JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

http://www.jmrr.org Volume: 6, Issue: 1, April 2025 | ISSN: 2708-9452

Customer Segmentation using Machine Learning

Ramesh K
Head of the Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode
Namakkal, Tamilnadu, India.
Aarthi R
PG Student, Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode,
Namakkal, Tamilnadu, India.

Abstract

This project presents a machine learning-based customer segmentation system designed to enhance personalized financial services such as savings plans, loans, and wealth management strategies. The system collects and preprocesses customer data including demographics, transaction history, and online behaviour to extract key features through exploratory data analysis. Clustering algorithms like K-Means and Spectral Clustering are employed to group customers into meaningful segments. The resulting model is deployed via a Streamlit-based web application that enables users to receive real-time, segment-driven financial recommendations aligned with their risk profiles and goals. Performance is evaluated using metrics such as Silhouette Score and Elbow Method to validate clustering effectiveness. The proposed solution demonstrates improved user engagement and model adaptability through feedback-driven refinement, offering financial institutions a scalable and intelligent platform for customer-centric service delivery.

Keywords: Customer Segmentation, Machine Learning, Python Programming, Streamlit, K-Means Clustering, Spectral Clustering, Financial Recommendation System.

1 Introduction

In the era of data-driven decision making, financial institutions are increasingly adopting intelligent systems to better understand and serve their customers[1]. Traditional one-size-fits-all approaches to financial planning are no longer sufficient in addressing the diverse goals, risk appetites, and spending behaviours of modern consumers. Customer segmentation the process of dividing a broad consumer base into distinct groups based on shared characteristics plays a crucial role in delivering targeted and personalized financial solutions[2].

This project proposes a machine learning-based segmentation model built using Python and deployed through the Streamlit framework. The system leverages customer data including demographics, transaction history, online interactions, and behavioural patterns to form customer segments via unsupervised learning techniques. By utilizing clustering algorithms such as K-Means and Spectral Clustering, the system identifies underlying structures in the data, enabling financial institutions to tailor offerings such as loan plans, savings strategies, and wealth management services to specific customer groups [3].

To make this system accessible and interactive, a web-based application is developed with Streamlit, allowing end-users to input personal data and receive real-time recommendations based on their segment profile. The platform also supports feedback-driven model refinement, ensuring improved accuracy and relevance over time [4]. This approach not only enhances customer satisfaction but also supports data-centric marketing and service strategies in the financial sector.

2 Literature Review

In today's competitive business landscape, the demand for AI-driven customer profiling systems is increasing [5]. As customer acquisition costs rise and retention becomes more challenging, businesses are shifting their focus from attracting new customers to strengthening relationships with existing ones. Maintaining long-term customer engagement provides several benefits, including increased repeat purchases, lower advertising costs, and organic growth through customer referrals [6].

This project aims to develop a machine learning-based consumer profiling system to enhance automated marketing, drive sales, and improve customer interaction. To achieve this, the following research tasks are addressed: collecting and processing customer data, exploring various machine learning techniques for segmentation, defining the structure of customer profiles, analyzing and organizing customer information, reviewing global approaches to profiling, and evaluating the most effective methods for business applications.

2.1 Customer Classification

Companies strive to meet the diverse needs of customers while expanding their market reach. Understanding individual customer preferences and behaviors is a complex task due to demographic and behavioral variations. A uniform strategy is no longer effective. Businesses therefore adopt segmentation strategies by grouping customers with shared characteristics, enabling personalized services and more effective marketing [7].

2.2 Big Data

Big Data refers to vast structured and unstructured datasets exceeding the capacity of traditional processing tools. Sources include sensors in smartphones, web logs, and operational databases. Big Data's five V's—Volume, Velocity, Variety, Veracity, and Value make it especially relevant in sectors like finance, healthcare, and fraud detection [8].

2.3 Data Repository

Data collection is a foundational step involving acquisition and measurement to assess trends and support critical analysis. A structured repository allows for effective storage, management, and processing of collected data, facilitating meaningful business insights. For this study, the dataset is sourced from the UCI Machine Learning Repository.

2.4 Clustering

Clustering involves grouping data points based on shared features. Various algorithms are suitable depending on data characteristics. As noted by Sulekha Goyat [9], no single algorithm fits all cases. In this project, three clustering algorithms are implemented using Python's scientific libraries.

2.5 K-Means

K-Means is a centroid-based clustering method where each point is assigned to the nearest cluster center. It identifies patterns that guide decision-making. The Elbow Method is used to determine the optimal number of clusters[10].

2.6 Spectral Clustering

Spectral clustering is a graph-based technique suitable for customer segmentation, relying on similarity matrices. Eigenvalues and eigenvectors are computed from the Laplacian matrix to define cluster boundaries in a transformed space.

3 Methodology

This system is designed to train a dataset for customer segmentation using machine learning algorithms to categorize users based on their attributes, behaviors, and preferences. The methodology optimizes marketing strategies and reveals behavioral insights.

3.1 Data Collection and Integration

Customer data is collected from various sources such as demographics, purchase history, social media, and feedback. It is cleaned and preprocessed to ensure accuracy and consistency. Feature engineering is applied to extract relevant features.

3.2 Machine Learning Model Selection

Clustering methods like K-Means, DBSCAN, and Agglomerative Clustering are considered along with dimensionality reduction techniques like PCA and t-SNE. Ensemble models such as Random Forests and Gradient Boosting are explored based on project needs.

3.3 Training and Validation

The dataset is split into training, validation, and testing subsets. Models are trained with optimized hyperparameters to ensure strong generalization. Performance is validated using internal metrics such as Silhouette Score.

3.4 Customer Segmentation

Trained models are applied to new or test data to identify distinct customer groups. Each group is analyzed for shared traits and behavior, helping define meaningful segments for financial targeting.

3.5 Interpretation and Analysis

Segmented results are interpreted by analyzing feature contributions. This enables the identification of high-value customer groups, potential leads, and strategic opportunities.

3.6 Visualization and Reporting

Segmentation results are visualized using graphs and scatter plots. Dashboards are built to allow stakeholders to interact with and explore insights easily.

3.7 Deployment and Integration

The system is deployed in a production-ready environment. Integration with CRM systems, marketing tools, and analytics platforms ensures real-world applicability and operational efficiency.

3.8 Continuous Improvement and Adaptation

The system supports continuous updates with new data and evolving business requirements. Feedback is collected for refining models and improving accuracy. Emerging machine learning techniques are monitored and incorporated to maintain relevance.

4 System Architecture

The architecture of the proposed customer segmentation system is designed to ensure seamless integration between data processing, machine learning models, and an interactive front-end application. The system follows a modular design, structured in a multi-tier architecture comprising data handling, model computation, and visualization layers.

4.1 1) Data Layer

This layer is responsible for collecting and storing structured and semi-structured data related to customers. Sources include CSV files, transactional logs, and web interaction data. The data layer handles cleaning, transformation, normalization, and feature selection before passing the processed data to the model layer.

4.2 2) Machine Learning Layer

This computational layer houses the clustering algorithms used for segmentation. K-Means is the primary algorithm applied for grouping customers based on feature similarity. The layer supports additional clustering techniques such as Spectral Clustering to accommodate nonlinear patterns. It includes evaluation functions such as Silhouette Score and Elbow Method for optimal model tuning and validation.

4.3 3) Application Layer (Streamlit Interface)

The front-end interface is developed using Streamlit to provide real-time interaction with the user. It enables users to upload data, visualize clusters using Matplotlib or Seaborn, and receive financial recommendations. This layer ensures responsiveness, scalability, and user-friendly access to the system's core functionality.

4.4 4) Integration and Feedback Loop

Post-deployment, a feedback mechanism is embedded to capture user responses and preferences. This loop is critical for iterative learning, allowing the model to evolve based on newly collected data and performance outcomes. Updates to clustering models are managed in a controlled environment to ensure system reliability.

4.5 System Workflow

The end-to-end system workflow begins with data upload and preprocessing, followed by unsupervised clustering of customers. The system then provides dynamic visualizations and displays tailored financial suggestions. As users interact with the system, their feedback contributes to model refinement, completing the cycle of continuous improvement.

5 Analysis

The effectiveness of the customer segmentation system is evaluated using computational complexity, clustering accuracy, and visualization metrics. This analysis validates the system's performance in terms of execution speed, clustering quality, and scalability for real-world applications.

5.1 K-Means Clustering Performance

The K-Means algorithm used for segmentation exhibits a time complexity of $O(n \times k \times i \times d)$, where:

- n = number of data points,

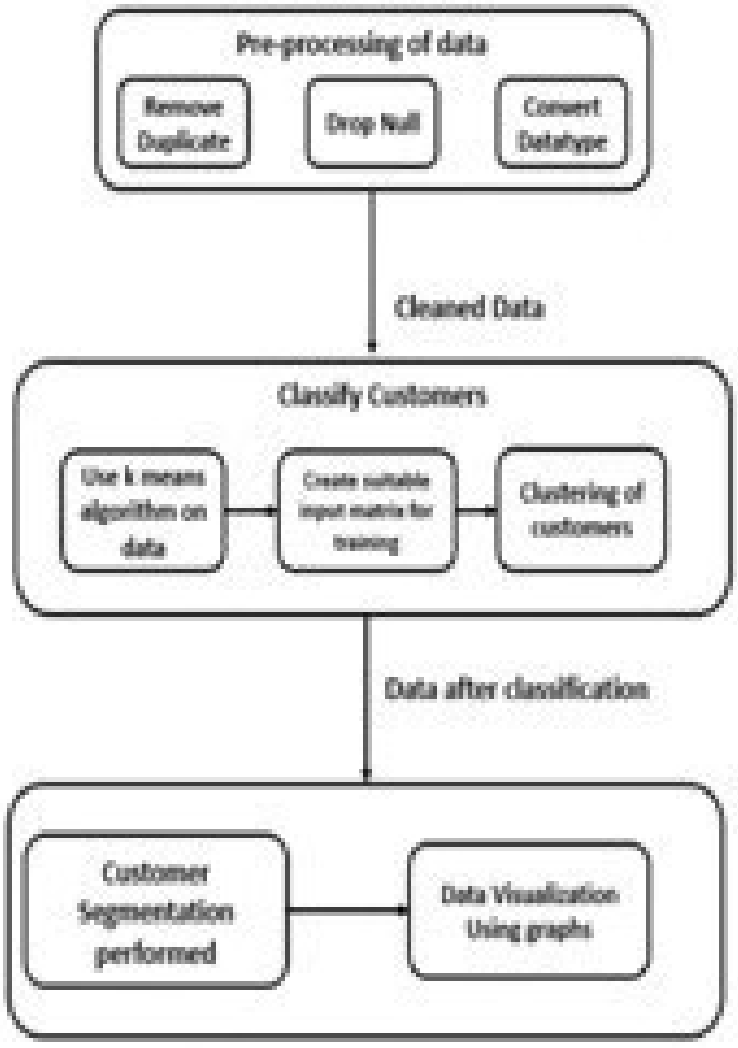

Fig 4.1 System Architecture for Customer Segmentation

Figure 1: System Architecture

- k = number of clusters,
- i = number of iterations until convergence,
- d = number of dimensions or features.

Given that k is typically small (e.g., 3–5), execution time primarily depends on the dataset size n. This makes K-Means suitable for small to moderately large datasets but less efficient for high-dimensional or massive data.

5.2 File Processing Efficiency

CSV file loading using `pandas.read_csv()` demonstrates linear complexity $O(n)$, where n is the number of rows. This operation is efficient for typical customer datasets but may require optimization techniques like chunking for large-scale deployments.

5.3 Visualization with Matplotlib

Visualization of clusters using `matplotlib.pyplot.scatter()` is also linear in complexity $O(n)$, making it feasible for displaying clusters of several thousand data points. However, interactive plotting tools may be preferred for more complex analyses involving higher dimensionality.

5.4 Clustering Validation Metrics

To validate the quality of clustering, the following metrics are employed:

Silhouette Score: Measures how similar each data point is to its own cluster compared to other clusters. A score close to 1 indicates well-defined, compact clusters.

Elbow Method: Plots the Within-Cluster Sum of Squares (WCSS) against varying values of k. The "elbow" point on the graph suggests the optimal number of clusters.

5.5 Example Evaluation

An example evaluation using the Silhouette Score in Python:

```
from sklearn.metrics import silhouette_score
score = silhouette_score(X, kmeans.labels_)
print(f"Silhouette Score: {score}")
```

Scores near 0.7–1.0 reflect strong clustering; scores near 0 indicate overlapping or ambiguous clusters, suggesting the need for alternative algorithms or refined features.

6 Output

The output of the proposed customer segmentation system includes visual insights and statistical validation of clustering performance. Two key outputs are generated as part of the system's evaluation and user interface.

6.1 Finding the Optimal Number of Clusters

The Elbow Method is used to identify the optimal number of clusters (k) by plotting the Within-Cluster Sum of Squares (WCSS) against varying k values. The point where the curve bends (elbow point) is considered the most appropriate k. This is visualized using a line plot that helps determine the ideal cluster count for segmentation.

- **Step 1:** Perform K-Means clustering for $k = 1$ to $k = 10$.
- **Step 2:** Calculate WCSS for each k.
- **Step 3:** Plot WCSS vs. k to identify the "elbow" point.

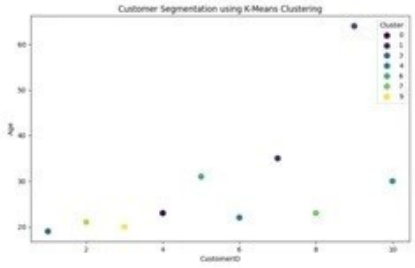

Figure 2: Finding optimal k using the Elbow Method

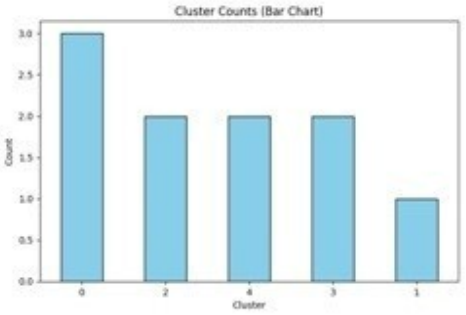

Figure 3: Visualizing segmented customer data

6.2 Cluster Visualization

After determining the optimal k, the dataset is clustered using K-Means, and the resulting clusters are visualized using a scatter plot. Each cluster is color-coded to show the distribution and separation of customer segments in the feature space. Centroids are marked to indicate the center of each cluster.

These outputs enable stakeholders to interpret the segmentation results effectively and validate the consistency and utility of the model before deploying it in production.

7 Conclusion

This project demonstrates the effectiveness of integrating machine learning techniques with Python and Streamlit to build an intelligent customer segmentation system for financial recommendation services. By leveraging clustering algorithms such as K-Means and Spectral Clustering, the system successfully groups customers based on shared attributes like demographics, transaction history, and behavioural patterns. These clusters allow financial institutions to deliver personalized recommendations in areas such as savings plans, loan options, and wealth management strategies.

The use of the Streamlit framework enables the deployment of an interactive, user-friendly web application that delivers real-time recommendations and visual insights. Through continuous feedback and model refinement, the system can evolve to meet changing customer behaviours and data dynamics. This approach enhances customer satisfaction, promotes trust, and ultimately contributes to improved engagement and business outcomes. Future enhancements may include the integration of deep learning

techniques, real-time data ingestion, and multi-channel user interfaces to further expand the system's capabilities.

References

[1] T. Blanchard, P. Bhatnagar, and T. Behera, Marketing Analytics Scientific Data: Achieve Your Marketing Objectives with Python's Data Analytics Capabilities. Birmingham, UK: Packt Publishing, 2019.

[2] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, "Sales business analysis: Customer categories use market basket data," Expert Systems with Applications, vol. 100, pp. 1–16, 2018.

[3] T. Hong and E. Kim, "It separates consumers from online stores based on factors that affect the customer's intention to purchase," Expert Systems with Applications, vol. 39, no. 2, pp. 2127–2131, 2011.

[4] Y. H. Hwang, Hands-on Advertising Science Data: Develop Your Machine Learning Marketing Strategies Using Python and R. Birmingham, UK: Packt Publishing, 2019.

[5] P. P. Premkanth, "Market classification and its impact on customer satisfaction with special reference to the Commercial Bank of Ceylon PLC," Global Journal of Management and Business Research, vol. 12, no. 1, 2012.

[6] S. Goyat, "The basis of market segmentation: A critical review of the literature," European Journal of Business and Management, vol. 3, no. 9, pp. 45–54, 2011.

[7] L. Welling and L. Thomson, PHP and MySQL Web Development, 5th ed. Boston, MA, USA: Addison-Wesley, 2016.

[8] J. Duckett, HTML and CSS: Design and Build Websites. Indianapolis, IN, USA: Wiley, 2011.

[9] M. Powell, Beginning Bootstrap: A Hands-On Guide to Building Websites, 2nd ed. New York, NY, USA: Apress, 2017.

[10] R. S. Pressman and B. R. Maxim, Software Engineering: A Practitioner's Approach, 8th ed. New York, NY, USA: McGraw-Hill, 2015.

Cite this article:

Ramesh K &Aarthi R, "Customer Segmentation using Machine Learning", Journal of Multidimensional Research and Review (JMRR), Vol.6, Iss.2, pp.108-116, 2025