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.

```
For Training Data:
        Number of Observations =  7726
        Proportion of Dataset =  0.7499514657348088

For Testing Data:
        Number of Observations =  2576
        Proportion of Dataset =  0.25004853426519125
```

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

```
For Training Data:
Count of target variable is:
 CAR_USE
Commercial    2842
Private       4884
dtype: int64
Proportion of target variable is:
 CAR_USE
Commercial    0.367849
Private       0.632151
dtype: float64
```

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

```
For Testing Data:
Count of target variable is:
 CAR_USE
Commercial     947
Private       1629
dtype: int64
Proportion of target variable is:
 CAR_USE
Commercial    0.367624
Private       0.632376
dtype: float64
```

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

```
Probability(observation is in the Training partition / CAR_USE = Commercial) =
0.7501144999138988
```

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

```
Probability(observation is in the Test partition / CAR_USE = Private) =
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?

```
Entropy value of the root node is: 0.9490060293033189
```

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

```
Entropy of CarType is : 0.7672354798857937
Entropy of Occupation is : 0.7184955941364275
Entropy of Education is : 0.9354337466111206
Best split : "OCCUPATION"
the values in the two branches are:
For Left branch:  ['Blue Collar', 'Student', 'Unknown']
For Right branch:  ['Manager', 'Professional', 'Doctor', 'Home Maker', 'Lawyer', 'Clerical']
```

From the above results we can see that the predictor “Occupation” has the lowest Entropy value among the three predictors. Hence, the first split is based on Occupation which has left and right branches as shown above.

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

```
For the first layer Entropy of the split of will be: 0.7184955941364275
Occupation contingency table:
CAR_USE   Commercial  Private   All
LE_Split
False            771     4062  4833
True            2071      822  2893
All             2842     4884  7726
```

Here we can see that the Entropy of the split for the first layer will be 0.7184955941364275

d) (5 points). How many leaves?

There will be total **Four** leaves.

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

```
Leaf One:
Decision rules are:  ['Blue Collar', 'Student', 'Unknown'] ['Below High School']
Counts:
      CAR_USE  COUNT  PROPORTION
0     Private    453    0.730645
1  Commercial    167    0.269355
Entropy is : 0.8405373462676067
```

For the leaf one, predicted class will be “**private**” as the count of private is more than the count of commercial.

```
Leaf Two:
Decision rules are :  ['Blue Collar', 'Student', 'Unknown'] ['High School', 'Bachelors',
'Masters', 'Doctors']
Counts:
      CAR_USE  COUNT  PROPORTION
0  Commercial   1904    0.837659
1     Private    369    0.162341
Entropy: 0.639879533017315
```

For the leaf two, predicted class will be “**Commercial**” as the count of commercial is more than the count of private.

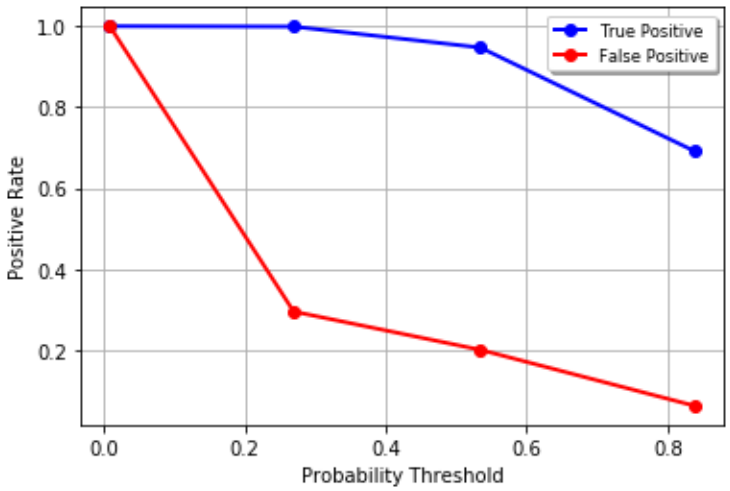
```
Leaf Three:
Decision rules are:  ['Professional', 'Clerical', 'Manager', 'Doctor', 'Home Maker', 'Lawyer']
['Pickup', 'Panel Truck', 'Van']
Counts:
      CAR_USE  COUNT  PROPORTION
0  Commercial    742    0.534197
1     Private    647    0.465803
Entropy: 0.9966230365790971
```

For the leaf three, predicted class will be "**Commercial**" as the count of commercial is more than the count of private.

```
Leaf Four:
Decision rules are  ['Professional', 'Clerical', 'Manager', 'Doctor', 'Home Maker', 'Lawyer']
['SUV', 'Minivan', 'Sports Car']
Counts:
      CAR_USE  COUNT  PROPORTION
0     Private   3415     0.99158
1  Commercial     29     0.00842
Entropy: 0.07012958082027575
```

For the leaf Four, predicted class will be "**Private**" as the count of private is more than the count of commercial.

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




```
The Kolmogorov Smirnov statistic is  0.7470789148375245
Event probability cutoff value 0.534197
```

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?

```
threshold value is: 0.3678488221589438
the Misclassification Rate in the Test partition is: 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?

```
Using Kolmogorov-Smirnov event probability cutoff value the Misclassification Rate
in the Test partition is: 0.15256211180124224
```

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

```
the Root Average Squared Error in the Test partition is: 0.30728850303889754
```

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

```
the Area Under Curve in the Test partition is: 0.9315819462837962
```

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

```
the Gini Coefficient in the Test partition is: 0.8631638925675924
```

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

```
the 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.