Customer Personality Analysis using clustering

Raunak Nandkumar More
Department of Computer Science
University of South Dakota

ABSTRACT:

The Customer Personality Analysis is used to analyze customer behavior and preferences by examining a rich dataset from an international retail company. The primary goal is to understand and analyze the ideal customers for our business. Through this analysis, we gain valuable insights that inform targeted marketing strategies and improve overall customer experiences.

INTRODUCTION:

In the competitive business landscape, companies use all means necessary to improve business. Practices that help them to tailor to specific customers, or to identify consumers better, or to customize the product to the audience are common. This paper considers one such method to leverage the power of data and algorithms to build models that can segmentalize customers based on different trends and behaviors of customers to optimize resource allocation by focusing efforts on high-potential customers. The paper tackles the need to perform clustering to summarize customer segments to better predict their behaviors like spending habits so companies can deduce different strategies to cater to different audience. Using a dataset of customer data consisting of their spending habits, household family data,their education,salaries, data about which discount offers they went for and applying k-means clustering, a unsupervised learning algorithm, we train a model to cluster the customers into groups so different strategies could be made for different clusters.The dataset is publicly available on the Kaggle[1].

The dataset consists of 2240 observations and 29 variables. The variables are a mix of date, categorical, and numerical types.

ID: Customer's unique identifier
Year_Birth: Customer's birth year
Education: Customer's education level
Marital_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
Complain: 1 if the customer complained in the last 2 years, 0 otherwise 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
MntSweetProducts: Amount spent on sweets in last 2 years
MntGoldProds: Amount spent on gold in last 2 years
NumDealsPurchases: Number of purchases made with a discount
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
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
Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
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

RELATED WORK:

lysis using machine learning:

1.Tsao, Hsiu-Yuan, et al. (2023): Predicting Consumer Personalities from What They Say. In: Applied Sciences. MDPI, Volume 13, Issue 10, Pages 6148. https://www.mdpi.com/2076-3417/13/10/6148

2.Chauhan, G. E. Deemed, and D. Dun (2021): Customer Segmentation Using Machine Learning. In: Elementary Education Online. Volume 20, Issue 3, Pages 3230–3237. https://ijisae.org/index.php/IJISAE/article/view/3864

3.Sarker IH (2021): Machine learning: algorithms, real-world applications and research directions. In: SN Computer Science. Springer, p 1–36. https://pubs.aip.org/aip/acp/article/2794/1/020016/2914509/Personality-prediction-using-machine-learning

METHODOLOGY:

We discuss the techniques used to preprocess the data, deriving new features from existing ones, training the k-means clustering model, and evaluating the segmented data to deduce insights from clusters.

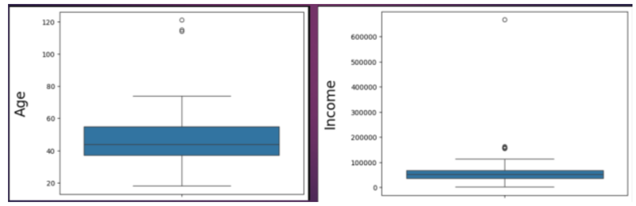
Python is used for the coding part and its libraries, such as numpy, pandas, seaborn, matplotlib, sklearn, etc .are imported to perform the data analysis and machine learning. Jupyter Notebook used as the interactive development environment. The code and the output can be found in the ipnb file.

The first step of our methodology was to import the necessary libraries as mentioned already. We used the pandas function read_csv to read the csv file containing the dataset and store it in a data frame called df. Handling missing values by finding what is missing and removing rows with NaN values which are seen in info.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

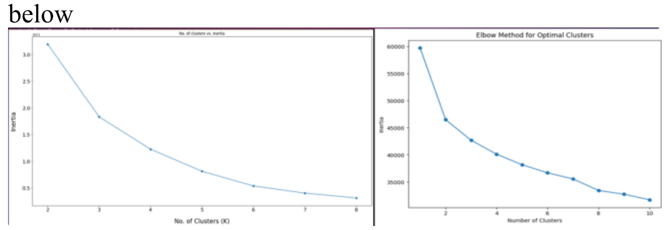
We noticed that some of the variables in our dataset were categorical, meaning that they had a finite number of possible values, These variables were represented by strings. However, machine learning algorithms usually require numerical inputs, so we changed these variables into numerical labels by assigning different numbers to different labels. We also noticed that one of the variables, Dt_Customer, in our dataset was a date, which is complicated to work with. So we derived a new feature Customer_For, which was the number of days the customer was enrolled with the company. We also derived a new column Age from the Birth_Year column present in the dataset by subtracting Birth Year values from 2014 as that is when the latest date of customer was recorded meaning it is the latest when the dataset was made. We also changed labels from 2 features Maritial_Status , which had multiple labels that meant the same thing to numerical representation, which is the same thing that we did for Education column that consisted of educational details of the customers.

We also handled outliers for the Income and Age columns as shown below



For the training part we went with training 2 different models, one having scaled data and other having unscaled data.

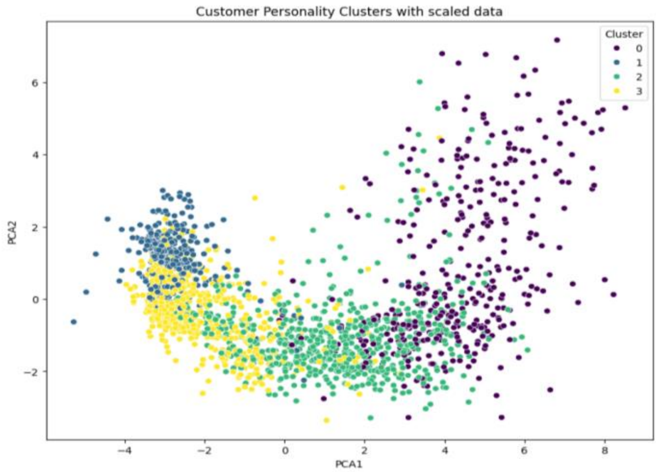
For finding the optimal number of clusters to use with k-means clustering we used the Elbow method which calculates the WCSS(Within-Cluster Sum of Squares) that measures the sum of distances between datapoints of same cluster which is sometimes called inertia to find the number of clusters after which the inertia did not decrease as much as it did before the optimal number of clusters called the elbow point, which was 4 as clusters as seen in the figure below



RESULTS:

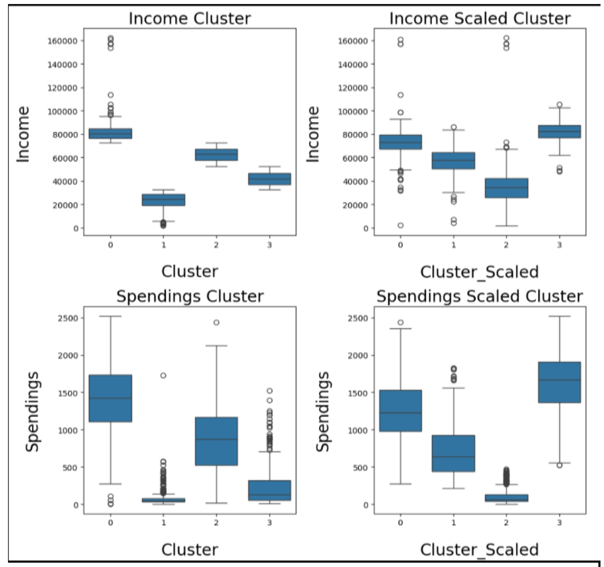
Visualizing the clusters formation using PCA dimensionality reduction.

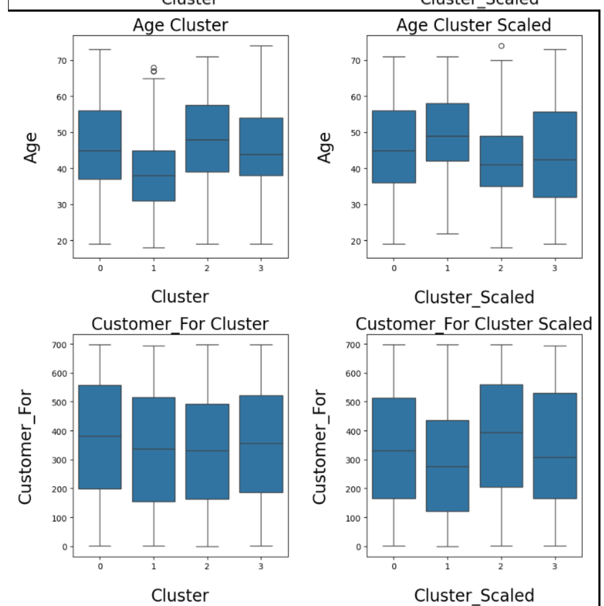
We visualize the clusters formed using PCA by reducing the dimensions to 2 as seen below for scaled and unscaled data, part of the reason was to see the clustering difference on scaled vs unscaled data, where we saw scaled data clusters being a bit difficult to visualize when compared to unscaled data, which is characteristic of dimensionality reduction which often leads to loss of data representation.



CONCLUSION:

When Checking variation of data using clusters vs different features using scaled and unscaled data we see that most notable feature of Clusters formed is in terms of spending. Higher spending clusters earn more too and the highest spenders are also generally enrolled with the company for longer. So the clusters are majorly formed based on the spending habits of customers. Since we trained 2 models with scaled and unscaled data, we plotted for both models and both of them while having different clusters have the same conclusions. The presence of outliers in some of the clusters mean that there is still need for more outlier handling,





REFERENCES:

[1]:https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis