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*.

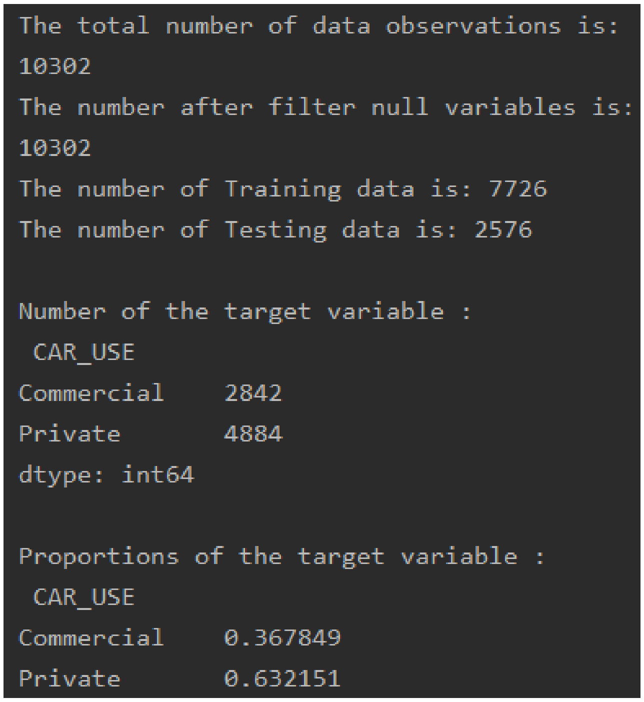
Analysis Specifications

- **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)

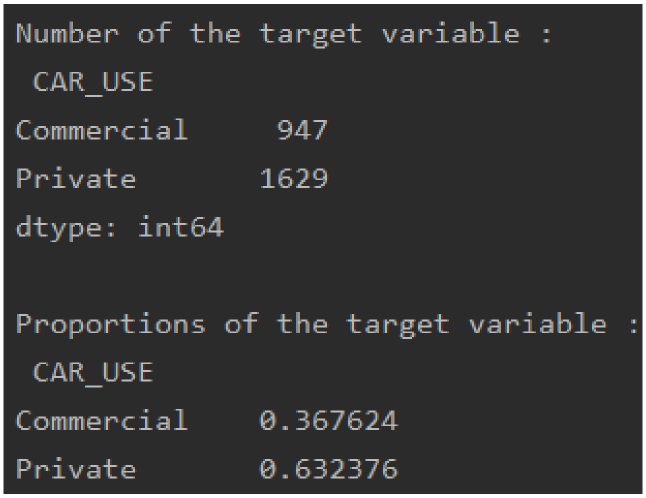
Please provide information about your Data Partition step. You may call the train_test_split() function in the sklearn.model_selection module in your code.

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



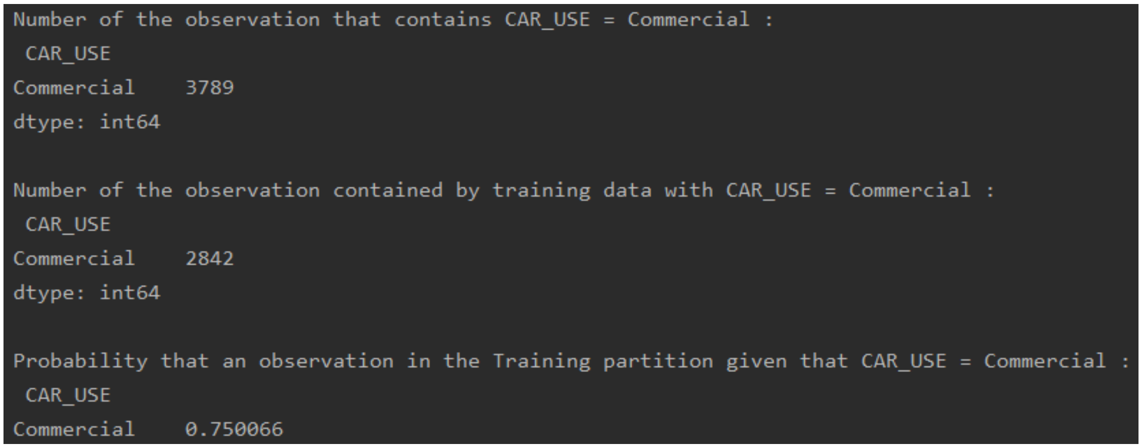
```
The total number of data observations is:
10302
The number after filter null variables is:
10302
The number of Training data is: 7726
The number of Testing data is: 2576

Number of the target variable :
 CAR_USE
Commercial    2842
Private       4884
dtype: int64

Proportions of the target variable :
 CAR_USE
Commercial    0.367849
Private       0.632151
```

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



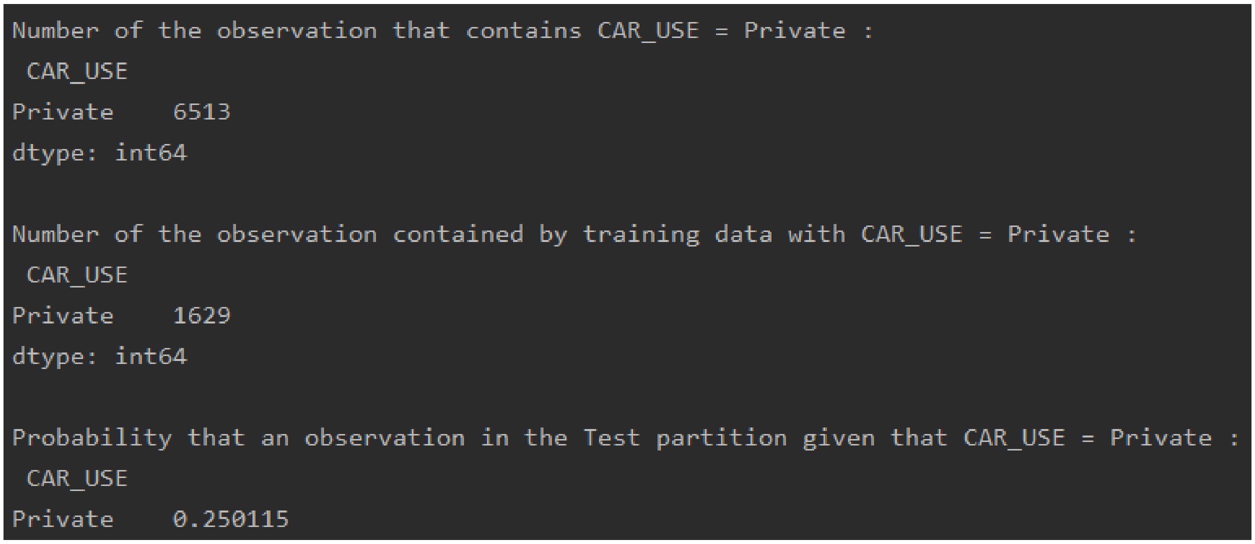
```
Number of the target variable :
 CAR_USE
Commercial     947
Private       1629
dtype: int64

Proportions of the target variable :
 CAR_USE
Commercial    0.367624
Private       0.632376
```

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



```
Number of the observation that contains CAR_USE = Commercial :
 CAR_USE
Commercial    3789
dtype: int64

Number of the observation contained by training data with CAR_USE = Commercial :
 CAR_USE
Commercial    2842
dtype: int64

Probability that an observation in the Training partition given that CAR_USE = Commercial :
 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*?

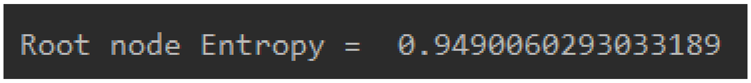


```
Number of the observation that contains CAR_USE = Private :
 CAR_USE
Private    6513
dtype: int64

Number of the observation contained by training data with CAR_USE = Private :
 CAR_USE
Private    1629
dtype: int64

Probability that an observation in the Test partition given that CAR_USE = Private :
 CAR_USE
Private    0.250115
```

Question 2 (40 points)

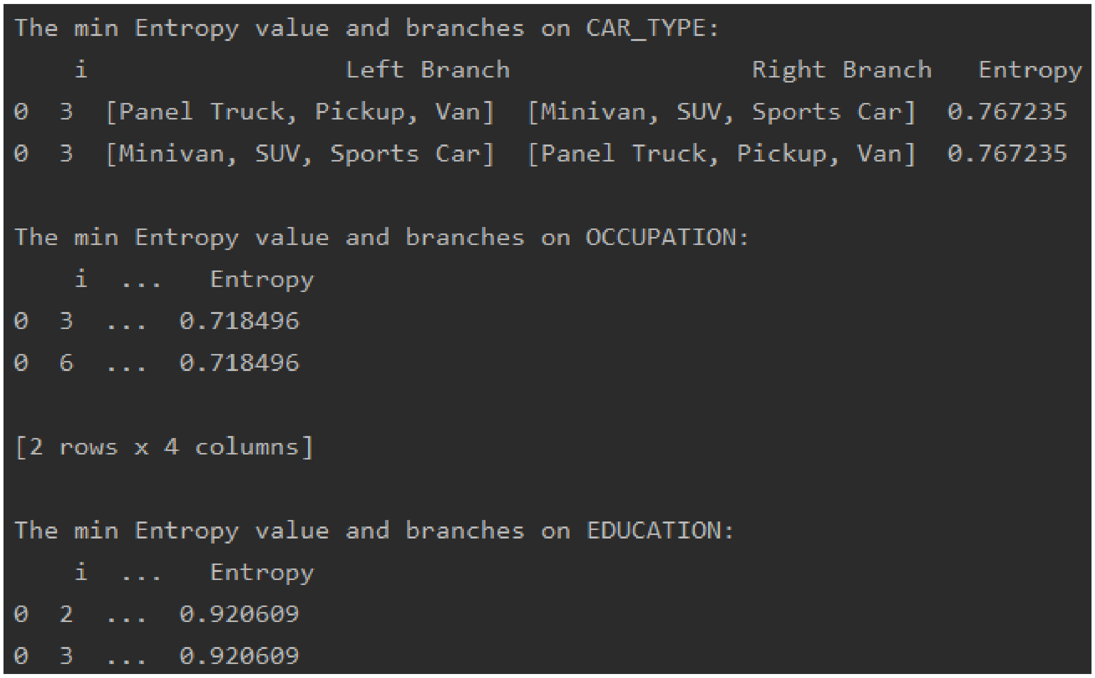
Please provide information about your decision tree. You will need to write your own Python program to find the answers.

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

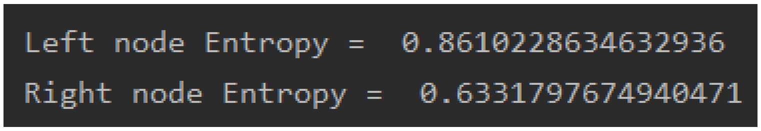


```
Root node Entropy =  0.9490060293033189
```

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

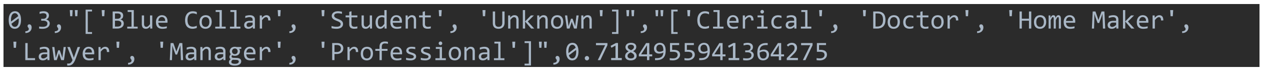


```
The min Entropy value and branches on CAR_TYPE:
    i                  Left Branch                  Right Branch   Entropy
0  3  [Panel Truck, Pickup, Van]  [Minivan, SUV, Sports Car]  0.767235
0  3  [Minivan, SUV, Sports Car]  [Panel Truck, Pickup, Van]  0.767235

The min Entropy value and branches on OCCUPATION:
    i  ...   Entropy
0  3  ...  0.718496
0  6  ...  0.718496

[2 rows x 4 columns]

The min Entropy value and branches on EDUCATION:
    i  ...   Entropy
0  2  ...  0.920609
0  3  ...  0.920609
```



```
Left node Entropy =  0.8610228634632936
Right node Entropy =  0.6331797674940471
```

According to these three result of the first level, I will choose the one split criterion which Entropy value is smallest. So The split criterion is:



```
0,3,"['Blue Collar', 'Student', 'Unknown']","['Clerical', 'Doctor', 'Home Maker',
'Lawyer', 'Manager', 'Professional']",0.7184955941364275
```

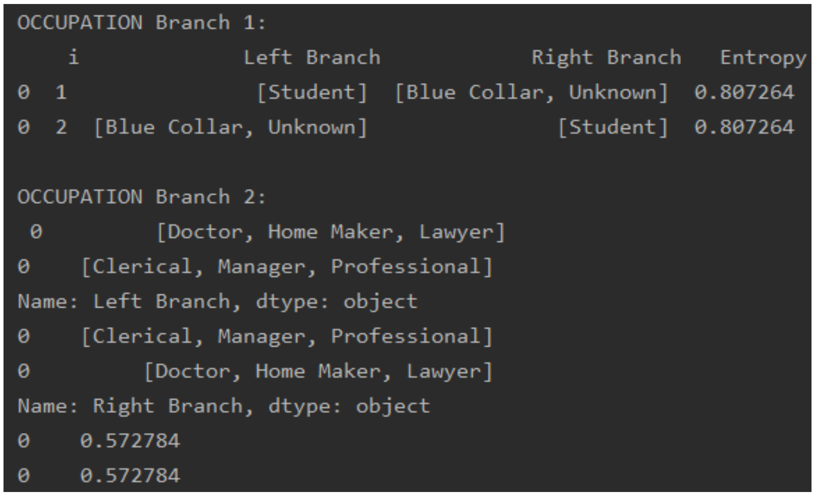
c) (10 points). What is the entropy of the split of the first layer?

The entropy of the split of the first layer is 0.7184955941364275.

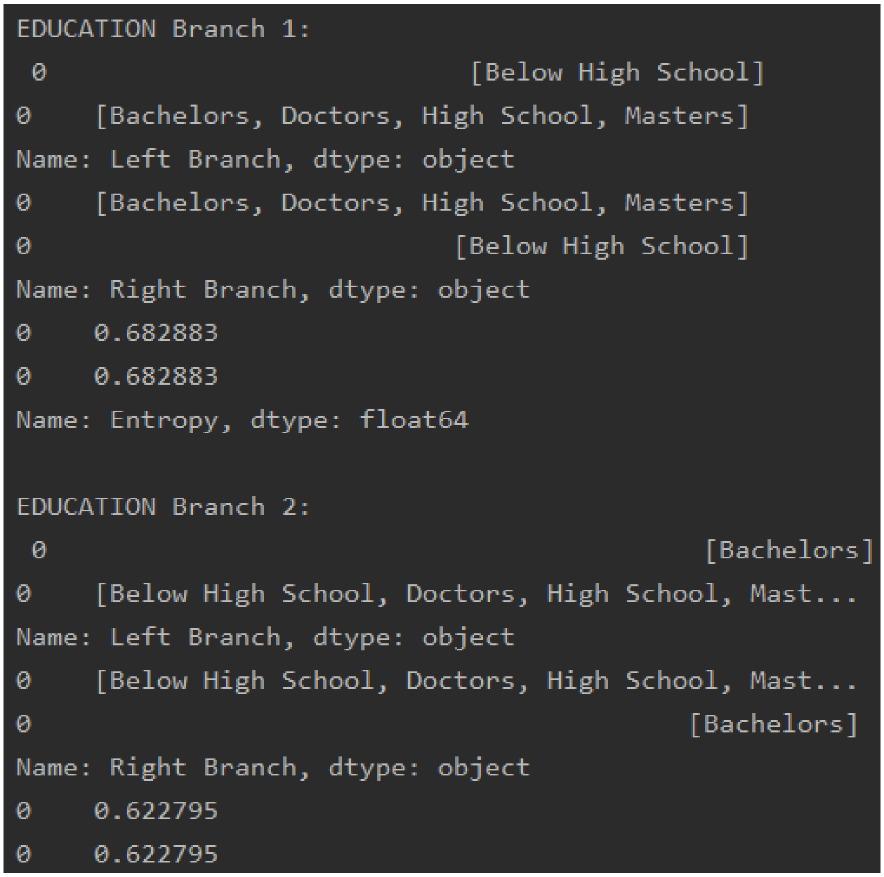
d) (5 points). How many leaves?

Because the maximum number of branches is two, and the maximum depth is two, so there are 2^2 = 4 leaves.

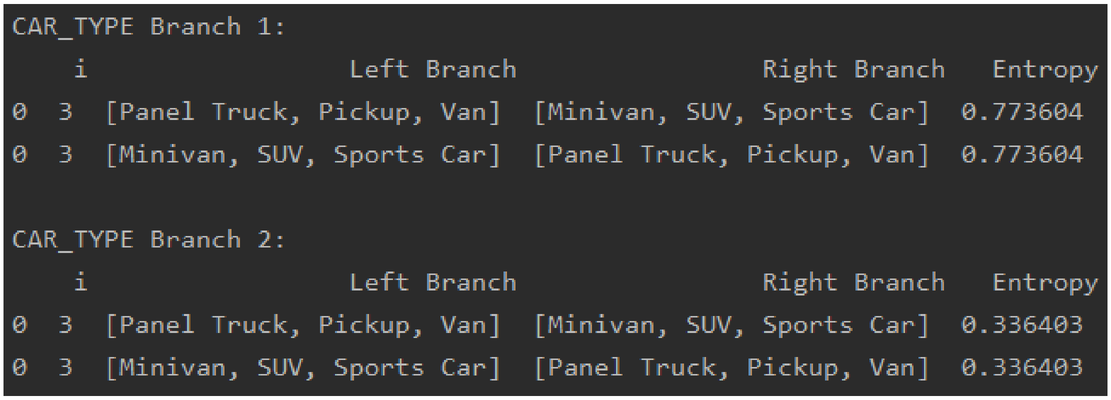
e) (10 points). Describe all your leaves. Please include the decision rules and the counts of the target values.



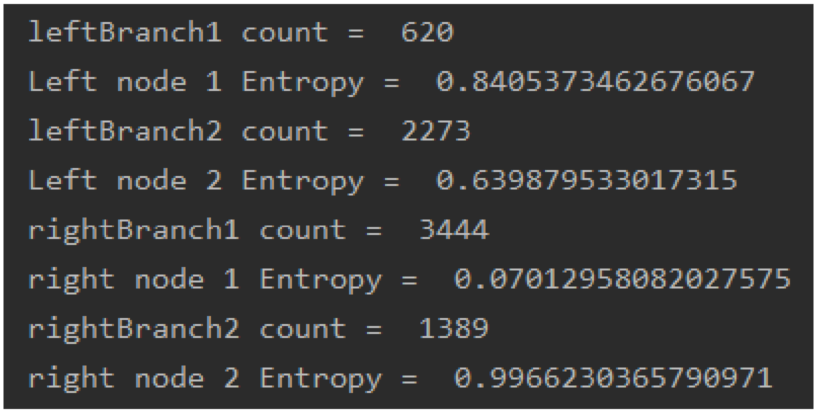
```
OCCUPATION Branch 1:
    i             Left Branch            Right Branch   Entropy
0  1              [Student]  [Blue Collar, Unknown]  0.807264
0  2  [Blue Collar, Unknown]              [Student]  0.807264

OCCUPATION Branch 2:
 0         [Doctor, Home Maker, Lawyer]
0    [Clerical, Manager, Professional]
Name: Left Branch, dtype: object
0    [Clerical, Manager, Professional]
0         [Doctor, Home Maker, Lawyer]
Name: Right Branch, dtype: object
0    0.572784
0    0.572784
```



```
EDUCATION Branch 1:
 0                           [Below High School]
0    [Bachelors, Doctors, High School, Masters]
Name: Left Branch, dtype: object
0    [Bachelors, Doctors, High School, Masters]
0                           [Below High School]
Name: Right Branch, dtype: object
0    0.682883
0    0.682883
Name: Entropy, dtype: float64

EDUCATION Branch 2:
 0                                          [Bachelors]
0    [Below High School, Doctors, High School, Mast...
Name: Left Branch, dtype: object
0    [Below High School, Doctors, High School, Mast...
0                                          [Bachelors]
Name: Right Branch, dtype: object
0    0.622795
0    0.622795
```



```
CAR_TYPE Branch 1:
   i                 Left Branch                Right Branch   Entropy
0  3  [Panel Truck, Pickup, Van]  [Minivan, SUV, Sports Car]  0.773604
0  3  [Minivan, SUV, Sports Car]  [Panel Truck, Pickup, Van]  0.773604

CAR_TYPE Branch 2:
   i                 Left Branch                Right Branch   Entropy
0  3  [Panel Truck, Pickup, Van]  [Minivan, SUV, Sports Car]  0.336403
0  3  [Minivan, SUV, Sports Car]  [Panel Truck, Pickup, Van]  0.336403
```



```
leftBranch1 count =  620
Left node 1 Entropy =  0.8405373462676067
leftBranch2 count =  2273
Left node 2 Entropy =  0.639879533017315
rightBranch1 count =  3444
right node 1 Entropy =  0.07012958082027575
rightBranch2 count =  1389
right node 2 Entropy =  0.9966230365790971
```

Decision Rule	Number Observations

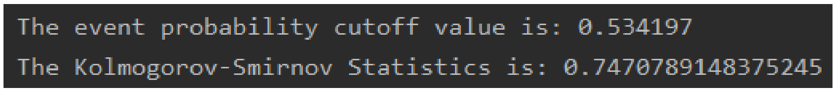
OCCUPATION = ['Blue Collar', 'Student', 'Unknown']
EDUCATION = ['Below High School']
	620

OCCUPATION = ['Blue Collar', 'Student', 'Unknown']
EDUCATION = ['Bachelors', 'Doctors', 'High School', 'Masters']
	2273

OCCUPATION = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']
CAR_TYPE = ['Minivan', 'SUV', 'Sports Car']
	3444

OCCUPATION = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']
CAR_TYPE = ['Panel Truck', 'Pickup', 'Van']
	1389

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

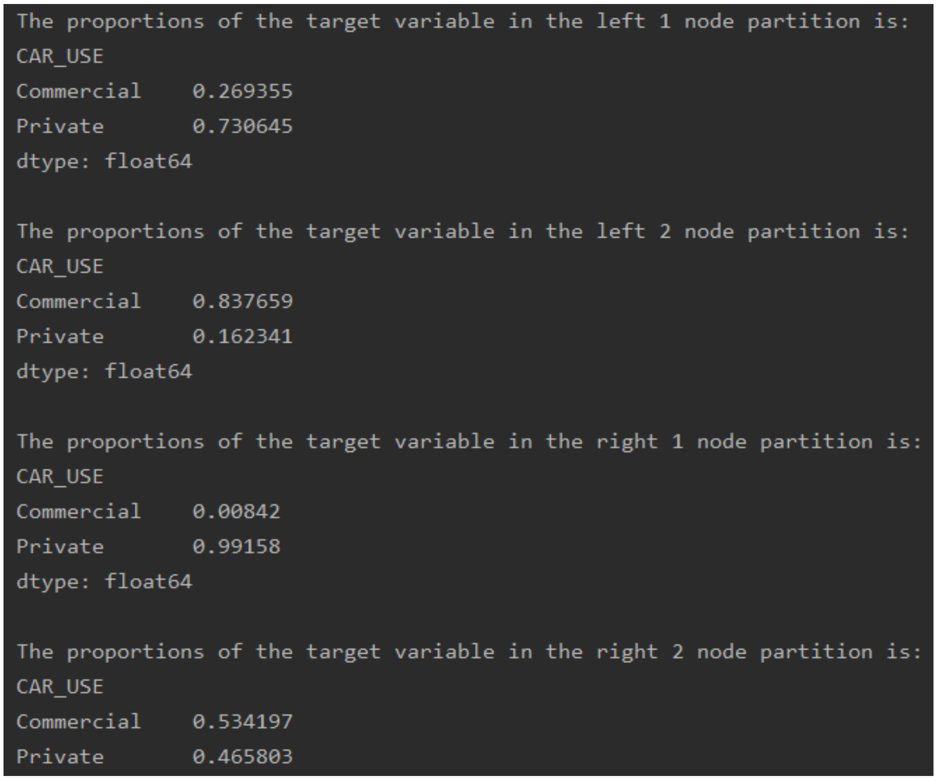


```
The event probability cutoff value is: 0.534197
The Kolmogorov-Smirnov Statistics is: 0.7470789148375245
```

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.

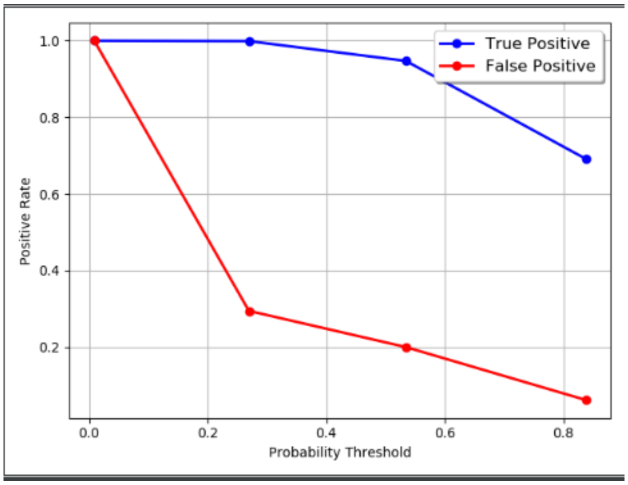
a) (5 points). Use the proportion of target Event value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?

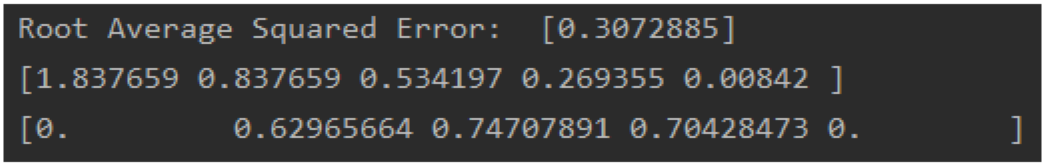


```
The proportions of the target variable in the left 1 node partition is:
CAR_USE
Commercial    0.269355
Private       0.730645
dtype: float64

The proportions of the target variable in the left 2 node partition is:
CAR_USE
Commercial    0.837659
Private       0.162341
dtype: float64

The proportions of the target variable in the right 1 node partition is:
CAR_USE
Commercial    0.00842
Private       0.99158
dtype: float64

The proportions of the target variable in the right 2 node partition is:
CAR_USE
Commercial    0.534197
Private       0.465803
```

Misclassification Rate: 0.1459627329193

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?

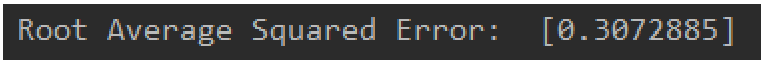




```
Root Average Squared Error:  [0.3072885]
[1.837659 0.837659 0.534197 0.269355 0.00842 ]
[0.         0.62965664 0.74707891 0.70428473 0.         ]
```

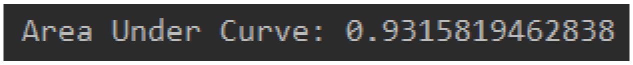
According to the result show above, the largest difference happens when the probability threshold is 0.53419726, and use it as the thresholds, then the leaf1 and leaf4 are nonevent, leaf2 and leaf3 are event('Commercial'), it is same as Q3a, so the Misclassification Rate is also 14.596%.

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



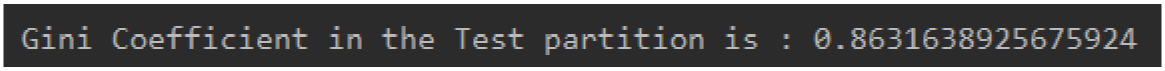
```
Root Average Squared Error:  [0.3072885]
```

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



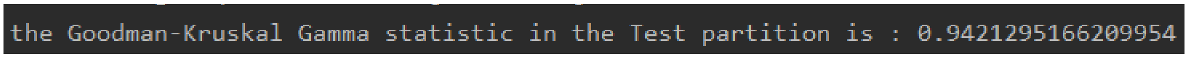
```
Area Under Curve: 0.9315819462838
```

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



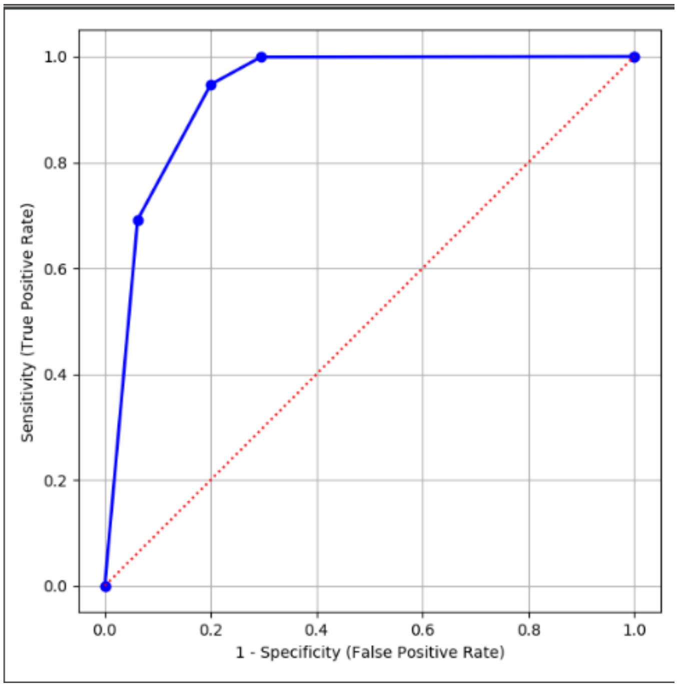
```
Gini Coefficient in the Test partition is : 0.8631638925675924
```

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.


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