IJMRT: Volume (5), Issue 7, 2023
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN(Online):2663-4007

Customer Segmentation using UMAP and HDBSCAN

Chandrasekaran K S[1] , Jeevanantham T[2*], Jemima Blessy R[3], Jesila foumiya Z[4]

[1]*Associate Professor, Department of Computer Science Engineering, Saranathan College of Engineering, Tamil Nadu, India.*
[2,3,4]*Student, Department of Computer Science Engineering, Saranathan College of Engineering, Tamil Nadu, India.*

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.7934909

Abstract

Customer segmentation is now a very common strategy for keeping customers happy and generating revenue for businesses. Customers from various organisations are categorised in this project based on behavioural traits including their ability to spend money and their income. In order to categorise the customers and create clusters, a computer algorithm known as UMAP and HDBSCAN is used. With the aid of these clusters, the business is better able to target specific clients and promote to them via advertising campaigns and social media platforms about content in which they have a genuine interest.Unsupervised machine learning techniques, such UMAP and HDBSCAN, have grown in popularity in recent years for client segmentation.Using UMAP, high-dimensional data can be reduced to low-dimensional space while retaining its structure and relationships.HDBSCAN is a density-based clustering technique that finds clusters of varied sizes and forms, and assigns each data point to a cluster or a noise category. Because they allow for quick and precise customer grouping, UMAP and HDBSCAN offer a potent combination for customer segmentation.

Keywords: Customer Segmentation, UMAP, HDBSCAN, Clustering, Machine Learning; Visualization; Preprocessing, Correlation of variables.

1. Introduction

Customer segmentation is a marketing technique that involves breaking a company's consumers into groups or segments based on shared features, habits, or needs. With the use

of customised marketing tactics and product offerings that better suit their requirements and preferences, particular groups of clients can be identified and targeted.Customer segmentation comes in a variety of forms, each based on a different set of factors. Customers are segmented based on demographics like age, gender, income, and education level. Psychographic segmentation is based on characteristics such as values, interests, and personality. While behavioural segmentation is based on consumer behaviour, such as purchase patterns or product consumption, geographic segmentation differentiates clients depending on their location.

Insights and patterns can be discovered by analysing customer data using machine learning approaches. Models using artificial intelligence are effective tools for decision-makers. They are able to accurately identify customer segments, which is much more difficult to do manually or using traditional analytical techniques.

The capacity of UMAP to handle noisy and complicated datasets is one of its advantages. It is also very flexible, letting users to change variables like the number of neighbours and the distance metric used to determine how similar data points are to one another.

UMAP's primary objective is to visualise highly dimensional data. The dimension is decreased by 2 or 3 to achieve this. However, because UMAP is a dimension reduction technique, it can be utilised to make the data more understandable and hence improve the outcomes of clustering algorithms.The HDBSCAN approach is flexible and may be used with a variety of datasets. It is particularly helpful when the number of clusters is unknown in advance. Based on the density and dispersion of the data, the programme automatically calculates the number of clusters.

Because it may identify groups of customers based on their behaviours, preferences, and other pertinent factors, HDBSCAN can be helpful in customer segmentation. The algorithm can recognise clusters of various densities and forms and can handle noisy and high-dimensional data, which is frequently the case with consumer data. This means that HDBSCAN can still find valuable client segments even when the customer data is complicated and heterogeneous.

Information on customers or users of a good or service is gathered in user data sets. These data sets are useful for many things and essential for client segmentation. The main characteristics of this project may include the client's income, the amount spent on each item, the client's year of birth, their educational background, and their marital status. Minor characteristics include the number of days since the customer's last purchase, the number of children and teenagers living in the family, the number of online transactions, and sales during deals and promotions.

2. Literature Survey

2.1. Customer Segmentation

The use of machine learning for customer segmentation was suggested by Nikhil Patankar et al. in 2021. The clients of the organization are segmented in this article based on behavioral (product categories ordered, annual income) and demographic (age, gender, and marital status) factors. Behavioral factors are a better technique for customer segmentation since they focus on individuality and allow for accurate segmentation, whereas demographic indicators do not emphasize the individuality of customers because the same age groups may have different interests.

The use case diagram of the suggested system, which includes the four users Data Analyst, Marketing Analyst, Data Warehouse Manager, and Customer, was highlighted in this article.

Additionally, it offers six use cases, including data analysis, data loading, segment identification, campaign tracking, and promotion sending. Based on the similarity determined using the Euclidean Distance, they create clusters using the K-Means algorithm.

This machine learning-based client segmentation was also suggested by Varad R. Thalkar (2021). He split the customer's data into categories based on a variety of variables, including geographic location, economic situation, demographics, and behavioural trends. demographic data, including gender, age, marital status, family situation, income, and employment. Geographical data that varies depending on the company's scope. If a firm is localized, this information may apply to particular towns or counties. Larger businesses may interpret this to signify a customer's city, state, or country of residence,psychological characteristics including social status, way of life, and personality features. Data about human behaviour, including patterns of consumption and expenditure, use of goods and services, and intended outcomes. According to V.Vijilesh et al., (2021), common segmentation types include demographic, RFM (Recency, Frequency, Monetary) analysis, HVCs (High-Value Customers), customer status, behavioural, psychographic, etc. Marketing strategy, promotion strategy, budget efficiency, product development, and others are some of the primary advantages of client segmentation. They used the fundamental analytics capabilities in this article to give the decision-makers—in this example, company investors—the data they needed to decide the right course of action. In this post, they provide a method for lowering risk variables and offer input on how to choose new company investments.

In her work, Banu Turkmena (2021) conducted a comparative analysis of several client segmentation methodologies based on online retail data. We compare the insights provided by

a few unsupervised machine learning (ML) clustering models, including the K-means clustering model, the hierarchical clustering model, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) model, and the recency, frequency, and monetary (RFM) clustering model. This post seeks to offer a fresh perspective on the long-standing consumer segmentation issue. Any business that wants to make informed judgements about pricing and demand forecasting must start by segmenting its customer base.K-means clustering is not the obvious or best option, so a different approach is investigated using this dataset. Jian Zhoua et al. (2020) proposed a novel market segmentation methodology that successfully combines two well-documented approaches and is tested for addressing our key concerns regarding how businesses should determine which customers are profitable and what the preferred market should be when they concentrate on those customers' needs and preferences in an effort to keep them on board with their business lines. In particular, the entropy approach is utilised to ascertain customer values, and the well-known Pareto principle is implemented to find the clients who can be more profitable for the company.This paper compares the BCBimax method to the approach based on the SKC algorithm. We can verify the accuracy of the outcomes in this way.

3. Proposed System Architecture

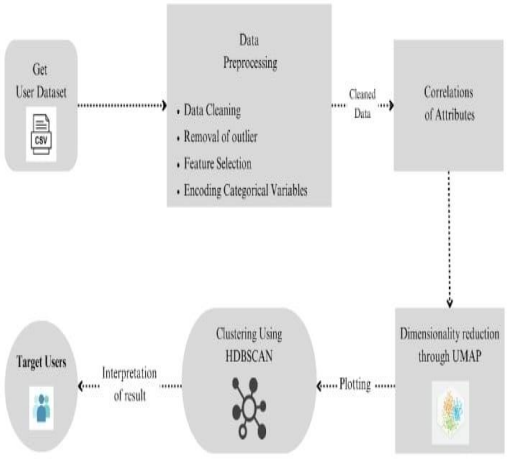

Figure.1. Proposed System Architecture

Customer segmentation is the process of dividing a company's customer base into distinct groups or segments based on certain criteria. System architecture for customer segmentation involves the design and implementation of a technical framework that enables effective segmentation of customers.The architecture starts from the process of getting the dataset as a CSV file ,preprocessing the datasets by data cleaning, removal of outliers, feature extraction and encoding the variables. Then The variables can be correlated and with that relation the clusters can be formed.It can be interpreted for final results and can be provided to the target users.

4. Modules Description

4.1. User Dataset

A user dataset is an assortment of information on patrons or clients of a company or organization. These details could include user demographics, purchasing history, website activity, and more. Customer segmentation in a user database entails classifying customers into several categories based on shared traits, interests, and behaviors. Its goal is to provide the raw data needed to create the segmentation model by storing and organizing consumer data in a systematic and consistent manner.The dataset was acquired via marketing campaigns as a CSV (comma separated values) file.This project's dataset was obtained from kaggle. In general, customer segmentation using a user dataset is a crucial method for firms to learn more about their clients and develop more specialized and successful marketing plans. Businesses can better understand their consumers' needs and adapt their products and services to match those needs by segmenting users based on their behavior and preferences. This increases customer happiness and loyalty.

The sample dataset are represented in the form of table with major attributes and values. The table is shown as follows,

Table.1. Sample Dataset

Year_birth	Income	Gender	KidHome
1987	80000	Male	4
2000	100000	Female	2
1990	75000	Female	1
1997	90000	Male	3
1970	60000	Male	2

4.2. Data Preprocessing

Data preprocessing, which involves cleaning and preparing the data for use in machine learning algorithms, is a crucial stage in customer segmentation. Modules for data preprocessing include a variety of functions for data cleansing, data transformation, data integration, feature selection, and data reduction, as well as the processes of data scaling and outlier elimination. Data cleaning modules work to make sure the data is accurate, consistent, and that any missing values are handled correctly. Data integration modules generate a comprehensive dataset by combining data from many sources, which is essential for consumer segmentation.

4.3 Correlation of Variables

In this phase of the module, data is gathered and analysed to find significant patterns and interactions between variables. Based on this data, targeted marketing strategies are then developed that are more likely to be accepted by each segment's target market. Machine learning may be used in a variety of ways to figure out how the association between the variables in a customer segmentation are related. Data preparation, feature selection, correlation analysis, visualisation, feature importance, dimensionality reduction, clustering

algorithms, and evaluation are some general procedures that can be applied. It explains how the attributes are related. Positive correlation is shown by a value over zero, whereas negative correlation is indicated by a value below zero

4.4 HDBSCAN

A clustering approach called HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is used in machine learning to segment customers. It is a module for density-based clustering that organises data points according to their density and separation from one another. In contrast to conventional clustering techniques, HDBSCAN does not require the number of clusters to be predefined, making it a more adaptable and versatile module.

4.5. Visualization

The data can be shown via visualisation modules in a variety of ways, such as scatter plots, bar charts, and heatmaps, making it simpler to spot groups of clients with similar purchasing patterns. either traits or actions. It is a data preservation technique that can maintain both global and local structure. strong customer segmentation tool. Another nonlinear system is UMAP. A technique for dimensionality reduction that can be applied to visualisation clustering.

5. System Requirements

5.1. Hardware Requirements

1. CPU: A multi-core processor with a clock speed of at least 2.5 GHz is recommended to handle the computational load of training and testing the model.

2. RAM: The amount of RAM required will depend on the size of the dataset and the complexity of the model. At least 8GB of RAM is recommended, but for larger datasets, you may need more.

3. GPU: A GPU can significantly speed up the training process of the model. A high-end GPU, such as NVIDIA GeForce GTX or AMD Radeon, with at least 4GB of video memory is recommended.

4. Storage: The dataset can be quite large, so it's essential to have enough storage space to store the data and trained models. A minimum of 100GB of storage space is recommended.

5. Other considerations: Cooling systems and power supplies may be necessary to support the hardware and prevent overheating.

5.2. Software Requirements

5.2.1. Programming Language

The system can be implemented using python programming language.

5.2.2. Machine Learning Libraries

The system will require machine learning libraries such as pandas, seaborn,matplotlib, HDBSCAN,UMAP.

5.2.3. Development Environment

The system can be developed using Jupyter.

6. System Implementation

6.1. Introduction

The proposed work entails defining customer segmentation criteria, such as transaction history, purchase behaviour, website activity, and more; collecting and preprocessing data, which includes scaling the data and data cleaning; applying the UMAP algorithm to reduce the dimensionality of the data; determining the optimal number of clusters using HDBSCAN; creating customer segments; developing marketing strategies for each segment; testing and refining the segment; and finally, implementing the proposed work. By using this algorithm, businesses can improve the results of their marketing initiatives and more closely meet the needs of their customers.

6.2. Proposed System Implementation

The following methods are used to carry out the Customer Segmentation Project:

6.2.1. Data Collection

Collecting pertinent information about your consumers, such as their demographics, purchasing patterns, and preferences, is the first stage in putting a customer segmentation strategy into practise. Numerous sources, including consumer surveys, purchase histories, and website analytics, can be used to get this information.The dataset is a 29-attribute.CSV file that was downloaded from Kaggle.

6.2.2. Algorithm for Data Collection

1. Identify segmentation criteria,These can include demographic information (age, gender, location), psychographic traits (interests, values, lifestyles), behavioral data (purchase history, engagement), or other relevant factors.

2. Gather existing customer data, This may include transactional data, customer profiles, feedback, or interaction history.

3. Organize this data and ensure its accuracy. This data may be stored in customer relationship management (CRM) systems, databases, or other sources.

4. Import the necessary libraries for the implementation of the project

5. Import numpy, an open source Python library. It can be utilized to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

6. Import pandas , a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

7. Import matplotlib, a comprehensive library for creating static, animated, and interactive visualizations.

8. Import seaborn, a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

9. Get the user dataset as a .csv file with the information of customers along with the attributed value.

10. After importing all the necessary libraries , write a line of code to read the dataset.

11. The function that can be used is data= pd.read_csv()

6.2.2. Data Preprocessing

To get rid of any discrepancies or inaccuracies, the data needs to be cleaned and preprocessed after it has been gathered. Data normalisation, outlier reduction, and missing value imputation may be involved in this.The Drop() technique is employed to eliminate the unnecessary data.

A. Algorithm for DataPreprocessing

1. Use the method of data.head ()

2. This is commonly used to display the first few rows of a DataFrame or dataset.

3. Use the data.shape attribute

4. It returns a tuple containing the number of rows and columns in the dataset.

5. Data Cleaning

5.1. Remove duplicate records, if any, from the dataset.

5.2. Handle missing values by either deleting the corresponding records, replacing them with mean or median values.

5.3 . Use the function data.isna().sum()

5.4. Replace the missing values and fill the empty cells using median() method

5.5. Create an instance as filler = data["Income"].median

5.6.Import the zscore to identify and handle the outliers .

6. Feature selection

6.1. Remove irrelevant or redundant features that do not contribute significantly to the segmentation process.

6.2. To remove the unwanted data like ID of the customer, Date of customer enrollment to the company, use the drop() method.

6.3.Call the method as data.drop(["Z_Revenue","Z_CostContact", 'Dt_Customer'], axis=1, inplace=True

7. Encoding categorical variables

7.1.Take the categorical variables such as marital status and education.

7.2. Encode it to the numericals by creating an encoder instance.

7.3. Use the following sample equation as

edu_encoder = {"Basic":0, "2n Cycle":1, "Graduation":2, "Master":3, "PhD":4}

data["Education"].replace(edu_encoder, inplace=True)

6.2.3. Correlation of Variables

The categorical variables and numerical values from the attributed data are presented as separate graphs alongside the records. The corr() technique is used to correlate the plotted variables for each characteristic.It determines how much one variable will change as a result of the other variable changing.

The first step is to utilise Matplotlib's plot function to build a new figure. Utilising Seaborn's heatmap() method, construct a heatmap next. The correlation matrix of the digit columns is shown in the heatmap. Create a function with the name cmap="coolwarm_r"

This changes the colour map to "coolwarm_r," a colormap with a blue (negative correlations) to red (positive correlations) colour spectrum. The very least value is -1. Values over zero denote positive correlations, whereas values below zero denote adverse correlations. There is no association when the value is zero.

6.2.4. UMAP Embedding

To visualize the clusters UMAP is used. It's like T-SNE but better at preserving global structure, and much faster. It reduces the dimensionality of feature space to 2 or 3 dimensions.

So that we can easily interpret the results.The UMAP is imported using the import function to the code.

A. Algorithm for UMAP

1. Import UMAP package by giving the comment 'import umap'

2. Use the umap package to perform UMAP dimensionality reduction on a dataset named scaled data

3. Initialize the UMAP class with the parameters random_state and n_components. random_state sets the random seed to ensure that the results are reproducible, while n_components Specifies the number of dimensions of the resulting UMAP embedding.

4. Then Call the fit transform method on the UMAP object to compute the low-dimensional embedding of the scaled data dataset.

5. umap_data=ump.fit_transform(scaled data)

6. Import the plotly express package to create a 3D scatter plot of the UMAP embedding stored in the umap_data variable.

7. Call The scatter_3d function with the parameters x, y, and z to specify the coordinates of the points in the scatter plot. Here, x represents the first dimension of the UMAP embedding, y represents the second dimension, and z represents the third dimension.

8. Set The title parameter to set the title of the scatter plot.

Now, a 3D scatter plot that shows the distribution of the data in the UMAP embedding can be visible. The plot can be interacted with to rotate and zoom in/out to examine the distribution of points from different angles.

6.2.5 HDBSCAN Clustering

The customer data is clustered using the hierarchical technique HDBSCAN based on their UMAP embeddings. Different clustering factors, such as the minimum cluster size and minimum sample size, are tested with.Similar to DBSCAN, but without the agonizing radius (epsilon) adjustment.Utilising "pip install hdbscan," it is imported.

A. Algorithm for HDBSCAN

1. Import the HDBSCAN module to perform hierarchical density-based spatial clustering of applications with noise (HDBSCAN).

2. Create an instance of the HDBSCAN class with the desired parameters (min_cluster_size and min_samples).

The instance is clustered= hdbscan.HDBSCAN(min_cluster_size=50, min_samples=150)

3. Call the fit method on the clusterer object to perform the clustering on the umap_data.

4. Use the scatter_3d function from the plotly_express module to create a 3D scatter plot.

5. Create the coordinates for the three axes (x,y,z) of the scatter plot. umap_data is a NumPy array or a similar data structure, and [:,0], [:,1], and [:,2] represent the values from the first, second, and third columns respectively.

6. Specify the color of each data point in the scatter plot as color=clusterer.labels_.astype(str)

7. Create an array of cluster labels as clusterer.labels , and convert the labels to strings using astype(str).

8. Set the title of the scatter plot to using the object title.

This phase creates a 3D scatter plot using the scatter_3d function, with data points positioned based on their coordinates from umap_data and colored according to the cluster labels from clusterer.labels_.

6.2.5. Visualization

To better appreciate their spatial linkages, the clusters are visualised in 2 or 3 dimensions. The clustered data that are imported at the beginning are plotted using programmes like matplotlib and seaborn.

The first step is to generate a figure with the necessary number of rows of subplots.The size of the figure is determined by the figsize option. For the subset of values in c where clusterer.labels_ equals the cluster values with the appropriate colour, make a histogram using,[seaborn.Sns.histplot(x=c[clusterer.labels_==0],color="purple", ax=axes[0])] is an example of a line of code.

The information obtained from the cluster analysis guides business decisions such as targeted marketing campaigns, product development, and client retention strategies.The results are interpreted to form the summary of the clusters.

7. Results and Discussion

The usage of UMAP and HDBSCAN for customer segmentation can produce data on consumer behavior and preferences that can be utilised to develop targeted marketing efforts and improve customer satisfaction. The findings of this project were interpreted as the following results:

7.1. Identification of Customer Segments

Customer segments are identified based on the attributes of the dataset. The graph is represented in the following Fig. 6.1 and 6.2.

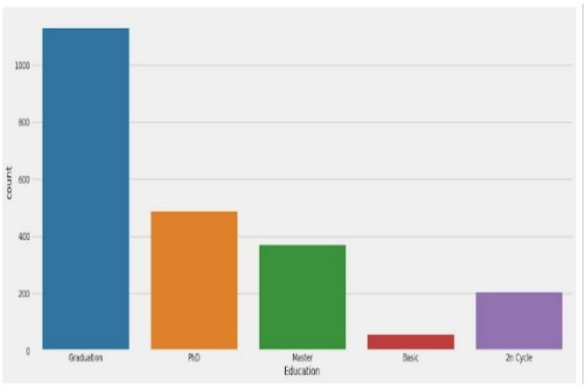

Figure.2. Identification of Customer Segments

7.2. Preprocessing of data

In order for machine learning models or other analytical tools to use the data efficiently, a number of approaches are used to clean, convert, and normalize the data. The accuracy and effectiveness of the analytical models are improved by preprocessing the data.The outliers are removed and the graph after the removal of outlier are shown in the figure.

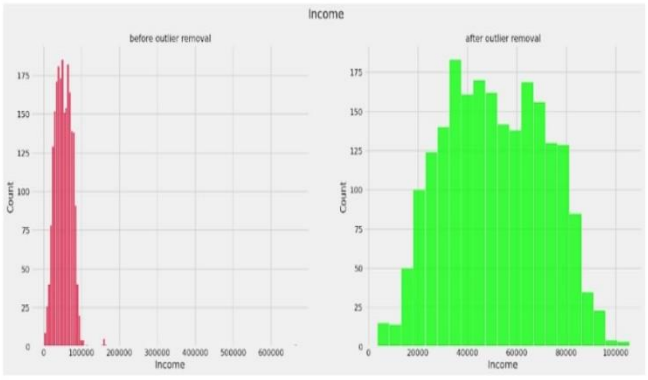

Figure.3. Representation of Outlier Removal

7.3. Correlation of Variables

Because it may be used to find patterns and correlations in the data, correlation is significant in data analysis. For instance, correlation analysis in business can be done to find out whether there is a connection between client age and purchasing behaviors.The correlation between all the attributes are shown as a heat map in the following figure.

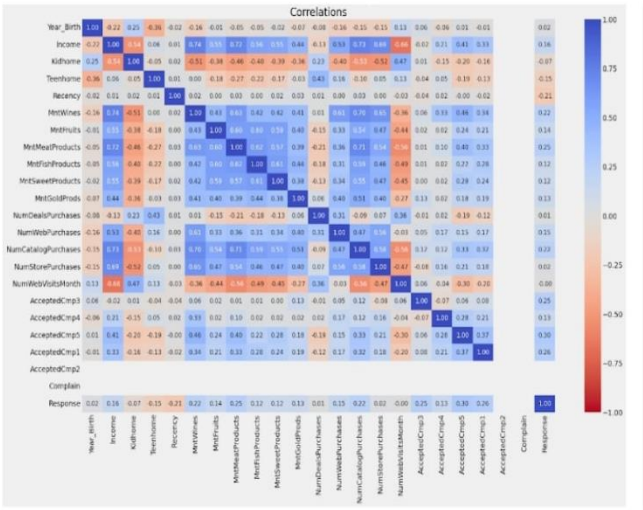

Figure.4. Correlation of Variables

7.4. Interpretation & Clustering using UMAP and HDBSCAN

UMAP and HDBSCAN can be used in conjunction to analyse large datasets and find patterns and connections between the data points. These methods can be used together to create data visualizations and exploratory analysis.The datasets are formed into different clusters and are interpreted.The formed clusters are shown in the below figure .

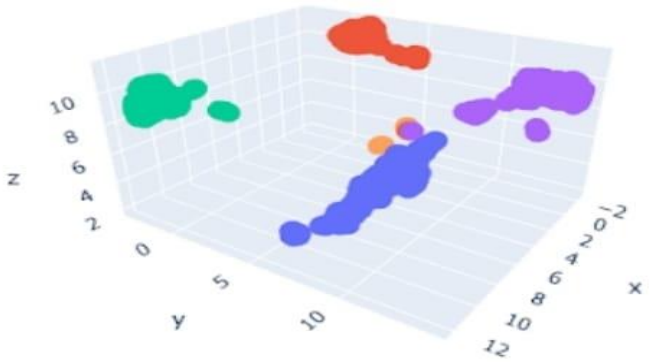

Figure.5. The Three Dimensional Clustering Structure

The results are interpreted as a graph for all the clusters, and based on the graph the strategies can be incorporated. The final interpretation is shown in the following figure

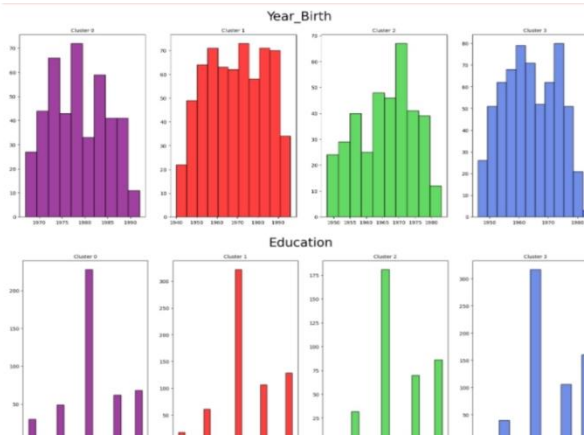

Figure.6. The interpretation of the Clusters

Summary

- People born in 1970 and graduate form the cluster and have the highest rate of purchases.
- The yellow cluster doesn't need to be taken into consideration.

8. Conclusion and Future Work

Strong machine learning algorithms that can be used in this project for client segmentation include UMAP and HDBSCAN. By examining consumer behaviour and preferences, businesses may raise overall profitability and boost customer happiness. While HDBSCAN can find groups of related data points in a dataset, UMAP can be used for dimensionality reduction, anomaly detection, and personalized recommender systems. We could anticipate even more creative applications in the area of customer analytics as these approaches develop, like real-time customer segmentation and integration with other data sources. Overall, UMAP and HDBSCAN have the power to change the way businesses view and interact with their clients.The accuracy, scalability, and application of consumer segmentation utilizing UMAP and HDBSCAN can be increased through method exploration in subsequent work, resulting in more successful marketing campaigns and higher levels of customer satisfaction.

REFERENCES

[1]. Blanco-Portals et al., (2021). Tailoring the Transport Properties of Mesoporous Doped Cerium Oxide for Energy Applications. The Journal of Physical Chemistry C, 125(30), 16451-16463.

[2]. Becht, A. et al.,(2022). Longitudinal associations between social media use,mental well-being and structural brain development across adolescence.Developmental Cognitive Neuroscience, 54, p.101088.

[3]. Bogensperger, A., Fabel. Y, (2021). A practical approach to cluster validation in the energy sector. Energy Inform Volume 4 (Suppl 3),Article 18

[4]. Jian Zhou, Linli Zhai, Athanasios A. Pantelous (2020), Market segmentation using high-dimensional sparse consumers data, Expert Systems with Applications, Volume 145.

[5]. Kovács et al.,(2021) "Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis"Journal of Big Data Volume 8, pp 1-25.

[6]. McInnes et al.,(2022). "A practical review of electroencephalography's value to consumer research." International Journal of Market Research 65.1 52-82.

[7]. Nikhil Patankar & Dixit, Soham & Bhamare, Akshay & Darpel, Ashutosh & Raina, Ritik. (2021). Customer Segmentation Using Machine Learning.10.3233/APC210200.

[8]. Om Atre, Shrinil Modhave, Prasad Torane, S.B.Jadhav (2022),"Customer Segmentation for Banking Strategy using Machine Learning" ,International Journal of Creative Research Thoughts (IJCRT) Volume 10, Issue 12 December.

[9]. Seshashayee. M and Srinivas Dileep (2022)," Customer segmentation using machine learning "International Research Journal of Modernization in Engineering, Technology and Science Volume:04/Issue:05/ Impact Factor- 6.752.

[10]. Shirole, Rahul, Laxmiputra Salokhe, and Sarasw(2021). "Customer Segmentation using RFM Model and K-Means Clustering." Int. J. Sci. Res. Sci. Technol 8 591-597.

[11]. Turkmen, Banu. "Customer Segmentation with Machine Learning for Online Retail Industry." The European Journal of Social & Behavioural Sciences(2022).

[12]. Varad R Thalkar (2021), "Customer Segmentation Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307,Volume 7, Issue 6, pp.207-211.

[13]. Vijilesh .V et al., (2021)" Customer segmentation using machine learning "International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056,Volume: 08 Issue: 05 , May.