MIT Applied Data Science Program

Marketing Campaign Customer Segmentation

September 23, 2022

Kristi Yamashita

Problem Definition

THE CONTEXT

This problem is important to solve because customer segmentation is a crucial aspect of marketing operations for businesses and organizations. By dividing a dataset of customers into groups of similar customers based on specific common characteristics, we can better understand the various needs and motivations of customers. Customer segmentation allows businesses and organizations to achieve efficient marketing efforts and obtain better return on investment because it allows them to know how to effectively use their money, time, and other resources to target the correct customer groups and create a personalized experience.

THE OBJECTIVES

The intended goal of customer segmentation is to divide the dataset of customers into groups based on certain common characteristics. The customer groups created will highlight the unique needs, motivations, and interests, which can then be used by businesses and organizations to implement an efficient marketing strategy.

THE KEY QUESTIONS

- Are there any issues with the data that we need to address before analyzing?
- Which variables should we use for clustering?
- What characteristics are common among the customers in the dataset to create customer groups?
- What variables can we create to further analyze the data?

THE PROBLEM FORMULATION

We are trying to create customer groups based on similar characteristics among the customers in the dataset so we can effectively target customer segments and create customized communications and offerings.

Data Exploration

DATA DESCRIPTION

This set of data was collected in the year 2016 and contains background information about customers and their spending habits. There are 2240 unique customers in this dataset. The dataset contains both categorical and numerical variables:

Demographics / Customer Background

- ID: Unique ID of each customer
- Year_Birth: Customer's year of birth
- Education: Customer's level of education
- Marital_Status: Customer's marital status
- Kidhome: Number of small children in customer's household
- Teenhome: Number of teenagers in customer's household
- Income: Customer's yearly household income in USD
- Recency: Number of days since the last purchase
- Dt_Customer: Date of customer's enrollment with the company
- Complain: 1 If the customer complained in the last 2 years, 0 otherwise

Spending Habits

- MntFishProducts: The amount spent on fish products in the last 2 years
- MntMeatProducts: The amount spent on meat products in the last 2 years
- MntFruits: The amount spent on fruits products in the last 2 years
- MntSweetProducts: Amount spent on sweet products in the last 2 years
- MntWines: The amount spent on wine products in the last 2 years
- MntGoldProds: The amount spent on gold products in the last 2 years
- NumDealsPurchases: Number of purchases made with discount
- NumCatalogPurchases: Number of purchases made using a catalog (buying goods to be shipped through the mail)
- NumStorePurchases: Number of purchases made directly in stores
- NumWebPurchases: Number of purchases made through the company's website
- NumWebVisitsMonth: Number of visits to the company's website in the last month

Campaign Interactions

- AcceptedCmp1: 1 if customer accepted the offer in the first campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the second campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the third campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the fourth campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the fifth campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

OBSERVATIONS & INSIGHTS

Key Patterns:

- All the columns are not missing any values except the Income column. The Income column is missing 24 values or 1.071429% of values.
- At least one customer accepted the offer of campaigns 1, 2, 3, 4, and 5. The offers from campaign 3 and 5 were accepted the most, and the offer from campaign 2 was accepted the least. More customers have accepted the offer in the last campaign compared to any of the offers from campaign 1, 2, 3, 4, and 5.
- 99.5% of the data for the Income variable is less than or equal to 102145.75000000003. This means that the outliers for the Income variable can be removed. After dropping the outliers the histogram for Income shows a symmetrical distribution.
- The amount customers spent on wine products, fruit products, meat products, fish products, sweet products, and gold products in the last 2 years show a right-skewed distribution and appear to have outliers.
- Most of the customers are Married (38.6%) followed by Together (25.8%) or Single (21.8%).
- A little more than half of the customers have a Graduation (50.4%) level of education. Customers with a Master (25.6%) or PhD (21.6%) level of education almost make up the other half.
- More than half (57.7%) of customers have 0 small children in their household and more than half (51.6%) of customers have 0 teenagers in their household.
- Almost all the customers in this dataset have not complained in the last 2 years (99.1%).
- There is a strong positive correlation (0.73) between income and amount spent on wine products and meat products in the last 2 years. This makes sense because the larger the income a customer has, the more they can spend on these types of products. This is further supported by the positive correlation between income and amount spent on fruit products, fish products, sweet products, and gold products in the last 2 years.
- There is also a strong positive correlation between income and number of purchases using a catalog (0.71) or directly in stores (0.69). Furthermore, there is a moderate positive correlation between income and number of purchases made through the company's website.
- There is a strong negative correlation between income and number of visits to the company's website in the last month.
- There is a strong positive correlation between amount spent on meat products in the last 2 years and number of purchases made using a catalog.
- Customers generally have the same income if they have a Graduation or Master level of education. However, customers with a PhD level of education have the highest income on average and customers with a Basic level of education have the lowest income on average.
- The customers generally seem to have a similar income on average regardless of their marital status.
- Customers with 0 small children in their household seem to have a higher income on average compared to those with 1 or 2 small children.
- Regardless of a customer's marital status, it seems that customers have either 0 or 1 small children in their household.
- The age variable seems to have a relatively symmetrical distribution. The ages of the customers seem to primarily fall between 39 to 57.
- The histogram for AmountPerPurchase shows a right-skewed distribution with many outliers. The average amount spent per purchase is 33.274270.

OBSERVATIONS & INSIGHTS (continued...)

Key Patterns:

- There is a strong positive association between Income and Expenses, which means that the higher a customer's income is, the higher their expenses are. This makes sense because if a customer has more money, they probably spend more money as well with the exception of some outliers.
- On average, families with one person have the highest income of approximately 65000. This is followed by families with 2 people. It seems like families with 3 or 4 people generally have a lower income on average.

Data Treatments / Pre-Processing:

- Drop the ID column because it has no null values and the number of unique values are equal to the number of observations. This column wouldn't provide any predictive power for our analysis.
- Combine categories
  - Replace "2n Cycle" with "Master" in Education
  - Replace "Alone", "Absurd", and "YOLO" with "Single" in Marital_Status
- Drop customers that are older than 115.
- Create new variables using the features we have
  - Age (2016 - Year_Birth)
  - Kids (Kidhome + Teenhome)
  - Family_Size (Status + Kids)
    - Replace "Married" and "Together" with "Relationship"
    - Replace "Divorced" and "Widow" with "Single"
    - Create "Status" variable to assign values 1 and 2 to categories Single and Relationship
  - Expenses (MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds)
  - NumTotalPurchases (NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases)
  - Engaged_in_days ((01-01-2015) - Dt_Customer)
  - TotalAcceptedCmp (AcceptedCmp1 + AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5 + Response)
  - AmountPerPurchase (Expenses / NumTotalPurchases)
- Drop observations that NumTotalPurchases equal 0.
- Impute the missing values for the Income variable with the median (51315).

Proposed Approach

POTENTIAL TECHNIQUES

- Visualize data using histograms, box plots, bar plots, correlation heat maps, and scatter plots
- Univariate analysis and Bivariate analysis
- Scale the data
- T-SNE
- Principal Component Analysis (PCA)
- Cluster Profiling
  - K-Means, K-Medoids, Hierarchical Clustering, DBSCAN, Gaussian Mixture Model

OVERALL SOLUTION DESIGN

- Check the dimensions of the dataframe in terms of rows and columns
- Check data types are correct and change data types accordingly if necessary
- Study summary of statistics
- Drop unnecessary variables
- Check and impute missing values
- Study correlation
- Detect outliers and remove if necessary
- Create new variables and groupings

MEASURES OF SUCCESS

- Elbow Method of showing the optimal k
- Silhouette Score
- Cophenetic Correlation

THANKS!

Any questions?

You can find me at:

- kristi.yamashita@yum.com
- www.linkedin.com/in/kristiyamashita