Data Mining Techniques

Spring Semester 2022-2023

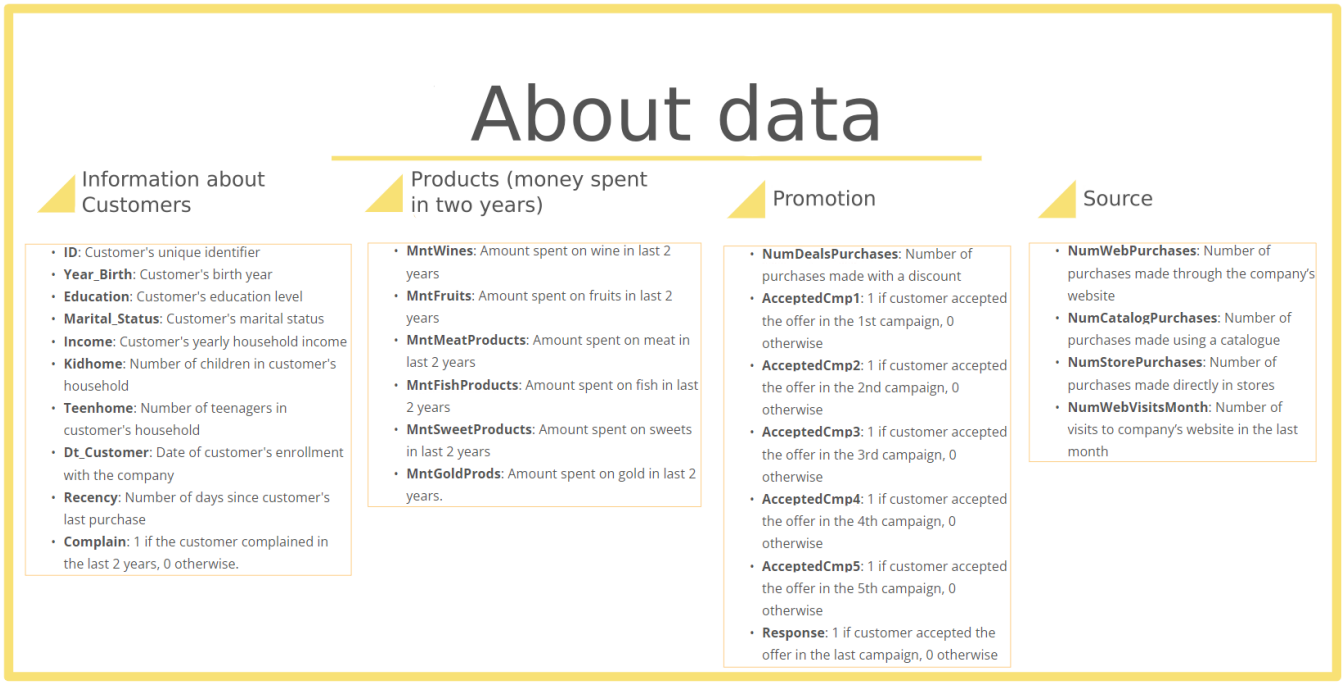
1st Assignment – Work individually or in groups of 2

Assignment purpose

The purpose of this assignment is to learn the basic stages of the procedure followed to apply data mining techniques, namely: data preprocessing / cleaning, conversion, application of data mining techniques and evaluation. The implementation will be done in Python using tools / libraries: jupyter notebook, pandas and SciKit Learn.

Description

Customer Profile Analysis is a detailed analysis of a company's ideal customers. It helps companies understand their customers better and make the modification of its products according to the specific needs, behaviors and concerns of different types of customers easier. For example, instead of spending money for the promotion of a new product to each customer, a company can analyze which customer category is most likely to buy that product and then promote it only to that particular category.



Questions

1. **Preprocessing / Cleaning**: Check if there are missing values in the data and handle them accordingly, convert colums about dates to DateTime objects and check for any dtype: object attributes that can be encoded / converted to numeric values **(5%)**.

2. Print the **unique values** of categorical attributes **Marital_Status** and **Education** to get a more clear picture of the data. Change values [Alone, Absurd, YOLO] of Marital_Status to 'Single'. Use any type of graph to show the number of values in each category **(5%)**.

3. **Create new attributes (10%)**:
   A. Create an attribute "**Customer_For**" that represents the number of days passed from the day the customers making purchases from the store in relation to the last recorded date (Recency).
   B. Extract the age "**Age**" of a customer based on "**Year_Birth**" which indicates the respective person's year of birth.
   C. Create another attribute, "**Spent**", which indicates the customer's total money spent in all categories in two years.
   D. Create an attribute "**Children**" that shows the number of children in a household, i.e. kids and teenagers.
   E. To get a better understanding of a household, create an attribute named "**Family_Size**" that shows the total number of people in a household.
   F. Create an attribute "**Is_Parent**" that indicates whether or not a customer is a parent.
   G. Create another attribute, "**Living_With**", using "**Marital_Status**" to extract the living situation of couples. Specifically, this attribute must have two values. "Partner" and "Alone".
   H. Create column "**Age_Group**" using column "**Age**", to group ages according to the following values: "21-30", "31-40", "41-50", "51-60", "61-70", "71-80", ">80".

4. Check for any **outliers** in the attributes and delete them from the data **(5%)**.

5. Then, examine the **correlation of the attributes** using a heatmap (excluding categorical attributes at this point) **(5%)**.

6. Answer the following questions with **graphs** – choose 10 of the following **(20%)**
    1. In which **Marital_Status** category does the largest percentage of the customers belong?
    2. How many customers have filled a **Complain**?
    3. What is the relationship between the number of purchases **Spent** and **Marital_Status**?
    4. What is the relationship between the number of purchases **Spent** and **KidHome** and **Family_Size** attributes?
    5. What is the relationship between **Age_Group** and **Spent** attributes?
    6. What is the relationship between **Income** and **Spent** attributes?
    7. What is the relationship between **Education** and **Income** attributes?
    8. What is the relationship between **Income** and **Family_Size** attributes?
    9. What is the relationship between **Income** and **KidHome** attributes?
    10. What is the relationship between **Income** and **Living_With** attributes?
    11. What is the relationship between **Income** and **Spent** attributes?
    12. What is the relationship between **NumWebPurchases** and **NumWebVisitsMonth** attributes?
    13. What is the percentage of customers that has accepted all deals from the store?
    14. Draw a histogram for **NumDealsPurchases** column.
    15. Draw a histogram for **Income** column.
    16. Draw a histogram for **KidHome** column.
    17. Draw a histogram for **Family_Size** column.
    18. Do customers with a masters degree spent more money on wine?

7. **Principal Component Analysis (PCA) (25%)**:
    In this problem, there are lots of factors on which a classification is made. These factors are basic attributes of features. The greater the number of attributes, the harder the problem is. Many of these attributes are correlated and thus, redundant. This is the reason why you are going to preform **dimension reduction** to specific attributes. Dimension reduction is the process of reducing the number of random variables that are under examination and leads to obtaining of a set of major variables.
    After the addition of attributes done in the previous questions, the dataset variables that are categorical and not numeric, are the following ['**Education**', '**Marital_Status**', '**Living_With**']. Use **LabelEncoder()** for these variables to convert them to numeric data (this process is called **one hot encoding**).

    Then, create a copy of the dataframe, which contains all numerical columns and delete columns related to offers and marketing campaigns, i.e. ['**AcceptedCmp1**', '**AcceptedCmp2**', '**AcceptedCmp3**', '**AcceptedCmp4**', '**AcceptedCmp5**', '**Complain**', '**Response**'].
    In this way, the data obtained contains attributes of various dimensions and variances. Different variances in attributes negatively affect the modeling of the dataset. The solution is to perform *Standardization*, so that each column / attribute / variable has $\mu = 0$ and $\sigma = 1$.

Finally, use Principal Component Analysis (PCA) to reduce the number of dimensions to n_components = 3. Draw the (three dimensional) graph of the result.

8. **Clustering (25%)**

   Steps
   - Elbow method to determine the number of clusters needed.
   - Agglomerative and K-Means clustering
   - Show the formed clusters with a graph (e.g. scatter plot).

9. **Customer Profile Analysis (Bonus)**

Try to sketch the profile of the formed clusters through graphs to arrive to a conclusion regarding who is the most "important" customer and who needs the most attention from the store's marketing team.

To achieve this, provide some of the attributes that are indicative of the personal characteristics of a customer depending on the cluster they belong to (e.g. **Age**, **Is_Parent**, **Family_Size** etc.). Finally, collect the main characteristics of each cluster.
e.g.

Cluster 0:	**Cluster 1:**
They spend the least	They spend the most
They have the lowest income	They have a large income
They have teenagers at home	Most of them aren’t parents
They are older	They have taken part in all 6 marketing campaigns