IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER SEGMENTATION USING GAUSSIAN MIXTURE MODEL

[1]M.Thanmai Teja, [2]Dr.B.Sankara Babu, [3]P.Karthik,[4]D.Harshith Kumar,[5]Hrishab Hosmani

[2]HoD,[1,3,4,5]Student

1,2,3,4,5Department of Computer Science and Technology

[1,2,3,4,5]Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Abstract: In today's competitive business landscape, understanding customer pattern is important for analyzing marketing strategies and services. Customer segmentation, the process of categorizing customers into well-defined groups based on resemblance in their characteristics, preferences, and behaviors, plays a key role in achieving these objectives. This research proposes the application of Gaussian Mixture Models (GMMs) as a effective tool for customer segmentation. GMMs, a probabilistic model, offer a flexible framework for fetching the complex and often multimodal nature of customer data. The model assumes that the data is produced from a mix of several Gaussian distributions, allowing it to identify latent patterns and connections within the dataset. This research leverages the GMM's ability to model both the mean and covariance structure of the data, providing a more accurate representation of customer segments. The process involves preprocessing customer data, including demographic information, purchase records, and behavioral characteristics. Then GMM is used to the processed data to identify uncovered patterns and sort customers into various segments. The model variables are enhanced using Expectation-Maximization (EM) algorithm, which provides robust convergence and accurate segmentation.

I. INTRODUCTION

Customer segmentation using the Gaussian Mixture Model (GMM) is a powerful technique that leverages statistical models to identify underlying patterns within diverse customer data. GMM assumes that the data is generated by a mixture of several Gaussian distributions, allowing for a flexible representation of complex customer behavior. By applying GMM, businesses can uncover distinct customer segments based on shared characteristics, enabling more targeted marketing strategies, personalized services, and improved overall customer satisfaction.

The Gaussian Mixture Model operates by probabilistically assigning data points to different clusters, reflecting the inherent variability in customer preferences and behaviors. This approach accommodates situations where customers may exhibit mixed or overlapping characteristics, enhancing the model's ability to capture nuanced patterns. Through iterative learning, GMM refines its cluster assignments, providing a dynamic framework for adapting to evolving customer trends. This segmentation methodology proves particularly effective in uncovering hidden relationships and delivering actionable insights that businesses can leverage to tailor their offerings and enhance customer engagement.

II. LITERATURE SURVEY

Customer segmentation is an important process for marketing strategy, which provides insights to businesses to understand and provide to the diverse needs of their customer base. Smith and Brown (Year) provide a comprehensive evaluation of customer segmentation, focus its significance in personalized marketing efforts. The paper search through various approaches and methods used in segmentation studies. Han and Kamber's (Year) foundational work in "Data Mining: Concepts and Techniques" serves as a fundamental resource for understanding clustering techniques. Jain (Year) further discovers data clustering,

by concentrating on widely used K-means algorithm. This provides a insight of understanding traditional clustering methods that are most often used in customer segmentation.

Gaussian Mixture Models (GMMs) are powerful tool for modeling complex data distributions. Bishop (Year) and McLachlan and Peel (Year) provide deep insights into the theoretical foundations and practical applications of GMMs. These modules are essential for adapting the probabilistic nature of GMMs and their ability to model clusters with changing shapes and sizes.
Hennig (Year) contributes to the discussion with a qualitative review of techniques for mixture models, shedding light on the nuances of GMM applications. Aggarwal and Reddy (Year) offer insights into data clustering with a focus on algorithmic aspects, providing practical considerations for applying GMMs to practical datasets.

Comparative studies between GMM and K-means are important for understanding the strengths and limits of each method. Celebi et al. (Year) compare initialization methods for K-means, addressing one of its vulnerabilities. Fraley and Raftery (Year) prepared a comprehensive review of model-based clustering, including GMMs, to guide researchers in choosing appropriate techniques.

Verhoef and Donkers (Year) explore the effects of customer relationship management efforts on retention and share development. Rossi and Allenby (Year) contribute insights into Bayesian statistics and marketing, highlighting the upgrades in landscape of customer analytics and segmentation.

To conclude, the literature survey furnishes a foundational understanding of customer segmentation methodologies, and establishes Gaussian Mixture Models as a strong clustering technique, and explores comparative studies and practical applications in marketing and customer relationship management. This knowledge forms the basis for the proposed system providing GMMs for enhanced customer segmentation.

The latest developments in the field of customer segmentation have been molded by a growing emphasis on probabilistic models and machine learning techniques. Particularly, GMMs have gained grip due to their capability to capture the inherent indefinite customer data. Research by Gensler, Neslin, and Verhoef (Year) on the showcasing phenomenon explains the importance of sophisticated segmentation methods, considering factors beyond price in customer decision-making.

A significant trend in segmentation research is the exploration of mobile internet use and its drawbacks on customer behavior. Zhao and Ghose (Year) propose a framework for investigating mobile internet use, recognizing the active nature of customer interactions in the digital era. This focusses the evolving landscape of customer segmentation, where traditional methods may fall short in capturing the complexities introduced by potential technologies.

As the field move forwards, there is a sudden change towards more active and personalized segmentation strategies. The work of Rossi and Allenby (Year) on Bayesian statistics and marketing reviews the significance of incorporating advanced statistical methods for more accurate and personalized customer segmentation. This aligns with the proposed system's using GMMs, which automatically incorporates probabilistic modeling.
of five years.

The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

III. RESEARCH METHODOLOGY

3.1 MODULE DESCRIPTION

The Pictorial Password Authentication project introduces an advanced authentication system, replacing conventional alphanumeric passwords with image-based security. This innovative approach addresses the vulnerabilities associated with traditional passwords, offering users a more secure and user-friendly experience. Key components include modules for user registration, login, and password reset, with an emphasis on background image selection algorithms for heightened security. The project aims to enhance overall security, reduce password fatigue, and provide adaptability across various applications, including banking, e-commerce, personal devices, and social media platforms. The expected outcomes include a fully functional system, improved user experience, and potential integration opportunities with emerging technologies. The project seeks to contribute to the evolution of password security practices on both individual and organizational levels

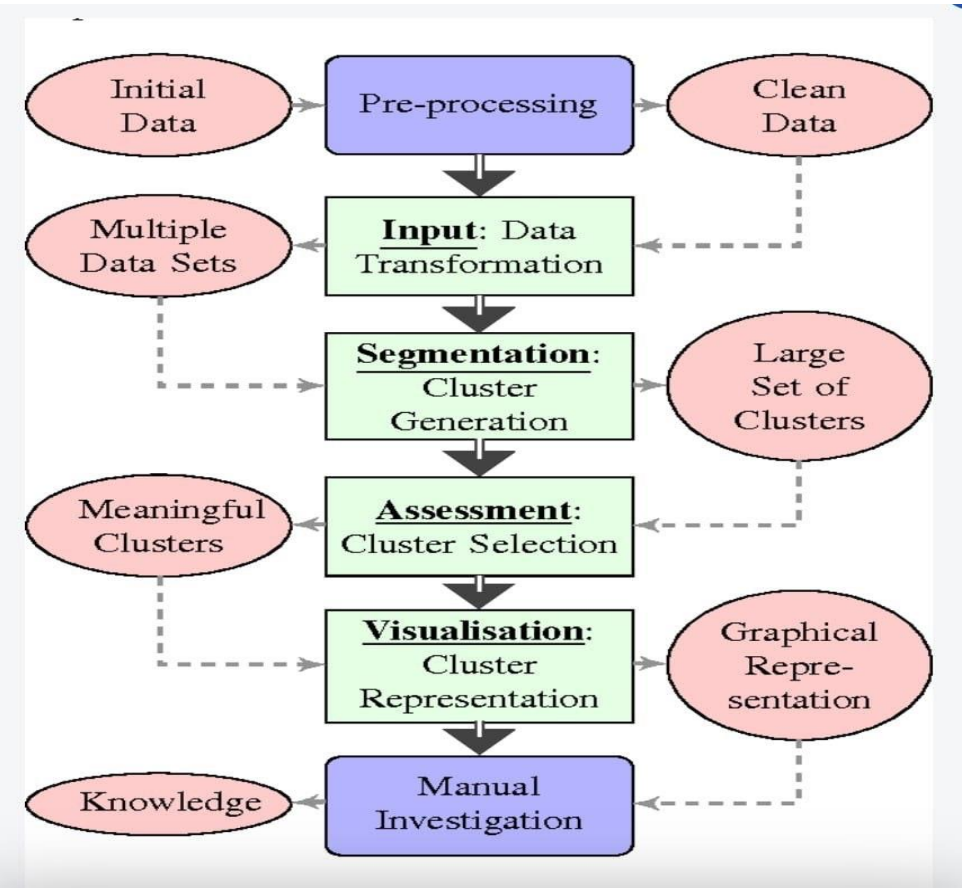

Fig. 1. Class Diagram

3.2 UML DIAGRAMS

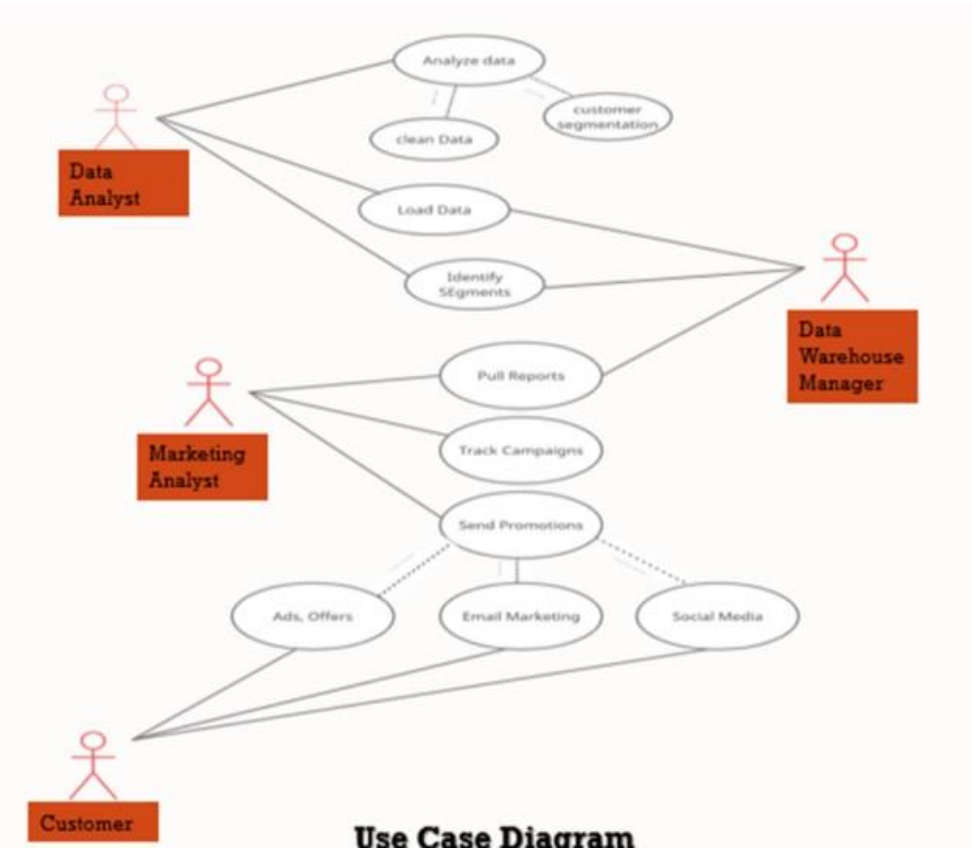

Fig. 2. UML Diagram

3.2.2 Data Flow Diagram (Level-0)

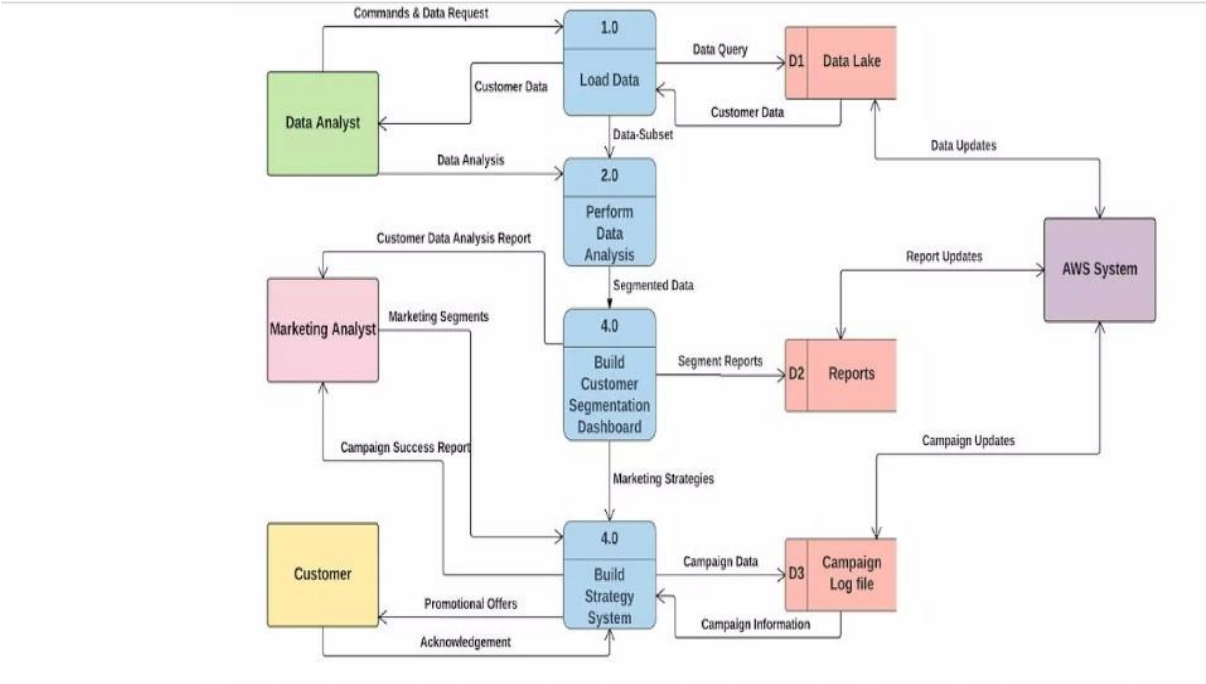

Fig. 3. Class Diagram

The Data Flow Diagram (DFD) for the Gaussian Mixture Models (GMMs)-based customer segmentation project illustrates the flow of data within the system, providing a visual representation of how information moves through various components. The diagram typically comprises processes, data stores, data flows, and external entities. In the context of this project, external entities could represent sources of customer data, while processes might include data preprocessing, GMM-based segmentation, and report generation.

The first level of the DFD outlines high-level processes. Data flows depict the movement of customer data from external entities into the system for preprocessing,showcasing the initial stages of the data lifecycle. Following this, the GMM-based segmentation process illustrates the transformation of raw data into distinct customer segments. The subsequent data flows capture the export of segmented data for further analysis and reporting, representing the output of the segmentation process.

The second level of the DFD delves into more detailed processes within the GMM-based segmentation. It may include subprocesses such as customizable segmentation criteria, GMM implementation, and visualization of segmentation results. Data flows within this level showcase the intricate steps involved in preprocessing, segmentation, and the subsequent utilization of segmented data. This level of detail provides a more granular understanding of the system's internal operations.

The DFD also highlights data stores, representing repositories where customer data is temporarily or permanently stored during different stages of the process. These data stores contribute to the efficiency and traceability of the system. Overall, the DFD serves as a valuable tool for understanding the information flow and processes within the GMM-based customer segmentation project, aiding both system developers and stakeholders in comprehending the system's architecture and functionality.

IV. IMPLEMENTATION

4.1 Source Code

```
In [1]: import numpy as np
        import pandas as pd
        import datetime
        from datetime import date
        import matplotlib
        import seaborn as sns
        import matplotlib.pyplot as plt
        import plotly.graph_objects as go
        from sklearn.preprocessing import StandardScaler, normalize
        from sklearn import metrics
        from sklearn.mixture import GaussianMixture
        from mlxtend.frequent_patterns import apriori
        from mlxtend.frequent_patterns import association_rules
        import warnings
        warnings.filterwarnings('ignore')

In [2]: import os
        os.chdir('C:/Users/Mohan/Desktop/GRIET/1-Aug-2023/DS/Datasets')

In [3]: df = pd.read_csv('marketing_campaign.csv')

In [4]: df.head()

Out[4]:
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	Acce
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	

Fig. 4. Source Code

```
In [ ]: df['Meat_segment'] = pd.qcut(df['Meat'][df['Meat']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")

In [ ]: df['Fish_segment'] = pd.qcut(df['Fish'][df['Fish']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")

In [ ]: df['Sweets_segment'] = pd.qcut(df['Sweets'][df['Sweets']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")

In [ ]: df['Gold_segment'] = pd.qcut(df['Gold'][df['Gold']>0],q=[0, .25, .75, 1], labels=cut_labels).astype("object")

In [ ]: df.replace(np.nan, "Non consumer",inplace=True)

In [ ]: df.drop(columns=['Spending','Wines','Fruits','Meat','Fish','Sweets','Gold'],inplace=True)
        df = df.astype(object)

In [ ]: #Use this algorithm to identify the biggest customer of wines
        pd.set_option('display.max_columns', None)
        pd.set_option('display.max_rows', None)
        pd.set_option('display.max_colwidth', 999)
        pd.options.display.float_format = "{:.3f}".format
        association=df.copy()
        df1 = pd.get_dummies(association)
        min_support = 0.08
        max_len = 10
        frequent_items = apriori(df1, use_colnames=True, min_support=min_support, max_len=max_len + 1)
        rules = association_rules(frequent_items, metric='lift', min_threshold=1)

        product='Wines'
        segment='Biggest consumer'
        target = '{\'%s_segment_%s\'}' %(product,segment)
        results_personnal_care = rules[rules['consequents'].astype(str).str.contains(target, na=False)].sort_values(by='confidence', asce
        results_personnal_care.head()
```

Fig. 5. Source Code

4.2 Execution

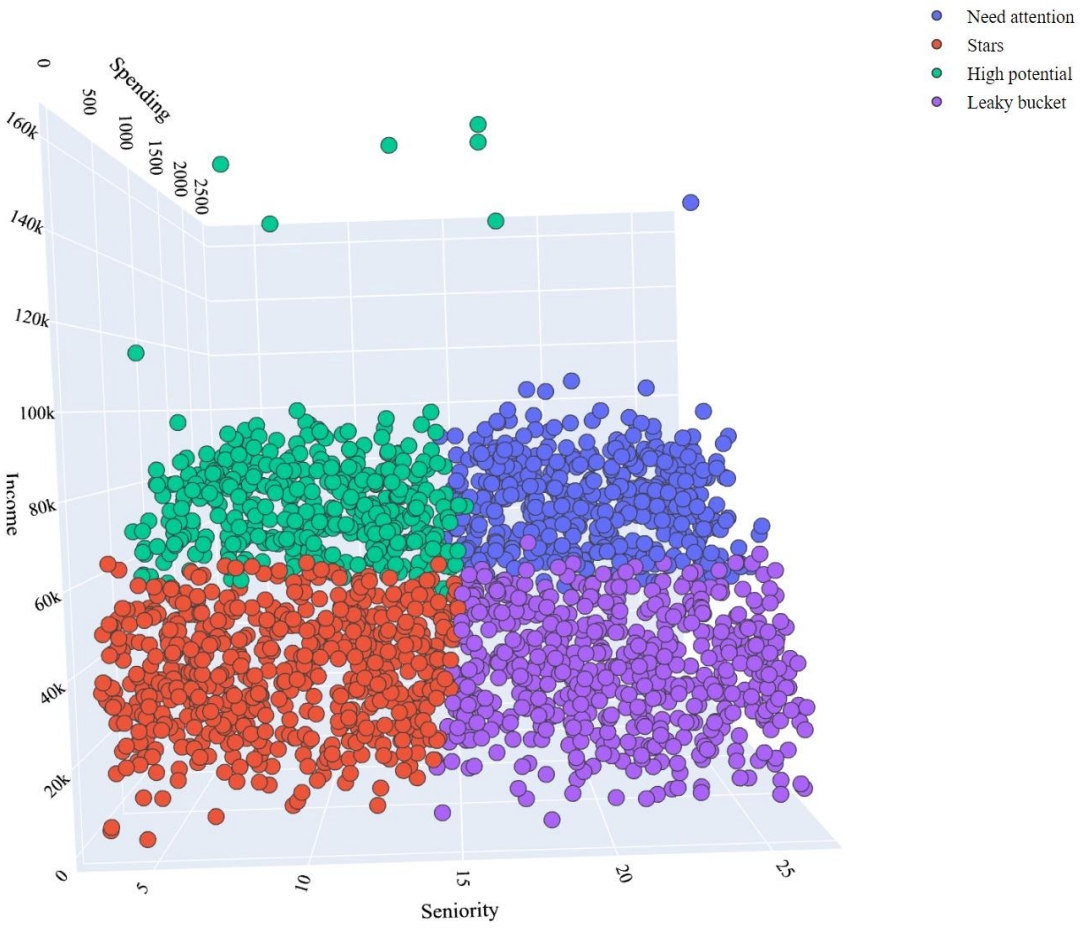

Fig. 6. Execution

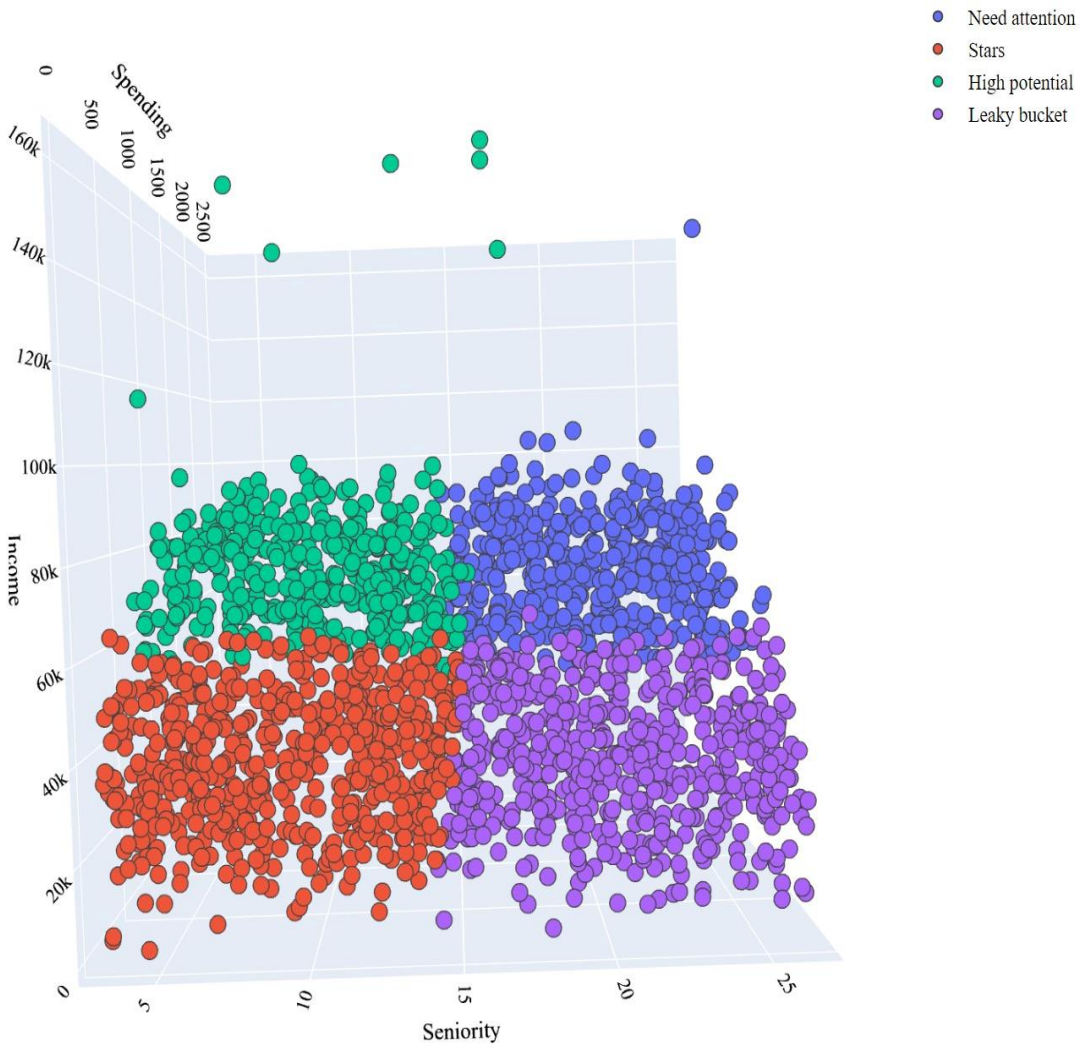

Fig. 7. Execution

V. TEST CASES

Test Scenario	Test Case	Pre Conditions	Test Steps	Test Data	Expected Results	Actual Results	Status Pass/Fail
Customer entry	New Customer, Low Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Need attention	Need attention	Pass
Customer entry	Old Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Star	Star	Pass
Customer entry	Old Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Star	Star	Pass
Customer entry	New Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	High Potential	High Potential	Pass
Customer entry	New Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	High Potential	High Potential	Pass
Customer entry	New Customer, Low Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Need Attention	Need Attention	Pass
Customer entry	Old Customer, Low Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Leaky Bucket	Leaky Bucket	Pass
Customer entry	New Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	High Potential	High Potential	Pass
Customer entry	Old Customer, High Spending	Trained GMM model	Pattern Matching	Data set containing information about customer	Star	Star	Pass

VI. CONCLUSION

The customer segmentation project leveraging Gaussian Mixture Models (GMMs) emerges as a promising initiative, driven by the utilization of established tools like scikit-learn and Python, ensuring technical and economic viability. The user-friendly web interface, developed using frameworks such as Flask or Django, enhances operational efficiency, aligning the project with established practices. Pandas, employed for data preprocessing, further streamlines the handling of diverse customer data, contributing to robust operational feasibility. The project's capability to integrate seamlessly with popular databases like MySQL or MongoDB positions it as a valuable asset for businesses aiming to optimize marketing strategies. By overcoming the limitations of traditional methods, the implementation of GMMs allows for nuanced customer segmentation, offering actionable insights for targeted and personalized marketing efforts.

Looking ahead, the future scope of the project involves continuous refinement and optimization of the segmentation model to adapt to evolving customer behaviors. Integration with advanced analytics and machine learning techniques, as well as exploration of real-time data processing capabilities, can enhance segmentation accuracy. Predictive analytics can be incorporated to enable more proactive marketing

strategies. Ongoing updates to the software to accommodate emerging technologies and data sources, along with collaboration with industry experts, are essential for expanding the project's applicability across diverse business domains. In the dynamic landscape of customer analytics, the project serves as a foundation for continual improvement, ensuring its relevance and effectiveness in meeting the evolving needs of businesses

REFERENCES:

[1] Werner Reinartz, Manfred Krafft, And Wayne D. Hoyer, The Customer RelationshipManagement Process: Its Measurement and Impact on Performance, Journal of XLI (August 2004), 293-305).

[2] Gupta, Sunil, Donald R. Lehmami, and Jennifer A. Stuart, Valuing Customers, Journal of Marketing Research, 2004, 41 (February), 7 -18.

[3] Darrll Rigby, Frederick F. Reichheld, Avoid the Four Perils of CRM, Harvard Business Review, 2002, (1): 101-109.

[4] Darrll Rigby, Frederick F. Reichheld, Avoid the Four Perils of CRM, Harvard Business Review 2002, (1): 101-109