A PROJECT REPORT

On

“Customer Segmentation based on their demographics,purchase pattern,behaviours”

Submitted to:

Celebal Technologies

Prepared By:

Name	Email
Rishabh Kumar	rishabhquasar23@gmail.com
Aman Raj	2005362@kiit.ac.in
Dipti Verma	2006173@kiit.ac.in
Sharad Kumar Agarwal	Skagarwal485@gmail.com
Simran Bhardwaj	Simranbhardwaj2607@gmail.com
Vardaan Khosla	Khoslavardaan1@gmail.com

ACKNOWLEDGEMENT

It gives us immense pleasure to present before you our project titled "**Customer Segmentation based on their demographics,purchase pattern,behaviours**". The joy and satisfaction that accompany the successful completion of any task would be incomplete without the mention of those who made it possible. We are glad to express our gratitude towards our prestigious Company **Celebal Technology** for providing us with utmost knowledge, encouragement and the maximum facilities in undertaking this project..

We express our deepest gratitude and special thanks to **Arpit Jain Sir** for all her guidance and encouragement.

.

RISHABH KUMAR
AMAN RAJ
DIPTI VERMA
SHARAD KUMAR AGARWAL
SIMRAN BHARDWAJ
VARDAAN KHOSLA

ABSTRACT

Customer Segmentation is a crucial data-driven approach that aims to divide a customer base into distinct groups based on their demographics, behaviors, and purchase patterns. In this project, we employ K-means clustering, a popular unsupervised learning algorithm, to cluster customers into meaningful segments. By understanding customer preferences and characteristics within each segment, businesses can design personalized marketing strategies, enhance customer experiences, and optimize resource allocation.

The project involves several stages, including data collection, preprocessing, and feature engineering to ensure data quality and relevance. Subsequently, the K-means algorithm is applied to group customers with similar attributes into distinct clusters. The clustering results are interpreted to gain valuable insights into customer behaviors, preferences, and loyalty.

The outcomes of this project offer significant benefits to businesses, enabling them to optimize marketing efforts, improve customer satisfaction, and make informed data-driven decisions. Customer segmentation facilitates personalized interactions, enhances customer loyalty, and empowers businesses to stay competitive in the dynamic market landscape.

Moreover, the project emphasizes the importance of data privacy and security, ensuring compliance with data protection regulations to safeguard customer information. Through continuous monitoring and iterative refinement, the clustering model and marketing strategies are updated to adapt to changing market dynamics and customer preferences.

CONTENTS

.

CHAPTER 1

INTRODUCTION

1.1 Purpose

The purpose of clustering customers into distinct groups based on their demographics, behaviors, or purchase patterns is to perform customer segmentation. Customer segmentation is a data analysis technique used in marketing and business to divide a large customer base into smaller, more manageable groups or segments with similar characteristics.

1.2 Product Scope

The product scope for customer segmentation involves the development and implementation of a data-driven solution or system that enables businesses to cluster their customers into distinct groups based on their demographics, behaviors, or purchase patterns. This solution aims to provide actionable insights and information for marketing, product development, and customer relationship management. The key components of the product scope include:

1.Data Collection
2. Data Preprocessing

3. Clustering Algorithms
4. Customer Segmentation
5. Visualization
6. Insights and Recommendations
7. Integration
8. Scalability and Performance

Overall, the product scope for customer segmentation aims to deliver a comprehensive solution that empowers businesses to understand their customers better, make informed decisions, and tailor their marketing efforts and product offerings to specific customer segments, ultimately leading to improved customer satisfaction and business growth.

CHAPTER 2

REFERENCES

Dataset link : https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Application Link : https://huggingface.co/spaces/amanr12/customer_segmentation

https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

https://www.javatpoint.com/clustering-in-machine-learning

https://analyticsindiamag.com/a-tutorial-on-various-clustering-evaluation-metrics/

CHAPTER 3

PROJECT DESCRIPTION

3.1 Algorithms Used

K-means: K-means is a popular clustering algorithm used in data analysis and machine learning to partition data points into K clusters based on similarity. The algorithm aims to group data points that are similar to each other while keeping the clusters as distinct as possible. K-means is an unsupervised learning algorithm, meaning it doesn't require labeled data for training.

Hierarchical Clustering: Hierarchical clustering is a popular unsupervised machine learning technique used for grouping data into a hierarchy of clusters. Unlike K-means clustering, which requires specifying the number of clusters (K) beforehand, hierarchical clustering creates a tree-like structure of nested clusters based on the similarity between data points. It is a bottom-up or agglomerative approach, where each data point initially represents its own cluster and is gradually merged with other similar clusters until all data points belong to a single cluster.

It is of two types:

Agglomerative Hierarchical Clustering

Divisive Hierarchical Clustering

BIRCH Algorithm: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an efficient and scalable hierarchical clustering algorithm designed for large-scale datasets. It was proposed by Tian Zhang, Raghu Ramakrishnan, and Miron Livny in 1996. BIRCH is particularly useful for datasets with a large number of data points, as it can handle such datasets efficiently without requiring multiple scans of the entire dataset. Overall, BIRCH is a useful algorithm for large-scale datasets, especially when memory constraints are a concern, and hierarchical clustering is desirable.

3.2 Libraries

NumPy: NumPy (Numerical Python) is a powerful library in Python used for numerical computations and handling multi-dimensional arrays and matrices efficiently. It is one of the fundamental libraries for scientific computing in Python and provides tools for working with large datasets and performing various mathematical operations**.** Pandas: Pandas is an open-source Python library that provides high-performance data manipulation and analysis tools. It is built on top of NumPy and is designed to work efficiently with structured and labeled data, making it an essential library for data wrangling and data analysis in Python.

Matplotlib: Matplotlib is a popular and widely-used Python library for creating static, interactive, and publication-quality visualizations and plots. It is designed to work seamlessly with NumPy and Pandas, making it an essential tool for data visualization in various scientific and data analysis projects.

Seaborn: Seaborn is a data visualization library in Python that is built on top of Matplotlib. It provides a high-level interface for creating informative and visually appealing statistical graphics. Seaborn is specifically designed for statistical data visualization and makes it easier to generate complex and aesthetically pleasing plots with minimal code.

Sklearn: scikit-learn, commonly abbreviated as sklearn, is a popular and widely-used Python library for machine learning. It is built on top of NumPy, SciPy, and Matplotlib and provides a simple and efficient set of tools for data mining and data analysis. scikit-learn is designed to be user-friendly and accessible to both beginners and experienced machine learning practitioners.

3.3 Characteristics of Data

The dataset used for customer segmentation consists of 2240 rows and 29 columns. The column details are as follows:

ID: A unique identifier for each customer in the dataset.

Year_Birth: Customer’s birth year

Education: Customer's education level

Martial_Status: Customer's marital status

Income: Customer's yearly household income

Kidhome : Number of children in customer's household

Teenhome: Number of teenagers in customer's household

Dt_Customer: Date of customer's enrollment with the company

Recency : Number of days since customer's last purchase

MntWines: Amount spent on wine in last 2 years

MntFruits: Amount spent on fruits in last 2 years

MntMeatProducts: Amount spent on meat in last 2 years

MntFishProducts: Amount spent on fish in last 2 years

MntSweetsProducts: Amount spent on sweets in last 2 years

MntGoldProds: Amount spent on gold in last 2 years

NumDealPurchases: Number of purchases made with a discount

NumWebPurchases: Number of purchases made through the company’s website

NumCatalogPurchases: Number of purchases made using a catalogue

NumStorePurchases: Number of purchases made directly in stores

NumWebVisitsMonth: Number of visits to company’s website in the last month

AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Z_CostContact:

Z_Revenue:

Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

CHAPTER 4

PROJECT WORKFLOW

4.1 Flow Chart

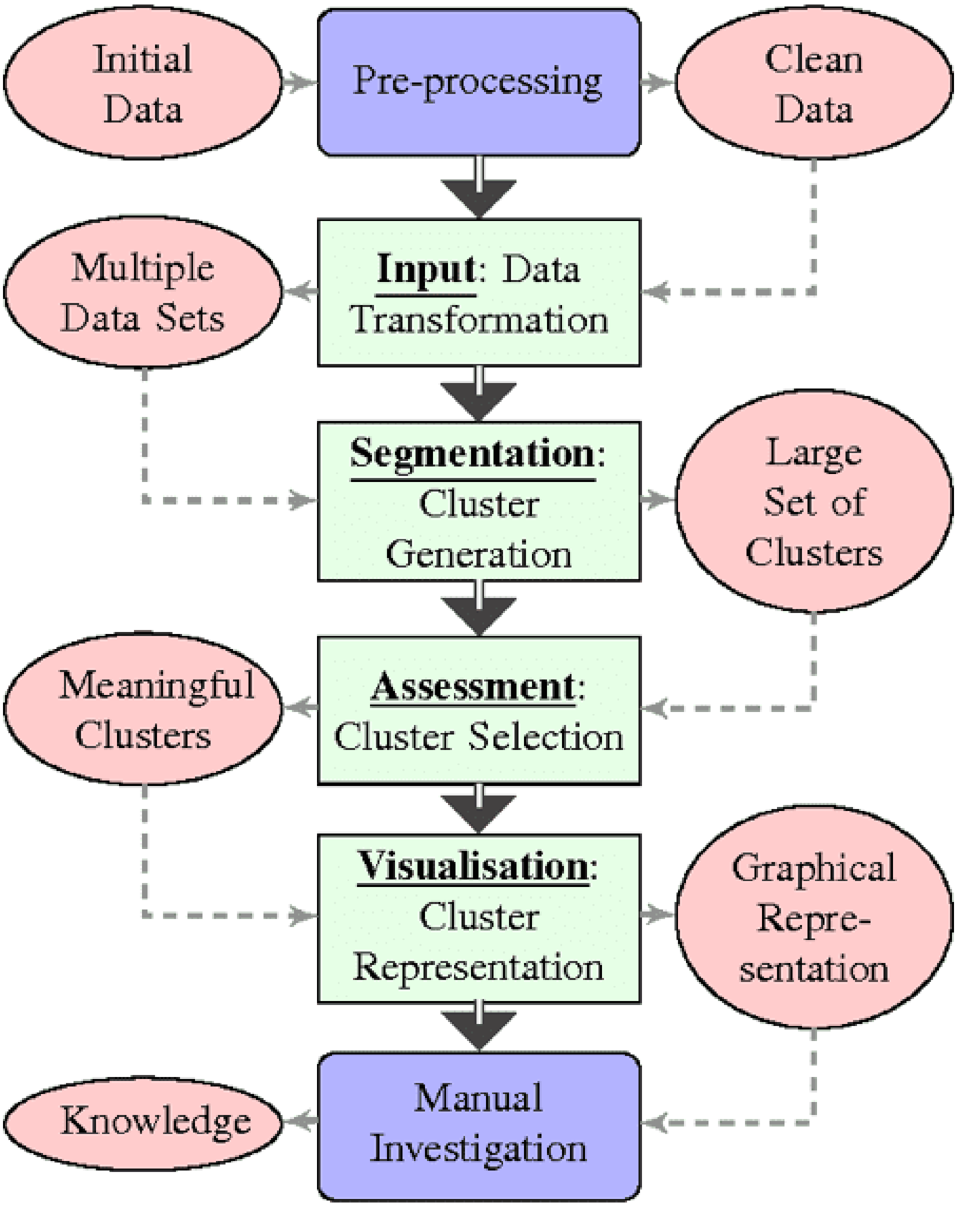
Initial Data
Pre-processing
Clean Data
Multiple Data Sets
Input: Data Transformation
Segmentation: Cluster Generation
Large Set of Clusters
Meaningful Clusters
Assessment: Cluster Selection
Visualisation: Cluster Representation
Graphical Repre-sentation
Knowledge
Manual Investigation

Fig. 4.1 Flow Chart

4.2 Steps Followed

Step 1. Data collection: Gather relevant customer data from various sources, such as demographics, purchase history, website interactions, and customer feedback.

Step 2 .Data PreProcessing: Clean and prepare the data by handling missing values, removing duplicates, and converting categorical variables into a numerical format using techniques like Label Encoding**.**

Step 3. Feature Engineering: Extract and create meaningful features from the data that can be used for customer segmentation. For example, calculate customer lifetime value, frequency of purchases, recency of interactions, etc**.**

Step 4.Exploratory Data Analysis(EDA): Perform EDA to gain insights into the dat and identify potential patterns or clusters that can guide the segmentation process.

Step 5.Apply Clustering Algorithm: Apply different clustering algorithm on the dateset.

Step 6.Comparing the Models: Compare the different models and select the appropriate machine learning algoritm.

Step 7.Evaluting the Clusters: After multiple clusters are formed, compare different clusters according to their demographic behaviours.

Step 8.Conclusion: Write down the conclusion that you have draw from the graph of different clusters.

CHAPTER 5

User Classes and Characteristics

Data Analysts and Data Scientists: These users have expertise in data analysis, machine learning, and statistical modeling. They are proficient in using algorithms like K-means clustering and have the technical skills to preprocess and analyze customer data.

Marketing Managers and Strategists: Marketing managers possess domain knowledge of marketing strategies, customer engagement, and business objectives. They understand the importance of targeted marketing and the potential impact of customer segmentation on marketing campaigns.

Data Engineers: Data engineers are skilled in data integration, data cleansing, and database management. They ensure that the customer dataset is accessible, accurate, and properly formatted for analysis. Data engineers collaborate with data analysts and data scientists to preprocess and prepare the customer data, enabling smooth data flow for the clustering process.

Business Executives and Decision Makers: Business executives hold a strategic role in the organization and focus on the project's alignment with overall business objectives. They prioritize data-driven decision-making to enhance business performance. Business executives use the customer segmentation insights to make informed decisions regarding resource allocation, marketing budget, and growth strategies, ensuring the project aligns with organizational goals.

CHAPTER 6

SWOT ANALYSIS

Strengths

- Improved Customer Understanding
- Personalized Marketing Strategies
- Enhanced Customer Experience

Weakness

- Data Quality Challenges
- Subjective K Selection
- Sensitive to Outliers

Opportunities

- Targeted Marketing Campaigns
- Product Customization
- Market Expansion Opportunities

Threats

- Data Privacy and Security Concerns
- Inaccurate Clustering Results
- Unforeseen Market Shifts

Chapter 7

MISCELLANEOUS

7.1 System Requirement

- 4GB RAM
- Operating Systems(Linux, Windows, macOS)
- Python Platform(Jupyter, IDLE)
- Following Python Packages
  - Numpy for performing mathematical operations.
  - Pandas for data manipulation and analysis.
  - Matplotlib for data visualisation.
  - Scikit-Learn for machine learning models

7.2 Technical Obstacles

- While Scaling , Clusters are overlapping with each other.
- When we applied DBScan algorithm, the number of clusters formed were only one and there was no segmentation.

7.3 Future Scope

The future scope of customer segmentation is promising and evolving, driven by advancements in technology, data analytics, and customer-centric marketing strategies. the future of customer segmentation holds great potential for businesses to better understand and engage with their customers. By leveraging advanced technologies and data analytics, businesses can enhance customer experiences, increase customer loyalty, and drive business growth in a customer-centric and data-driven world.

INDIVIDUAL CONTRIBUTION

AMAN RAJ: Cleaning data, Feature Engineering, Clustering, Evaluation metrics, Model application development & deployment.

VARDAAN KHOSLA: EDA, Models Deployment & Comparison **,** Evaluation metrics calculation and Conclusions.

SIMRAN BHARDWAJ: EDA, Clustering, Model Comparison, Conclusions.

RISHABH KUMAR: Documentation.

DIPTI VERMA: Powerpoint Presentaion.

SHARAD KUMAR AGARWAL: Powerpoint Presentaion.