Journal of Statistics, Computing and Interdisciplinary Research
ISSN (P): 2707-7101

J. Stat. Comp. & Interdis. Rese.
Volume 5(2) (2023), 203-221
http://scir.wum.edu.pk/index.php/ojs

Utilizing the K-Means Clustering Algorithm for Advanced Visualization of Credit Card Data in a Banking Context

Zarsha Nazim[1]*, Maria Tariq[2], Fatima Tariq[3], Anila Barkat[4], Tehreem Fatima Rai[5]

[1]Department of Software Engineering, Lahore Garrison University, Lahore, Pakistan.
[2]Department of Computer Science, Lahore Garrison University, Lahore, Pakistan. & Department of Computer Science, NCBA&E, Lahore Pakistan.
[3]Department of Software Engineering, Lahore Garrison University, Lahore, Pakistan.
[4]Department of Software Engineering, Lahore Garrison University, Lahore, Pakistan.
[5]Department of Software Engineering, University of Central Punjab, Lahore, Pakistan.
***Corresponding Author:** zarsha.nazim@lgu.edu.pk

Abstract

Due to the emergence of numerous entrepreneurs with startup ideas and competitors, the new businesses are in significant need of exploring tools and technologies to figure out new buyers and at the same time to keep the older ones as well. Customer segmentation using k mean clustering is an imperative technique to separate the customers into targets segments which can help the businesses to apply marketing strategies accordingly. This can also help in providing exceptional customer services. The research is based on analyzing the dataset of a bank to estimate the customer segmentation of a credit card by proposing a model to help a company define its marketing strategies. K-mean algorithm has been used for dividing the group of customers into segments in the form of clusters by determining the value of k through a silhouette technique. For better visualization Principal Component analysis has been used for dimensionality reduction and to achieve better results by implementing better visualization in Jupiter notebook.

Key Words: *mean algorithm, customer segmentation, clustering.*

Introduction:

The practice of examining data to discover patterns and valuable information is known as data mining. Data mining is used to analysis data and to make its utilization accessible for mining technologies. This data is being used for a variety of applications such as customer segmentation & systems, production management & control, medical & healthcare, market analysis, manufacturing, engineering, scientific discoveries, decision making and others.[3]

a) Background:

Data mining is researched for various database categories such as relational databases as per objects, simple relational databases, data warehousing, and multimedia related databases, among others. Many applications, such as market-basket analysis, rely heavily on data mining. Frequent item sets play an important impact on data mining, which is used to discover correlating relations between database fields. The relation rule depends on the discovery of frequent item sets that are frequently used by retail stores. [6]

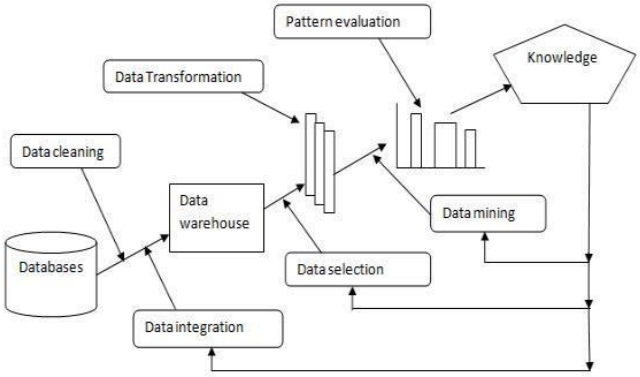


Fig. 1: Correlations in data mining with often occurring item sets.

Mining data progresses for extraction of data that is in indirect manner and such data can be requested from databases.

The customer segmentation is used to explore the characteristics of Customers. It helps to differentiate groups of customers according to their properties and helps in categorizing them by means of their similar attributes in the form of segments. Segmenting involves different aspects and is divided according to behavior, demographic, geographical and psychographic segmentation. It is one of the marketing approaches to enhance your business. Businesses

require to analyze data to know their targeted market. Data mining is used in market segmentation to forecast new trends. It will be significant in exploring market trends and making explorations. The data can be analyzed by identifying patterns and unseen relationships. Hence it can be improved to run the business in a better way.[1][9]

b) Clustering in Data Mining:

Cluster analysis is widely utilized in a variety of applications such as market research, pattern identification, data analysis, and image processing.[16] Clustering in business can assist marketers in discovering client interests based on purchase patterns and characterizing customer groupings. In biology, it may be used to construct plant and animal taxonomies, classify genes with comparable functioning, and obtain insight into population patterns.

Clustering can be used by geologists to recognize areas of similar lands, relative houses in the town, and so on. Clustering of information may also be useful for categorizing records on the Web to facilitate information discovery.[2][4]

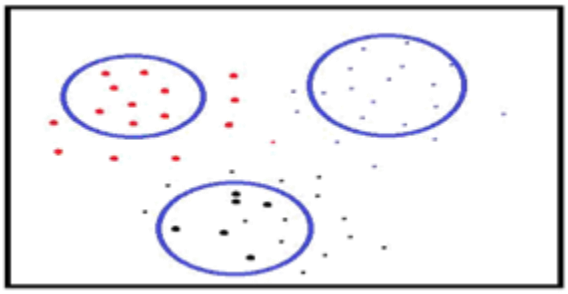


Fig. 2: insights in data mining with data clustering.

Data clustering is not listed as supervised classification strategy that seeks to classify things into groups, or clusters which are represented in a manner that items in the similar clusters are almost the same and items in separate clusters are quite diverse. Cluster analysis is a well-known concept in the field of data mining. It is the initial step toward fascinating valuable information discovery. Clustering is the process of categorizing data items into a collection of distinct classes known as clusters. Items within a class have a strong similarity to one another, whereas objects from different classes are more distinctive.[10]

c) Problem statement:

To understand the business needs of a bank there is a need to extract knowledge from the data of a particular bank and to develop bank customer segmentation which will help to define marketing strategy. The dataset that has been analyzed includes the observations of approximately 9000 credit card users that are active during the tenure of 6 months. The data includes 18 behavioral variables.

Literature Review:

a) **Segmentation:**

The question that arises while implementing the model for a customer segmentation is what exactly segmentation is and why do we require it? Basically, Segmentation is an important constituent for the initialization of marketing strategies and objectives to further generate the significant analysis of how the product could be sold or developed. That strategy will be based on the current customer segments.[7]

b) **Need of segmentation:**

Sometimes for a company segmentation is critical because of resource constraints and other influencing factors and in such cases, there is a chance of loss in context to the customers. Therefore, there is always a need to identify the customers for the particular business and to provide them with suitable marketing strategies. Similarly, when we talk about effective segmentation our focus should be to help highlight the most suitable group of customers that should be served by a company and to best position product and services for each group. [15]

c) **Customer Segmentation:**

Customer characteristics data can be found out by applying segmentation on data. There are a lot of ways to apply segmentation. We can differentiate between the customers and the organizations based on the treatment provided. For this we can simply use some data mining techniques which will be applied in the customers' data set. Further in details, we will use clustering mechanism. In clustering mechanism, we can create new informative data segments by using relevant customer's data and customer's metadata. Then we can do cluster analysis to dig further in depth. Based on the analysis on these clusters, we can group them by relevant similarities.[14] Later we can apply business models on these groups of similar clusters, for example we can run a marketing campaign on each group based on customer

interests as per their group. These campaigns can be product based or service based or simply email, or messages based. We can make multiple clusters using segmentation. These clusters can be based on customers' behavior, how much the customers spends weekly, monthly, yearly etc., what kind of products a certain age group of customers buys, which customers spends the most and much more cases like these. Now after grouping customers based on analysis, we can treat customers accordingly, for example, high spending customers can be benefit to extra free services like discounts, gift cards etc. [18]

d) **K-Means Clustering:**

In K-means we split data into clusters based on number of groups provided at initial level. These number of groups are defined by an initial centroid value. This is a very important phase as while applying K-means clustering algorithm we can end up in unstable data sets. The initial cluster is called the centroid cluster. The initial cluster center can be changed and affect the over data set. The initial value of the cluster's center point represents the base case from where we can determine a cluster. K-means clustering algorithm splits the given data set into k number of clusters based on the distances.[11]

1. We need to find out the number of clusters and the maximum number of iterations that will be performed.
2. We need to do calculations to determine the initial cluster midpoint.

$$C_j = \frac{1}{M}\sum_{j=1}^{M} x_j \qquad (1)$$

3. Now we can link clusters with each other using this base case. We simply need to connect our base case cluster with the nearest cluster. We can use Euclidean distance equation to calculate the distance.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (2)$$

4. Now we will regroup data based on the distance between them.

$$a_{ij}\begin{cases}1 & d = \min\{D(x_i, c_i)\} \\ 0 & otherwise\end{cases} \qquad (3)$$

5. Now we will calculate the central point or mid points for each cluster. In the above equation 'd' represents the smallest distance from starting point 'x' to the center of a particular group 'c'. 'a' represents the value of membership. We can also find out the objective using an equation,

$$J = \sum_{i=1}^{n}\sum_{l=1}^{k} a_{ic} D(x_i, c_1)^2 \qquad (4)$$

This is the MacQueen equation, 'a' represents the membership value. 'x' is the starting position, 'c' is the center or midpoint of a cluster. If data is from a set of anngota group, then value 'a' is 1 else the value of 'a' is 0.

6. At the end if there is midpoint or center value updates or somehow gets number of iterations gets less than max number of iterations then in that case continue again form step 3 else return the result.

e) Elbow Criterion:

The Elbow Criterion method can be helpful in finding the best value of k at a cluster which is most suitable. By using Elbow Criterion method if we increment the k value then the graph will slowly decrease resulting in stability, we will eventually get to a point where the value of k is stable in graph.[13][1] To find the value of K at best cluster,

1. Determine the value of k as the clusters are created. Now select the k number clusters and apply grouping on them using k-means algorithm. The elbow method is determined by the net sum of the Squared errors.

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \|X_i - C_k\|_2^2 \qquad (5)$$

In the above equation, k is the number of clusters, x is the data present in each cluster.

2. The center point or midpoint of cluster is determined at the beginning. Then we can calculate the next i[th] cluster using the following formula,

$$v = \frac{\sum_{i=1}^{n} x_i}{n}; i = 1,2,3, \ldots\ldots, n \qquad (6)$$

3. Calculate the equation between objects using.

   Euclidian Distance formula,

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\ ; i = 1,2,3, \ldots\ldots, n \qquad (7)$$

4. Select the object with nearest centroid and connect them.
5. Connecting the object with each cluster and iterate with k means such that each member of the cluster has proximate distance from the midpoint of the cluster.
6. Iterate through the entire graph and find the position of centroid.
7. If the new centroid value does not match with the previous value of centroid, then repeat from step 3, otherwise return the value.

f) Data Analytics:

Data analytics is the science of studying raw data to extract valuable information from which to derive conclusions. Many businesses and organizations may benefit from data analytics to make valuable business decisions. Such analytics of data focuses on inference, which is the act of drawing a conclusion based purely on what the researcher is informed and knows well. Data analytics is classified into two types: They are as follows:

1. Data Classification and its analysis.
2. Data Prediction and its analysis.

Classification of data models anticipate class tags as per categories, whereas prediction models is involved in the prediction continuous valuable functions and data. While prediction model predicts possible consumers' spending on computer equipment in dollars based on their total household salary or income and profession relation information.[6]

Data management, talents, information technology, and modelling are all brought together by Prediction Analytics. Analytics that is predictive is a data science, blended discipline data skill that is critical for organizations that are actually non-profit, commercial level achievement, and government level achievement. Marketing sales forecasts or market share, a solid retail site possibility has been discovered. Predictive analytics is also used to identify customer groups and target marketing as well as dangers connected with existing goods.

Classification of data and information is a data mining approach that falls under machine learning and is used to predict data instance group membership. The classification approach, which can analyze a broader range of data than regression, is gaining favor.[17].

Methodology:

To cluster comparable types of data for data analysis based on prediction model of data set, the k mean clustering method is used. The frequency of the most relevant function is estimated in the k-mean clustering procedure, and the functions are grouped using the Euclidian distance formula. In this study, we will modify the Euclidian distance formula to improve the all over cluster. Normalization will serve as the foundation for the improvement. Two new features will be implemented as part of the improvement. The first step is to compute normal distance measurements based on normalization. The functions will be clustered in the second point based on majority voting. Jupiter Notebook is used to implement the proposed approach.[18] [4]

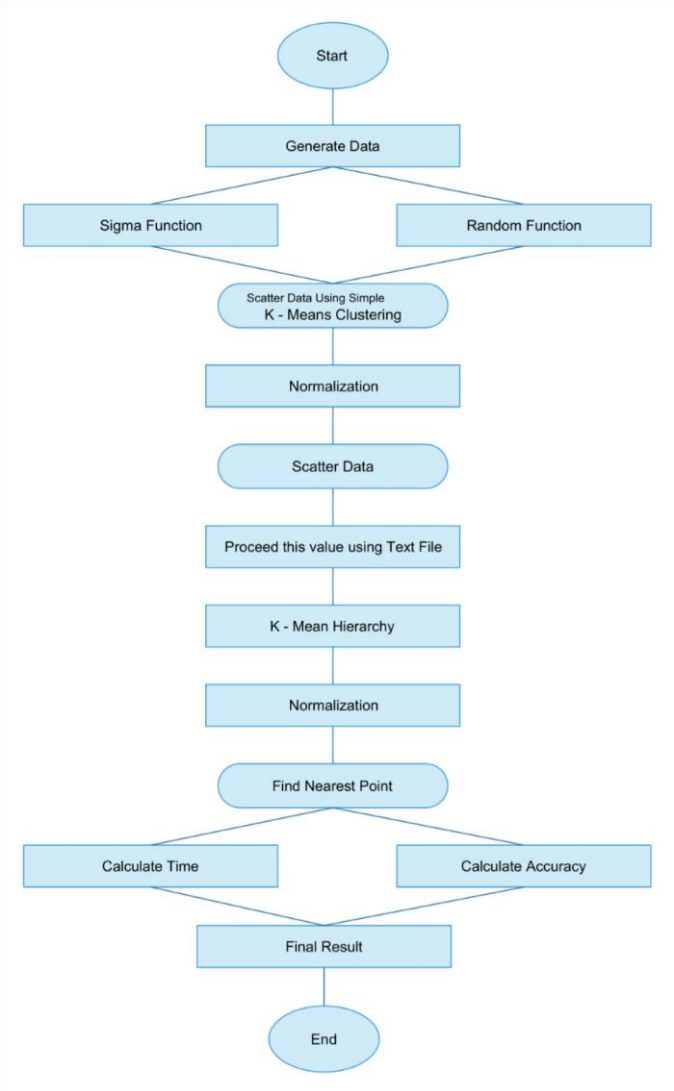


Fig. 3: Methodology to generate results, comprising data generation, k-means clustering, normalization, and iterative refining.

The following steps are followed to generate the results:

1. Firstly, we began the process by generating information from the user side, in these processes we provide a variety of inputs based on data created by sigma and functions that are random.
2. Later the entire data set is produced, k-means is then used, and the result is shown in the sub part.
3. We produced scatter data in the second sub part. after normalizing the data.

4. Now we performed normalization, which was preceded by reading text file data from that produce data and then applying hierarchy k-means before normalizing, which resulted in a different shape than the first subplot.
5. Following the completion of this phase, the iterations process is completed.
6. This method is repeated until and unless we don't acquire a closest point that is near or very close to data generation.
7. Finally, we calculated their total duration and obtained findings that suggest an improvement in cluster accuracy.[6]

Results and Discussion:

a) Data Understanding:

The research being done to define customer segments while using data mining techniques. Customer segmentation is one of the significant applications in Data Mining. The goal is to build a segmentation model by applying descriptive analysis of our data. K-mean clustering will be performed which is an important algorithm for clustering of unlabeled dataset. Firstly, the dataset will be downloaded, and data will be analyzed by using Python language. Firstly, it will be checked whether the data is metric or not. Then segmentation variables are selected. Furthermore, clustering and segmentation will be done.

Customer segmentation is very important for every type of business as it helps to narrow down and identify the target market for such a business. In this paper I am going to use data mining techniques and tools to make customer segmentation by taking a dataset of bank to target customer segmentation. Customer segmentation is a practice done by a company where they divide their customers into different segments to predict which customer will purchase their product after launch. This is also done by new businesses to identify the target customer for their services or products. Through customer segmentation the business analyst finds how their product will impact their customer lives.

Analyst’s goal is to maximize the profits of their business by analyzing the report they create by collecting the data using segmentation. One of the ways to make the reports is by using data mining tools and data sets.

Initial Steps and Procedure: Firstly, understand the problem statement that is based on business case banking segmentation. Following Unsupervised machine learning algorithm for customer segmentation. We will divide out different objects to make groups. Machine

learning algorithm will be applied to perform bank customer segment. It can help customers to segment in groups. Also, we will Import libraries to obtain optimal no of clusters and to categorize no of customers in many segments.

CUST ID	Identification of Credit card holder (Categorical)
BALANCE	Balance amount left in customers account to make purchases
BALANCE FREQUENCY	How frequently the balance is updated.
PURCHASES	No of purchases made from account
ONEOFFPURCHASES	Maximum purchase in one-go
INSTALLMENT PURCHASES	Amount of purchase done in installment
CASH ADVANCE	Cash in advance given by the user
PURCHASESFREQUENCY	How frequently the purchases are being made
ONEOFF PURCHASES FREQUENCY	How frequently purchases are happening in one go
PURCHASE INSTALLMENT FREQUENCY	How frequently purchases in installments are being done
CASH ADVANCE FREQUENCY	How frequently the cash in advance being paid
CASH ADVANCE TRX	Number of transactions made with 'cash in advance'
PURCHASES TRX	Number of purchase transactions made
CREDIT LIMIT	Limit of card for user
PAYMENTS	Amount of payments done by user
MINIMUM PAYMENTS	Minimum amount of payments made by user
PRC FULL PAYMENTS	Percent of full payment made by user
TENURE	Tenure of credit card service for user

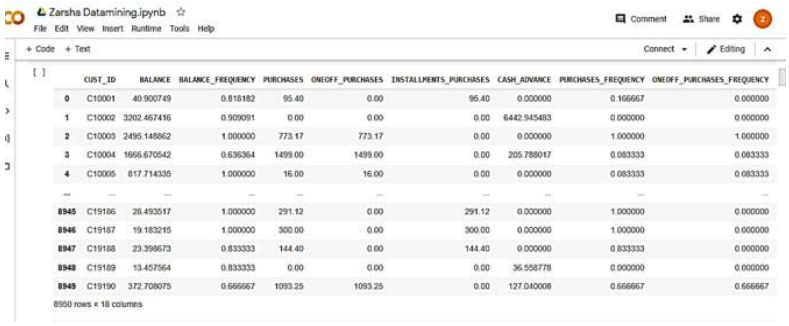
Fig. 4: The extensive data dictionary of the credit card dataset.

b) Defining Data:

The following table highlights the data dictionary for credit card dataset and explains the purpose of each categorical data attribute.

c) Pre-processing Data:

The initial process is data pre-processing which includes checking missing data, type of data, importing libraries and packages. Also selecting features.


Zarsha Datamining.ipynb

File Edit View Insert Runtime Tools Help

Comment Share

+ Code + Text

Connect Editing

[ ]

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY
0	C10001	40.900749	0.818182	95.40	0.00	95.40	0.000000	0.166667	0.000000
1	C10002	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	0.000000	0.000000
2	C10003	2495.148862	1.000000	773.17	773.17	0.00	0.000000	1.000000	1.000000
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.00	205.788017	0.083333	0.083333
4	C10005	817.714335	1.000000	16.00	16.00	0.00	0.000000	0.083333	0.083333
...	...	...	...	...	...	...	...	...	...
8945	C19186	28.493517	1.000000	291.12	0.00	291.12	0.000000	1.000000	0.000000
8946	C19187	19.183215	1.000000	300.00	0.00	300.00	0.000000	1.000000	0.000000
8947	C19188	23.398673	0.833333	144.40	0.00	144.40	0.000000	0.833333	0.000000
8948	C19189	13.457564	0.833333	0.00	0.00	0.00	36.558778	0.000000	0.000000
8949	C19190	372.708075	0.666667	1093.25	1093.25	0.00	127.040008	0.666667	0.666667

8950 rows × 18 columns

Fig. 4: Data pre-processing, early analysis, and missing data check I.

As we can see from the table there are 8950 rows and 18

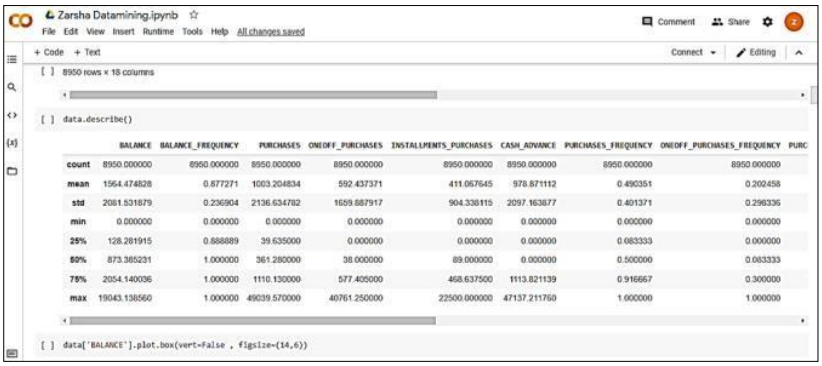


Fig. 5: Dataset overview: 8950 rows, 18 columns, key statistics-II

Above data shows max, min count, mean, standard deviation and interquartile range. We can examine from the data that there are many outliers as we can see the max value. Outliers can be taken as extreme values here. Now we need to fill out the credit limit column since we are in a part of data cleaning, we need data to be consistent and not missing. We will replace missing values with the mean value of minimum payment column. We will use loc to access column. We can tell how many null values and then we are replacing it with current average values Inside loc we provide row comma column, so we are providing two things to access value. We provide rows and columns inside loc.

```
[ ] bankseg_df.isnull().sum()

CUST_ID                          0
BALANCE                          0
BALANCE_FREQUENCY                0
PURCHASES                        0
ONEOFF_PURCHASES                 0
INSTALLMENTS_PURCHASES           0
CASH_ADVANCE                     0
PURCHASES_FREQUENCY              0
ONEOFF_PURCHASES_FREQUENCY       0
PURCHASES_INSTALLMENTS_FREQUENCY 0
CASH_ADVANCE_FREQUENCY           0
CASH_ADVANCE_TRX                 0
PURCHASES_TRX                    0
CREDIT_LIMIT                     1
PAYMENTS                         0
MINIMUM_PAYMENTS               313
PRC_FULL_PAYMENT                 0
TENURE                           0
dtype: int64
```

Fig. 7*:* Data cleaning; Filling missing values for consistency.

```
[ ] bankseg_df.isnull().sum()

CUST_ID                          0
BALANCE                          0
BALANCE_FREQUENCY                0
PURCHASES                        0
ONEOFF_PURCHASES                 0
INSTALLMENTS_PURCHASES           0
CASH_ADVANCE                     0
PURCHASES_FREQUENCY              0
ONEOFF_PURCHASES_FREQUENCY       0
PURCHASES_INSTALLMENTS_FREQUENCY 0
CASH_ADVANCE_FREQUENCY           0
CASH_ADVANCE_TRX                 0
PURCHASES_TRX                    0
CREDIT_LIMIT                     1
PAYMENTS                         0
MINIMUM_PAYMENTS                 0
PRC_FULL_PAYMENT                 0
TENURE                           0
dtype: int64
```

Fig. 8*:* Data types: Dropping 'CUST_ID' column - *IV.*

Now further we will be checking the data types. We have examined the data types and after the observations it can be seen from the table below that the data type of CUST_ID is object so it will be dropped.

Data Normalization:

The next part would be to apply normalization on data. Data Normalization is a technique that is applied to prepare data by using Machine learning. Data scaling is a part of data normalization where the data values of numeric columns are changed to common scale. Normalization is only required when there are different ranges otherwise every dataset doesn't need normalization.

```
[ ] bankseg_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   CUST_ID                           8950 non-null   object
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64
 12  PURCHASES_TRX                     8950 non-null   int64
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64
dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB
```

Fig. 9: PCA: Reducing dataset to 2 dimensions for visualization.

Using PCA for Dimension Reduction We have applied PCA for dimensionality reduction on our dataset to divide data into 2 dimensions as the data cannot be visualized in 17 dimensions. The purpose of PCA is to divide a large set of data variables into smaller ones that consists of most of the information in the large set.

Clustering:

We have opted for clustering technique keeping in consideration our dataset as this technique is significant in giving information about the structure of the data and identifying subgroups in the data where data points in similar clusters are considered similar while in different clusters data points are considered different.

K-Means Clustering:

K means algorithm will be used to analyze our data. We will visualize data in various ways. We cluster data according to the groups. Unsupervised learning involves clustering and association we use unlabeled data in unsupervised learning. Clustering algorithm is a part of unsupervised machine learning. Firstly, get some data points. We will segment or group these data points. To obtain the optimal no of clusters we have applied the

Will be a centroid of a particular cluster each cluster will have a centroid. It groups the data that have similar values by applying Euclidean distance mean works by grouping some data points together in an unsupervised way. It groups data by measuring Euclidean distance between the points. First, we need to specify no of clusters. We randomly select them we need to segment our customers into n groups. We cluster using elbow method and then we will calculate distance between each data point and cluster. We select cluster centers. Then calculate distance. Clusters and centroids are equal. Point near to centroid will fall under that centroid group. We recalculate centroid for each cluster. We reassign data point to centroid until it is stable.

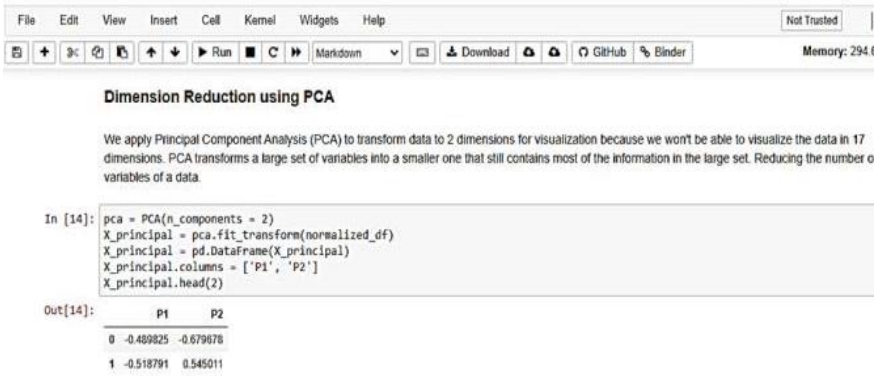


Fig. 10: Data Normalization: Scaling data for machine learning.

Elbow method to select optimal no of clusters.in elbow method will experiment with different no of clusters and then we will plot a graph. There will be a graph like an elbow if graph is linear than this is optimal value.so no of clusters will vary and then we will calculate distance. We measure the distance from every point and then some of the distance of every point in that cluster. Same is for single cluster. Calculate the distance of every point to this centroid, if the number of clusters is less and distance is less.

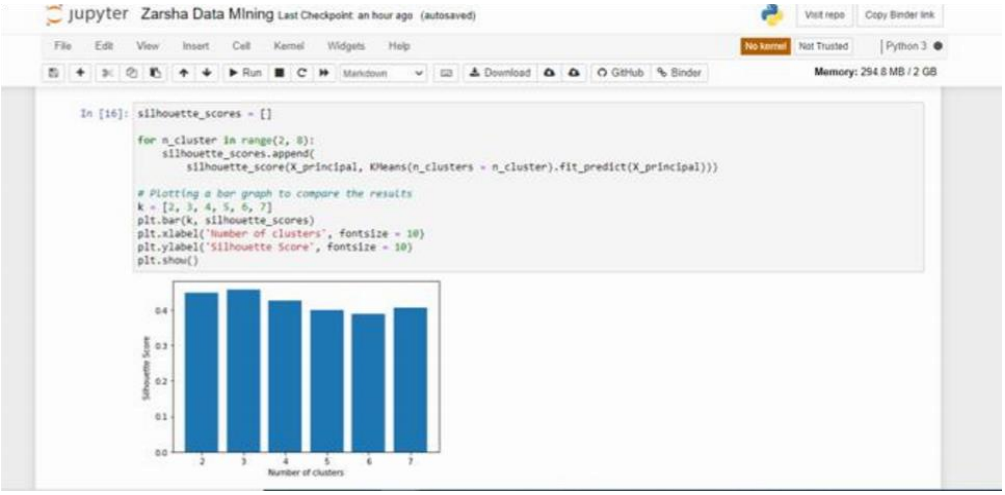


Fig. 11: Elbow Method Analysis: Bar graph to identify optimal number of clusters.

Data scaling is a part of data normalization where the data values of numeric columns are changed to common scale. The normalization is only required when there are different ranges otherwise every dataset doesn't need normalization in implementing this model, we have used k-means clustering algorithm for our proposed dataset as it's an iterative algorithm which helps to divide the dataset into subgroups where each data point belongs to one group. To apply k-mean clustering, firstly we need to specify k by using elbow method. Elbow method is a simple technique to know the number of clusters or the optimal number of k. While applying elbow method the results were difficult to determine as it was hard to find the curve so silhouette.

Specifying K using the elbow method:

Specifying K using the elbow method while applying elbow method it was difficult to figure out the elbow point of the curve. Instead, the silhouette method was simply used as it was easier to find k. As can be seen, taken against the number of clusters.

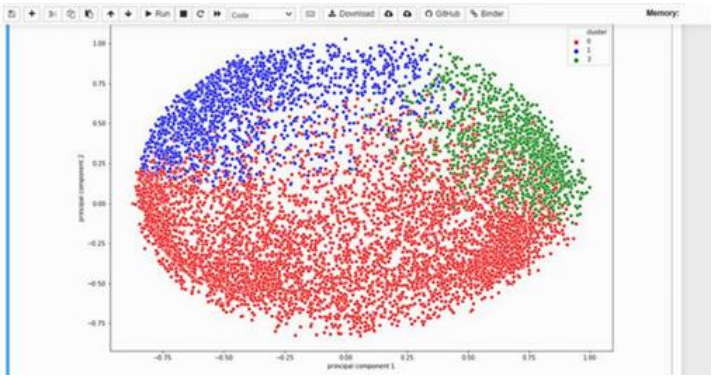


Fig. 12: Cluster Visualization: Customer groups based on attributes.

The value of silhouette coefficient ranges from -1 to 1. These two ranges signify different information. Silhouette coefficient for 1 depicts that the data point is closest to its cluster to which it belongs and far away from other clusters. Similarly, value of -1 seems to be the worst value where we can highlight the values near 0 belongs to overlapping clusters. From the figure above we can see that the most suitable value for k is equal to 3. This means that this value includes the best no of clusters.so accordingly k=3 was assigned to K means model. There is total k=3 clusters Cluster 0,1 and 2. From the figure below it can be depicted from the data that the Cluster 0 contain group of customers with low purchase, balance, purchase frequency, cash advance, minimum payment, credit limit and high Balance frequency which tells that in this cluster exists small group of customers with low credit limit. From the extracted data of Cluster 1 we can conclude that this group belongs to customers having high credit limit.

These are the large group of customers having high balance, low purchases, high cash advance and high minimum payment. According to the data, this group of customers denoted as cluster 2 have very low purchase frequency, high balance, low purchases, high cash advance and a high credit limit. This group of customers use their credit cards for loan purposes.

The findings of the process of identifying the best number of clusters using the elbow and K-Means approaches show that the process of selecting the best number of clusters may yield the same number of clusters K on the amount of varied data. Based on the case study, the outcome of finding the optimal number of clusters with the elbow approach will be the default for the characteristic procedure.

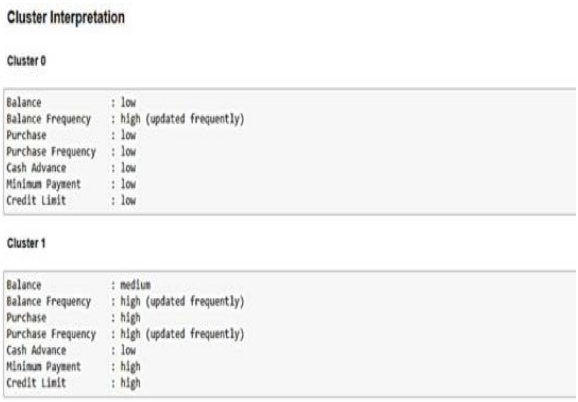


Fig. 13: Cluster Distribution Analysis-I

Fig. 14: Credit Usage Analysis -II

Conclusion:

This research concludes that clustering is a strategy for dividing huge datasets into discrete data collections known as clusters. The cluster contains data set that has been converted into meaningful and reliable information. The data clusters differ from one another since they have distinct values. There are a variety of techniques that can partition a dataset into clusters and perform well for clustering data.

Recommendations:

Cluster 0:

After the analysis of data, we recommend a silver credit card as it has the lowest credit card limit. The benefit of this card is that the limit of this card is not too high.

Cluster 1:

For cluster 1 gold credit card is suitable as the benefit of this card is that the limit is high. It will help the customers to own expensive items faster. It helps to buy electronics and motorbikes, but it has a higher annual fee as the credit limit is high.

Cluster 2:

For cluster 2, platinum credit card is suitable, but it’s owned by very few people due to constraints and is not that easy to own it.

In this research, a strategy for modifying the K means Cluster is suggested. In this suggested improvement, the clustering will be free of the two key shortcomings of the clustering method, which are the accuracy level and the computation time required to cluster the dataset.

Future Work:

In future work, we will apply the DBSCAN (Density Based Spatial Clustering of Applications with Noise) Algorithm, which finds high-density core samples and forms clusters from them. It is useful for data that comprises clusters of comparable density. When K-Means cannot form arbitrarily shaped clusters, this is when the DBSCAN algorithm is used to compare k means clustering to see what results we achieve.

References:

[1] A Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia databases with noise", Proc. of the 4th Int'l Conf. on Knowledge Discovery and Data Mining (KDD'98), 1998.

[2] Asiabi T P K and Tavoli R 2015 A Review of Different Data Mining Techniques in Customer Segmentation J. Adv. Comput. Res. 6 51–63

[3] ALN Fred and JMN Leitão, "Partitional vs hierarchical clustering using a minimum grammar complexity approach", Proc. of the SSPR & SPR 2000. LNCS 1876, pp. 193-202, 2000, [online] Available: http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.htm. [4] Yoseph, Fahed et al. 'The Impact of Big Data Market Segmentation Using Data Mining and Clustering Techniques'. 1 Jan. 2020: 6159 – 6173.

[5] C Ding and X. He, "K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization" in Proc. of the ACM Symp. on Applied Computing, Nicosia:ACM Press, pp. 584-589, 2004.

[6] K. Tsiptsis and A. Chorianopoulos, "Data Mining Techniques in CRM: Inside Customer Segmentation", John Wiley and Sons, Ltd., Publication, 2009.

[7] Cai, Z. Technical aspects of data mining. PhD thesis, Cardiff University, Cardiff, 2001.

[8] He X and Li C 2016 The Research and Application of Customer Segmentation on E-commerce Websites 2016 6th International Conference on Digital Home pp 203–9

[9] Castro, V. E., Yang, J. A fast and robust general purpose clustering algorithm. In Proceedings of the Fourth European Workshop on Principles of Knowledge Discovery in Databases and Data Mining (PKDD 00), Lyon, France, 2000, pp. 208–218.

[10] F Yuan, Z. H Meng, H. X Zhang and C. R Dong, "A New Algorithm to Get the Initial Centroids", Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, 2004.

[11] Halkidi, M., Batistakis, Y., Vazirgiannis, M. Cluster validity methods. Part I. SIGMOD Record, 2002, 31(2); available online http://www.acm.org/sigmod/record/.

[12] Moghaddam S Q, Abdolvand N and Harandi S R 2017 A RFMV Model and Customer Segmentation Based on Variety of Products J. Inf. Syst. Telecommun. 5 155–61

[13] Han, J., Kamber, M. Data Mining: Concepts and Techniques, 2000 (Morgan Kaufmann, San Francisco, California).

[14] K.A. Abdul Nazeer and M.P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceeding of the World Congress on Engineering, vol. 1, 2009.

[15] Kerr, A., Hall, H. K., Kozub, S. Doing Statistics with SPSS, 2002 (Sage, London).

[16] Maulina N R, Surjandari I and Rus A M M 2019 Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering 2019 16th International Conference on Service Systems and Service Management (ICSSSM) (IEEE)

[17] Lindeberg, T. Scale-space Theory in Computer Vision, 1994 (Kluwer Academic, Boston, Massachusetts).

[18] Pena, J. M., Lazano, J. A., Larranaga, P. An empirical comparison of four initialisation methods for the K-means algorithm. Pattern Recognition Lett., 1999, 20, 1027–1040.

[19] R Gelbard and I. Spiegler, "Hempel's raven paradox: A positive approach to cluster analysis", Computers and Operations Research, vol. 27, no. 4, pp. 305-320, 2000.