Data Mining and Visualization (83676)

Final Project

Part 1

Due: 31/3/2022

General

This project will have two submissions that require to apply in Python the concepts you have learned in the course. The submission is in pairs and should include a notebook file (.ipynb) with the full code and a report (word or PDF) with the project description and explanations. The first submission constitutes 40% of the project final grade. You will be given a dataset, and in this part, you will investigate the dataset and apply preprocess. The output of this part will be an input for the next part, which will be a classification task.

The grade evaluation takes into consideration:

- Techniques used: Did you select appropriate and diverse techniques and justify why?
- The process you followed: Is it correct (given the techniques you used), did you describe it well?
- Interpretation of results: Did you correctly understand and interpret the results you obtained?
- Quality of writeup: Did you present your work well, in an understandable and usable manner?

Problem Description

Marketing campaigns are one of the most effective ways to reach out to people for selling a product or service. However, they require large investments to actually execute these campaigns. Furthermore, the increasingly vast number of marketing campaigns over time has reduced its effect on the general public. All of these as well as economic pressures and competition have led marketing managers to invest in directed campaigns with a strict and rigorous selection of contacts.

The Goal

The goal is to increase the efficiency of the direct marketing campaign by predicting who will respond to an offer for a product or service.

Data Information

The data in this project is related to a direct marketing campaign for selling products. It contains costumers information from previous campaigns and combines personal information and purchase history.

Attribute Information

1. ID
2. Year_Birth: Customer's year of birth
3. Education: Customer's level of education
4. Status: Customer's marital status
5. Income: Customer's yearly household income
6. Num_of_kids: Number of small children in customer's household
7. Num_of_Teen: Number of teenagers in customer's household
8. Registration_date: Date of customer's enrolment with the company
9. Recency: Number of days since the last purchase
10. Mnt_Fruits: Amount spent on fruits products in the last 2 years
11. Mnt_Meat: Amount spent on meat products in the last 2 years
12. Mnt_Sweet: Amount spent on sweet products in the last 2 years
13. Mnt_Wines: Amount spent on wines products in the last 2 years
14. Mnt_Gold_Products: Amount spent on gold products in the last 2 years
15. Mnt_Fish: Amount spent on fish products in the last 2 years
16. Num_Web_Purchases: Number of purchases made through company's web site in the last month
17. Num_Store_Purchases: Number of purchases made directly in stores in the last month
18. Num_Deals_Purchases: Number of purchases made with discount in the last month
19. Num_Catalog_Purchases: Number of purchases made using catalogue in the last month
20. Num_Web_Visits: Number of visits to company's web site in the last month
21. Response_Campaign_1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
22. Response_Campaign_2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
23. Response_Campaign_3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
24. Response_Campaign_4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
25. Response_Campaign_5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
26. Complain: 1 if customer complained in the last 2 years, 0 otherwise
27. Cost_Contact: Cost to contact a customer
28. Revenue: Revenue after client accepting campaign
29. Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Instructions

In this first part, you will initially explore and understand the given data set using statistic method

and visualization tools and then apply the preprocess that will be used for the classification task in the following project part.

You should implement the following sections and add more necessary actions, analytics and visualizations to enrich your work.

- Show the data information, e.g., types of attributes, the attributes values etc.
- Show the data statistics, e.g., distribution, skewness, median and more.
- Show and explain attributes correlations.
- Show and explain visualizations that present interesting insights from the data, e.g., identify relations, trends, the effect of an attribute on the target variable etc.
- Data cleaning - check for each one of the problems and take care of them properly, e.g., missing values, inconsistent etc.
- If necessary, add and/or delete attributes.
- Data reduction - apply at least one of the methods we learned.
- Data transformation - apply the appropriate methods to the required attributes, e.g., normalization, discretization etc..

Additionally, you should use methods that have not been demonstrated in class. You should explain in the report all the steps you made.

Good Luck!