CS 584: Machine Learning

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?

Ans: Frequency Table of the target variable in Training Partition:

```
     CAR_USE  COUNT  PROPORTION
0    Private   4884    0.632151
1 Commercial   2842    0.367849
```

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

Ans: Frequency Table of the target variable in Test Partition:

```
     CAR_USE  COUNT  PROPORTION
0    Private   1629    0.632376
1 Commercial    947    0.367624
```

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

Ans: Probability that an observation is in the Training partition given that CAR_USE = Commercial is: 0.7501144999138988

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

Ans: Probability that an observation is in the Test partition given that CAR_USE = Private is: 0.25006661142240155

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?

Ans: Root Node Entropy value: 0.9490060293033189

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

Ans:

Split criterion for first layer

predictor name: OCCUPATION

predictor value:

Left Subset: ('Blue Collar', 'Student', 'Unknown')

Right Subset: ('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')

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

Ans:

Entropy for Occupation 0.7184955941364275

Cross Table for Occupation

```
 CAR_USE   Commercial  Private   All
LE_Split
False              771     4062    4833
True               2071    822     2893
All                2842    4884    7726
```

d) (5 points). How many leaves?

Ans: There are 4 leaves

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

Ans:

Leaf 1:

Decision rules are 1.('Blue Collar', 'Student', 'Unknown') and 2.['Below High School']:

Entropy: 0.8405373462676067

Total Count: 620

Commercial Count: 167

Private Count: 453

Commercial Probability: 0.2693548387096774

Private Probability: 0.7306451612903225

Class: Private

Leaf 2:

Decision rules are 1.('Blue Collar', 'Student', 'Unknown') and 2.['High School', 'Bachelors', 'Masters', 'Doctors']:

Entropy: 0.639879533017315

Total Count: 2273

Commercial Count: 1904

Private Count: 369

Commercial Probability: 0.8376594808622966

Private Probability: 0.16234051913770348

Class: Commercial

Leaf 3:

Decision rules are 1.('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') and 2.('Minivan', 'SUV', 'Sports Car'):

Entropy: 0.07012958082027575

Total Count: 3444

Commercial Count: 29

Private Count: 3415

Commercial Probability: 0.008420441347270616

Private Probability: 0.9915795586527294

Class: Private

Leaf 4:

Decision rules are 1.('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') and 2.('Panel Truck', 'Pickup', 'Van'):

Entropy: 0.996623036579097

Total Count: 1389

Commercial Count: 742

Private Count: 647

Commercial Probability: 0.5341972642188625

Private Probability: 0.4658027357811375

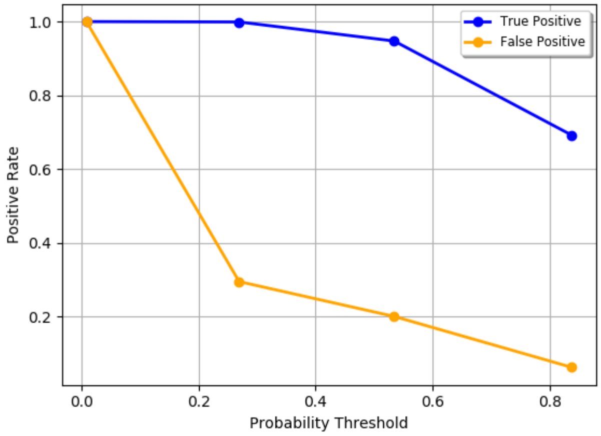
Class: Commercial

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

Ans:

The Kolmogorov-Smirnov statistic is 0.7470789148375245

Event probability cutoff value 0.5341972642188625



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?

Ans:

Accuracy: 0.8540372670807453

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?

Ans:

Accuracy: 0.8474378881987578

Misclassification Rate: 0.15256211180124224

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

Ans: Root Average Squared Error in the Test partition is 0.307288496016368

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

Ans: Area Under Curve in the Test Partition is 0.9315819462837962

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

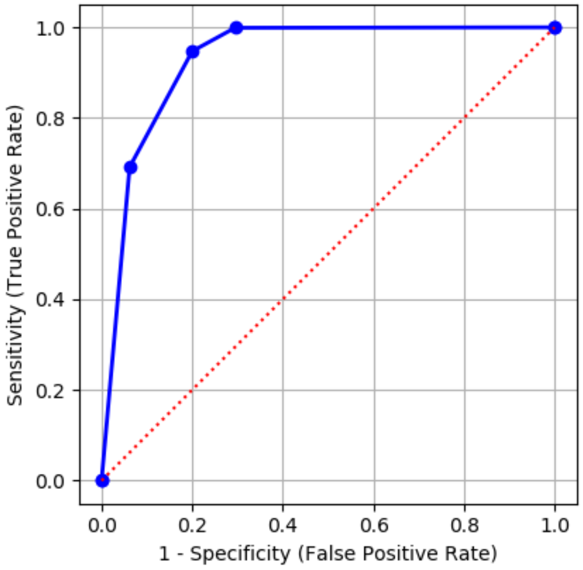
Ans: Gini coefficient in Test Partition is 0.8631638925675925

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

Ans: Goodman-Kruskal Gamma statistic in the Test partition is 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.

Ans:


Sensitivity (True Positive Rate)
1 - Specificity (False Positive Rate)