CSX4202/ITX4202 Data Mining

Mini-project 1: Data preprocessing, exploratory and analysis

The **GOALs** of this project are 1) to explore, preprocess, and analyze the given dataset and 2) to provide audiences with useful information/insight to be used for their business strategy. The main project's outcomes include statistical summary of the datasets so that they understand more about the characteristics of dataset obtained, and the analyzed results, which help the audiences accomplishing the project's goals.

Synopsis:

Suppose that you work as a data scientist in an agency company and be assigned to work on one data exploratory and analysis project for a restaurant franchise in order to analyze the survey data gathered from people in the area of franchising restaurants and to suggest 3 new strategies/ways to give coupons to the potential customers in car so that it can increase the chances that these customers will redeems their coupons, which results in increasing sale volume. The data includes demographic and contextual attributes specified as follows.

Data description (Attribute Information):

1. **destination:** No Urgent Place, Home, Work
2. **passanger:** Alone, Friend(s), Kid(s), Partner (who are the passengers in the car)
3. **weather:** Sunny, Rainy, Snowy
4. **temperature:** 55, 80, 30
5. **time:** 2PM, 10AM, 6PM, 7AM, 10PM
6. **coupon:** Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50)
7. **expiration:** 1d, 2h (the coupon expires in 1 day or in 2 hours)
8. **gender:** Female, Male
9. **age:** 21, 46, 26, 31, 41, 50plus, 36, below21
10. **maritalStatus:** Unmarried partner, Single, Married partner, Divorced, Widowed
11. **has_Children:** 1, 0
12. **education:** Some college - no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School
13. **occupation:** Unemployed, Architecture & Engineering, Student, Education&Training&Library, Healthcare Support, Healthcare Practitioners & Technical, Sales & Related, Management, Arts Design Entertainment Sports & Media, Computer & Mathematical, Life Physical Social Science, Personal Care & Service, Community & Social Services, Office & Administrative Support, Construction & Extraction, Legal, Retired, Installation Maintenance & Repair, Transportation & Material Moving, Business & Financial, Protective Service, Food Preparation & Serving Related, Production Occupations, Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry
14. **income:** $37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 - $87499, $50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999, Less than $12500

15. **Bar:** never, less1, 1~3, gt8, nan4~8 (feature meaning: how many times do you go to a bar every month?)
16. **CoffeeHouse:** never, less1, 4~8, 1~3, gt8, nan (feature meaning: how many times do you go to a coffeehouse every month?)
17. **CarryAway:** n4~8, 1~3, gt8, less1, never (feature meaning: how many times do you get take-away food every month?)
18. **RestaurantLessThan20:** 4~8, 1~3, less1, gt8, never (feature meaning: how many times do you go to a restaurant with an average expense per person of less than $20 every month?)
19. **Restaurant20To50:** 1~3, less1, never, gt8, 4~8, nan (feature meaning: how many times do you go to a restaurant with average expense per person of $20 - $50 every month?)
20. **toCoupon_GEQ15min:** 0,1 (feature meaning: driving distance to the restaurant for using the coupon is greater than 15 minutes)
21. **toCoupon_GEQ25min:** 0, 1 (feature meaning: driving distance to the restaurant for using the coupon is greater than 25 minutes)
22. **direction_same:** 0, 1 (feature meaning: whether the restaurant is in the same direction as your current destination)
23. **direction_opp:** 1, 0 (feature meaning: whether the restaurant is in the same direction as your current destination)
24. **Y:** 1, 0 (whether the coupon is accepted)

Data file: received_coupon_in_car_10K.csv

REQUIRED TASKS: you have to perform the following tasks:

1. (10 points) Preprocess the data, e.g.,

- missing values

- typos, if any

- discretization, if needed

2. (10 points) Summarize your finding (e.g., statistics and attributes relationships via 6 visualizations (especially relating to the class attributes))

3. (10 points) Suggest 3 new strategies/ways to give coupons to the potential customers (state the reason to support your answers, in particular, how they are related to your data analysis/exploration?

4. (10 points) prepare PowerPoint presentation to be presented in the class. (The presentation file must be ***submitted before the deadline*** in *MS Team's assignment* – ***Mini-project1***).