UCB ML/AI Certification: Assignment 5.1

Problem Statement

This assignment seeks to answer the question, “Will a customer accept the coupon?” The goal of this project is to use what you know about visualizations and probability distributions to distinguish between customers who accepted a driving coupon versus those who did not.

The next sections of this README are written according to the CRISP model.

Data Understanding

This data set is from the UCI Machine Learning repository and was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios, including the destination, current time, weather, passenger, etc., and then asks people whether they will accept the coupon if they are the driver. Answers are given that the users will drive there “right away” or “later before the coupon expires” are labeled as “Y = 1”, and answers “no, I do not want the coupon” are labeled as “Y = 0”. There are five different types of coupons—less expensive restaurants (under $20), coffee houses, carry-out and takeaway bars, and more expensive restaurants ($20–$50).

The initial data set had 12684 entries in 26 columns.

The data was of three types:

1. User attributes: Gender, Age, Marital Status, Number of children, Education, Occupation, Annual income, the number of times the user frequented bars, bought takeaway food, went to a coffee shop, ate at a less expensive restaurant, and ate at an expensive restaurant
2. Contextual attributes: driving destination, the relative location of the user, weather, temperature, time, passenger,
3. Coupon attributes: the time before it expires

Data Preparation

The initial data set had 12684 entries in 26 columns. 18 data series have data type "object" and 8 data series have data type "int64".

Null Data

Null data check showed six data series had null data values.

```
car                     12576
Bar                       107
CoffeeHouse               217
CarryAway                 151
RestaurantLessThan20      130
Restaurant20To50          189
```

Of these the series "car" had 12576 out of 12684 entries as null. A check of the remaining 108 data points in the "car" series showed that they are not statistically significant. So the data series "car" was dropped.

Fewer Null Entries

The remaining null entries were total of 794 null data points across 604 rows in five data series. A quick analysis of these five data series indicated that these data points are not statistically significant. So the 604 rows with null data entries were dropped.

With this, the data frame has 12097 rows in 25 columns.

Duplicate entries

The data set had 72 duplicated rows and these were dropped. This final data frame had 12007 entries in 25 series.

See future work #1 and #2 for improvements in data preparation.

Modeling

No significant modeling was performed for this assignment. Most of the work involved direct analysis of the data set.

Evaluation

The assignment included specific questions, the answers to which are given in the Jupyter Notebook. One specific analysis is described here.

The problem was to explore of of the other coupon groups to determine the characteristics of users who accept the coupons. For this, I chose to determine which drivers accept the coupon to the value restaurants based on their age and income

a. go to restaurants more than once a month, age <=25 and age > 25, income < 50,000
b. go to restaurants based on marital status (Single, Married, other) and income < 50,000 and income => 50,000

A new dataframe was created with only the data series needed to address this problem statement.

For the problem "go to restaurants more than once a month, age <=25 and age > 25, income < 50,000", the findings are:

415 Coupons offered to Age <= 25
237 Coupons accepted by Age <= 25
57.11 Percentage Coupon Accepted by drivers Age <= 25

984 Coupons offered to Age > 25
783 Coupons accepted Age > 25
79.57 Percentage Coupon Accepted by drivers Age > 25

For the problem "go to restaurants based on marital status (Single, Married, other) and income < 50,000 and income => 50,000", the findings are:

2646 Total coupons issued for Restaurant<20
1878 Total coupons accepted for Restaurant<20

Percentage coupons accepted by drivers with income <$50,000
18.37 Single
9.03 Married
10.32 Others
Percentage coupons accepted by drivers with income =>$50,000
9.56 Single
18.59 Married
5.1 Others

Deployment

The Jupyter Notebook for this assignment is checked into the Git Hub repository at https://github.com/rvraj26/Assignment5.1_RajVarada

Results

Three main conclusions from the (limited) analysis performed on the data set are below.

1. The bar coupon acceptance rate (at 76.17%) is best amongst the drivers who frequent the bars more than one times a month.
2. Factoring among the three most common vectors (
   Age, company you keep, and presence of kids), the most accepted coupons are in the following groups:
   i. drivers that are with friends
   ii. drivers who frequent the bars more than one to three times a month
   iii. drivers who frequent the bars more than four to eight times a month
3. Coupons to less expensive restaurants that cost < $20 are most accepted by users of Age > 25

More coupons should be issued to these demographics to improve the coupon acceptance rate.

See future work #3 for arriving at a more comprehensive set of analysis

Future Work

1. Convert the data series with type “object” to numerical values. This will enable simpler and better analyses like correlation
2. For the data series that had a few rows with null values, replace the null data values with “reasonable” (mean) data values
3. Perform analysis on all data series to find a more comprehensive set of recommendations on the user groups to target for the coupons