IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 5, November 2024

Customer Personality Analysis

Sumit A. Hirve[1], Nilesh N. Thorat[2], Sakshi Tambe[3],
Vaishnavi Naidu[4], Aditya Upadhye[5], Shivraj Yadav[6]
Students, Department of Computer Science Engineering[1-5]
Guide, Department of Computer Science Engineering[6]
MIT Arts Design and Technology University, Pune, India

Abstract*: One kind of machine learning technique that can be used to examine customer data and find recurring themes and characteristics within a collection of consumers is called unsupervised learning. Businesses can build client groups with unique personality traits by employing clustering techniques to group like consumers together. This enables them to better target each segment with their marketing and sales activities. Large volumes of data may be used to train unsupervised learning algorithms, which makes them effective tools for businesses trying to better understand and cater to their clientele.*

Keywords: machine learning.

I. INTRODUCTION

The process of determining and comprehending the distinctive qualities and attributes that comprise each individual client's personality is known as customer personality analysis. Businesses can utilize this data to better target and cater to the unique requirements and preferences of each consumer by customizing their marketing and sales intiatives. Marketing and sales teams have historically conducted customer personality analyses by hand, using their knowledge and experience to find recurring themes and trends among clients. But because to the development of data mining and machine learning, this procedure can now be automated with algorithms that can examine vast volumes of data and find recurring patterns and characteristics among clients.

One kind of machine learning algorithm that can be used for customer personality analysis is unsupervised learning, which is especially well suited for customer personality analysis because it can find subtle differences and trends that may not be immediately obvious to humans.

The dataset for this project is a public dataset from Kaggle provided by Dr. Omar Romero- Hernandez. This dataset has 2,240 rows of observations and 28 columns of variables. Among the variables, there are 5-character variables and 23 numerical variables. In this paper, we will discuss the use of unsupervised learning algorithms for customer personality analysis, and how they can be used by companies to better understand and serve their customers. We will also provide an overview of the different types of unsupervised learning algorithms and how they work, as well as discuss some of the challenges and limitations of using unsupervised learning for customer personality analysis.

II. METHODOLOGY

Data Preprocessing:

Data cleaning refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Data cleansing can be done in batches using scripting or a data quality firewall, or it can be done interactively with data wrangling tools. Data cleaning is an essential step in the customer segmentation process, as it helps to ensure that the data used for analysis is accurate and reliable. To clean data for customer segmentation, you will need to identify and address any errors, inconsistencies, and missing values in your dataset. This can be done by manually reviewing the data, using visualizations, or applying statistical tests.

Steps taken for preprocessing:

1. The data frame's income column contains 24 missing values; since these are fewer, the missing numbers are being removed.

DOI: 10.48175/IJARSCT-22441

2. We are developing a feature that will use the "Dt Customer" data to show how long a customer has been in the company's database.
3. To make the data more consistent and clear, I added a new column called "Living With," which has the same meaning as an existing column with a different name.
4. The "Is parent" feature has been put into place, and it returns a value of 1 for parents and 0 for non-parents.
5. Include a new feature called "Spent" that shows the total amount spent by the consumer over a two-year period in all categories.
6. Removing some unnecessary features.
7. Plotting for "Income" and "Age" reveals outliers, which will be eliminated.
8. Performing label encoding for descriptive attributes, such as "Education" and "Living With."

The data is quite clean, and the new features have been included

PCA:

One method for lowering the dimensionality of these datasets is principal component analysis (PCA), which improves interpretability while minimizing information loss.

This challenge's final categorization will be determined by a variety of factors. In essence, these criteria are these aspects or characteristics. The more features it contains, the harder it is to work with. Some of these features are superfluous because of their association. Because of this, I will lower the dimensionality of the characteristics before putting them through a classifier.

The dimensions are being shrunk to three. Plotting these data frames will come next.

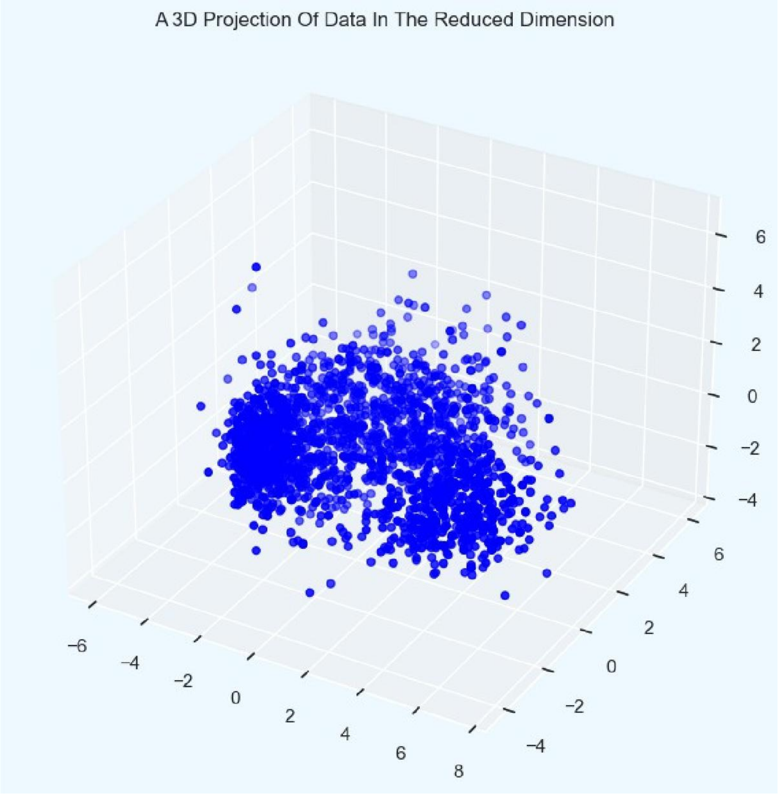

Figure 1: 3D plot of data after PCA

After principal component analysis (PCA) processing, the data is shown in three dimensions in the previously described picture. This projection allows us to see the data in a way that is easier to understand and analyze.

III. METHODS

Hierarchical Clustering

For this project, hierarchical clustering will be used. Agglomerative clustering is the most common hierarchical clustering technique that groups objects according to how similar they are. AGNES (Agglomerative Nesting) is another term for it. The program first treats each object as a singleton cluster. Pairs of clusters are progressively integrated when all clusters have been combined into a single, sizable cluster that contains every item. The result is a tree-based representation of the objects called a dendrogram. Identifying different client personality types and focusing marketing and communication efforts on particular customer groups would both benefit from this.

K-Means Clustering

K-means is a centroid-based clustering algorithm where each data point is assigned to a cluster based on the distance between it and a centroid. Finding the K number of groups in the dataset is the objective. The goal of K-means clustering, a vector quantization technique that originated in signal processing, is to divide n observations into k clusters, where each observation belongs to the cluster that has the closest mean, which acts as the cluster prototype. With the use of k-means we will be finding clusters. In total we will be selecting 4 clusters.

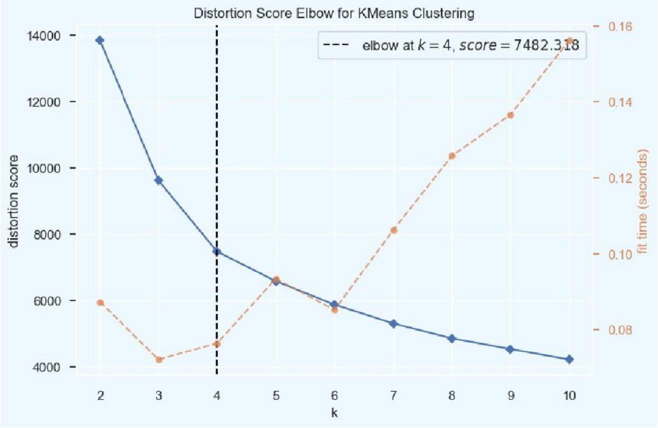

Figure 2: Elbow method visualization

We will use agglomerative clustering to carry out the clustering after the attributes have been reduced to three dimensions. Agglomerative clustering is a hierarchical clustering approach. Examples are combined until the desired number of clusters is obtained. First, we'll count the number of clusters using the elbow apporach.

IV. RESULTS

Cluster formed using Hierarchical Clustering:

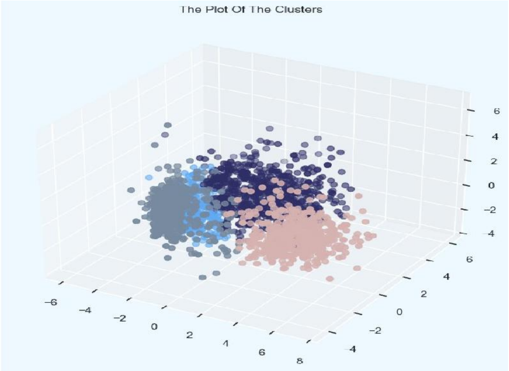

Figure 3: 4 cluster formed

Four clusters have been formed as a result of applying hierarchical clustering to the data, as seen in the plot above.

DOI: 10.48175/IJARSCT-22441

Count of each cluster:

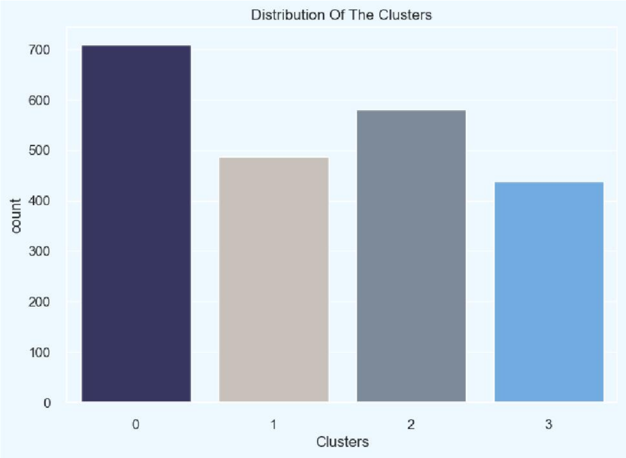

Figure 4: Count of each cluster

The plot above shows the distribution of data points within each cluster, as determined by a clustering algorithm Profile of a group of individuals based on their income and spending habits:

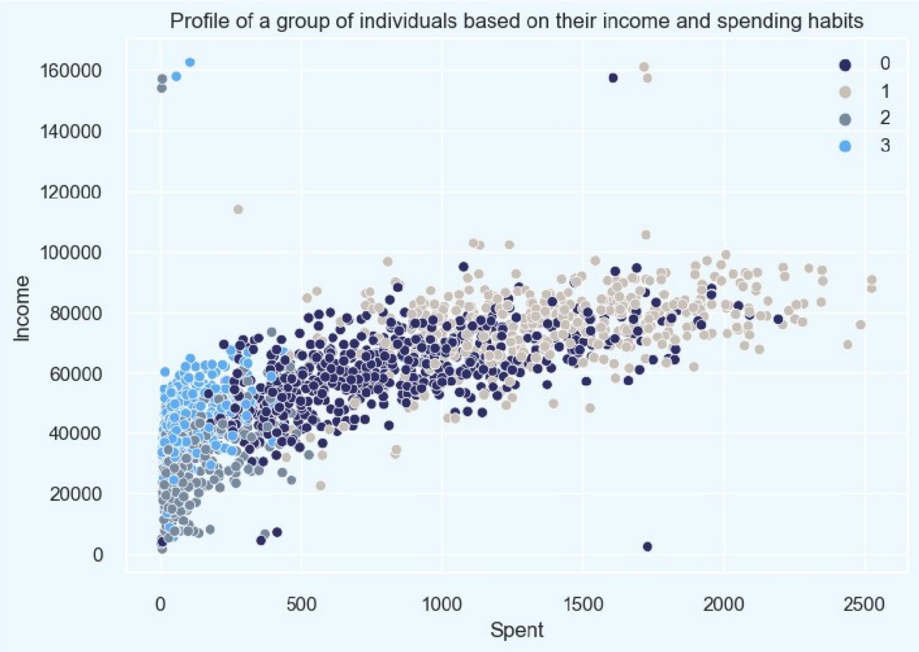

Figure 5: Income vs Spent

Income vs spending plot shows the clusters pattern:
Group 0: those with average income and high spending
Group 1: those with high income and high spending
Group 2: those with low income and low spending
Group 3: those with low income and high spending

DOI: 10.48175/IJARSCT-22441

Profiling Customers:

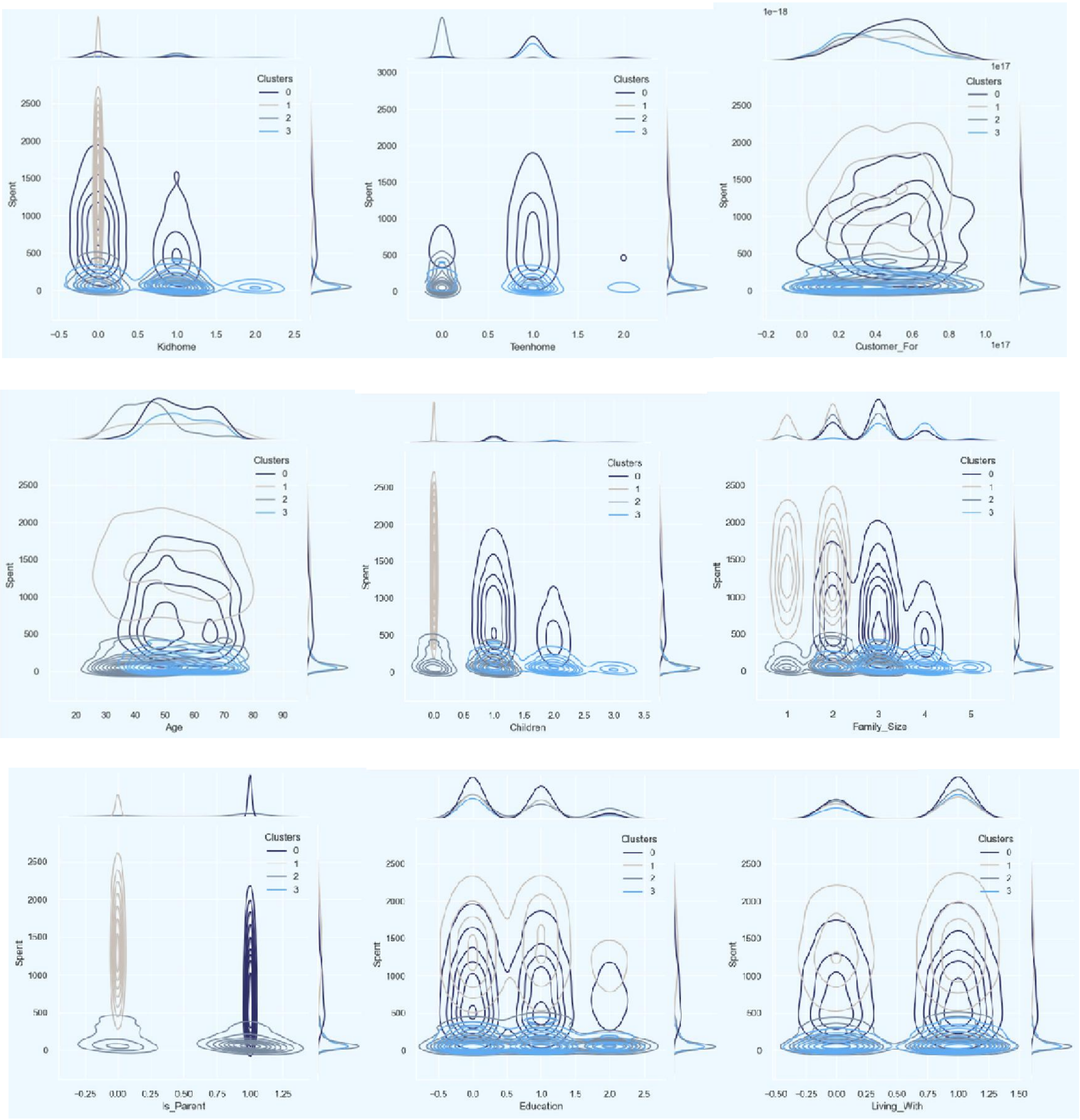

Figure 6: Profiling Customers

Below is the Analysis of Cluster Numbers 0-3

About Cluster Number: 0

Based on the chart, it can be inferred that the group under discussion consists of parents who are relatively elderly and have a family of at least two people and no more than four. Single parents are a subset of this group, and the majority of parents in this group have an adolescent at home. Please be aware, though, that this is merely a judgment drawn from the data presented and might not always be true. Any information should be checked and validated before being used to guide judgments or decisions.

About Cluster Number: 1

It is clear from the data in the charts that the group under discussion consists of non-parents with no more than two family members. Rather than being single, couples make up the bulk of this category. This group's members are from a

DOI: 10.48175/IJARSCT-22441

high-income demographic and range widely in age. Please be aware, though, that this is merely a judgment drawn from the data presented and might not always be true. Any information should be checked and validated before being used to guide judgments or decisions.

About Cluster Number: 2
According to the charts, the group under discussion consists of parents who are comparatively younger and have families of little more than three people. Usually, these parents have a single child who is not a teenager. Note that this is merely a judgment drawn from the data presented, and it might not always be correct. Before drawing any conclusions or making judgments based on information, it is crucial to confirm and validate it.

About Cluster Number: 3
Based on the information in the charts, it can be inferred that the group under discussion consists of parents who are comparatively older and have a family of at least two people and a maximum of five. The majority of these parents belong to a lower-income category and have an adolescent at home. Please be aware, though, that this is merely a judgment drawn from the data presented and might not always be true. Any information should be checked and validated before being used to guide judgments or decisions.

V. DISCUSSION AND CONCLUSION

Developing and implementing a machine learning model that can reveal hidden insights and patterns in customer behavior data is the aim of a study on consumer personality analysis using unsupervised learning. In order to enhance the overall customer experience and propel commercial success, the project would employ this model to gain a deeper understanding of the personalities and preferences of its customers. In order to find patterns and trends in the data, the model may need to be trained on a sizable collection of customer behavior data using methods like clustering or dimensionality reduction. After that, the model may be applied to new data analysis and consumer personality prediction.

To sum up, unsupervised learning has the potential to be a useful technique for analyzing the personalities of customers. You may anticipate customer personality traits based on their behavior and features and find patterns and links in customer data by employing dimensionality reduction techniques and clustering algorithms. This can give you important information about your clientele and assist you in customizing your marketing and customer support tactics to better suit their requirements and tastes. Chatbots and virtual assistants are two possible applications of unsupervised learning in consumer personality profiling. These systems could offer more efficient and tailored interactions with customers by integrating the findings of customer personality research into their algorithms, which could increase customer happiness and loyalty.

Customer personality presents a number of difficulties. The crucial element is the quality of the data. High-quality data that is devoid of biases and errors is necessary for accurate and trustworthy consumer personality predictions. Customer personality analysis may have limitations. Among them is reliance on data. The quality and volume of data available for analysis have a significant impact on the accuracy and dependability of customer personality predictions made with unsupervised learning algorithms.

The personality of the customer poses several challenges. The quality of the data is the most important factor. Accurate and reliable forecasts of consumer personality require high-quality data free from biases and errors.

Analyzing customer personalities may have drawbacks. Reliance on data is one of them. The reliability and accuracy of customer personality forecasts using unsupervised learning algorithms are strongly influenced by the quantity and quality of data available for analysis.

REFERENCES

[1]. Julien Ah-Pine (2018). An Efficient and Effective Generic Agglomerative Hierarchical Clustering Approach, 19(42):1−43, 2018. URL: https://www.jmlr.org/papers/volume19/18-117/18-117.pdf

DOI: 10.48175/IJARSCT-22441

[2]. Margareta Ackerman and Shai Ben-David (2016). A characterization of linkage-based hierarchical clustering. Journal of Machine Learning Research, 17:1–17, 2016. URL: https://www.jmlr.org/papers/volume17/11-198/11-198.pdf

[3]. D T Pham, S S Dimov, and C D Nguyen (2004). Selection of K in K-means clustering. URL: https://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf

[4]. Data Set: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?datasetId=1546318&sortBy=voteCount