**Name:** Garima Sharma

**Student ID:** 300343374

**Title:** Predict whether a person will accept the coupon recommended to him in different driving scenarios

1. **Introduction and discovery**

The dataset is taken from UCI, a Machine Learning Repository .It is a centre for machine learning and intelligent systems at UC Irvine. Because of research in field of machine learning and intelligent systems, these systems cop with the problem of developing the computer algorithms that can take and use the data in an intelligent way to solve a variety of real-world problems.

This data was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios including the destination, current time, weather, passenger, Time etc., and then ask the person whether he will accept the coupon if he is the driver. This data studies whether a person will accept the coupon recommended to him in different driving scenarios.

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. Here, we want to predict, which driving scenarios are affecting the person to choose to accept or reject the coupon.

1. **Data Preparation**

2.1**) Data inventory** – This data studies whether a person will accept the coupon recommended to him in different driving scenarios. The dataset has 12684 number of instances along with 23 number of attributes. This data was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios including the destination, current time, weather, passenger, Time etc., and then ask the person whether he will accept the coupon if he is the driver. I found that dataset on UCI machine learning repository and is donated on 2020-09-15. The data set is categorical which is suitable for classification analysis. It has 23 attributes that are mentioned below:

**Destination-**  No Urgent Place, Home, Work

**Passanger-**  Alone, Friend(s), Kid(s), Partner (who are the passengers in the car)

**Weather-** Sunny, Rainy, Snowy

**Temperature-** 55, 80, 30

**Time-**  2PM, 10AM, 6PM, 7AM, 10PM

**Coupon-**  Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50)

**Expiration-**  1d, 2h (the coupon expires in 1 day or in 2 hours)

**Gender-**  Female, Male

**Age-**  21, 46, 26, 31, 41, 50plus, 36, below21

**maritalStatus-**  Unmarried partner, Single, Married partner, Divorced, Widowed

**has\_Children-** 1, 0

**Education-**  Some college - no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School

**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

**Income-**  $37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 - $87499, $50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999, Less than $12500

**Bar-**  never, less1, 1~3, gt8, nan4~8 (feature meaning how many times do you go to a bar every month?)

**CoffeeHouse-** never, less1, 4~8, 1~3, gt8, nan (feature meaning how many times do you go to a coffeehouse every month?)

**CarryAway-** n4~8, 1~3, gt8, less1, never (feature meaning how many times do you get take-away food every month?)

**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?)

**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?)

**toCoupon\_GEQ15min-** 0,1 (feature meaning driving distance to the restaurant/bar for using the coupon is greater than 15 minutes)

**toCoupon\_GEQ25min-** 0, 1 (feature meaning driving distance to the restaurant/bar for using the coupon is greater than 25 minutes)

**direction\_same-** 0, 1 (feature meaning whether the restaurant/bar is in the same direction as your current destination)

**direction\_opp-** 1, 0 (feature meaning whether the restaurant/bar is in the same direction as your current destination)

**Y-** 1, 0 (whether the coupon is accepted)

2.2) **Data Processing:** In this dataset, it has 12684 observations where features car, Bar, CoffeeHouse, CarryAway, RestaurantLessThan20 and Restaurant20To50 has 12576, 107, 217, 151 and 130 null values respectively. Then, all the rows having null value are dropped. After that, feature car is also dropped as it has more than 75% of the null values. After the removal of this feature, any row having null value is dropped and at the end we are left with 12079 non-null values. Then we did the summary analysis on our data and we found that, our response variable has two levels as 0 and 1 where 0 represents that driver will not accept the coupon and 1 represents that the driver will accept the coupon. Then labels of the feature are changed as all of the column names has lower case first character which is capitalized.

From the summary of our dataset, we can see that feature Temperature has maximum value 80 and minimum value as 30 having median as 80. Most of the features has values as 0 and 1. As we can see, feature income as special character “$”, which is removed from the dataset and then unique values are checked for the income, where we found that it has 9 levels, where income’s lower level is when it is less than 12500 and maximum level is 100000 or more. Then the feature time is converted into AM as it has time in AM and PM format. The time is changed in 24-hour format for that. Moreover, column Expiration also represents value in days and in hours. So, data for this feature is converted in days instead of hours.

As we know, feature occupation has so many features. They need to be combined. So, different categorical values are combined under different labels for this feature such as category containing the Sales\_&\_Related, Education & Training, Library, Management, Office\_&\_Administrative\_Support and Business\_&\_Financial to the business & education category. Similar for other levels.

After this, feature Occupation’s whitespace is replaced with “\_”. From our data, we can see that there are three different features that show the driving distance to the restaurant/bar for using the coupon is greater than 5 minutes, 15 minutes and 25 minutes respectively. I combined these columns into one column having 3 different levels such as 1, 2, 3 described as driving distance to the restaurant/bar for using the coupon is greater than 5 minutes, 15 minutes and 25 minutes respectively. After this, all the unwanted columns are removed along with the duplicate values.

After all the data cleaning, univariate and multivariate analysis is done, which shows unique values along with count plot for most of the features such as :

Chart, bar chart

Description automatically generated

From the destination plot, we can see that we have higher counts for destination when it is not an urgent place followed by home and then work.

Chart, bar chart

Description automatically generated

From the Passanger plot, we can see that we have higher counts for Passenger when it is alone followed by friends, partner and then kids.

Chart, bar chart

Description automatically generated

From the Weather plot, we can see that we have higher counts for weather when it is sunny followed by snowy and then rainy weather.

|  |
| --- |
| Chart, bar chart  Description automatically generated  From the temperature plot, we can see that we have higher counts for temperature when it is 80. |

Chart, bar chart

Description automatically generated

From the graph, the graph shows that young people, specifically under 35 are highly likely to interact with coupons. They are more likely to accept the coupons as well as to not accept them as compared to the people above 35 years of age.

Chart, bar chart

Description automatically generated

From the graph, the graph shows that people with no urgent destination, are highly likely to interact with coupons. They are more likely to accept the coupons as well as to not accept them as compared to the people having destination as home or work.

Chart, bar chart

Description automatically generated

From the graph, the graph shows that people having income, specifically between 25000-37499 are highly likely to interact with coupons. They are more likely to accept the coupons as well as to not accept them as compared to the people having income from 12500-24999. People having income from 75000-87499, are somewhat likely to not to accept the coupons.

In our case, dummy coding transform is used and it transforms one feature into m-1 features. Here, categorical variables are converted into dummy variables. After that conversion, we got 70 columns in total. Then feature selection technique is used. I used variance threshold and selectk best with f\_classification features. Due to variance threshold technique, it selected 28 features whereas selectKbest technique selected 15 features, whereas Expiration, ToCoupon\_inMins, destination which is no urgent place and work along with other features are common among both of these techniques. After selection of the features, standard scaling and robust scaling technique is used on the selected features. In scaling features technique, only those features are scaled who are not in range of -1 to 1. After that, cleaned data is saved.

1. **Model Planning**:

According to my dataset, as my data contains categorical variables, I prefer to use Logistic Regression, Random forest Classifier, CatBoost Classifier, Decision Tree classifier among some other classifiers. Perhaps I will use Clustering as well. I will fit all the models using pipe and then see which model is giving me the best result. I will look for the accuracy along with confusion matrix. Then I will choose that model for further prediction analysis. I will see which features should be considered. At the end, best model will be used for analysis.