Bank Customer Segmentation Using Data Science

Mrs. Shraddha Sandip Said ,
Mrs. Asmita Mohan Phatangare
Department of Master of Computer Application,
JCEI's Jaihind Institute of Management and Research, Vadgaon Sahani Email: saidshraddha99@email.com,

asmitafatangare41@gmail.com

ABSTRACT

Given the high number of transactions, deposits, withdrawals, loans, cards, and other financial operations that a bank undertakes, it accumulates a significant amount of data about its customers. Information assets can be utilized in various ways, including analyzing the behavior of customers and understanding their financial needs. This paper aims to propose a data science approach for the segmenting of bank customers. It will undertake the task of exploring and preparing the data for analytical procedures. The study will attempt to engineer several features that can help identify segments with similar characteristics. The research will engage such methods as K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Model to operationalize the task of customer segmentation. The paper will also try to incorporate a temporal component into the analysis. The proposed approach would allow banks to better comprehend the customer base and facilitate more precise marketing strategies.

Keywords

Customer Segmentation, Banking Analytics, Data Science, Machine Learning, Clustering Algorithms, K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Model (GMM), Feature Engineering, Customer Behavior Analysis, Temporal Analysis, Financial Data Analysis, Customer Profiling, Targeted Marketing.

***

1. INTRODUCTION

The banking industry plays a role in a countrys growth. It offers people services such as savings accounts, loans, credit, investments and online banking. As digital technology expands banks gather amounts of customer data from activities, account usage and online services. Examining this data helps banks understand how customers behave discover what customers need financially and improve services. When banks treat all customers as a group it becomes difficult to develop effective strategies for people with different preferences and financial needs. Customer segmentation is a way to split a group of customers into categories based on things they have in common. These things can be age, income, money in their account how often they spend, how much they use their credit and how they like to bank. When banks find out what customers have in common they can figure out which group each customer is in and create services, for each group. For example some customers like to handle their banking while others need help with money planning saving money or getting credit.

The data science field presents techniques and approaches to analyze customer data and extract valuable information. Machine learning is particularly useful for customer segmentation as it can discover groups based on data similarities without predefining those groups. Clustering methods such as K-Means,

Hierarchical Clustering and DBSCAN can be applied to customer data to reveal behavior patterns. Techniques such as cleaning data scaling features inspecting data closely and reducing the number of features improve the analysis and interpretation of the resulting groups.

In this study data science and machine learning methods are employed to segment bank customers according to age, balance and behavior. The study examines customer data identifies groups and evaluates the quality of these groups using metrics. The Elbow Method and Silhouette Score help determine the number of groups and assess how distinct the groups are. The identified groups can then be analyzed to understand variations in customer behavior and financial needs. The proposed method can assist banks in creating marketing plans increasing customer satisfaction, identifying customers who require attention and offering products that fit better.

Throughout this process customer data must be handled carefully respecting privacy regulations and addressing bias in the groupings. Overall this study demonstrates how data science can transform customer information into insights and help banks make better decisions.

Originality note: This introduction was written for your topic and is not copied from a published paper. It cannot be guaranteed as new without reviewing other sources. If you incorporate ideas, data or results from authors, in your final paper be sure to give proper credit.

1.1 Problem Statement:

It is common knowledge that each bank customer may have a divergent financial pattern, transaction behavior, and banking service demand. Such complexity cannot be understudied by the traditional segmentation approaches. This paper, therefore, seeks to apply data science and machine learning tools to identify customer groups, evaluate clustering algorithms' efficacy, and investigate temporal shifts in customers' financial patterns to offer a personalized approach to the bank's performance.

1.2 Objectives:

To analyze the customers' demographics, finances, and behavior.

To identify the customer segments using the clustering algorithms.

To compare the different clustering methods.

To determine the key characteristics of the customer groups.

To evaluate and interpret the results of the segments identified.

To analyze how the customers' behavior changed over time.

To personalize the bank services and make decisions based on the analysis results.

2. LITERATURE REVIEW

Customer segmentation plays an important role in modern banking as it allows institutions to learn more about their clients' behaviors and financial expectations. Researchers are now trying to come up with innovative ways to improve customer segmentation. The traditional approach is based on demographical information, while the more recent trends propose considering transactional and behavioral traits. It is worth mentioning that most previous studies aim at simply determining the customer segments using either K-Means, Hierarchical, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), or Gaussian mixture models. Others try to evaluate RFM scores to assess the customers' value.

However, there are few articles that combine several clustering algorithms and analyze the results of customer segmentation in depth to reveal the key differences between the segments and explain the patterns of their behavior over time. This research is intended to fill this gap by providing more accurate segmentation and conducting a thorough analysis of the obtained results.

3. METHODOLOGY

This study utilizes a data-driven approach to identify unique segments of bank customers. It consists of

collecting customer data, preparing and engineering the data, performing clustering, analyzing the results, and exploring the obtained segments.

3.1 Data Collection

Customer data of a bank is gathered from an appropriate public database or an authorized source. The database may comprise demographic, economic, and transaction data of the customers, such as age, income, account balance, deposits, withdrawals, transaction frequency, loan information, transaction dates, etc.

3.2 Data Preprocessing

The gathered data is cleaned and preprocessed. In this stage, null values, duplicates, inconsistencies, and outliers are identified and removed. Moreover, numerical columns are normalized or standardized if needed to ensure that columns with large ranges do not distort the results.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to explore the data. It consists of statistically describing the data, visualizing the data using appropriate plots, and identifying relationships between variables.

3.4 Feature Engineering

The features to be used in the clustering process are chosen. It may consist of extracting new features from the available database, such as annual transaction value, average transaction value, total deposits, total withdrawals, and transaction frequency.

3.5 Customer Segmentation

The clustering process is carried out using appropriate clustering algorithms, such as K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Model.

3.6 Cluster Evaluation

The results of the clustering process are evaluated using appropriate clustering evaluation metrics, such as Silh ouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

4. FUTURE SCOPE

Data science techniques are used in this research to understand customer characteristics and to classify bank customers into groups. Banks serve customers who have financial goals, spending habits and service preferences. Identifying these differences can help banks develop customer management strategies. This research examines how data analysis and machine learning techniques can be used to extract meaningful insights from customer data. The research covers the collection or selection of a banking dataset that contains relevant customer information.

The selected data can include details, income levels, account balances, transaction records, product usage and customer activity. These attributes are analyzed to find similarities and differences among customers. Data cleaning, transformation and normalization are performed when needed to prepare the dataset for analysis. The study also examines the use of algorithms to divide customers into groups based on their characteristics. K- Means clustering is a technique because it can quickly group customers according to similarities in their data. Other methods, such as Hierarchical Clustering and DBSCAN may also be explored to compare their suitability for customer segmentation. Appropriate evaluation techniques are used to examine the quality and usefulness of the groups that are generated.

5. EXISTING RESEARCH

Several researchers have examined customer segmentation in the banking and financial sector with data analysis and machine learning. These studies mainly aim to locate groups of customers that share financial traits and transaction habits. In research people mainly used factors such, as age, gender, income, occupation and location to segment customers. Later studies added transaction- related factors including transaction frequency, transaction amount, account balance, deposits, withdrawals and loan activities.

K- Means clustering has been widely applied because it is simple and fits well for grouping customers with characteristics. Other researchers have used Hierarchical

Clustering to explore how customer groups relate to one another.

Density- based methods such as DBSCAN have also been tried to spot customer groups and unusual patterns. RFM analysis has been employed to study customer activity based on recency, frequency and monetary value. Some researchers have merged RFM features with algorithms to build more detailed customer segments.

Recent research has focused on improving customer segmentation by engineering features using clustering techniques and analyzing behavior. These methods give a picture of customer groups and their traits. While existing research offers ways, for customer segmentation many studies rely on a limited set of features or a single clustering algorithm. This gap offers an opportunity to create an approach that blends demographic, financial, transaction and time- based customer data together.

6. HYPOTHESIS

H0 : Data science and machine learning methods are not effective in identifying meaningful groups of bank customers based on their behavior.

H1: Data science and machine learning methods are effective in identifying meaningful groups of bank customers based on their behavior. The hypotheses will be evaluated based on the clustering outcomes and analysis obtained from the selected dataset

7. TOOLS AND TECHNOLOGIES

Python Programming: Utilized for the construction of the customer segmentation system and for the purpose of conducting data analysis.

Jupyter Notebook: Facilitates the process of developing the analysis and offers an interactive environment.

Pandas Library: This library is employed in the organization, processing, and simplification of the customer data that will be analyzed.

Numpy Library: It supports mathematical calculations and efficient processing of numerical data.

Matplotlib: Its purpose is the construction of graphics and visualization of the analyzed data results.

Seaborn: It is utilized for the demonstration of statistical patterns and connections between the characteristics of the customers

Scikit-learn: It is used for the implementation of machine-learning methods and measurement of the model's performance.

K-Means Method: It is used to group customers according to similar characteristics.

DBSCAN Method: It is used for the purpose of separating customer groups characterized by similar attributes and identifying outliers.

PCA (Principal Component Analysis): This technique reduces the number of variables by retaining the most significant information, making the dataset simpler to analyze.

Cluster Validation Measures: They are applied for the purpose of determining the efficiency of the customer clustering.

Open Dataset: An open dataset that includes an appropriate open license is utilized for the purpose of experimentation and research.

8. IMPLEMENTATION/EXPERIMENTAL SETUP

1. Development Environment: The proposed system is implemented using Python in a Jupyter Notebook environment. Library such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn are used for the data processing, visualization, and machine-learning operations .

2. Dataset Collection: A publicly available and openly licensed banking dataset is selected for the experiment. The dataset contains customer-related information such as age, occupation, account balance, loan details, contact history, and customer response.

3. Data Preprocessing: The collected dataset is examined for missing values, duplicate records, incorrect entries, and unnecessary attributes.

4. Feature Selection: Important customer attributes are selected according to their relevance to customer grouping. Attributes such as age, balance, loan status, contact frequency, and previous customer interactions are considered.

5. Data Transformation: Categorical attributes are converted into numerical representations using suitable encoding methods. Numerical attributes are transformed into a common scale to prevent high-value variables from dominating the analysis.

9. CONCLUSION

The Bank Customer Segmentation Using Data Science study can benefit the banks by providing them with the knowledge of how customers behave, what are their needs, and how they spend their money. The demographic and financial data of the customers can be analyzed and divided into segments based on specific criteria. The segments can be identified using the data science methods, such as K-Means, Hierarchical Clustering, DBSCAN, or Gaussian Mixture Models. The information gained from the analysis can be used to find the marketing strategies, which would be beneficial for the bank, its clients, and shareholders. Despite the potential usefulness of data science, there are some considerable threats and difficulties, including the privacy and accuracy concerns regarding the data and the complexity of choosing the right model for the analysis.

10. REFERENCES

[1] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*(3), 264–323. https://doi.org/10.1145/331499.331504

[2] Niyagas, W., Srivihok, A., & Kitisin, S. (2006). Clustering e-Banking Customer Using Data Mining and Marketing Segmentation. *ECTI Transactions on Computer and Information Technology, 2*(1), 63–69. https://doi.org/10.37936/ecti-cit.200621.53272

[3] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* https://aaai.org/papers/kdd96-037-a-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with-noise/

[4] Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[5] Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A., & Alasgarov, E. (2020). Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning. *arXiv preprint.* https://arxiv.org/abs/2008.08662

[6] Rizkyanto, H., & Gaol, F. L. (2023). Customer Segmentation of Personal Credit Using Recency, Frequency, Monetary (RFM) and K-Means on Financial Industry. *International Journal of Advanced Computer Science and Applications, 14*(4). https://doi.org/10.14569/IJACSA.2023.0140417

[7] (2018). Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. *2018 International Conference on Information Management and Technology (ICIMTech).* https://doi.org/10.1109/ICIMTech.2018.8528086

[8] (2018). Customer Segmentation Based on RFM Model and Clustering Techniques With K-Means

Algorithm. *2018 Third International Conference on Informatics and Computing (ICIC).* https://doi.org/10.1109/IAC.2018.8780570

[9] Bhuiyan, M. S. M. (2024). Optimizing Customer Segmentation in the Banking Sector: A Comparative Analysis of Machine Learning Algorithms. *Journal of Computer Science and Technology Studies.* https://doi.org/10.32996/jcsts.2024.6.4.1

[10] Grigorova, I., Efremov, A., & Karamfilov, A. (2025). An Automated Machine Learning Framework for Interpretable Customer Segmentation in Financial Services. *International Journal of Financial Studies, 13*(4), 243. https://doi.org/10.3390/ijfs13040243

[11] Hammami, S., Haddad, M., & Touzri, A. (2025). Customer Segmentation in Banking with RFM, CLV and K-Means: Insights from the BIAT Trading Room. *2025 IEEE 4th International Conference on Signal, Control and Communication (SCC).* https://doi.org/10.1109/SCC66964.2025.11424735

[12] Scikit-learn Developers. (n.d.). Clustering: User Guide. *Scikit-learn Documentation.* https://scikit-learn.org/stable/modules/clustering.html

[13] Scikit-learn Developers. (n.d.). Gaussian Mixture Models. *Scikit-learn Documentation.* https://scikit-learn.org/stable/modules/mixture.html

[14] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999).Data Clustering: A Review. ACM ComputingSurveys,31(3),264323.https://doi.org/10.1145/331499.331504

[15] Rousseeuw, P. J. (1987).Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7