Data Mining

First Assignment

Kourosh Parand
Computer Science, Shahid Beheshti University

November 11, 2022

For the first Data Mining course assignment you should implement all statistical theory which you know in a practical problem. Dataset of customers of a chain grocery stores. Sales team wants to send some customized offers to their customers which suits best to them. There are many features in this dataset which describe a specific customer. The dataset's name is "marketing_campaign.csv". In this step, you should follow your statistical knowledge to preprocess this dataset.

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
Income	Customer's yearly household income in USD
Kidhome	Number of small 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 the last purchase
MntWines	The amount spent on wine products in the last 2 years
MntFruits	The amount spent on fruits products in the last 2 years
MntMeatProducts	The amount spent on meat products in the last 2 years
MntFishProducts	The amount spent on fish products in the last 2 years
MntSweetProducts	Amount spent on sweet 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
NumWebPurchases	Number of purchases made through the company's website

NumCatalogPurchases	Number of purchases made using a catalog (buying goods to be shipped through the mail)
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to the company's website in the last month
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
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
Complain	1 If the customer complained in the last 2 years, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

List of Tasks:

A) Load Data

B) Check the shape of the data

C) Observe first five rows

D) Observe last five rows

E) Check data integrity and print missing values of each column

F) Drop “ID” column as a not useful feature

G) Summary statistics of numerical variables

H) Explore more in all categorical variables and unique observations in each category

I) Replace “2n Cycle” with “Master” in “Education” column

J) Replace [“Alone”, “Absurd”, “YOLO”] with “Single” in “Martial_Status” column

K) Plot histogram and box-plot of customers Income and detect outliers visually

L) Remove all outliers with respect to IQR (Inter Quartile Range) and plot cleared data

M) Draw heat map correlation of dataset

N) Draw Income vs Education, Marital Status vs Income, Kidhome vs Income

O) Feature engineer Age column and calculate Age of customer and remove outliers and plot histogram of Age feature

P) Create feature Kids = Kidhome + Teenhome

Q) Replace “Married” and “Together” with “Relationship”

R) Replace “Divorced” and “Widow” with “Single”

S) Replace “Single” with integer 1 and “Relationship” with integer 2 and put it in “Status” column

T) Create column “Family Size” = “Status” + “Kids”

U) Create column “Expenses” which equals add the amount spent on each product 'MntWines', 'MntFruits', ‘MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', ‘MntGoldProds’

V) Create column “Total Purchases” = “NumDealsPurchases” + “NumWebPurchases” + “NumCatalogPurchases” + “NumStorePurchases”

W) Change “Dt_Customer” to pd.to_datetime and get min and max of date

X) Add column “Engaged_in_days” = Today - “Dt_Customer”

Y) Create “TotalAcceptedCmp” column which equals “AcceptedCmp1" + “AcceptedCmp2" + “AcceptedCmp3" + “AcceptedCmp4" + “AcceptedCmp5" + “Response"

Z) Create column “AmountPerPurchase” = “Expenses” / “NumTotalPurchases”

AA) Get “AmountPerPurchase”’s max value. What is wrong? Drop problematic rows and get a describe from “AmountPerPurchase”

BB) Fill missing values of “Income” column with the median of this feature

CC) Plot Income vs Expenses scatter plot and fit a curve and return coefficients of fitted curve.

DD) Print transpose of data frame

In each section, you should write an observation. Your assignment file should be an jupyter notebook file. You are allowed to use all available modules like numpy, pandas and ….

This assignment is active till 29 Aban.

Good Luck