

Let's look more closely at how each RFM factor works, and how companies might strategize on the basis of it.

(1) **Recency:** The more recently a customer has made a purchase with a company, the more likely they will continue to keep the business and brand in mind for subsequent purchases. Compared with customers who have not bought from the business in months or even longer periods, the likelihood of engaging in future transactions with recent customers is arguably higher.
Such information can be used to remind recent customers to revisit the business soon to continue meeting their purchase needs. In an effort not to overlook lapsed customers, marketing efforts could be made to remind them that it has been a while since their last transaction while offering them an incentive to rekindle their patronage.
(2) **Frequency:** The frequency of a customer's transactions may be affected by factors such as the type of product, the price point for the purchase, and the need for replenishment or replacement. If the purchase cycle can be predicted, for example when a customer needs to buy new groceries, marketing efforts could be directed towards reminding them to visit the business when items such as eggs or milk have been depleted.
(3) **Monetary Value:** Monetary value stems from the lucrativeness of expenditures the customer makes with the business during their transactions. A natural inclination is to put more emphasis on encouraging customers who spend the most money to continue to do so. While this can produce a better return on investment in marketing and customer service, it also runs the risk of alienating customers who have been consistent but have not spent as much with each transaction.

These three RFM factors can be used to reasonably predict how likely (or unlikely) it is that a customer will do business again with a firm or, in the case of a charitable organization, make another donation.

3.3 Clustering Algorthims

Cluster analysis, or clustering, is an unsupervised machine learning task.

It involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modeling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space.

3.3.1 K-Means Clustering. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. It is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. The way kmeans algorithm works is as follows:

(1) Specify number of clusters K.
(2) Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
(3) Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

3.3.2 DBSCAN Clustering Algorithm. Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers. The DBSCAN algorithm uses two parameters:

- **minPts :** The minimum number of points (a threshold) clustered together for a region to be considered dense.
- **eps :** A distance measure that will be used to locate the points in the neighborhood of any point.

3.3.3 Hierarchical Clustering. Hierarchical Clustering is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. For e.g: All files and folders on our hard disk are organized in a hierarchy. The algorithm groups similar objects into groups called clusters. The endpoint is a set of clusters or groups, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. This clustering technique is divided into two types:

(1) **Agglomerative Hierarchical Clustering** - Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.
(2) **Divisive Hierarchical Clustering** - Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

3.3.4 Gaussian Mixture Model Selection. A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

The GaussianMixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion to assess the number of clusters in the data.

The GaussianMixture comes with different options to constrain the covariance of the difference classes estimated: spherical, diagonal, tied or full covariance.

3.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning. High dimensionality means that the dataset has a large number of features. The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set. Richard Bellman described this phenomenon in 1961 as the Curse of Dimensionality where "Many algorithms that work fine in low dimensions become intractable when the input is high-dimensional."

- The ability to generalize correctly becomes exponentially harder as the dimensionality of the training dataset grows, as the training set covers a dwindling fraction of the input space. Models also become more efficient as the reduced feature set boosts learning rates and diminishes computation costs by removing redundant features.
- PCA can also be used to filter noisy datasets, such as image compression. The first principal component expresses the most amount of variance. Each additional component expresses less variance and more noise, so representing the data with a smaller subset of principal components preserves the signal and discards the noise.
- PCA makes maximum variability in the dataset more visible by rotating the axes. PCA identifies a list of the principal axes to describe the underlying dataset before ranking them according to the amount of variance captured by each.

3.5 Performance Metrics for Clustering Algorithms

3.5.1 Silhouette Coefficient. If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters.

3.5.2 Calinski-Harabasz Index. If the ground truth labels are not known, the Calinski-Harabasz index , also known as the Variance Ratio Criterion - can be used to evaluate the model, where a higher Calinski-Harabasz score relates to a model with better defined clusters.

3.5.3 Davies-Bouldin Index. If the ground truth labels are not known, the Davies-Bouldin index can be used to evaluate the model, where a lower Davies-Bouldin index relates to a model with better separation between the clusters.

4 RESULTS AND OBSERVATIONS

4.1 RFM Modelling

- Platinum group represents the most loyal customers
- Gold: Are recent customers and spends more than silver but not frequently.
- Silver is those are Spends less than the gold and not much frequent to visit the platform.
- The bronze group represents the group that has not purchased anything for quite long.

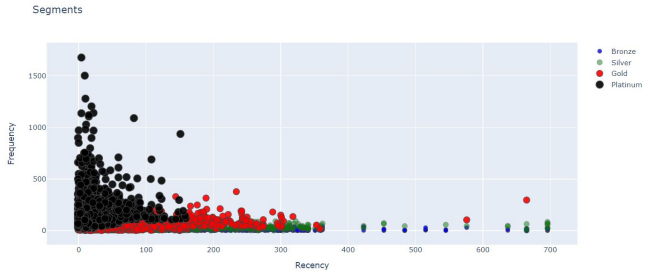


Figure 10: Scatter plot depicting the graph between Recency and Frequency

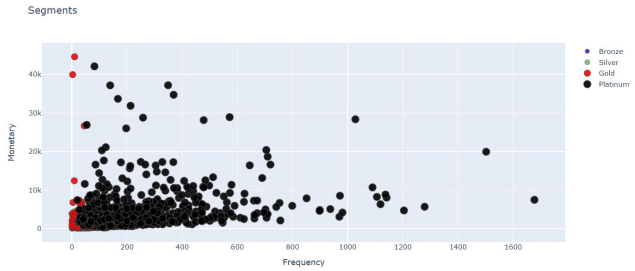


Figure 11: Clusters of four loyalties with respect to frequency and Monetary.

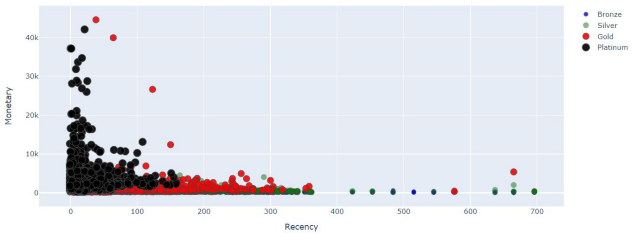


Figure 12: Clusters of four loyalties with respect to Monetary and Recency

Loyality	RFM Score		
Platinum	3	5	4
Gold	7	8	6
Silver	10	9	
Bronze	11	12	

4.2 K-Means Clustering

4.2.1 Determination for optimal Clusters. ***Elbow Method*** - Running the algorithm multiple times over a loop with an increasing no. of clusters choice and then plotting a clustering score as a function of no. of clusters when K increases the centroids are closer to cluster points. Improvements will decline at some point rapidly giving the elbow shape.

Fig 13 shows that the curve dramatically decreases at 3 giving the optimal value for K.

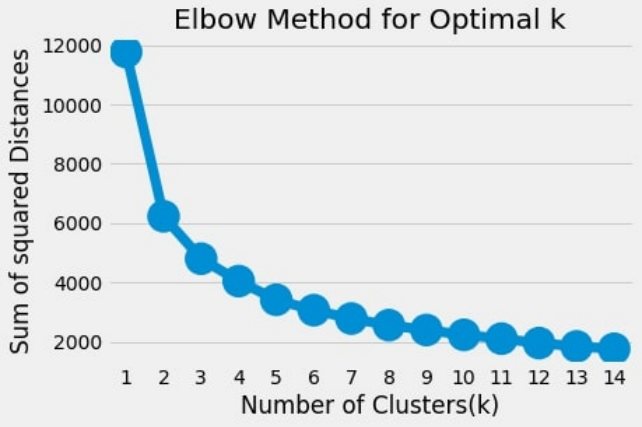


Figure 13: Elbow Method in K-Means Clustering

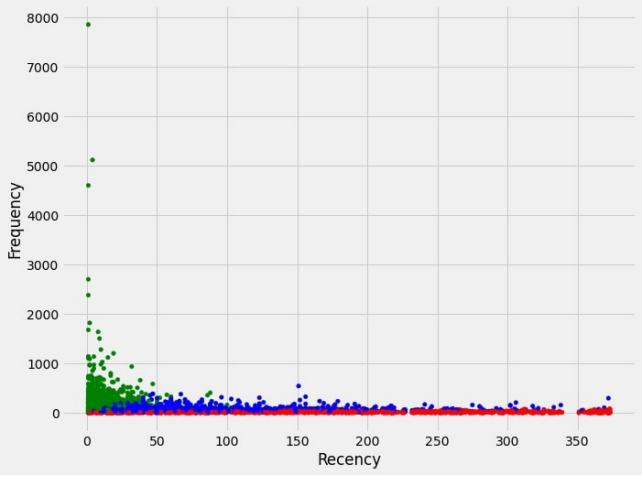


Figure 14: Recency vs Frequency Clustering Graph

4.2.2 Total Clusters. Fig 15 describes the following clusters :

(1) **Cluster 0 - Red (Gold)** - Spends less than the platinum group and not much frequent to visit the platform.
(2) **Cluster 1 - Green (Silver+Bronze)** - One who hasn't purchased from the brand from quite long and he or she may be on the verge of churning out.
(3) **Cluster 2 - Blue (Platinum)** - The most valuable and loyal customer that they don't want to lose.

Clusters	Platinum	Gold	Silver	Bronze
Cluster 0	328	1106	230	0
Cluster 1	0	36	665	705
Cluster 2	812	39	0	0

4.2.3 PCA applied to K-Means. PCA applied to in K-Means clsuetring to observe a more clear picture of clusters where dimensionality is reduced from 3 to 2.

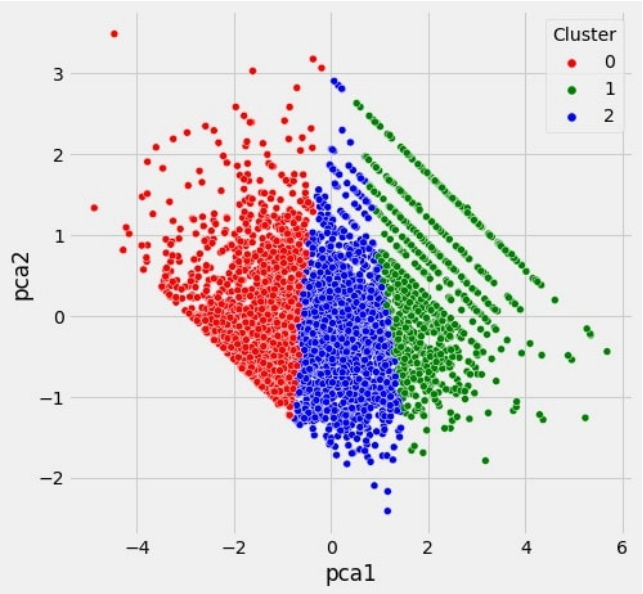


Figure 15: PCA for K-Means

4.3 DBSCAN Clustering Algorithm

4.3.1 Determination for the values for minPts and epsilon.

- **MinPts** = 2*dim, where dim= the dimensions of data set (By Thumb Rule)
- **eps** = Calculated the average distance between each point and its k nearest neighbors, where k = the MinPts value we selected. The average k-distances are then plotted in ascending order on a k-distance graph. We find the optimal value for eps at the point of maximum curvature (i.e. where the graph has the greatest slope).

Fig 16 shows the value for epsilon as near to 0.4

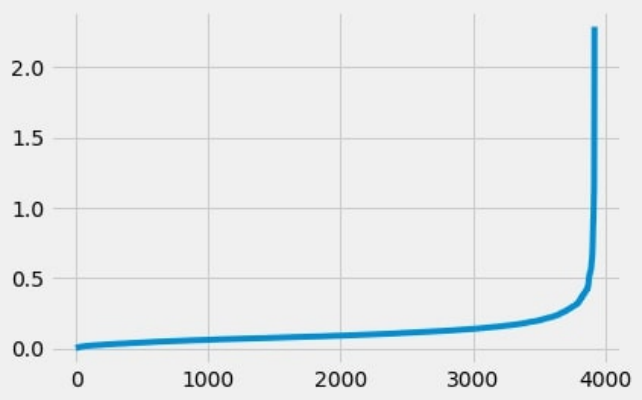


Figure 16: Graph to determine the value for epsilon

4.3.2 Number of Clusters. Fitting the values of minPts and epsilon, we get the number of clusters as 3 shown in Fig 17 below.

- **Cluster 0 - Orange** - It includes most of those who are less loyal than platinum(most loyal) customers and less frequent in recent times along with most loyal customers.
- **Cluster 1 - Green** - It includes 90 perc. most loyal customers.
- **Cluster 2 - Blue** - It has bronze customers who are at the churning out phase.

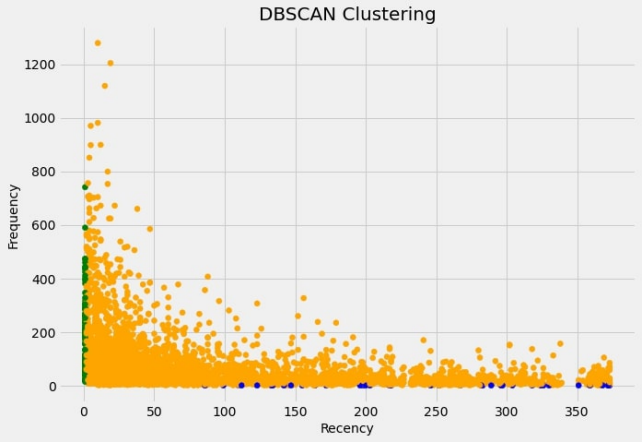


Figure 17: DBSCAN Clustering

Clusters	Platinum	Gold	Silver	Bronze
Cluster 0	995	1128	847	648
Cluster 1	87	10	0	0
Cluster 2	0	0	1	36

4.3.3 PCA applied to DBSCAN. Fig 18 describes the clustering where ***CLuster= -1 Red*** shows the outliers present in the data.

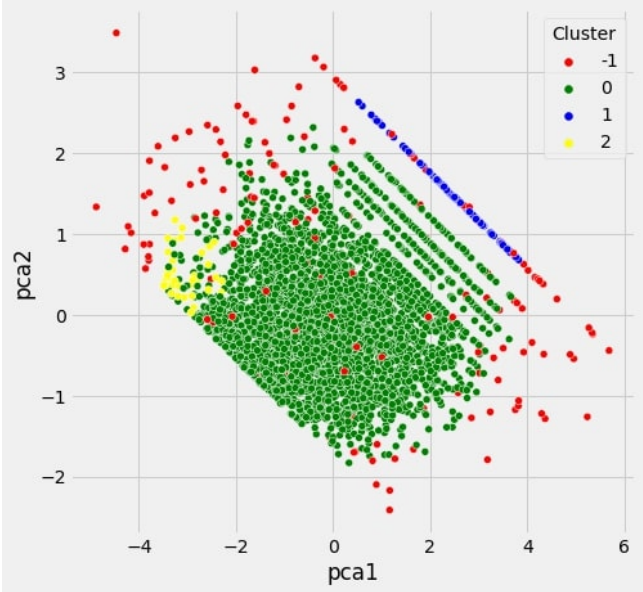


Figure 18: Caption

4.4 Agglomerative Hierarchical Clustering

We can use a ***DENDROGRAM*** to visualize the history of groupings and figure out the optimal number of clusters.

(1) Determine the largest vertical distance that doesn't intersect any of the other clusters
(2) Draw a horizontal line at both extremities.
(3) The optimal number of clusters is equal to the number of vertical lines going through the horizontal line.

Fig 19 is build with the Linkage criteria as *Ward* which uses the distance metrics as *Euclidean.*

Figure 19: Dendogram giving the the optimal no of 2 clusters

4.4.1 Number of Clusters. Fig 20 shows the no of clusters.

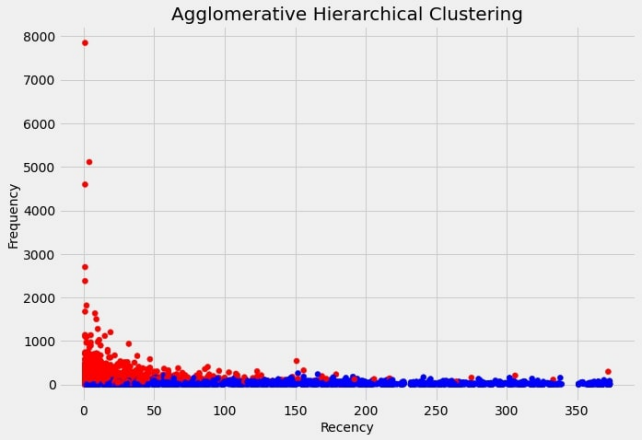


Figure 20: Agglomerative Hierarchical Clustering

- **Cluster 0 - Red** - It includes most of those who are least loyal customers(silver) and not frequent in recent times along with more loyal customers(gold) and those who are at churning out phase(bronze) or simply includes all types except the most loyal ones.
- **Cluster 1 - Blue** - It includes approx 80 perc. most loyal customers along with less loyal ones i.e. gold.

Clusters	Platinum	Gold	Silver	Bronze
Cluster 0	26	833	892	705
Cluster 1	1114	348	3	0

Figure 21: PCA applied to Agglomerative Hierarchical Clustering

4.4.2 PCA applied.

4.5 Gaussian Mixture Model Selection

The BIC criterion can be used to select the number of components in a Gaussian Mixture in an efficient way. In theory, it recovers the true number of components only in the asymptotic regime (i.e. if much data is available and assuming that the data was actually generated i.i.d. from a mixture of Gaussian distribution). ***The model with the lowest BIC is selected***

Fig 22 shows the minimum score of 4 in full model, therefore, number of clusters are 4.

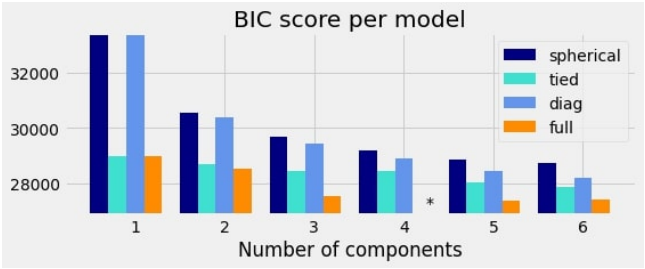


Figure 22: BIC Score Per Model

4.5.1 Number of Clusters. Fig 23 shows 4 clusters

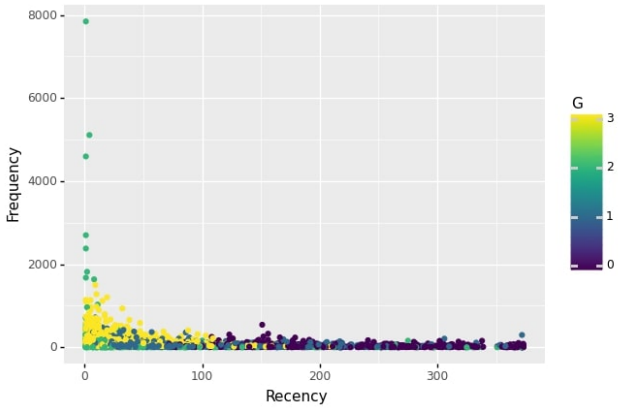


Figure 23: Gaussian Mixture Model

Clusters	Platinum	Gold	Silver	Bronze
Cluster 0	0	46	90	149
Cluster 1	364	165	78	3
Cluster 2	774	448	1	0
Cluster 3	2	522	553	726

- **Cluster 0 - Purple** - It includes most of those at churning out phase(bronze) or simply includes all types except the most loyal ones with highest number of churning out customers.
- **Cluster 1 - Blue** - It includes approx 80 perc. most loyal customers along with less loyal ones i.e. gold and very less number of silver and bronze types.

- **Cluster 2 - Green** - It includes approx 80 perc. most loyal customers along with less loyal ones i.e. gold and very less number of silver type.
- **Cluster 3 - Yellow** - It includes most un-loyal customers(at churning out phase) along with less loyal ones i.e. gold and silver types.

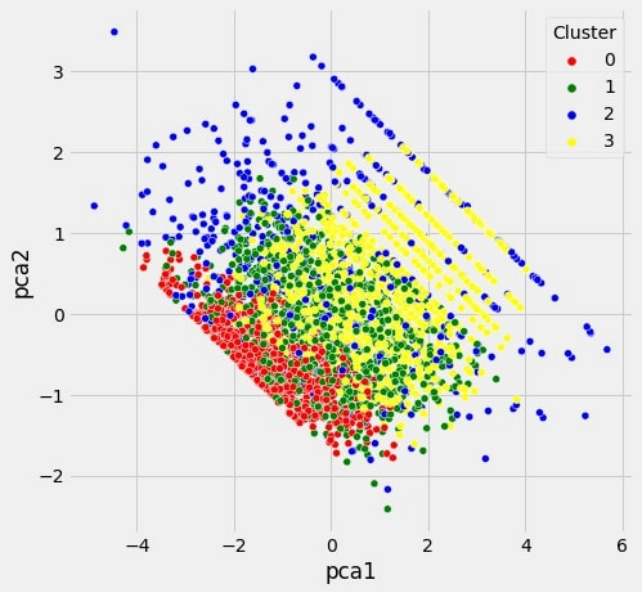


Figure 24: PCA applied on Gaussian Mixture Model

4.5.2 PCA applied.

4.6 Clustering Performance Evaluation

Table 1: Comparative analysis between prior research works.

S.No.	Clustering Algo.	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index	Clusters
1.	K-Means	0.3068	2844.800	1.0871	3
2.	DBSCAN	-0.0443	70.9740	1.4509	3
3.	Hierarchical Agglomerative	0.3889	3290.3554	0.9520	2
4.	Gaussian mixture Model	0.0310	492.6882	2.5921	4

After analysing the performance metrics of different clustering algorithms, we observed that Hierarchical Agglomerative Clustering is preferred over other as it has the highest Silhouette Coefficient, highest Calinski-Harabasz Index and lowest Davies-Bouldin Index.

5 FUTURE WORK AND CONCLUSION

Future Work of this project includes customer churn out procedure. Customer churn (or customer attrition) is a tendency of customers to abandon a brand and stop being a paying client of a particular business. The percentage of customers that discontinue using a company's products or services during a particular time period is called a customer churn (attrition) rate. In a subscription-based business, even a small rate of monthly/quarterly churn will compound quickly over time. Just 1 percent monthly churn translates to almost 12 percent yearly churn. Given that it's far more expensive to acquire a new customer than to retain an existing one, businesses with high churn rates will quickly find themselves in a financial hole as they have to devote more and more resources to new customer acquisition.

Segmentation allows businesses to make better use of their marketing budgets, gain a competitive edge over rival companies and, importantly, demonstrate a better knowledge of your customers' needs and wants.