ESSENTIAL OF DATA ANALYTICS

CSE3506

BANK CUSTOMER SEGMENTATION

SUBMITTED BY

Akanksha Singh 20BPS1003

Poulami Bera 20BCE1305

Mehali Samanta 20BCE1983

BACHELOR OF TECHNOLOGY

IN

School of Computer Science

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi**, School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan**, Dean, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our Head of the Department for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report entitled "Bank Customer Profiling and Segmentation" is a bona-fide work of **Akanksha Singh (20BPS1003), Poulami Bera (20BCE1305) , Mehali Samanta (20BCE1983)** carried out the "J"-Project work under my supervision and guidance for CSE3506 – Essentials of Data Analytics.

Dr. R. Rajalakshmi

SCOPE

TABLE OF CONTENTS

1. ABSTRACT

The goal of businesses is to acquire a more in-depth understanding of the customer that they are aiming for. Therefore, their goal needs to be particular, and it should be adapted so that it can meet the needs of each and every one of their customers individually. In addition, businesses can obtain a more in-depth understanding of client preferences as well as the needs for locating profitable customer segments that would bring them the greatest amount of profit by making use of the data that they have collected. They will be able to maximise the effectiveness of their marketing strategies and reduce the likelihood that their investment will be at risk as a result of this.

One of the key goals of every banking sector for a more durable existence is to achieve profitability. The customer satisfaction index identifies the consistency in the relationship between the customer and the bank and, as a result, gives the concept for developing new policies and strategies for a positive relationship between consumers and the bank. Understanding the consumer is necessary in order to provide him with services and goods based on his preferences and wants. The two primary goals of CRM (Customer Relationship Management), namely customer development and retention, can only be achieved with the help of client segmentation and profiling.The main aims of customer profiling and segmentation include expanding customer base, design of tailor made products, micro targeting of sales, aligning right channels for right products, increasing effectiveness of cross selling and up-selling, enhanced customer experience by focused customer relationship, prioritizing relationship with high value customers, effectively managing cost with low value customers based on the profiling and segmentation of customers. In this project, we will be implementing K-means and Hierarchical clustering along with PCA and also DBSCAN method for customer segmentation and profiling.

2. INTRODUCTION

Bank customer profiling and segmentation is a common data analysis task in the banking industry. It involves grouping customers with similar characteristics together for targeted marketing campaigns, personalized services, and risk management purposes.

Unsupervised learning is one of the most essential applications, and one of those applications is customer segmentation. Companies are able to identify the many subgroups of clients, which enables them to better target the possible user base. Clustering techniques are used to do this. K-means clustering, which is the most important approach for clustering unlabeled datasets, will be utilised in this particular project that we are working on. Two popular clustering algorithms for customer profiling and segmentation are k-means clustering and hierarchical clustering.

K-means clustering is a centroid-based clustering algorithm that partitions a dataset into k clusters. Each data point is iteratively assigned to the closest cluster centre by the algorithm, which then recalculates the cluster centre as the average of all the data points in the cluster. Up to convergence, this process keeps on. K-means clustering has the drawback of being sensitive to the initial selection of cluster centres, which might lead to inconsistent results.

Hierarchical clustering, on the other hand, does not require the number of clusters to be pre-specified. Instead, it creates a hierarchy of clusters by iteratively merging the two closest clusters until all data points belong to a single cluster. Hierarchical clustering can be performed using either agglomerative or divisive methods.

Both k-means clustering and hierarchical clustering can be combined with Principal Component Analysis (PCA) to reduce the dimensionality of the data. PCA is a technique for linear dimensionality reduction that transforms the original high-dimensional data into a lower-dimensional space while retaining most of the variance in the data. This can improve the clustering results by reducing the noise and highlighting the most important features.

Another clustering algorithm that we use in our project for customer profiling and segmentation is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).DBSCAN is a density-based clustering algorithm that groups together data points that are closely packed together and separates outliers as noise. It is more resistant to outliers and noise than k-means clustering or hierarchical clustering and does not require the number of groups to be pre-specified. The selection of hyperparameters, such as the minimum number of points in a cluster and the radius of the area surrounding each point, might affect how DBSCAN performs.

Thus, we implement various clustering algorithms to see their performance and predict the most efficient algorithm for customer profiling and segmentation.

3. RELATED WORK

Evolutionary multi-objective customer segmentation approach based on descriptive and predictive behaviour of customers: application to the banking sector

Citation-Ben Ncir, Chiheb-Eddine, et al. "Evolutionary multi-objective customer segmentation approach based on descriptive and predictive behaviour of customers: application to the banking sector." Journal of Experimental & Theoretical Artificial Intelligence (2022): 1-23.

Building homogeneous segments of clients based on their shared traits and behaviours is the difficult task of customer segmentation in marketing. Due to the need to evaluate a number of variables, including the descriptive and predictive qualities of customers, this challenge is regarded as multi-objective. The identification of homogeneous consumer segments in terms of both predictive and descriptive factors, however, becomes a significant problem given that the majority of current segmentation approaches are focused on the optimization of a single-objective function. Typically, descriptive and predictive traits are viewed as two distinct and independent goals that cannot be optimised combined. We provide a multi-objective segmentation strategy to address this issue that is built on the three conceptual axes of descriptive, predictive, and quality-validation. Our suggested approach has the specificity of directly optimising the multi-objective problem using a customised genetic algorithm that directly approximates a set of Pareto-optimal solutions, in addition to the specificity of the design of the multi-objective model. In an empirical study that intends to segment bank credit card consumers using their descriptive traits and their predictive behaviour, we have used and evaluated the proposed approach. Results obtained have demonstrated the capability of the suggested approach to identify useful homogenous groups and assist decision-makers in formulating more specialised marketing strategies.

Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review

CITATION- Peker, Serhat, and Özge Kart. "Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review." Journal of Data, Information and Management (2023): 1-21.

The amount of customer transactional data that has been made available to businesses has grown in recent years as digital transformation has gained momentum. Customer segmentation has drawn significant attention from various industries thanks to the use of such enormous amounts of transactional data and the application of various data mining techniques. Significant research effort has also been devoted to this topic, and the body of literature has started to grow. This paper's goal is to give a thorough assessment of the literature on transactional data-based customer segmentation in order to identify various field features, examine the use of data mining techniques, and propose key areas for more study. Three significant internet databases were used to assess the literature already published in the topic, and 84 pertinent articles from journals from reputable

publishers were ultimately chosen. The discovered articles were then thoroughly examined in accordance with the steps of the CRISP-DM (CRoss Industry Standard Process for Data Mining) framework's various criteria, and the findings were presented. By offering a thorough overview of research work on consumer segmentation using data mining and suggesting advice for future study in this field, this systematic literature review can be highly helpful for academics and practitioners.

Customer Segmentation Based on Mobile Banking User's Behaviour

CITATION: Mamashli, Zahra, and Sarfaraz Hashemkhani Zolfani. "Customer Segmentation Based on Mobile Banking User's Behavior."

This study was done to incorporate customer behaviour modelling and data mining into the RFMT model for working with mobile banking users in Iranian private banks. The public now needs the internet to conduct activities in a wide range of fields because it has grown so large. The financial or banking segment is one of them. In order to provide high-quality services, banks must guarantee client happiness. This segmentation model was expanded to include customer groupings according to transaction history, frequency, regency, amount, and time context. Customers of mobile banking are divided into six clusters. This study demonstrated how utilising behavioural ratings to identify clients makes it easier to choose a marketing plan.

CUSTOMER SEGMENTATION MODEL BASED ON RFM+B APPROACH

CITATION:Firdaus, Uus, and D. Utama. "development of bank's customer segmentation model based on rfm+ b approach." Int. J. Innov. Comput. Inf. Cont 12.1 (2021): 17-26.

Analysis of recency-frequency-monetary (RFM) is an analytical method that focuses on customer behavior. Fundamentally, R shows the last transaction, F is the number of transactions, and M represents a total amount of expenses. It has often been applied and provides an effective analysis for decision makers to promote their product strategies. However, this is considered not able to accommodate the segmentation needs of banking customers; thus customer balance should be involved in the analysis process theoretically. The customer balance (B) is potentially able to be functioned for the customer segmentation process and fruitful in marketing strategies. The developed model is called the recency-frequency-monetary-balance (RFM+B) model. It is a segmentation model of bank's customer considering four aspects: recency, frequency, monetary, and balance, where it is developed by using main method K-Means. The constructed model is applied successfully in one bank's 65 thousand customers coming from 147 thousand transaction data in period of the first half of 2017: cash payments, cash deposits, overbooking, and transactions through ATMs. The result shows that clusters 0 until 3 are dominantly filled by customers with high R (with average 113.17 times), high B (with average 3,487,790,000 Indonesian rupiah), high F (with 315.73 times), and high M (with average 5,000,000,000 Indonesian rupiah) respectively

New Methods of Customer Segmentation and Individual Credit Evaluation Based on Machine Learning

CITATION:Yuping, Zhou, et al. "New methods of customer segmentation and individual credit evaluation based on machine learning." "New Silk Road: Business Cooperation and Prospective of Economic Development"(NSRBCPED 2019). Atlantis Press, 2020.

The internet has made it possible for consumers' behaviour and perceptions of e-commerce to change fundamentally. The following article's primary goal is to describe the most recent developments in client segmentation practises related to individualised credit evaluation using machine learning. The first section addresses the existing environment and advancements in omnichannel payment methods. We discuss how society has changed as a result of the absence of physical money, how consumer purchasing behaviour has evolved, and what this transformation means for the digital economy and marketing. The traditional personal credit evaluation approach of the commercial bank faces a serious challenge in the evaluation of personal

credit against the backdrop of the quick development of big data and Internet technology. The second section explores the necessity of research on the personal credit evaluation based on the machine learning method and then delves into the comprehensive personal credit evaluation dimension and the sophisticated data acquisition method of the Internet finance company based on the limitations of the existing personal credit evaluation method. After that, the dynamic desensitisation technique was used to perform the LOF test and data desensitisation. The random forest approach and the abnormal value of the tested data fill in the gaps left by the missing value of the data. The scorecard model based on logical regression generates the personal credit evaluation score after screening the importance index using the gradient boosting decision tree approach. The model is then evaluated by the BP neural network, and the degree of personal credit is forecasted. Segmentation of the client market is encouraged by personal credit standing.

4. PROPOSED METHODOLOGY

1. Performed Data Analysis

1.1 Minimum and maximum amount in Balance Attribute-

Depending on the precise context and data being studied, a report's balance property may have a minimum or maximum value.

For a certain account or collection of accounts, the balance property typically indicates the difference between the total debits and credits. As a result, the minimum balance, which represents the account or group of accounts with the lowest aggregate balance, would be the most negative amount in the report. The account or set of accounts with the highest total balance would be represented by the maximum balance, which would be the most positive value in the report.

The customer with the largest outstanding debt, for instance, might be represented by the lowest amount in a report of customer account balances in a retail business, while the customer with the highest credit balance might be represented by the maximum balance.

You would need to construct the report and evaluate the data in order to ascertain the precise minimum and maximum balance amounts in a report.

2. Data pre processing-

Customer segmentation is a technique used by businesses to group their customers into distinct segments based on their shared characteristics such as demographics, buying behavior, interests, and needs. Data pre-processing is an essential step in customer segmentation as it helps to clean, transform, and prepare the data for analysis.

The following are the main steps involved in data pre-processing for customer segmentation:

Data Collection: The first step in data pre-processing is to collect the data that will be used for customer segmentation. The data can be collected from various sources such as customer surveys, transaction data, web analytics, social media, and other relevant sources.

Data Cleaning: The collected data may contain errors, inconsistencies, missing values, or outliers, which can affect the accuracy of the segmentation analysis. Therefore, data cleaning is done to remove or correct such data issues. Data cleaning involves tasks such as removing duplicates, filling in missing values, correcting inconsistent values, and removing outliers.

Data Tansformation: The collected data may also need to be transformed into a format that is suitable for analysis. This involves converting categorical data into numerical data, scaling the data to a common range, and normalizing the data to remove any biases.

Data Reduction: When dealing with large datasets, it may be necessary to reduce the data to a manageable size. This can be done through techniques such as sampling, feature selection, and dimensionality reduction.

Data Integration: In some cases, the data may be collected from different sources, and therefore, it needs to be integrated into a single dataset before analysis. This involves combining the data from different sources and resolving any inconsistencies in the data.

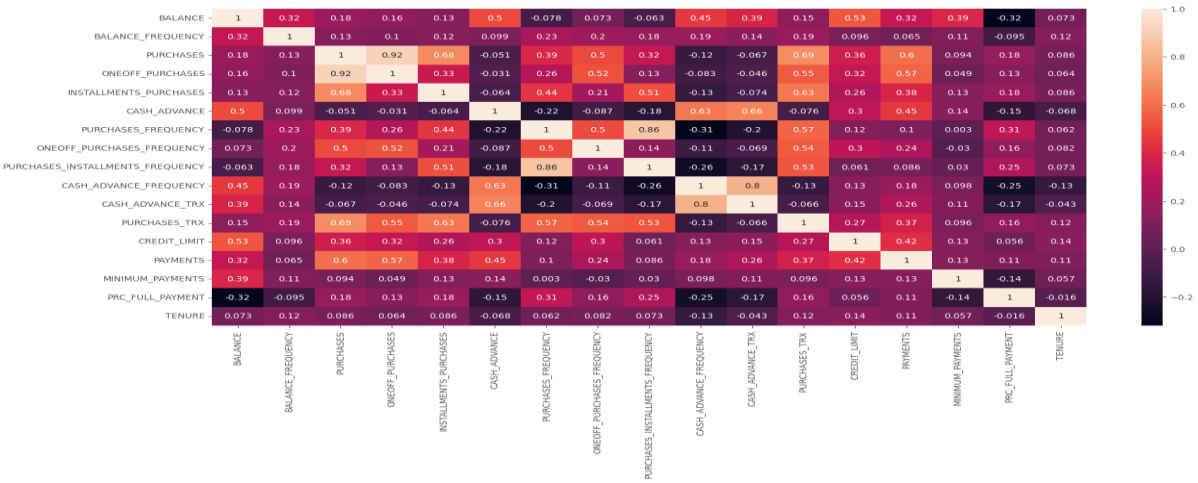
One the data has been pre-processed, it is ready for analysis, which involves applying customer segmentation techniques such as clustering, decision trees, and regression analysis to identify meaningful segments of customers. These segments can then be used to develop targeted marketing strategies, improve customer satisfaction, and increase sales.

3. Data Visualization

In customer segmentation, displot graphs are often used to identify patterns and subgroups within the customer base based on their behavior, preferences, or demographics. For example, a displot graph of customer age might reveal that the majority of customers fall within a certain age range, such as 25-40, while there are smaller subgroups of older and younger customers. This information can be used to create targeted marketing campaigns or product offerings for each group.

Overall, displot graphs are a useful tool for understanding the distribution of customer data and identifying patterns and subgroups within that data, which can inform effective customer segmentation strategies.

By using a heat map correlation graph in customer segmentation, businesses can gain insights into which customer attributes are most strongly correlated with each other. This information can help to identify groups of customers with similar characteristics, which can be used to create more targeted marketing campaigns and tailored customer experiences. Additionally, it can help businesses to identify areas of weakness in their customer segmentation strategy and make data-driven decisions to improve their approach.



4. Feature Engineering-

Customer segmentation is the process of dividing customers into smaller groups based on shared characteristics such as demographics, behavior, or preferences. One of the key steps in creating customer

segments is feature engineering, which involves selecting and transforming variables that are relevant to the segmentation task.

In customer segmentation, feature engineering typically involves identifying variables that are most useful in distinguishing between different customer groups. For example, variables such as age, income, and gender may be important in segmenting customers based on demographics, while variables such as purchase history, website activity, and social media engagement may be important in segmenting customers based on behavior.

Once relevant variables have been identified, feature engineering may involve transforming them in various ways to make them more useful for segmentation. For example, variables may be standardized or normalized to make them comparable across different scales, or they may be transformed using statistical techniques such as principal component analysis (PCA) to reduce their dimensionality.

Feature engineering is an iterative process, and the specific techniques used will depend on the data and the segmentation task at hand. Ultimately, the goal of feature engineering in customer segmentation is to identify the variables that are most informative for distinguishing between different customer groups, and to transform them in a way that maximizes their utility for segmentation.

5. Applying K-Means and Hierarchical clustering-

Customer segmentation is the process of dividing customers into different groups based on their characteristics or behavior. K-means clustering and hierarchical clustering are two popular methods for customer segmentation.

K-means clustering is a simple and efficient algorithm that groups similar customers together based on their similarities in a set of variables or features. The algorithm starts by randomly selecting k centroids, which represent the centers of each cluster. Then, each customer is assigned to the cluster whose centroid is closest to their features. The algorithm iteratively updates the centroids until the within-cluster sum of squares is minimized. The result is k clusters, where each cluster represents a group of customers with similar characteristics.

Hierarchical clustering is a method that groups customers based on their similarities and differences, building a hierarchy of clusters. The algorithm starts by treating each customer as a separate cluster and then iteratively combines the most similar clusters, creating a tree-like structure called a dendrogram. The algorithm can be either agglomerative, starting with singletons and successively merging them into larger clusters, or divisive, starting with all customers in a single cluster and successively dividing them into smaller clusters. The result is a hierarchy of clusters, where each level of the hierarchy represents a different level of aggregation, from individual customers to the entire dataset.

In customer segmentation, both methods can be used to identify customer groups with similar behavior or characteristics. K-means clustering is often used when the number of clusters is known or pre-determined, and when the data has a relatively simple structure. Hierarchical clustering is often used when the number of clusters is unknown, and when the data has a more complex structure. Ultimately, the choice between these methods depends on the specific characteristics of the data and the goals of the analysis.

6. DBSCAN Algorithm-

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a popular clustering algorithm used for customer segmentation. It works by grouping data points based on their density in a given space.

The DBSCAN algorithm takes two parameters: epsilon (ε), which represents the radius of a data point's neighborhood, and minPts, which represents the minimum number of data points required to form a dense region.

The algorithm begins by randomly selecting a data point and checking if it has enough neighboring points within a distance ε to form a dense region. If the point does not have enough neighbors, it is labeled as noise and excluded from any cluster. If the point has enough neighbors, it is assigned to a new cluster, and all of its neighbors are checked for additional points to add to the cluster. This process continues until all points have been assigned to a cluster or labeled as noise.

The result of DBSCAN is a set of clusters and a set of noise points. Clusters are formed by points that have a sufficient number of neighboring points within a distance ε, while noise points are those that do not belong to any cluster.

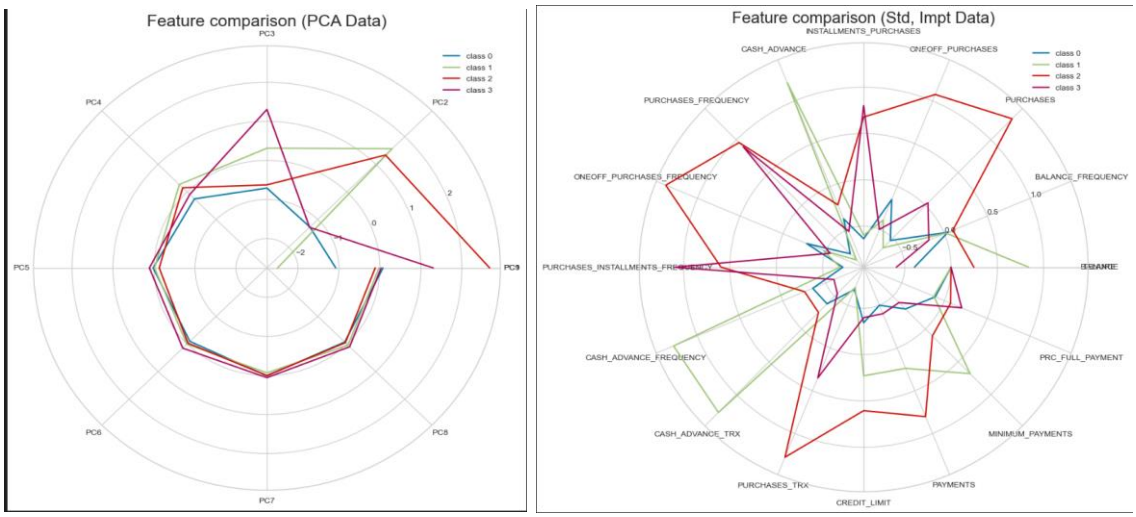
In customer segmentation, DBSCAN can be used to group customers based on their purchasing behavior, such as frequency of purchases and total amount spent. By clustering customers with similar buying patterns, companies can tailor their marketing strategies and improve customer satisfaction.

5. EXPERIMENT AND RESULTS

The results of the clustering algorithms applied to the bank customer dataset using both normal dataset and PCA implemented dataset were evaluated using two metrics, the Davis-Bouldin Index and the Silhouette Coefficient.

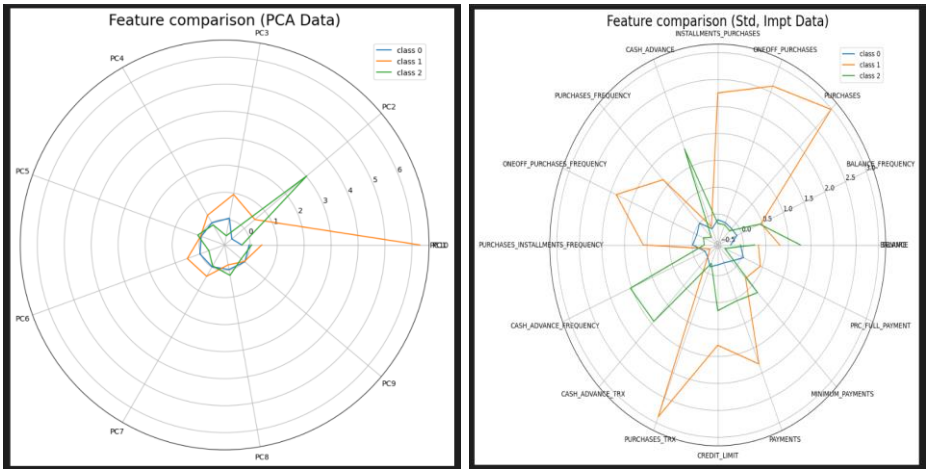
The Davis-Bouldin Index measures the average similarity between each cluster and the closest cluster. Lower values of the index indicate better clustering. The Silhouette Coefficient measures the tightness and separation of the clusters. Higher values of the coefficient indicate better clustering.

For K-means clustering, 4 clusters were formed using the normal dataset, with a Davis-Bouldin Index of 1.687 and a Silhouette Coefficient of 0.174. On the other hand, when PCA was applied to the dataset, 4 clusters were formed with a lower Davis-Bouldin Index of 1.492 and a higher Silhouette Coefficient of 0.257. Therefore, the clustering performance was improved by using PCA.



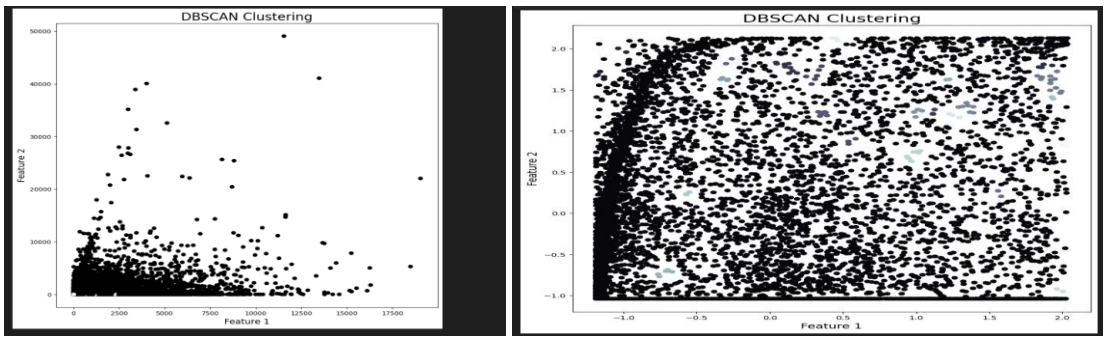
K-means Clustering for PCA and Normal dataset

For hierarchical clustering, 5 clusters were formed using the normal dataset, with a Davis-Bouldin Index of 1.538 and a Silhouette Coefficient of 0.177. When PCA was applied, 3 clusters were formed with a lower Davis-Bouldin Index of 1.384 and a higher Silhouette Coefficient of 0.317. Therefore, PCA improved the clustering performance.



Hierarchical Clustering for PCA and Normal dataset

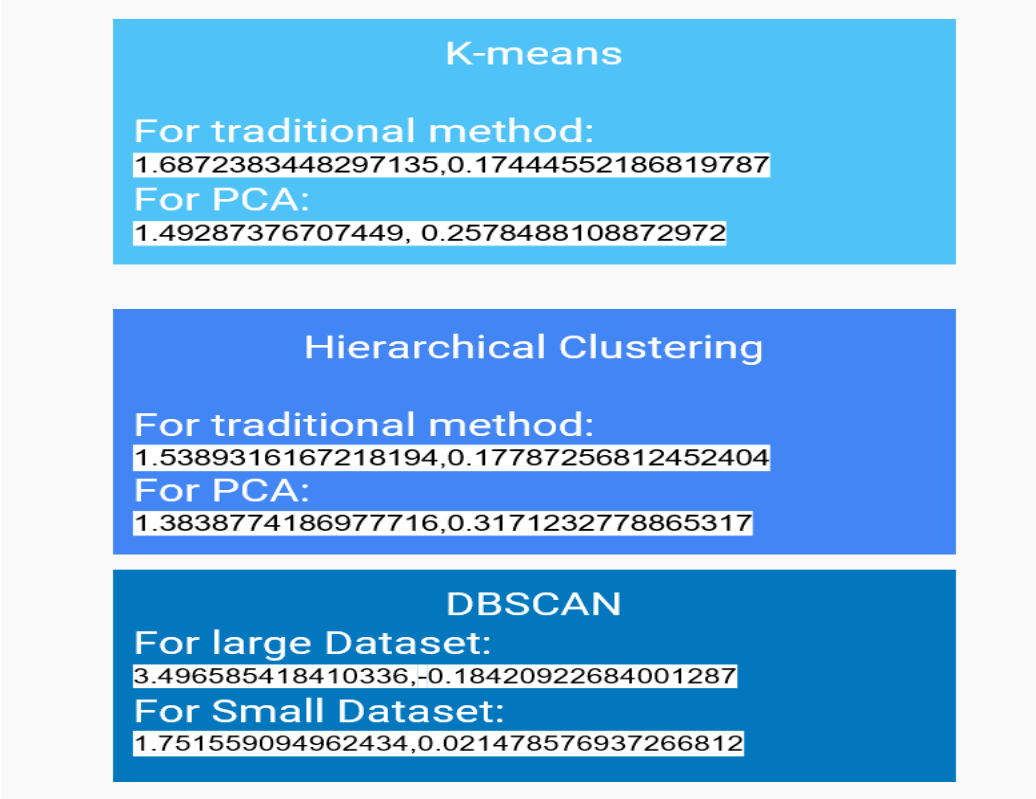
For DBSCAN, using the normal dataset, the Davis-Bouldin Index was 3.497 and the Silhouette Coefficient was -0.184. When PCA was applied, the Davis-Bouldin Index decreased to 1.752, and the Silhouette Coefficient increased to 0.021. Therefore, PCA improved the clustering performance for DBSCAN as well.



DBSCAN clustering for large and small dataset

In conclusion, the results show that PCA implementation improved the clustering performance of all three algorithms on the bank customer dataset. The Davis-Bouldin Index decreased, indicating better cluster separation, and the Silhouette Coefficient increased, indicating tighter clusters.

6. PERFORMANCE ANALYSIS



These Score are the Davis-Bouldin Index and Silhouette Coefficient for each of the model. Clearly from the scores its visible that implementing PCA has improved the model accuracy to cluster the data in the proper way and give better results . Also for our dataset we can see DBSCAN performs better on small dataset i.e when the dataset is broken down into training and testing dataset and forms better cluster as per Davis Boulding Index.

7. CONCLUSION

The conclusion of this study is that the use of Principal Component Analysis (PCA) improved the clustering performance of K-means, hierarchical clustering, and DBSCAN algorithms on the bank customer dataset. The results of the evaluation metrics, Davis-Bouldin Index and Silhouette Coefficient, showed that PCA implementation resulted in better cluster separation and tighter clusters compared to the clustering results obtained from using the normal dataset.

PCA is a powerful technique for reducing the dimensionality of high-dimensional datasets. It transforms the data into a lower-dimensional space while preserving the most important information in the data. In this study, PCA reduced the number of features in the bank customer dataset and eliminated the redundancy and noise in the data, which resulted in improved clustering performance.

The improved clustering performance has important practical implications in the banking industry, where customer segmentation is a critical task for developing targeted marketing strategies, identifying potential customers for new products, and detecting fraudulent activities. Accurate clustering of bank customers helps in identifying the customers' needs, preferences, and behaviors, which can be used for improving customer satisfaction and loyalty.

Overall, the use of PCA in clustering algorithms can be a useful technique for improving clustering performance on high-dimensional datasets, and can be applied in various industries for customer segmentation, marketing, and fraud detection purposes.

8. REFERENCE

- Ben Ncir, Chiheb-Eddine, et al. "Evolutionary multi-objective customer segmentation approach based on descriptive and predictive behaviour of customers: application to the banking sector." Journal of Experimental & Theoretical Artificial Intelligence (2022): 1-23.
- Mamashli, Zahra, and Sarfaraz Hashemkhani Zolfani. "Customer Segmentation Based on Mobile Banking User's Behavior."
- Firdaus, Uus, and D. Utama. "development of bank's customer segmentation model based on rfm+ b approach." Int. J. Innov. Comput. Inf. Cont 12.1 (2021): 17-26.
- Yuping, Zhou, et al. "New methods of customer segmentation and individual credit evaluation based on machine learning." "New Silk Road: Business Cooperation and Prospective of Economic Development"(NSRBCPED 2019). Atlantis Press, 2020.