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.

	Count	**Proportion**
Car_Use		
Private	**4884**	**0.632151**
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.

	Count	Proportion
Car_Use		
Private	**1629**	**0.632376**
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.

	Count
Car_Use	
Commercial	**0.750066**

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

Ans.

	Count
Car_Use	
Private	**0.250115**

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. **0.94900603**

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

Ans. **Predictor Name: CAR_TYPE**

Left Child: [('Blue Collar', 'Student', 'Unknown')]

Right Child: [('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')]

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

Ans. **Entropy of split of First Layer: 0.9489832368663098**

d) (5 points). How many leaves?

Ans. **Number of Leaves: 4**

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

Ans. **Description of Leaf Nodes**

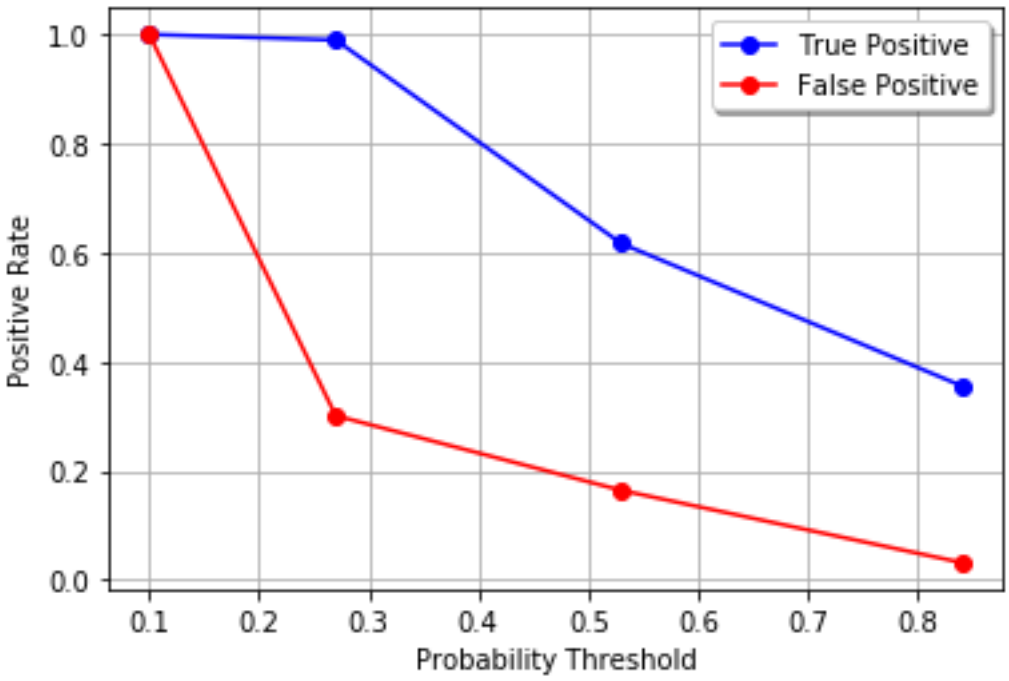
Decision Rules
Leaf Node 1:
{'Education': [('Below High School',)], 'Occupation': [('Blue Collar', 'Student', 'Unknown')]}
Leaf Node 2:
{'Education': [('High School', 'Bachelors', 'Masters', 'Doctors')], 'Occupation': [('Blue Collar', 'Student', 'Unknown')]}
Leaf Node 3:
{'Car_Type': [('Minivan', 'SUV', 'Sports Car')], 'Occupation': [('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')]}
Leaf Node 4:
{'Car_Type': [('Panel Truck', 'Pickup', 'Van')], 'Occupation': [('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')]}

	Entropy	No of Observations	% of Commercial
Index			
Leaf Node 1	**0.8405373462676066**	**620**	**27.0**
Leaf Node 2	**0.6398795330173151**	**2273**	**84.0**
Leaf Node 3	**0.07012958082027576**	**3444**	**1.0**
Leaf Node 4	**0.9966230365790971**	**1389**	**53.0**

	Predicted Class	Commercial	Private
Index			
Leaf Node 1	**Private**	**167**	**453**
Leaf Node 2	**Commercial**	**1904**	**369**
Leaf Node 3	**Private**	**29**	**3415**
Leaf Node 4	**Commercial**	**742**	**647**

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

Ans.



Kolmogorov-Smirnov

From the Graph,

The KS Statistic : (0.65-0.18) = 0.47

Commercial Probability Cut-off = 0.65

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. **MissClassification 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. **MissClassification Rate with Kolmogorov-Smirnov event probability cutoff value as Threshold: 0.15256211180124224**

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

Ans. **Root Mean Squared Error for Test Partition: 0.3142158483891002**

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

Ans. **Area Under Curve in Test Partition: 0.9315819462837962**

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

Ans. **GINI Coefficient in the Test Partition: 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: 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.