Customer Behaviour Analysis using Machine Learning

Problem Statement

Customer Behavior Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers.

Customer Behavior analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

Dataset Description

Customer Dataset

Table showing the metrics of the shopping company that is used to analyze the customer buying behavior

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	88	546	172
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	1	6	2
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	49	127	111
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	4	20	10
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	43	118	46

Dataset Description

Customer Dataset

People

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

Products

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

Promotion

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

Scenario 1

You are working as a Machine Learning Engineer to segment the dataset on the basis of customer behaviour.

The objective of this scenario is to extract the dataset and perform cleaning operations post understanding of the basic properties of the dataset.
In this scenario you need will be also required to perform datetime tasks so kindly think of the logic before applying it in Task2

You are required to complete all the tasks as shared below.

Scenario 1

Task 1

- Importing the relevant packages (Packages example: numpy, pandas....)
- Extract the dataset from the file named "dataset.csv" and save it in customer variable
- Is the dataset successfully called? Can you check with the top 10 records?
- Check for the structure and dimensions of the dataset
- Show the column names

Scenario 1

Task 2

- Create a new column "Age" by subtracting the column Year_Birth from 2015
- Check the statistics of the dataset
- Work on analysing Missing values and show the output
- Display the missing values using heatmap and remove the y labels while plotting
- Drop the missing values

Scenario 2

Congratulations on getting the dataset in workable format. You will be learning exploratory data analysis here that can assist us to further clean data.
We have dataset with 2240 records and 30 columns.
Over here you will be creating few new columns and clean the data by learning hidden insights

Feature Engineering
There is a lot of information given in the dataset related to the customers. In some cases we can group some columns together to create new features and in some cases we can create new columns based on the existing one's to create new features. This would help to better explore the data and draw meaningful insights from it.

Scenario 2

Task

- From the enrolment date of customers, let's calculate the number of months the customers are affiliated with the company with name "Month_Customer". The mathematical equation is sum of number of months between enrolment year to 2015 AND one month less to the enrolment date months data.
- Create a column named as "TotalSpendings". This is sum of amount spent on products.
- On the basis of Age let's divide the customers into different age groups and create a column "AgeGroup". The logic for that is;
    - Age Group is Teen for age less than 19
    - Age Group is Adults for age between 20 and 39
    - Age Group is Middle Age Adults for age between 40 and 59
    - Age Group is Senior for age more than 60

Scenario 2

Task

- Information is given separately for kids and teens at home for every customers. Let's sum them up, as they can be better represented together as the number of children at home, with column name "Children".
- The Marital Status column has different string values: Together, Married, Divorced, Widow, Alone, Absurd, YOLO. Most of them fall under the same category. So let's represent the marital status of customers based on 2 main categories i.e. Married and Single
- There seems to be some outliers in the Age and Income columns. Let's check them. Use boxplot for each individual columns. Update the dataset by removing the records that are outliers. (recheck the same with Boxplots)

Scenario 3

So far you have completed tasks such as extracting the dataset, creating new columns, cleaning the dataset and removing the outliers.
We have dataset with 2205 records and 34 columns.

Exploratory Data Analysis (EDA)
You have to complete below tasks further to answer the questions that are based on your analysis and make sure you keep a note of your insights or assumptions while performing it.
Requirements:- Need Matplotlib and seaborn packages

Scenario 3

Task

- Univariate analysis of each variable
- Bivariate Analysis of categorical vs numerical variables (Take target variable as fixed variable here)
- Multivariate Analysis of categorical and numerical variables
- Check distribution of variables

Scenario 4

Congratulations on completing the exploratory data analysis. You have successfully understood the dataset and now is the time to segregate the customers on the basis of some specific columns.

You are going to learn to apply unsupervised machine learning here and find the best number of groups that define the clusters.

Once the number of clusters have been predicted, you will be asked a couple of questions on the basis of your dataset "customer" that you have provide your answers on the basis of the labels (clusters) predicted by model
We have dataset with 2205 records and 34 columns.

Scenario 4

KMEANS Algorithm

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Scenario 4

The way kmeans algorithm works is as follows:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Scenario 4

Task

- Requirements:- sklearn, cluster, metrics
- Drop the columns as mentioned below and save the result in new variable "X":
- ('ID', 'Year_Birth', 'Education', 'Marital_Status', 'Kidhome', 'Teenhome', 'MntWines', 'MntFruits','MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds','Dt_Customer', 'Z_CostContact', 'Z_Revenue', 'Recency', 'NumDealsPurchases', 'NumWebPurchases','NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Response', 'AgeGroup')

- Use Kmeans algorithm on "X" dataset and find best number of clusters using Elbow Method
- Predict the labels for that best number of clusters
- Create a new dataset "customer_kmeans" that has all columns of "X" and also include predicted labels.

Conclusion

Always leave a final conclusion.

💡 **Task for the Presentation:**

Conclude your presentation with analysis done so far.

Thank You, All the best 👍