Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 5s (2024)

A Study on Customer Segmentation for Banking Sector Through Cluster Analysis: Ethical Implications

Dr. L. Kuladeep Kumar[1], Dr. D.Venkatesh[2], Dr.J. Katyayani[3], Dr.Sreenivasulu Sunkara[4], Dr. Gowthami. K[5]

[1]Associate Professor, School of Commerce and Management, Mohan Babu University, kuladeep79@gmail.com

[2]Assistant Professor, School of Commerce and Management, Mohan Babu University, drduvvuri1112@gmail.com

[3]B.Tech., MBA.,Ph.D., M.Tech, Professor, Department of Business Management, Sri Padmavathi Mahila Vishwavidyalaya-Tirupati-AP, India, Email id: jkatyayani@gmail.com

[4]Assistant Professor, Dept. of Master of Business Administration, The Oxford College of Engineering, Bangalore

[5]Technical Facility Manager, KJR Bio-Science Pvt. Ltd, Tirupati

Article History:

Received: *29-04-2024*

Revised: *10-06-2024*

Accepted: *28-06-2024*

Abstract:

The banking industry, inherently customer-focused, relies on understanding and fulfilling the diverse needs of its clientele for success. Customization of offerings is crucial for banks serving individuals, families, or businesses across various financial stages. Client segmentation stands out as a primary strategy in achieving this customization. By categorizing customers based on shared characteristics, banks can deploy targeted marketing efforts, allocate resources efficiently, and deliver tailored banking experiences. In the contemporary banking landscape, where vast amounts of data are generated daily, thorough analysis is indispensable. Customized business strategies have become increasingly vital amidst intensifying industry competition. Customer segmentation serves as a pivotal aspect of market research, facilitating the grouping of customers based on common characteristics and behaviors. This segmentation enables banks to tailor marketing campaigns to suit the distinct requirements and preferences of each segment. This study focuses on employing cluster analysis, a statistical technique for organizing data points, to achieve efficient client segmentation in the banking sector. Specifically, we utilize the K-means algorithm, a popular clustering method, to categorize clientele into discrete groups based on transaction history, banking preferences, and demographic data. To ensure the accuracy and robustness of our segmentation methodology, sophisticated machine learning techniques like the Elbow and Silhouette methods are employed. These techniques enable the evaluation of clustering effectiveness and determination of the optimal number of clusters. Our objective is to utilize machine learning to identify meaningful and actionable customer groups, guiding banks' strategic decision-making processes. The segmentation approach outlined in this study empowers banks to optimize services and elevate customer satisfaction levels. By aligning offerings with the specific needs of each segment, banks can cultivate stronger customer relationships, drive revenue growth, and gain a competitive advantage in the dynamic banking landscape.

Keywords: Banking industry; Customer segmentation; Cluster analysis; Machine learning techniques; Strategic decision-making.

1. Introduction:

In the dynamic landscape of business, adaptability to marketing strategies is imperative for both existing and emerging enterprises. The axiom "change or perish" holds truer than ever in today's

competitive marketplace, where businesses strive to meet the evolving demands of a burgeoning customer base. In this context, leveraging data mining techniques becomes crucial for uncovering latent patterns within a company's database. Customer segmentation emerges as a pivotal data mining approach, facilitating the categorization of customers into distinct groups based on various attributes such as demographics, preferences, and purchasing behavior. Through segmentation, customers are stratified based on shared characteristics, enabling businesses to tailor their marketing strategies more effectively. Segmentation not only directly influences marketing strategies but also opens up avenues for discovering novel insights. By identifying segments that align with specific products or services, businesses can customize marketing initiatives, offer targeted discounts, and decipher nuanced customer-product relationships previously obscured. A well-crafted customer segmentation strategy empowers firms to allocate marketing resources more efficiently, fostering opportunities for cross-selling and upselling. Moreover, personalized communications tailored to specific customer segments enhance customer satisfaction and loyalty, consequently bolstering customer retention. Clustering, a form of unsupervised learning, has proven instrumental in customer segmentation. Techniques such as K-means clustering, hierarchical clustering, and DBSCAN clustering offer avenues to identify clusters within unlabeled datasets. This study aims to employ data mining methodologies, specifically the K-means clustering algorithm, to delineate customer groups within the banking sector. Leveraging the silhouette method, which yields optimal cluster delineation, this research endeavors to unveil distinctive customer segments critical for enhancing marketing strategies and improving customer-centric initiatives in the banking industry.

This study seeks to explore the application of cluster analysis in customer segmentation within the banking industry. Its objective is to provide insights into different customer segments, aiding in the customization of marketing strategies, optimization of operational efficiency, fostering innovation, and ensuring regulatory compliance. The study findings are anticipated to be valuable for policymakers, financial authorities, and investors, contributing to informed decision-making and sustainable growth in the banking sector. The scope of the study on customer segmentation for the banking sector through cluster analysis involves identifying distinct customer segments based on demographic, behavioral, and transactional data, analyzing their impact on marketing strategies, product development, and operational efficiency, while addressing regulatory compliance, data privacy, and ethical considerations. The study aims to provide actionable insights for policymakers, financial authorities, banking executives, and investors to inform decision-making and promote sustainable growth within the banking industry.

2. Methodology:

In this study, the methodology involves data collection, data preprocessing, data analysis and exploration, cluster analysis, and data visualization. The process is detailed below:

Data collection: The data collection process involved administering questionnaires directly to bank customers. The questionnaires gathered information on various aspects such as demographic details, income, and spending behavior. The responses from 200 customers were collected and compiled into the dataset for analysis.

Data preprocessing: Python Libraries (pandas, NumPy) were employed to cleanse and manipulate the extracted data. This involved handling missing values, formatting inconsistencies, and outlier detection to ensure the quality and integrity of the dataset for subsequent analysis.

Data analysis and exploration: Python Libraries (pandas, seaborn, matplotlib) were utilized to analyze and visualize customer data to identify patterns and trends. Histograms, scatter plots, and boxplots were created to understand the distribution of variables, explore relationships between variables, and identify potential clusters within the data.

Cluster analysis: Python Libraries (scikit-learn) were employed for cluster analysis. K-Means clustering, a popular choice for customer segmentation, was utilized to group customers into pre-defined clusters based on their similarity in terms of features such as demographics, income, and spending behavior. Hierarchical clustering was also explored for a more exploratory approach to cluster analysis.

Data visualization: python libraries (seaborn) were utilized to visualize the resulting customer segments using scatter plots or dimensionality reduction techniques like PCA (Principal Component Analysis) for better interpretation. Visualizations aided in understanding the distinct clusters identified through cluster analysis and provided insights into the characteristics of each segment. This methodology allowed for a comprehensive analysis of customer segmentation in the banking sector, from data collection to visualization of the resulting clusters. The utilization of Python libraries facilitated efficient data processing, analysis, and visualization, enabling meaningful insights to be drawn from the dataset.

3. Results And Discussion

I. Customer Segmentation In The Banking Sector:

In the fiercely competitive landscape of the banking sector, companies must continually enhance their profitability and expand their customer base to align with evolving client expectations and attract new clientele. However, addressing each customer's unique needs is a daunting and time-consuming task due to the diverse array of customer aims, interests, and preferences. Instead of employing a generic "one-size-fits-all" approach, customer segmentation emerges as a vital marketing strategy, dividing the customer base into distinct, homogeneous groups. Through the utilization of data encompassing various factors such as regional circumstances, economic patterns, demographic trends, and behavioural attributes, customer segmentation facilitates the classification of customers into categories. By employing cluster analysis, this study aims to delve into customer segmentation within the banking sector, providing insights into effective marketing resource allocation and enhancing overall strategic decision-making processes.

II. Clustering:

Cluster analysis serves as a valuable technique for identifying similar groups within a vast dataset, where members within each group share greater similarities compared to those in other groups. This approach has gained prominence in data analysis, especially in marketing within the banking sector, since the 1970s. However, it's important to note that clustering is not a standardized data analysis method; its effectiveness heavily relies on the specific dataset or sample used. In our investigation of

customer segmentation through cluster analysis in the banking sector, we employ a statistical approach known as the "tandem technique." This technique combines factor analysis and cluster analysis to derive insights. Despite its utilization, the tandem technique has faced criticism, primarily due to its potential to disrupt existing cluster formations through early factor analysis. To mitigate this issue, hierarchical cluster analysis, particularly with binary variables, may serve as an alternative to the tandem approach. While the validity of this method was questioned in the past, non-hierarchical methodologies have since gained prominence in research within the banking sector.

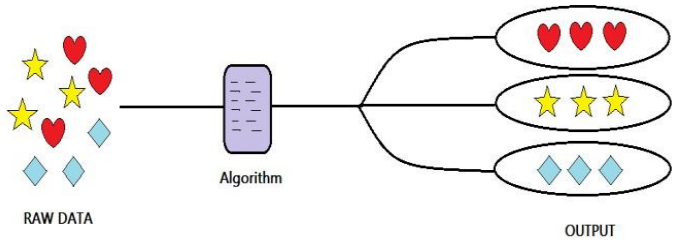
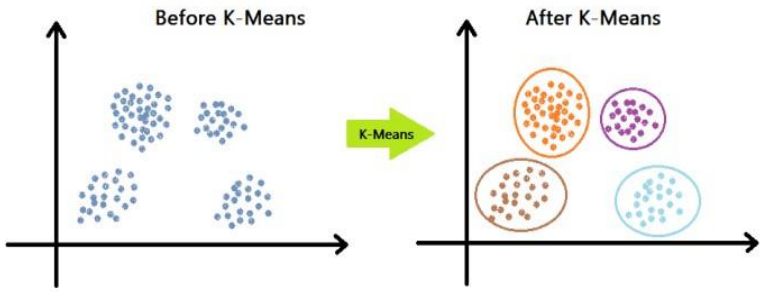


Figure.1. Clustering

III. Customer Segmentation Using K-Means Clustering:

The most well-known unsupervised partitioning clustering approach is K-Means Clustering. This clustering approach, commonly known as the centroid-based technique, divides data into non-hierarchical categories.

The dataset is separated into a collection of k groups in this sort of partitioning, where K is the number of pre-defined groups or



clusters. When compared to another cluster centroid, the cluster centre is built so that the distance between data points in one cluster is as low as possible

IV. Algorithm

Step 1: Select the number K to determine the number of clusters.

Step 2: At random, select K locations or centroids. (It's possible that it's not the same as the incoming dataset.) Step 3: Form the preset K clusters by assigning each data point to the centroid that is closest to it.

Step 4: Calculate the variance and move the centroid of each cluster.

Step 5: Reverse the previous three steps, reassigning each datapoint to the cluster's new closest centroid.

Step-6: Go to step-4 if there is a reassignment; otherwise, go to FINISH.

Step 7: The model is now complete.

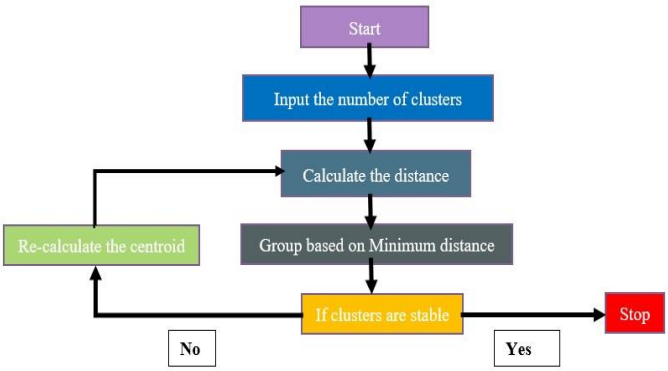
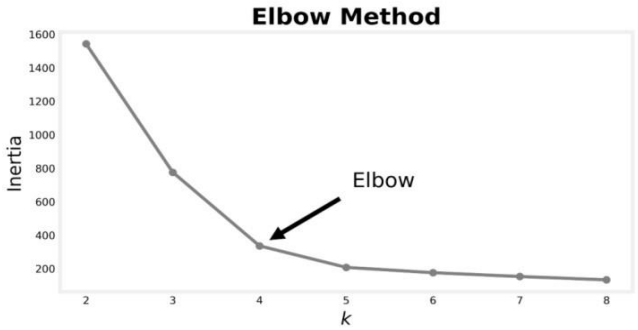


Figure.2. Flow of K-Means algorithm

V. Outline Of Existing Model:

1.The existing base paper uses "Elbow method" to find out the minimum optimal clusters for the K-means clustering

2.But elbow method does not work effectively in a few cases, for instance take the below scatter plot



VI. Outline Of Proposed Model:

The average silhouette coefficient across all dataset occurrences is used to calculate the silhouette score. The silhouette coefficient, which ranges from -1 to 1, measures how close points in one

cluster are to points in nearby clusters.

The silhouette coefficient is calculated as follows:

$$\frac{b-a}{\max(a,b)}$$

VII. Libraries Used:

Scikit-learn: It is a free Python machine learning software, sometimes known as sklearn. It is meant to interact with the Python numerical and scientific libraries NumPy and SciPy, and features support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering algorithms.

Seaborn: Seaborn is a matplotlib-based open-source Python library. It's used for exploratory data analysis and data visualization. Seaborn makes using data frames and the Pandas library a breeze. The graphs that are generated may also be readily changed.

NumPy (numerical Python):is a package that contains multidimensional array objects and tools for manipulating them. NumPy is a Python library that allows us to perform mathematical and logical operations on arrays.

NumPy is widely used in combination with SciPy and Matplotlib (Scientific Python) (plotting library). This combination is frequently used as a substitute for MATLAB, a prominent technical computing platform. The Python counterpart to MATLAB, on the other hand, is today regarded as a more contemporary and comprehensive programming language.

Pandas: is a Python toolkit for data science, data analysis, and machine learning that is open-source. It is based on NumPy, a multi-dimensional arrays-supporting library. Pandas, being one of the most widely used data manipulation tools, works well with a variety of other Python data science modules.

Matplotlib: For 2D array charts, Matplotlib is a superb Python visualization library. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the entire SciPy stack. The ability to show vast volumes of data in simple images is one of the most essential advantages of visualization. Line, bar, scatter, histogram, and more graphs are available in Matplotlib.

The Platform that was utilized was:

Jupyter Notebook is a server-client program that allows you to edit and run notebook documents, code, and data using a web browser. The Jupyter Notebook App can be operated locally on a PC with no internet connection (as described in this article) or remotely on a server with internet access. Users can build and organize processes in data science, scientific computing, computational journalism, and machine learning using the versatile interface.

. Data:

The data collection process involved administering questionnaires directly to bank customers. The questionnaires gathered information on various aspects such as demographic details, income, and spending behaviour. The responses from 200 customers were collected and compiled into the dataset for analysis.

Customer	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77

Table 1. Dataset

To begin, we'll need to figure out what kind of data we'll be working with (see table 1 for the dataset). We employ a straightforward yet comprehensive dataset that contains customer ID, gender, age, yearly income, and purchase score. The dataset's structure has been correctly displayed, and there are no null values.

If a dataset contains null values, duplicates, or other noisy data, data cleaning must be performed. Data cleansing ensures that information is reliable, usable, and available for analysis. When we have the data, we may visualize it by comparing the annual income and spending score, which is gender-specific. According to the study, there are five different types of plots that illustrate groups of customers who engage in the following activities, as well as customer behaviours linked to yearly income and expenditure scores:

1.Score of High Income/Low Spending

2.Low Income- A high score for spending

3.A high score for spending-despite Low Income

4.Average Income- Average Spending Score

5.High income – High spending score

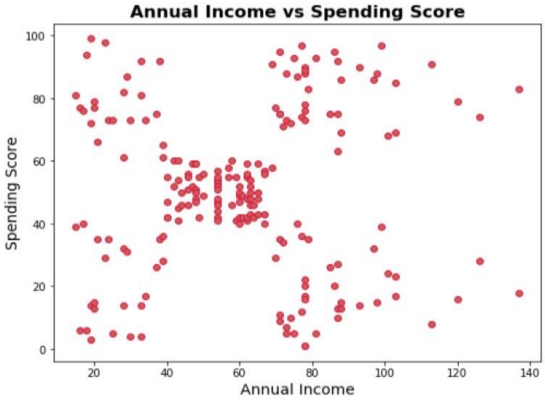


Figure.6. Annual Income vs Spending Score we can now build a K-means model based on the fact that there are a lot of groups, but not in great detail. The silhouette coefficient approach is used to do Clustering using k-means for a range of k clusters (let's say 1 to 10) and estimate the sum of square distances from each point to its assigned centre for each value. Decide on the number of clusters that will give you the best silhouette score. This defines how the silhouette score is calculated. We noticed that once K=5 is reached, there is no rapid movement in WCSS (Within Cluster Sum of Squares). And, given the number of clusters we have now, K=5 will be the correct number of clusters.

```
silhouette scores:
2 : 0.297
3 : 0.468
4 : 0.493
5 : 0.554
6 : 0.540
7 : 0.526
8 : 0.458
9 : 0.457
10 : 0.459
11 : 0.438
maximum silhouette score for 5 clusters:  0.554
```

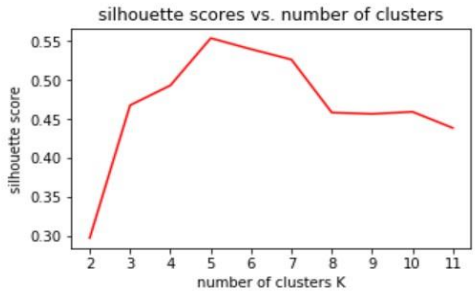


Figure.7. Silhouette approach result.

We can divide the plot into various groups, determine cluster can be prioritized, and then assign a label to each using the method stated above. The K-means approach can be used to decide which of the five clusters should be targeted, namely clients with Moderate Income- Moderate Spending Score, High Income- High Spending Score, and Low Income- High Spending Score. The required consumers have been located, as shown in Figure 8

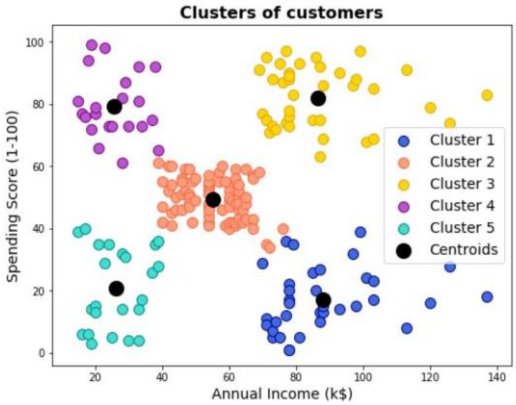


Figure.8. Final cluster of customers

4. Experiment Results:

1. Cluster 1 - Blue Group (High Income, Low Spending): Despite their high incomes, customers in this segment demonstrate relatively low spending habits. Banks can explore reasons behind their low spending and implement strategies to encourage greater engagement, possibly through personalized offerings or incentives.

2. Cluster 2 - Orange Group (Average Income, Average Spending): This segment consists of customers with moderate incomes and expenditures. While they may not significantly contribute to revenue, they still represent a considerable portion of the customer base. Banks can utilize data analysis techniques to better understand their spending patterns and devise strategies to increase engagement.

3. Cluster 3 - Yellow Group (High Income, High Spending): Customers in this segment have high incomes and exhibit high spending behaviours, making them prime targets for banks. Their frequent engagement with banking services offers an opportunity for targeted marketing efforts to capitalize on their profitability.

4. Cluster 4 - Violet Group (Low Income, High Spending): Customers in this segment exhibit high spending behaviours despite having low incomes. This behavior could be attributed to satisfaction with banking services or other motivations. Banks can leverage this by providing personalized services tailored to their specific preferences to enhance customer satisfaction and loyalty.

5. Cluster 5 - Green Group (Low Income, Low Spending): Customers in this segment have low incomes and demonstrate poor spending habits. Although they may have limited budgets, they still present potential opportunities for banks. Implementing tailored financial literacy programs or offering budget-friendly banking solutions can help encourage greater engagement from this group.

Implications:

By analysing customer segmentation within the banking sector, banks can better understand and predict customer behaviour based on their income and spending patterns. This information can inform targeted marketing strategies aimed at maximizing profitability and enhancing customer satisfaction. For example, targeting high-income customers with personalized wealth management services or offering promotions and discounts to customers with lower incomes to encourage greater engagement with banking services. Additionally, understanding the needs and preferences of different customer segments can drive product development and service enhancements tailored to specific customer segments, ultimately strengthening customer relationships and competitiveness in the banking sector.

5. Conclusion:

This study highlights the feasibility and benefits of customer segmentation within the banking sector through cluster analysis. While this machine learning application proves to be highly valuable in the market, it is essential for bank managers to utilize the insights gained from customer segmentation effectively.

Banking managers must prioritize understanding the needs and preferences of different customer segments identified through cluster analysis. By analysing customer purchasing behaviours and establishing regular interactions with customers, banks can tailor their services to meet specific demands effectively.

Furthermore, it is crucial for banking managers to focus on providing personalized experiences that make customers feel valued and comfortable. This includes offering tailored financial products, providing relevant promotions and discounts, and ensuring seamless customer interactions across various channels.

In conclusion, by leveraging customer segmentation through cluster analysis, banks can better understand their customers' requirements and preferences, leading to more effective marketing strategies, enhanced customer satisfaction, and ultimately, increased competitiveness in the banking sector.

References:

[1] “Customer segmentation based on survival character,” IEEE, Jul. 2003.

[2] “Customer Segmentation Using K Means Clustering,” Towards Data Science, Apr. 2019.

[3] Ruhul Reddy, “Who’s who: Understanding your business with customer segmentation,” INTERCOM.

[4] Kristen Baker, “The Ultimate Guide to Customer Segmentation: How to Organize Your Customers to Grow Better,” Hunspot.

[5] Tim Ehrens, “customer segmentation,” TechTarget.

[6] V.Vijilesh, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 05, May 2021.

[7] Expert Systems with Applications, vol. 100, Feb. 2018, "Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data."

[8] “Cluster analysis.”, Wikipedia.

[9] "CUSTOMER SEGMENTATION USING MACHINE LEARNING," IJCRT, AMAN BANDUNI and ILAVENDHAN A, vol. 05, 2018.

[10] Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury, “Customer Segmentation using K-means Clustering,” IEEE, Jul. 2019.

[11] I. S. N. Chinedu, S. O. C. Kalu, E. & C. E. D. U. of U. U. A. S. O. C. Kalu, E. & C. E. D. U. of U. U. A. S. O. C. Kalu, E. & C "Application of the K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," vol. 4, no. 10, 2015, by Pascal Ezenkwu, International Journal of Advanced Research in Artificial Intelligence.

[12] Author Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning [Beginners Guide],” neptuneblog, Dec. 13, 2021.