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?

Target Variable in Training Partition

CAR_USE	Count	Proportions
Commercial	2851	0.369014
Private	4875	0.630986

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

Test Variable in Training Partition

CAR_USE	Count	Proportions
Commercial	983	0.36413
Private	1683	0.63587

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

Probability of an observation present in Training with Car is Commercial: **0.7524898516284976**

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

Probability of an observation present in Test with Car is private: **0.2514484970903517**

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?

Based on Target Variable in Training Partition

CAR_USE	Count	Proportions
Commercial	2851	0.369014
Private	4875	0.630986

Probability of **Commercial** car in train data set = **0.369014**
Probability of **Private** car in train data set = **0. 630986**
Entropy = -(probTrainCommCar * log(probTrainCommCar) + probTrainPrivateCar * log(probTrainPrivateCar))

Entropy value of the root node **0.9499120060532422**

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

EDUCATION

Left Child	[Below High School]
Right Child	[High School, Bachelors, Masters, Doctors]
Entropy	**0.938232**

OCCUPATION
Left Child [Blue Collar, Student, Unknown]
Right Child [Clerical, Doctor, Home Maker, Lawyer, Manager, Professional]
Entropy **0.713872**

CAR TYPE
Left Child [Minivan, SUV, Sports Car]
Right Child [Panel Truck, Pickup, Van]
Entropy **0.768504**

In my observation for predictor **OCCUPATION,** getting lowest entropy value **0. 713872**. So this will be our split criterion. And Left branch elements are [**Blue Collar, Student, Unknown**] and Right branch elements are [**Clerical, Doctor, Home Maker, Lawyer, Manager, Professional**]

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

Entropy of the first layer is **0.7138723890228705**

d) (5 points). How many leaves?

In the next layer, Splits of lowest Entropy with each predicates

Left Side if layer 1
CAR TYPE
Left Child [Minivan, SUV, Sports Car]
Right Child [Panel Truck, Pickup, Van]
Entropy 0.769665

Occupation
Left Child [Student]
Right Child [Blue Collar, Unknown]
Entropy 0.802303

Education
Left Child [Below High School]
Right Child [High School, Bachelors, Masters, Doctors]
Entropy **0.673644**

- From the left branch we found Education will be our second layer.
- Left branch consists “Below High School”
- Right branch consists “High School, Bachelors, Masters, Doctors”

Right Side if layer 1

CAR TYPE

Left Child	[Minivan, SUV, Sports Car]
Right Child	[Panel Truck, Pickup, Van]
Entropy	**0.328087**

Occupation

Left Child	[Doctor, Home Maker, Lawyer]
Right Child	[Clerical, Manager, Professional]
Entropy	0.564225

Education

Left Child	[Below High School, High School, Bachelors]
Right Child	[Masters, Doctors]
Entropy	0.617822

- From the right branch we found Car Type will be our second layer.
- Left branch consists “Minivan, SUV, Sports Car”
- Right branch consists “Panel Truck, Pickup, Van”

In this way we got the **4** leaves in the depth 2 tree.

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

```
Counts:
CAR_USE  Commercial  Private
Leaf
0             173      460
1            1920      356
2              23     3409
3             735      650
```

```
probabilityCAR_USE:
CAR_USE  Commercial   Private
Leaf
0        0.273302  0.726698
1        0.843585  0.156415
2        0.006702  0.993298
3        0.530686  0.469314
```

		Target: CAR_USE	
Leaf	**Decision Rule**	Commercial	Private
0	OCCUPATION in ['Blue Collar', 'Student', 'Unknown'] and EDUCATION in ['Below High School']	173 (0.273302)	460 (0.726698)
1	OCCUPATION in ['Blue Collar', 'Student', 'Unknown'] and EDUCATION in ['High School', 'Bachelors', 'Masters', 'Doctors']	1920 (0.843585)	356 (0.156415)
2	OCCUPATION in ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] and 'CAR_TYPE in ['Minivan', 'SUV', 'Sports Car']	23 (0.006702)	3409 (0. 993298)
3	OCCUPATION in ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] and 'CAR_TYPE in ['Panel Truck', 'Pickup', 'Van']	735 (0.530686)	650 (0.46931)
	Overall	**2851** **(0.3690137)**	**4875** **(0.6309862)**

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

Threshold: **0.3690137199068082**

```
[1.84358524,  0.84358524,  0.53068592,   0.27330174,   0.00670163]
[0.,              0.60042227,    0.72489322,  0.69121471,   0.             ]
```

Kolmogorov-Smirnov statistic: **0.72489322**

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?

I took Event as a Commercial, and its proportion in train data sets is **0.369014.**
Threshold of the event: **0.369014.**

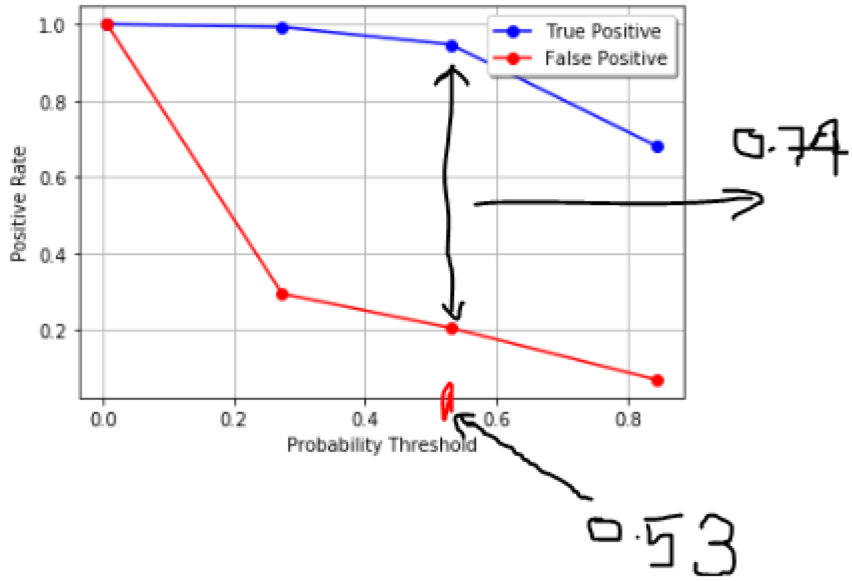
Observed CAR_USE	Predicted CAR_USE	
	Commercial (EVENT)	Private(Non Event)
Commercial (EVENT)	888 (TP)	50 (FN)
Private(Non Event)	336 (FP)	1302 (TN)

Misclassification Rate = (Number of misclassified observations) / (Number of observations)
Misclassification Rate = FN+FP / TP+TN+FN+FP => (50+336)/(888+1302+50+336)
MR = 386/2576 => 0.1498447204
Misclassification Rate : 0.1498447204 OR 14.98 %

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?

[1.84358524, 0.84358524, **0.53068592**, 0.27330174, 0.00670163] ------ Threshold
[0, 0.6116396, **0.74156689**, 0.69766552, 0.] ------ TP-FP

Cut off value: **0.53068592**
KS Statistic: **0.74156689**



Accuracy: **0.8501552795031**

Misclassification Rate **(1- Accuracy)**: **0.1498447204969**

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

Root Average Squared Error = **0.3144866942298878**

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

Area Under Curve: **0.9235377273757;** Acceptable model because **AUC > 0.5**

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

Gini Coff = 2*AUC - 1
Gini Coefficient: **0.8470754547514.**
Gini Coefficient say that our model will be accepted if we have **Gini Coff > 0**

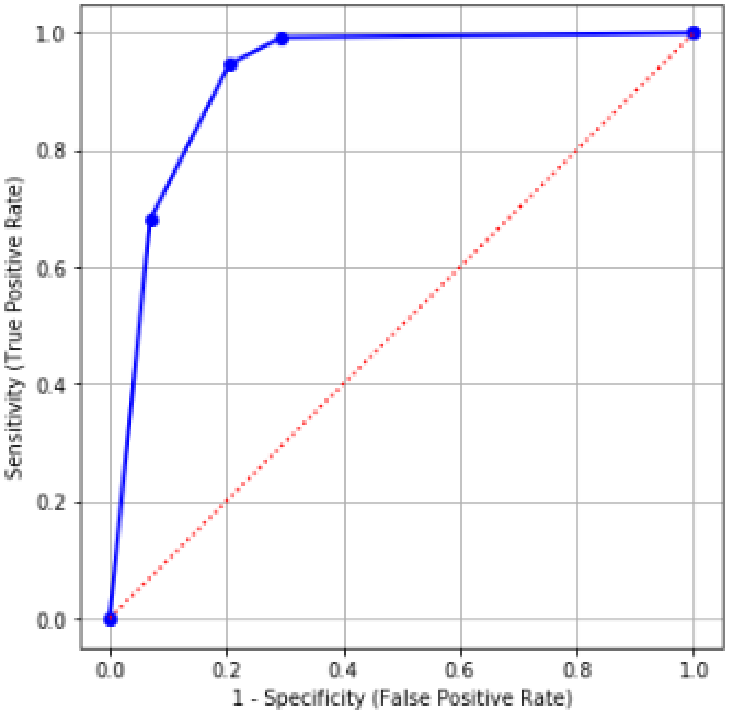
f) (5 points). What is the Goodman-Kruskal Gamma statistic in the Test partition?

Number of Pairs: 1536444
Number of Concordant (C) pairs: 1347699
Number of Discordant (D) pairs: 46215
Number of Tied (T) pairs: 142530

Area = 0.5 + 0.5*((Concordant - Discordant)/pair) = 0.9235377273756804
Gini: (Concordant - Discordant)/pair) = 0.8470754547513609

Goodman-Kruskal Gamma statistic: **(Concordant - Discordant)/ (Concordant + Discordant)**
Gamma statistic: **0.9336903137496287**

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.


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