CS 584: Machine Learning

NAME : SOURAV YADAV

AID : A20450418

Spring 2020 Assignment 3

You are asked to use a decision tree model to predict the usage of a car. The data is the claim_history.csv which has 10,302 observations. The analysis specifications are:

Target Variable

- **CAR_USE**. The usage of a car. This variable has two categories which are *Commercial* and *Private*. The *Commercial* category is the Event value.

Nominal Predictor

- **CAR_TYPE**. The type of a car. This variable has six categories which are *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
- **OCCUPATION**. The occupation of the car owner. This variable has nine categories which are *Blue Collar*, *Clerical, Doctor*, *Home Maker*, *Lawyer*, *Manager*, *Professional*, *Student*, and *Unknown*.

Ordinal Predictor

- **EDUCATION**. The education level of the car owner. This variable has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

Analysis Specifications

- **Partition**. Specify the target variable as the stratum variable. Use stratified simple random sampling to put 75% of the records into the Training partition, and the remaining 25% of the records into the Test partition. The random state is 60616.
- **Decision Tree**. The maximum number of branches is two. The maximum depth is two. The split criterion is the Entropy metric.

Question 1 (20 points)

Please provide information about your Data Partition step. You may call the train_test_split() function in the sklearn.model_selection module in your code.

a) (5 points). Please provide the frequency table (i.e., counts and proportions) of the target variable in the Training partition?

```
Frequency Table of Target Varibale in Train
Counts
Private       4884
Commercial    2842
Name: CAR_USE, dtype: int64
Proportion
Private       0.632151
Commercial    0.367849
Name: CAR_USE, dtype: float64
```

b) (5 points). Please provide the frequency table (i.e., counts and proportions) of the target variable in the Test partition?

```
Frequency Table of Target Varibale in Test
Counts
Private       1629
Commercial     947
Name: CAR_USE, dtype: int64
Proportion
Private       0.632376
Commercial    0.367624
Name: CAR_USE, dtype: float64
```

c) (5 points). What is the probability that an observation is in the Training partition given that CAR_USE = *Commercial*?

The probability that an observation is in the Training partition given that CAR_USE = Commercial ::

0.7500659804697809

d) (5 points). What is the probability that an observation is in the Test partition given that CAR_USE = *Private*?

The probability that an observation is in the Training partition given that CAR_USE = Private ::

0.25011515430677106

Question 2 (40 points)

Please provide information about your decision tree. You will need to write your own Python program to find the answers.

a) (5 points). What is the entropy value of the root node?

Entropy of Root Node:

0.9490060293033189

b) (5 points). What is the split criterion (i.e., predictor name and values in the two branches) of the first layer?

Split Criterion:-

Predicator Name : OCCUPATION

Left Node ['Blue Collar', 'Student', 'Unknown']

Right Node ['Lawyer', 'Manager', 'Professional', 'Home Maker', 'Clerical', 'Doctor']

c) (10 points). What is the entropy of the split of the first layer?

Split Entropy :0.7184955941364273

d) (5 points). How many leaves?

Four Leaves

e) (10 points). Describe all your leaves. Please include the decision rules and the counts of the target values.

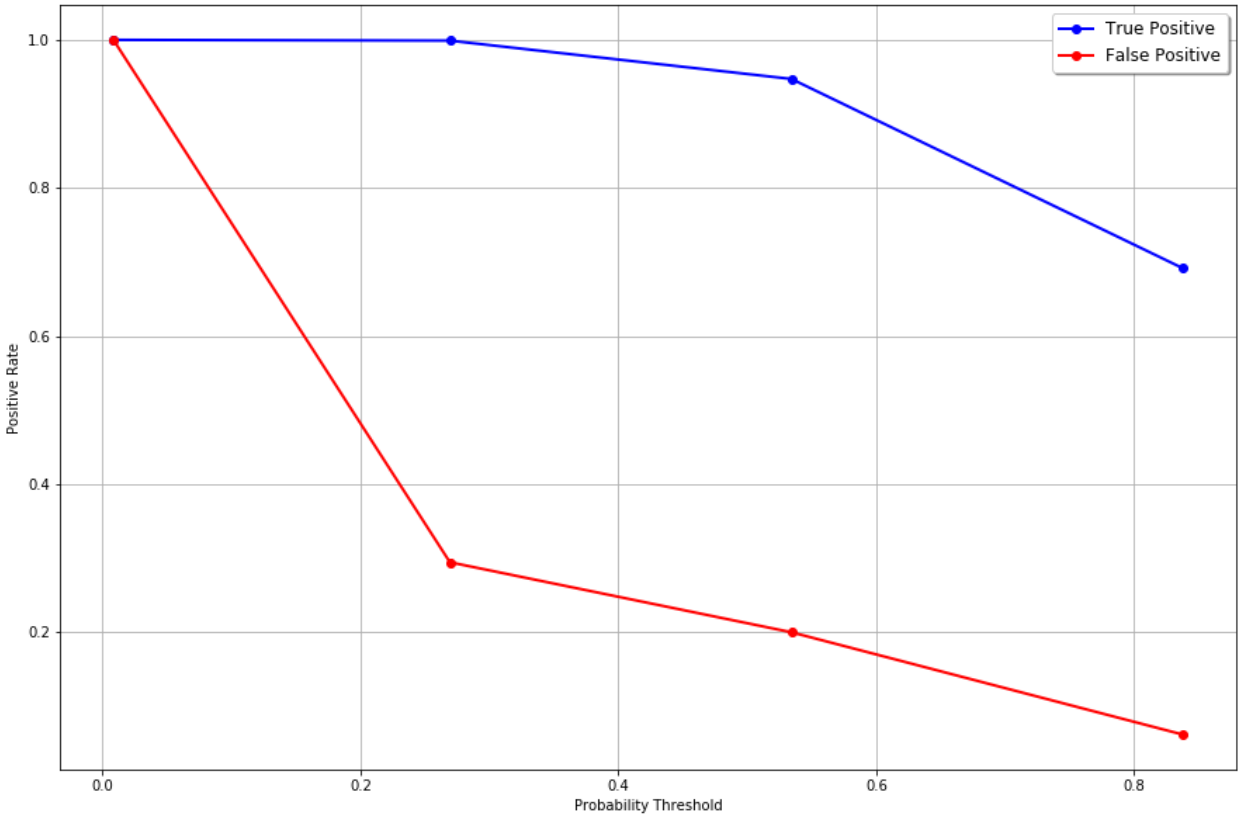
Leaves	**Decision Rule**	**Values**
Left_Subtree_Left_Node	If Occupation ['Blue Collar', 'Unknown', 'Student'] AND Education ['Below High School'] Then Data in Left_Subtree_Left_Leaf	Commercial 167 Private 453 Class: Private
Left_Subtree_Right_Node	If Occupation ['Blue Collar', 'Unknown', 'Student'] AND Education ['Bachelors', 'Masters', 'High School', 'Doctors'] Then Data in Left_Subtree_Right_Node	Commercial 1904 Private 369 Class : Commercial
Right_Subtree_Right_Node	If Occupation ['Lawyer', 'Home Maker', 'Clerical', 'Manager', 'Doctor', 'Professional'] AND Car_Type ['Minivan', 'SUV', 'Sports Car'] THEN Data in Right_Subtree_Left_Node	Commercial 29 Private 3415 Class: Private

Right_Subtree_Left_Node	If Occupation ['Lawyer', 'Home Maker', 'Clerical', 'Manager', 'Doctor', 'Professional'] AND Car_Type ['Van', 'Pickup', 'Panel Truck'] Then Data in Right_Subtree_Right_Leaf	Commercial 742 Private 647 Class: Commercial

f) (5 points). What are the Kolmogorov-Smirnov statistic and the event probability cutoff value?

KS Cut Off Probabilty:0.5341972642188625

Kolmogorov Smirnov statistic 0.7470789148375245



Question 3 (40 points)

Please apply your decision tree to the Test partition and then provide the following information. You will choose whether to call sklearn functions or write your own Python program to find the answers.

a) (5 points). Use the proportion of target Event value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?

Train Proportion Misclassification Rate using Threshold : 0.367849

Misclassification Rate: 0.14596273291925466

b) (5 points). Use the Kolmogorov-Smirnov event probability cutoff value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?

KS Cut Off Probabilty:0.5341972642188625

Corresponding Misclassification Rate: 0.15256211180124224

c) (5 points). What is the Root Average Squared Error in the Test partition?

Root Average Squared Error: 0.307288496016368

d) (5 points). What is the Area Under Curve in the Test partition?

Area Under Curve: 0.9315819462837962

e) (5 points). What is the Gini Coefficient in the Test partition?

Gini Coeffiecient:

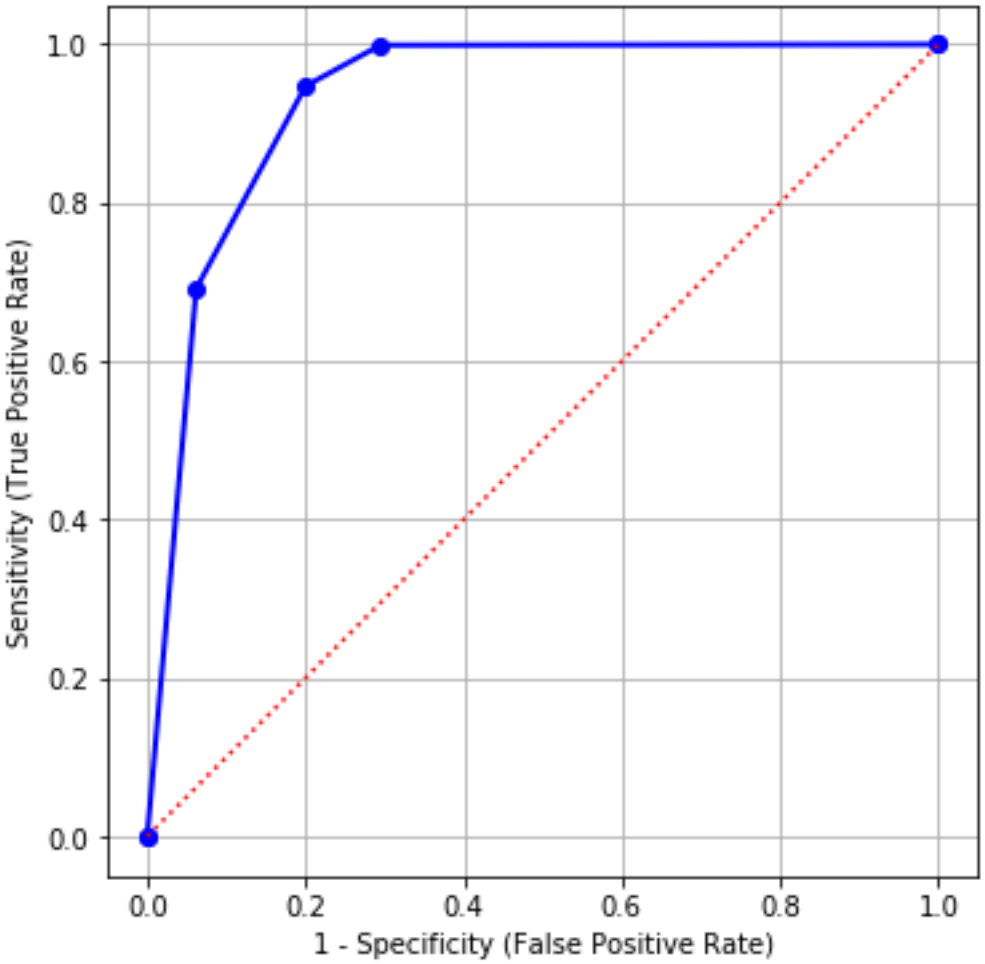
0.8631638925675925

f) (5 points). What is the Goodman-Kruskal Gamma statistic in the Test partition?

Goodman-Kruskal Gamma statistic :

0.9421295166209954

g) (10 points). Generate the Receiver Operating Characteristic curve for the Test partition. The axes must be properly labeled. Also, don't forget the diagonal reference line.


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity (True Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity (False Positive Rate)