Bank Customer Segmentation

Sparsh Jain

***Abstract*— This research uses K-means clustering to give a thorough study of bank customer segmentation. The dataset includes a variety of bank client financial characteristics, such as account balances, purchase histories, and credit card usage trends. The first steps of the study are data exploration and cleaning. Appropriate methods, including imputation and isolation forest, are used to address missing values and outliers. Investigative Data Analysis (EDA) is used to visualise distributions, correlations, and trends in order to obtain understanding of the connections between various variables. In addition, dimensionality reduction is accomplished by the use of Principal Component Analysis (PCA), which permits data visualisation in a lower-dimensional space. The data is then compressed into a lower-dimensional representation using a basic autoencoder neural network, which is used for additional dimensionality reduction.Using the reduced-dimensional representation derived from PCA and autoencoder, K-means clustering is then used to divide the clientele into groups according to their financial behaviour. After being visualised and analysed, the clustering results offer important insights into various client categories and their traits. The bank can make better decisions thanks to this analysis, which also helps with customised client care and targeted marketing tactics.**

Keywords: **PCA,K-means,Autoencoder,Heirarichical Clustering,Elbow Plot,Dendogram**

link of Overleaf project: `https://www.overleaf.com/project/6601a4d401c3c558ed501031`
link of Colab Notebook: `https://colab.research.google.com/drive/1Y8oJNZynU9DWkSjKACghgfbs3RMkTLSM?usp=sharing`

I. INTRODUCTION

Banks today rely heavily on advanced analytics and customer segmentation to understand their diverse clientele. By categorizing customers based on demographics, behavior, and preferences, banks can tailor their offerings to meet specific needs. This personalized approach enhances customer satisfaction, loyalty, and profitability. Segmentation allows banks to move away from a one-size-fits-all model, offering targeted services that improve overall performance. In a competitive market, effective segmentation provides insights into customer behavior, enabling banks to optimize resources and mitigate risks. Overall, customer segmentation is essential for banks to thrive by enhancing relationships, driving revenue, and maintaining a competitive edge.

K-means clustering is one such method that has gained a lot of interest because to its capacity to divide data into coherent clusters, each of which represents a different client group. Through the utilisation of K-means clustering on extensive datasets that comprise various financial parameters, financial institutions can obtain significant insights into the spending habits, risk assessments, and preferences of their clientele. Equipped with this understanding, financial institutions may craft focused promotional strategies, customise product lines, and enhance customer support to cater to the distinct requirements of every market niche.
Using a variety of financial attribute datasets, our project seeks to investigate the use of K-means clustering in bank customer segmentation in this particular setting. In an ever-changing financial market, we aim to provide banks with actionable insights that drive strategic decision-making and strengthen customer relationships by identifying the complex patterns hidden inside client data.

II. DATASET

The dataset used in the project offers a comprehensive glimpse into the behaviors and transactional patterns of bank customers. Comprising *18 columns and 8950 Entries* of Customers, it serves as a rich repository of data for conducting thorough analysis and segmentation. Each column provides specific insights into various aspects of customer behavior, ranging from account balances to purchase frequencies and credit limits.

Column Details:

BALANCE: This column indicates the total account balance of individual customers, revealing insights into their financial standing and liquidity.
BALANCE FREQUENCY: Reflects how often customers update their account balances, indicating their engagement with their finances.
PURCHASES: Represents the total amount spent by customers on purchases, shedding light on their spending habits and preferences.
ONEOFF PURCHASES: Specifies the amount spent by customers on singular, one-time purchases, showcasing their inclination towards single-payment transactions.
INSTALLMENT PURCHASES: Denotes the amount spent by customers on purchases paid in installments, indicating their preference for staggered payment options.
CASH ADVANCE: Indicates the total cash advances taken by customers, revealing their usage of credit facilities for immediate cash needs.
PURCHASES FREQUENCY: Reflects the frequency of purchase transactions made by customers, offering insights into their purchasing behavior.
ONEOFF PURCHASES FREQUENCY: Specifies the frequency of one-time purchase transactions, indicating customers' tendency for occasional significant expenditures.
PURCHASES INSTALLMENTS FREQUENCY: Denotes the frequency of installment purchase transactions,

showcasing customers' preference for spreading payments over time.
CASH ADVANCE FREQUENCY: It indicates how often customers use their credit cards for cash transactions, revealing their dependence on credit for obtaining cash.
CASH ADVANCE TRX:Specifies the number of transactions for cash advances, offering a quantitative measure of customers' utilization of credit facilities.
PURCHASES TRX: Denotes the number of purchase transactions, providing insights into the frequency and volume of customers' purchasing activities.
CREDIT LIMIT: Indicates the credit limit assigned to each customer, reflecting their creditworthiness and borrowing capacity.
PAYMENTS: Represents the total payments made by customers, offering insights into their repayment behavior and financial discipline.
MINIMUM PAYMENTS: Specifies the minimum payments made by customers, indicating their compliance with payment obligations.
PRC FULL PAYMENT: Reflects the percentage of full payments made by customers, offering insights into their repayment behavior and credit utilization patterns.
TENURE: Denotes the number of months a customer has been associated with the bank, providing insights into customer loyalty and long-term relationships.

III. METHODOLOGY

*Objective:*The main aim of the "Bank Customer Segmentation" project is to use smart ways like PCA ,etc and grouping to sort out bank customers based on how they handle their money. This helps in making the data simpler, finding important patterns, and then splitting customers into different groups. The end goal is to understand and separate different types of customers, so the bank can create special plans and services just for them. This way, the bank can make its customers happier and more engaged by offering things that suit them better

Methodology compromises of the following steps:

1]Fundamental Data Pre-processing
2]Advanced Data Pre-processing
3]Clustering Methods

A. Fundamental Data Pre-processing

This is the first and essential step in getting raw data ready for analysis and modelling is data preparation. To prepare data for more investigation and analysis, it undergoes cleaning, converting, and organising it. Some fundamental data preparation methods employed in this project are listed below:

a.*Dealing with Missing Values*:
Data points that are absent or undefined in the dataset are referred to as missing values.
Effective methods for handling missing data include imputation, which replaces missing values with a statistical measure like mean, median, or mode, and removal, which gets rid of rows or columns that have missing values.In our project, missing values are identified using the *isnull()* function from the Pandas library.There were 313 values missing in 'MINIMUM_PAYMENTS' and 1 value missing in 'CREDIT_LIMIT' Column. Then, various strategies are employed to handle them:
For the 'MINIMUM_PAYMENTS' column, missing values are filled using the minimum value in the column with fillna() function since the values are left-skewed.
For the 'CREDIT_LIMIT' column, rows with missing values are dropped using dropna() function since only 1 value was missing in the dataset.

b.*Normalisation of Data:*

Data normalization involves scaling numerical feature values to a standard range, typically between 0 and 1, or with a mean of 0 and a standard deviation of 1. This process helps prevent certain features from dominating others and ensures that all features contribute equally to the analysis. In the dataset, z-score normalization, implemented using the StandardScaler class from the scikit-learn library, is applied. This technique scales numerical features to have a mean of 0 and a standard deviation of 1, effectively standardizing the data and facilitating fair comparison between different features.

B. Advanced Data Pre-processing

a.*Correlational Analysis*:
To determine the direction and degree of a relationship between two variables, correlation analysis is utilised. Correlation analysis is useful in determining which features in a dataset are highly connected to one another when it comes to data preparation. Comprehending these relationships plays a crucial role in feature selection, redundant feature identification.
Correlation analysis is carried out in the project using methods like *correlation heatmaps* and the *Pearson correlation coefficient*. The Pearson correlation coefficient quantifies the linear relationship between two continuous variables, ranging from -1 to 1. Coefficients close to 1 indicate strong positive correlations, while coefficients near -1 suggest strong negative correlations. A coefficient close to 0 indicates no linear association between the variables. To visualize these relationships, a correlation heatmap is generated from the correlation matrix, where colors represent the direction and strength of correlations between pairs of features.

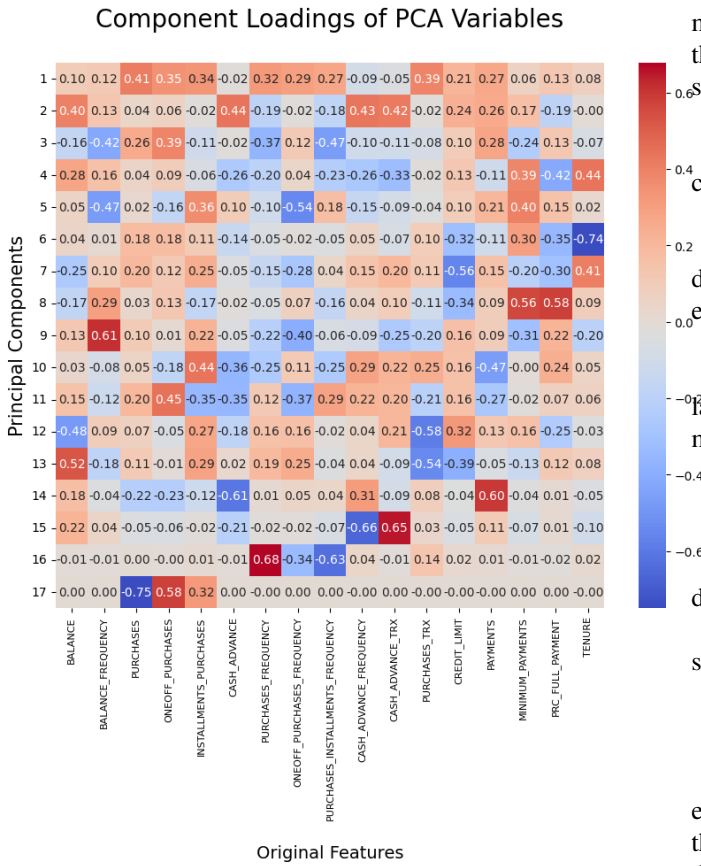

Fig. 1: Components Loading of PCA Variables

b.*PCA Analysis*:

A dimensionality reduction method called principal component analysis (PCA) converts high-dimensional data into a lower-dimensional representation while keeping the most crucial information. The data is projected onto the directions, or main components, that PCA determines capture the most variance in the data. PCA streamlines the dataset by minimising its number of dimensions, so preserving as much variance as feasible and facilitating analysis and visualisation.

Feature Extraction with PCA:
A dimensionality reduction method called principal component analysis (PCA) is used to convert high-dimensional data into a lower-dimensional space while keeping the majority of the data's variation. Principal components are linear combinations of the basic features that PCA finds and ranks according to how much variance they explain. When working with datasets that have a large number of features, PCA is very helpful for clustering because it can lower computing cost and enhance clustering performance.

The following are the steps in PCA:
1.Standardisation: To guarantee that every feature has a mean of zero and a standard deviation of one, standardise the features by removing the mean and dividing by the standard deviation.

2.Covariance Matrix: Determine the standardised data's covariance matrix.

3.Eigenvalue Decomposition:Conduct eigenvalue decomposition on the covariance matrix to derive the eigenvectors and eigenvalues..

The mathematical formulation of PCA involves the calculation of eigenvectors and eigenvalues from the covariance matrix:

$$\Sigma = \frac{1}{n-1} X^T X$$

where Σ is the covariance matrix, X is the standardized data matrix, and n is the number of data points.

The eigenvectors (v_i) and eigenvalues (λ_i) are obtained by solving the eigenvalue equation:

$$\Sigma v_i = \lambda_i v_i$$

4.Principal Components: Select the top $k = 2$ eigenvectors (principal components) corresponding to the largest eigenvalues(pca1, pca2) where k is the desired dimensionality of the reduced space.

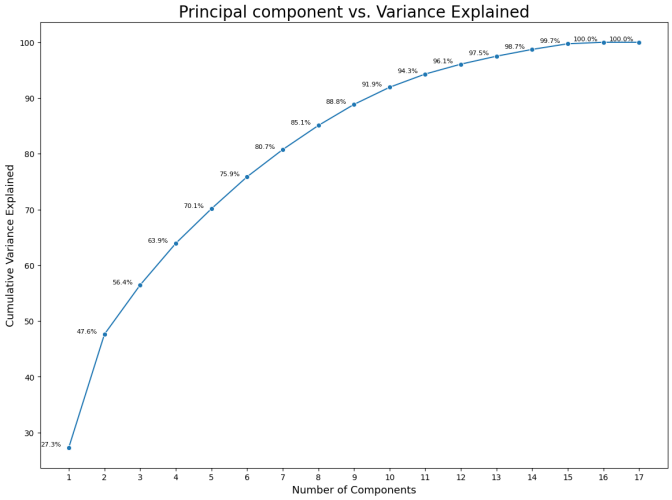

Fig. 2: Principal Component vs Variance Explained

5.Projection: Project the original data onto the selected principal components to obtain the reduced-dimensional representation.

c .Dimensionality reduction using autoencoder

Neural network architectures called autoencoders are specifically made for unsupervised learning applications, like

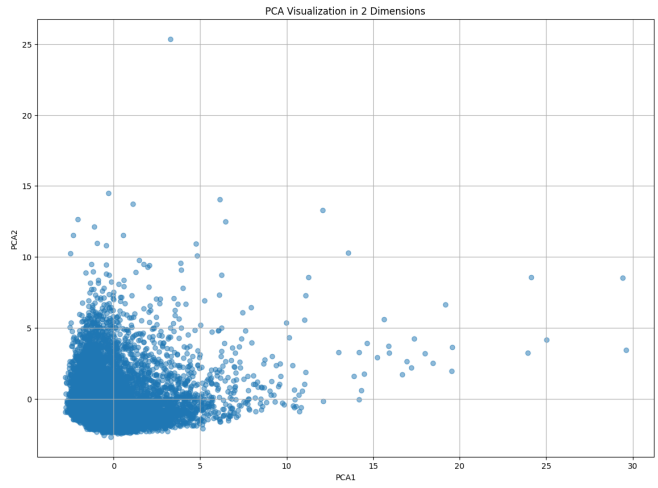

Fig. 3: PCA Visualization In 2D

data compression and dimensionality reduction. A decoder network reconstructs the original input data from the latent space representation, whereas an encoder network maps the input data to a lower-dimensional latent space representation. The autoencoder develops a compressed representation of the input data that captures the most significant features by training it to minimise the reconstruction error.

Autoencoder Architecture:
An encoder and a decoder are the usual components of an autoencoder neural network architecture.
The input data is compressed into a lower-dimensional representation by the encoder network.
The compressed form is used by the decoder network to reassemble the original input data.
ReLU and other activation functions are frequently employed in the encoder and decoder's hidden layers.
In the project, the following steps are involved in implementing dimensionality reduction using autoencoder:
Training:
To reduce the reconstruction error, the autoencoder is trained using the preprocessed data.
Using optimisation methods like Adam or stochastic gradient descent (SGD), the network parameters (weights and biases) are optimised throughout the training phase.
The loss function is utilised to quantify the disparity between the input and reconstructed data, such as mean squared error (MSE).
Dimensionality Reduction:
The encoder part of the trained autoencoder is then utilized to transform new input data into this reduced representation, enabling efficient dimensionality reduction. This reduced-dimensional representation, with dimensions reduced from 18 to 10 , retains essential information while discarding redundant or less important features

C. Clustering

Unsupervised machine learning relies heavily on clustering, a basic approach for finding patterns and putting related data points in one group. In order to understand the underlying structure of the data, we analyse a dataset using a variety of clustering algorithms in this study. We examine the data, use PCA to reduce dimensionality, and employ a variety of clustering strategies, including Agglomerative Clustering, Hierarchical Clustering, and K-Means. An outline of each clustering methodology is included in the study, along with a discussion of its applications and an interpretation of the outcomes of each method.

K-Means Clustering:
One of the most commonly employed clustering methods is K-Means. It partitions the data into K clusters by minimizing the sum of squared distances between data points and the centroids (mean points) of each cluster. The algorithm follows these steps:

1. Initialization: Set K centroids in the feature space using the Elbow plot.

2. Assignment: Using the Euclidean distance as a guide, assign each data point to the closest centroid.
3. Update: Determine the centroids once more using each cluster's mean data point.
4. Iteration: Until convergence (when centroids no longer move appreciably or a predetermined number of iterations is reached), repeat steps 2 and 3 again.

The goal of the K-Means algorithm can be expressed through the following objective function:

$$\underset{C}{\text{minimize}} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Here, C represents the set of clusters, C_i signifies the ith cluster, μ_i denotes the centroid of cluster C_i, and $\|\cdot\|$ represents the Euclidean distance. The objective is to minimize the sum of squared distances between each data point x and its respective centroid μ_i within their assigned clusters.
Results:

The comparison graph above illustrates significant variations in trends. It suggests that the optimal number of clusters might be 2, 4, or 5 based on the elbow method. These points represent where the silhouette scores or the cost, as indicated by the elbow plot, show distinct changes in trend, indicating potentially suitable cluster numbers.Here cost and scores 2 are inertia.(sum of squared distances of samples to their closest cluster center.)
Elbow plot:

So,choosing No. of clusters=4

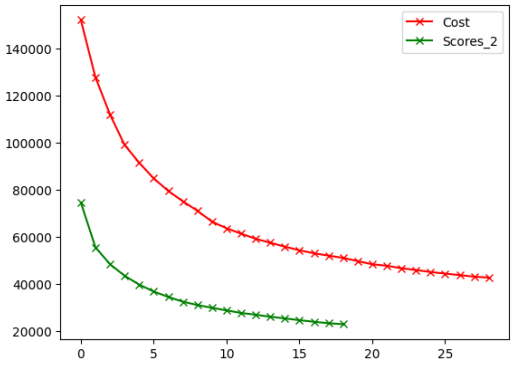

Fig. 4: Elbow Plot

K-means Clustering on PCA :

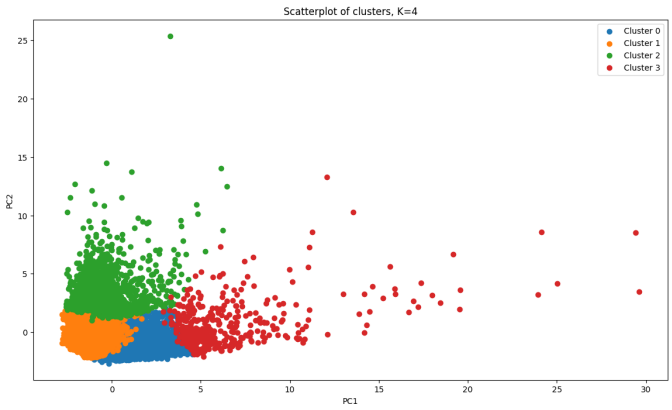

Fig. 5: K-means Clustering On PCA

4 Clusters of Customers:

There are four distinct clusters of customers based on their credit card usage patterns:

1.Transactors: These customers are cautious with their spending and typically pay off their balances in full. They have the lowest balances and cash advances, with only 23% of them paying their full balance each month.

2. Revolvers: This group utilizes their credit cards as a form of loan, maintaining high balances and frequently taking cash advances. Despite their high cash advance frequency and transactions, they have a low percentage of full payments, indicating they often carry over balances.

3.VIP/Prime: Customers in this category have high credit limits and consistently pay off their balances in full. They are prime targets for increasing their credit limits and encouraging higher spending.

4.Low Tenure: These customers have relatively short tenures with the credit card company and maintain low balances. They represent a distinct segment with potential for targeted marketing or retention strategies.

K- Means Clustering on Dimensionally Reduced Data using Autoencoder(K=4)

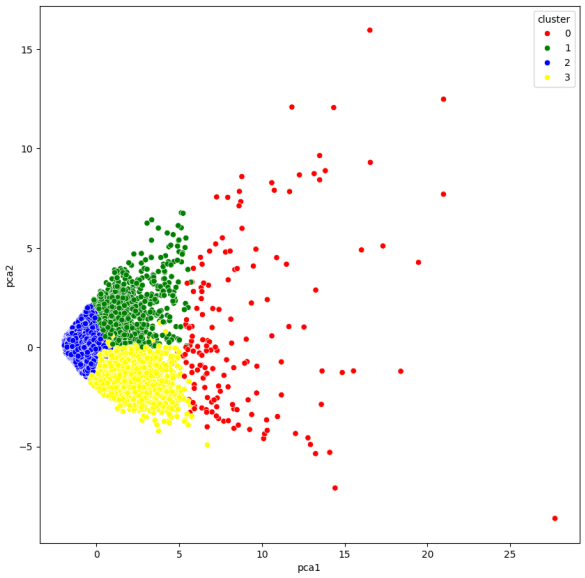

Fig. 6: K-Means on Dimensionally Reduced Data

Clearly better clustering is performed than the previous one.

Hierarchical Clustering:

Hierarchical clustering is a bottom-up approach that builds a hierarchy of Clusters by the iterative joining or dividing of clusters according to a distance measure. The technique produces a structure that resembles a tree and is known as a dendrogram; it does not require a set number of clusters. Two primary varieties of hierarchical clustering exist but in this project we have used only one type of it.

Agglomeration Clustering:
It involves merging the closest pairs of clusters iteratively until only one cluster remains. Initially, each data point is treated as a singleton cluster. Various linkage techniques, such as single, complete, and average linkage, can be employed to compute the distance between clusters.

Agglomerative hierarchical clustering begins by treating each data point as a separate cluster and progressively merges the most similar clusters until a stopping criterion is reached. The similarity between clusters is determined using a linkage method, which specifies how the distance between clusters is calculated.Linkage methods used are :

Single Linkage: Single linkage clustering defines the distance between two clusters as the shortest path between any two points within the clusters.

Complete Linkage: In complete linkage clustering, the distance between two clusters is determined by the maximum distance between any two points in each cluster.

Ward Linkage: Ward linkage clustering calculates the distance between two clusters based on the increase in the total squared error when the clusters are merged.

Steps in Implementation:
Initialisation: Bring in the required libraries and packages, such as scikit-learn, pandas, and numpy.
Preprocessing the Dataset: Load and prepare the dataset. The data may be subjected to normalisation or standardisation.
The Agglomerative Clustering algorithm can be used to do hierarchical clustering with certain parameters, such as the number of clusters and linking method.
Visualisation: To comprehend the cluster formation and hierarchy, visualise the dendrogram.
Interpretation: To comprehend the connections between data points and clusters, interpret the dendrogram and the clustering results.

Single Linkage: The shortest path between locations in various clusters is used to build clusters. It is prone to extended cluster formation and outlier sensitivity.

	Cluster Label	No of Observation
0	0	8946
3	1	1
1	2	1
2	3	1

Fig. 7: Single Linkage

Complete Linkage: The greatest distance between points in various clusters is used to create clusters. It is less susceptible to outliers and has a tendency to create compact clusters.

Ward Linkage: When combining clusters, the goal is to reduce variation. It usually results in clusters that are comparable in size and form.

In short in a down to up method called hierarchical clustering, each data point begins in its own cluster, and clusters are progressively combined according to distance-based standards.

Cluster Label	No of Observation
0	8918
1	23
2	1
3	7

Fig. 8: Complete Linkage

Cluster Label	No of Observation
0	2858
1	4668
2	1400
3	23

Fig. 9: Ward Linkage

The interpretation of results and the cluster structure are influenced by the linkage method selection.
Understanding the clusters' hierarchical structure is aided by the visualisation of dendrograms.

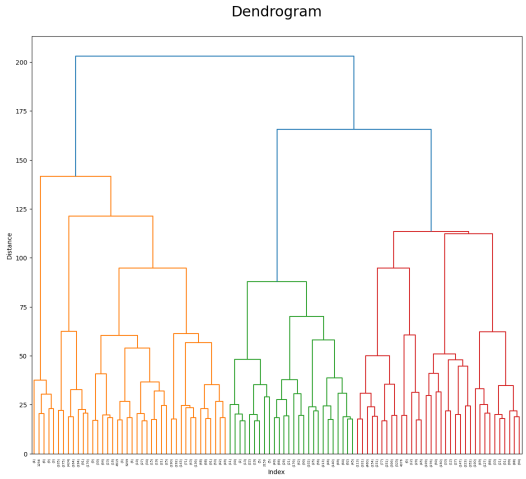

Fig. 10: Dendogram of Ward Linkage

The implementation makes it possible to investigate several hierarchical clustering models and comprehend how they affect the dataset.

Clustering Method	Silhouette score
K-means on PCA	0.31453
K-means on Dimensionality reduced data	0.49238
Single Linkage Hierarchical clustering	0.81210
Complete Linkage Hierarchical clustering	0.77900
Ward Linkage Hierarchical clustering	0.093760

TABLE I: Silhouette Scores

IV. Evaluation Metrics:

The silhouette score serves as a widely adopted metric for evaluating clustering algorithms. It gauges both the distance between clusters and the compactness within clusters, offering a numerical assessment of cluster quality. Elevated silhouette scores signify improved cluster separation and compactness, with a maximum score of 1 indicating dense, well-separated clusters. Conversely, scores nearing 0 imply either overlapping or poorly separated clusters.

Here's how the silhouette score is calculated for each sample in the dataset:

1. *Calculate the Cluster Separation (a):*
For each sample, compute the average distance between that sample and all other points in the same cluster. This represents how well-separated the sample is from other clusters.

2. *Calculate the Cohesion Within the Cluster (b)*:
For each sample, compute the average distance between that sample and all other points in the nearest neighboring cluster (the cluster to which it does not belong). This represents how well the sample is assigned to its own cluster.

3. Calculate the Silhouette Score (s):
The silhouette score (s) for each sample is given by:
$s = \frac{b-a}{max(a,b)}$

The silhouette score ranges from -1 to 1:

- A score close to 1 indicates that the sample is well-clustered and appropriately assigned to its cluster.
- A score close to 0 indicates that the sample lies close to the decision boundary between two clusters.
- A score close to -1 indicates that the sample may have been assigned to the wrong cluster

4. *Calculate the Average Silhouette Score:*
Finally, the average silhouette score is computed over all samples in the dataset to obtain a single metric representing the overall quality of the clustering.

Application in Bank Customer Segmentation:

Clustering techniques like K-Means and hierarchical clustering are employed in the context of bank customer segmentation to group clients based on their credit history, spending patterns, demographics, and other pertinent data. Banks can customise their marketing tactics, product offerings, and customer service to each segment's unique needs and preferences by grouping their consumers into homogeneous groups.

1.*Targeted Marketing:* To promote pertinent goods and services like credit cards, loans, or investment accounts, banks can identify high-value client segments and create tailored marketing campaigns.

2.*Risk management:* Banks can apply suitable risk management techniques, such establishing credit limits, interest rates, and loan terms, by using clustering to evaluate the credit risk connected to various client segments.

3.*Customer Retention:* By understanding the characteristics and preferences of different customer segments, banks can develop personalized retention strategies to improve customer loyalty and reduce churn rates.

A. *Conclusion*

Customer segmentation is a crucial component of any business strategy, particularly within the banking sector where tailored services are essential for retaining and satisfying clients. To effectively categorize bank customers according to their financial behaviors and characteristics, our project employed a range of clustering methodologies, including K-means clustering, hierarchical clustering, and dimensionality reduction techniques such as Principal Component Analysis (PCA) and autoencoder neural networks.

General Insights:

1. The significance of customer segmentation is in its ability to help banks divide their clientele into discrete categories according to their needs, tastes, and financial habits. Banks may enhance customer satisfaction and loyalty by customising their offerings to better cater to the unique needs of each group by comprehending these categories and developing tailored products, services, and marketing tactics.

2. *Improved Personalisation:* Banks can offer tailored discounts, product recommendations, and communication channels based on individual interests by segmenting their customer base. Customers' whole banking experience is improved and stronger relationships are fostered by this personalised approach.

3. *Operational Efficiency*: Banks can optimise resource allocation and improve operational operations by identifying discrete consumer categories. This involves the cost-effective deployment of resources for risk management,

product development, and customer service.

4. *Risk Management:* Banks can more effectively evaluate and manage any risks connected to various customer groups by segmenting their clientele based on risk profiles. Banks can protect their financial interests by taking proactive steps to limit risks, such as credit default or fraudulent operations, by identifying high-risk areas.

Technical Insights:

1. Clustering Techniques: Based on similarities in their financial attributes, we divided our consumers into different groups using a variety of clustering algorithms, such as K-means clustering and hierarchical clustering. Hierarchical clustering creates a dendrogram to show the hierarchical relationships between data points, whereas K-means clustering is an iterative method that divides data into K groups by minimising the within-cluster variation.

2. Dimensionality Reduction: To reduce the dimensionality of the data while maintaining critical information, dimensionality reduction techniques such as PCA and autoencoder neural networks were used. Complex datasets can be visualised and interpreted with the use of principal component analysis (PCA), which finds the main components that capture the most variance in the data. With the ability to learn compressed representations of input data, autoencoder neural networks enable high-dimensional feature encoding and decoding with efficiency.

3. Evaluation Metrics: To ascertain the ideal number of clusters and rate the calibre of clustering outcomes, evaluation metrics including the elbow method and silhouette score were applied. The silhouette score measures the cohesion and separation of clusters, whereas the elbow approach assists in determining the point of decreasing returns in terms of clustering performance.

4. Interpretation and Visualisation: To comprehend client segments and derive useful insights, interpretation and visualisation of clustering results are crucial. To visualise clusters, find patterns, and investigate interactions between variables, tools like scatter plots, dendrograms, and heatmaps were used. This allowed for the creation of strategies and the making of well-informed decisions.

To sum up, customer segmentation is a complex process that combines data analytics, technology innovation, and subject expertise to reveal important information about the behaviour and preferences of customers. In an increasingly competitive market scenario, banks may successfully personalise their services, promote customer engagement, and ultimately achieve sustainable development by implementing a comprehensive approach that includes both general and technical components of client segmentation.

B. Future Directions

Incorporating External Data: To improve consumer segmentation models and prediction capacities, future research should investigate the integration of external data sources, such as sociodemographic data, transactional data from other industries, and customer feedback.

Dynamic Segmentation: Banks may gain real-time insights and be able to implement proactive engagement strategies catered to the specific requirements and preferences of each individual client by implementing dynamic segmentation techniques that adjust to changing market trends and customer behaviours.

3Advanced Machine Learning Techniques: Banks may be able to find hidden patterns and more accurately forecast future customer behaviour by utilising advanced machine learning techniques like deep learning and reinforcement learning. These techniques could also improve the accuracy and robustness of customer segmentation models.

C. Refrences

link of Research paper 1 : **https://www.researchgate.net/publication/356756320_Customer_Segmentation_Using_Machine_Learning**

link of Research paper 2 : **https://www.researchgate.net/publication/328533235_A_customer_segmentation_approach_in_commercial_banks**

APPENDIX

Add the first page of Plagiarism Report here, after I provide the report to you (must have less than 15% similarity). Each team member should add their signature on the report page

project_report .pdf

ORIGINALITY REPORT

7%	4%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Aston University Student Paper	1%
2	Submitted to Imperial College of Science, Technology and Medicine Student Paper	1%
3	www.dellemc.com Internet Source	1%
4	docs.oracle.com Internet Source	1%
5	Submitted to CSU, San Jose State University Student Paper	1%
6	Submitted to UNITEC Institute of Technology Student Paper	1%
7	docplayer.net Internet Source	<1%
8	www.i-scholar.in Internet Source	<1%