CUSTOMER PERSONALITY ANALYSIS USING UNSUPERVISED LEARNING

Woodi Raghavendra Varun
Graduate student
Department of Data Science
University of Massachusetts Dartmouth
Dartmouth MA 02747
Email: wvarun@umassd.edu

Pradyoth Singenahalli Prabhu
Graduate student
Department of Data Science
University of Massachusetts Dartmouth
Dartmouth MA 02747
Email: psingenahalliprabhu@umassd.edu

Abstract

Unsupervised learning is a type of machine learning algorithm that can be used to analyze customer data and identify common patterns and traits among a group of customers. By using clustering techniques to group similar customers together, companies can create customer segments with distinct personality characteristics, allowing them to tailor their marketing and sales efforts to better target each segment. Unsupervised learning algorithms can be trained on large amounts of data, making them powerful tools for companies looking to better understand and serve their customers.

Introduction and Background

Customer personality analysis is the process of identifying and understanding the unique characteristics and traits that make up an individual customer's personality. This information can be used by companies to tailor their marketing and sales efforts to better target and serve each customer's specific needs and preferences.

Traditionally, customer personality analysis has been done manually by marketing and sales teams, who would use their expertise and experience to identify common patterns and trends among customers. However, with the advent of data mining and machine learning, it is now possible to automate this process using algorithms that can analyze large amounts of data and identify common patterns and traits among customers.

One type of machine learning algorithm that can be used for customer personality analysis is unsupervised learning. Unsupervised learning algorithms are trained on a large amount of data and can automatically detect patterns and similarities among customers without being explicitly told what to look for. This makes them particularly well suited for customer personality analysis, as they can uncover subtle differences and trends that might not be immediately apparent 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.

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 income column of the data frame has 24 missing values. The missing numbers are being dropped because it is fewer.
2. We are creating a feature that displays the length of time a customer has been in the company's database. This feature will be based on the "Dt Customer" data.
3. In order to improve the clarity and consistency of the data, I added a new column called "Living With" which has the same meaning as an existing column with a different name. The same approach was used for the "Education" data as well.
4. The "Is parent" feature has been implemented and it returns a value of 1 for parents and 0 for non-parents.
5. Add a new feature called "Spent" that displays the customer's overall spending across all categories over a two-year period.
6. Dropping some of the redundant features.
7. After plotting for “Income” and “Age”, outliers are present which will be deleted
8. Doing label encoding for categorical features i.e., ‘Education’ and ‘Living With’.

9. Dropping the columns of deals accepted and promotions, then scaling the remaining features using “Standard scaler”.

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

PCA:

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

The final classification in this challenge will be based on a wide range of variables. These aspects or characteristics are essentially these factors. Working with it becomes more challenging the more features it has. The correlation between several of these features makes them unnecessary. For this reason, before running the features through a classifier, I shall reduce their dimensionality.

We are reducing the dimensions to 3. After this we will be plotting these data frame.

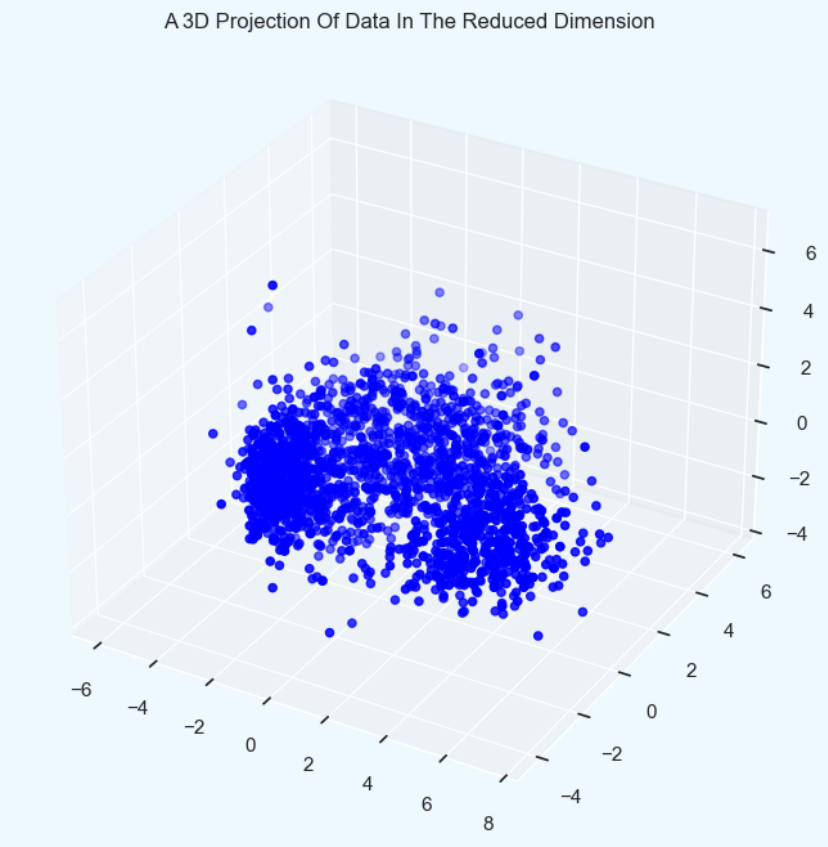


Figure 1: 3D plot of data after PCA

The plot above shows the data in three dimensions after it has been processed using principal component analysis (PCA). This projection allows us to visualize the data in a way that is easier to understand and interpret.

Methods

Hierarchical Clustering

We will be using hierarchical clustering in this project. The most typical hierarchical clustering method used to put objects in clusters based on their similarity is called agglomerative clustering. Another name for it is AGNES (Agglomerative Nesting). Each object is first treated as a singleton cluster by the algorithm. Once all clusters have been merged into a single large cluster containing all items, pairs of clusters are gradually combined. The outcome is a dendrogram, which is a tree-based representation of the objects. This would be useful for identifying distinct customer personality types and for targeted marketing and communication efforts to specific groups of customers.

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.

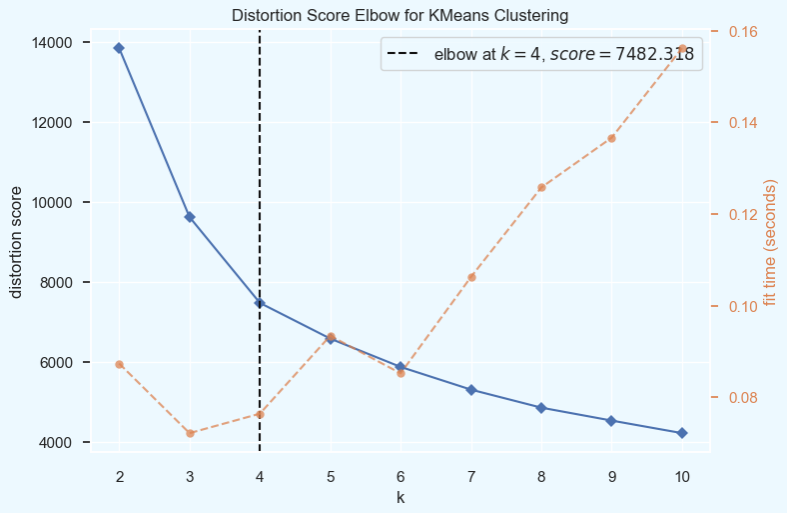


Figure 2: Elbow method visualization

After reducing the attributes to three dimensions, I will apply Agglomerative clustering to perform the clustering. A hierarchical clustering technique is agglomerative clustering. Up until the appropriate number of clusters is reached, examples are merged. First, we will be using elbow method to determine the number of clusters. Once we have got the clusters then we will apply hierarchical clustering based on the PCA dataset what we had created earlier.

Finally, clusters will be formed then we will plot the result.

Results

Cluster formed using Hierarchical Clustering:

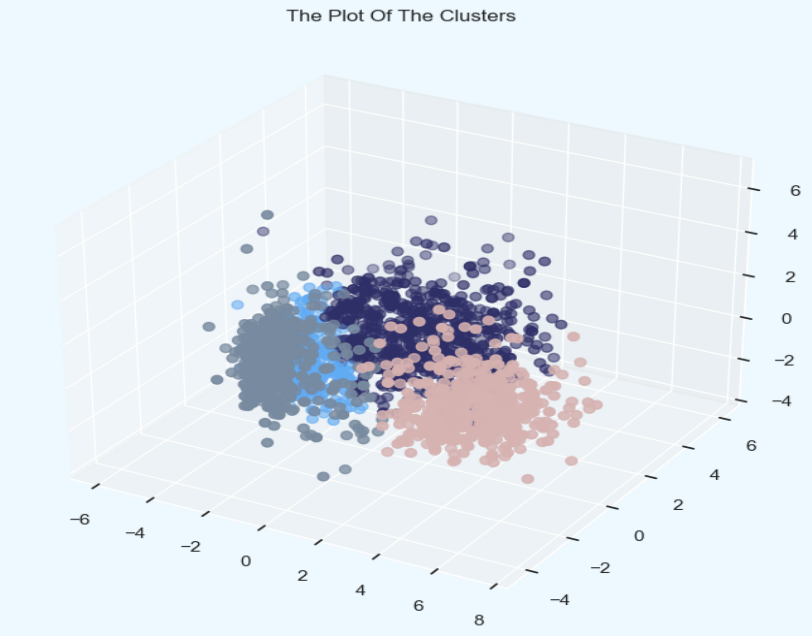


Figure 3: 4 cluster formed

The plot above demonstrates the results of applying hierarchical clustering to the data, which has resulted in the formation of four clusters.

Count of each cluster:

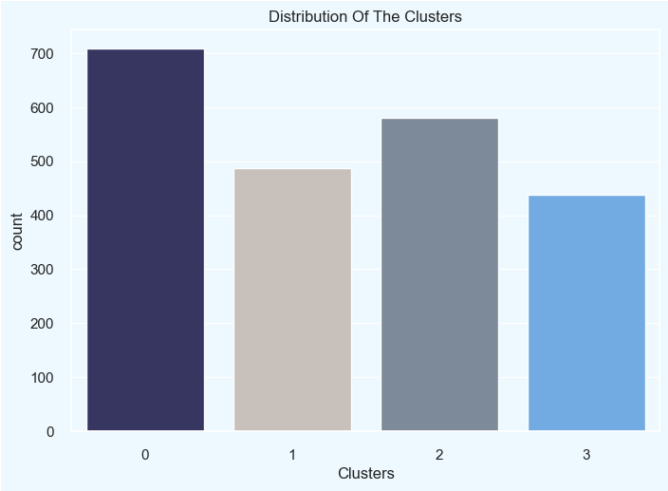


Figure 4: Count of each cluster

The plot above shows the distribution of data points within each cluster, as determined by a clustering algorithm.

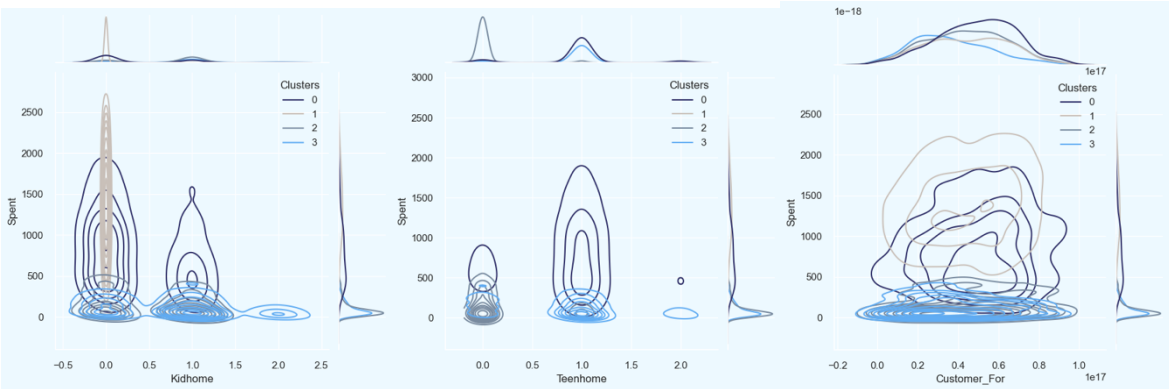
Profile of a group of individuals based on their income and spending habits:

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

Profiling Customers:



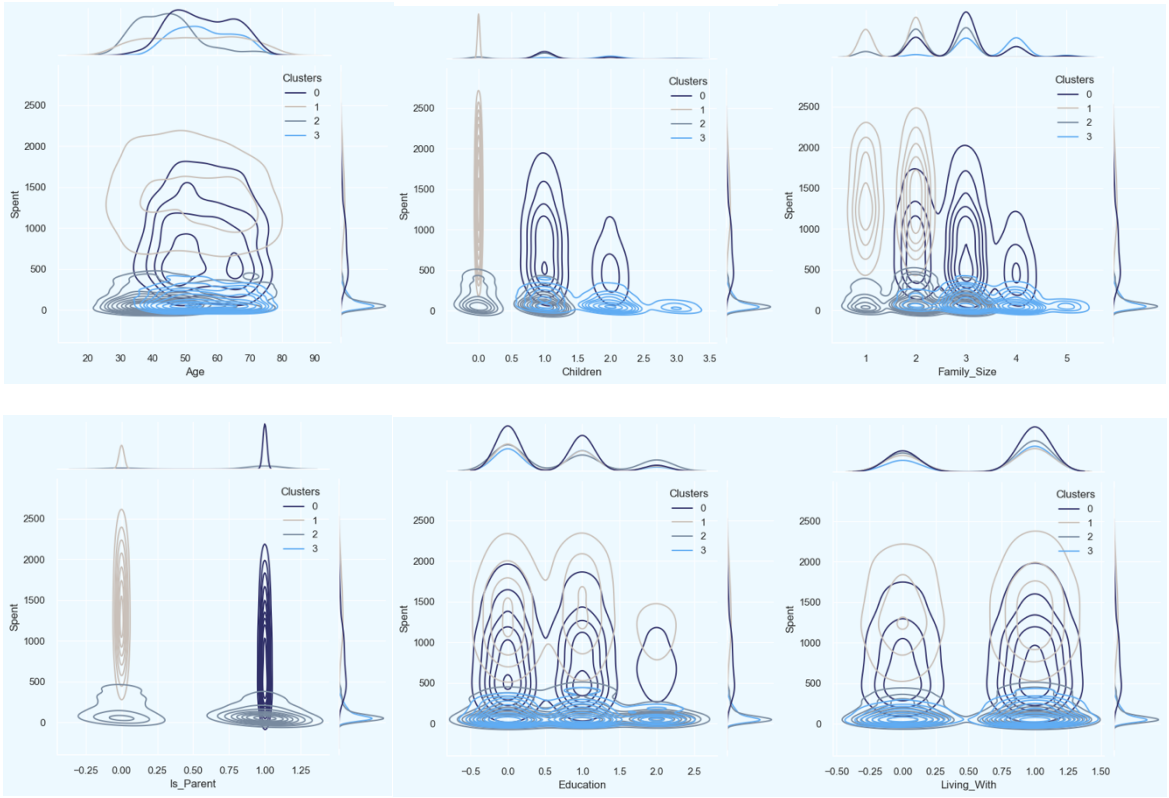


Figure 6: Profiling Customers

Below is the Analysis of Cluster Numbers 0-3

About Cluster Number: 0

As shown in the chart, it can be concluded that the group being discussed is a group of parents who are relatively old and have a family with a maximum of four members and at least two members. Most of the parents in this group have a teenager at home, and single parents are a subset of this group. However, please note that this is only a conclusion based on the information provided, and it may not be accurate in all cases. It is important to verify and confirm any information before making conclusions or decisions based on it.

About Cluster Number: 1

According to the data in the charts, it can be concluded that the group being discussed is a group of non-parents who have a maximum of two members in their families. The majority of this group consists of couples, rather than single people. The members of this group span a wide range of ages and are part of a high-income group. However, please note that this is only a conclusion based on the information provided, and it may not be accurate in all cases. It is important to verify and confirm any information before making conclusions or decisions based on it.

About Cluster Number: 2

The charts indicate, it can be concluded that the group being discussed is a group of parents who are relatively younger and have families with a maximum of three members. Most of these parents have one child, who is typically not a teenager. However, please note that this is only a conclusion based on the information provided, and it may not be accurate in all cases. It is important to verify and confirm any information before making conclusions or decisions based on it.

About Cluster Number: 3

The data in the charts suggests, it can be concluded that the group being discussed is a group of parents who are relatively older and have a family with a maximum of five members and at least two members. Most of these parents have a teenager at home, and they are part of a lower-income group. However, please note that this is only a conclusion based on the information provided, and it may not be accurate in all cases. It is important to verify and confirm any information before making conclusions or decisions based on it.

Discussion and Conclusion

The objective of a project on customer personality analysis using unsupervised learning would be to develop and implement a machine learning model that can uncover hidden insights and patterns in customer behavior data. The goal of the project would be to use this model to better understand customer personalities and preferences, and to use this information to improve the overall customer experience and drive business success. This could involve training the model on a large dataset of customer behavior data, using techniques like clustering or dimensionality reduction to identify patterns and trends in the data. The resulting model could then be used to analyze new data and make predictions about customer personalities. The project could also involve developing tools and processes for integrating the model into business operations, and for using the insights it generates to inform marketing and sales efforts.

In conclusion, unsupervised learning can be an effective method for customer personality analysis. By using clustering algorithms and dimensionality reduction techniques, you can identify patterns and relationships in customer data and predict customer personality traits based on their behavior and characteristics. This can provide valuable insights into your customer base and help you tailor your marketing and customer service strategies to better serve their needs and preferences. Potential application of unsupervised learning in customer personality analysis is in the use of virtual assistants and chatbots. By incorporating the results of customer personality analysis into their algorithms, these systems could provide more personalized and effective interactions with customers, potentially leading to improved customer satisfaction and loyalty.

There are many challenges associated customer personality. Data quality is the important factor. Accurate and reliable customer personality predictions require high-quality data that is free from errors and biases. Limitations of customer personality analysis are possible. Dependence on data is one of them. The accuracy and reliability of customer personality predictions using unsupervised learning algorithms are highly dependent on the quality and quantity of data available for analysis.

In some cases, the data may be limited or incomplete, which could limit the ability of the algorithm to make accurate predictions.

Overall, the future of customer personality analysis using unsupervised learning is likely to involve the development of increasingly sophisticated algorithms and techniques for uncovering hidden patterns and relationships in customer data, and the integration of these insights into a wide range of business applications.

References:

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

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

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

Other References:

- Python 3: https://www.python.org/
- PCA: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- Pandas: https://pandas.pydata.org/
- Numpy: https://numpy.org/
- Seaborn: https://seaborn.pydata.org/
- Matplotlib: https://matplotlib.org/
- Elbow: https://www.scikit-yb.org/en/latest/
- Label Encoding: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
- Standard Scaler: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
- K-Means Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Hierarchical Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
- Data Set: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?datasetId=1546318&sortBy=voteCount