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*.

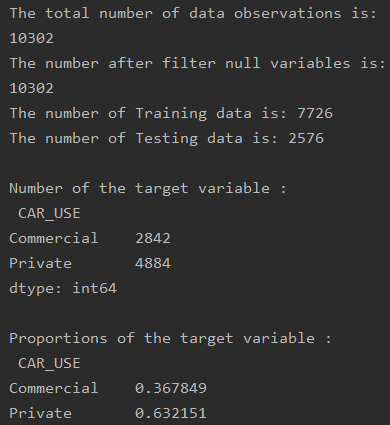
**Analysis Specifications**

* **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.

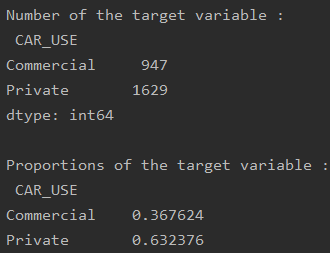
# Question 1 (20 points)

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.

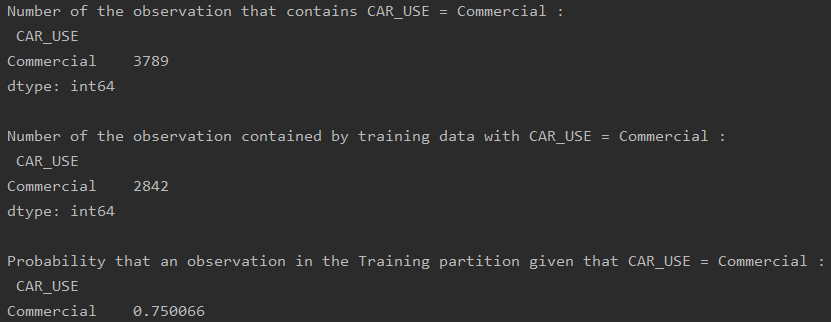
1. (5 points). Please provide the frequency table (i.e., counts and proportions) of the target variable in the Training partition?



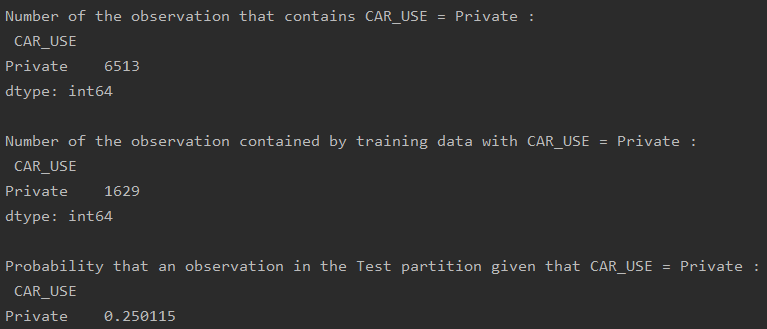
1. (5 points). Please provide the frequency table (i.e., counts and proportions) of the target variable in the Test partition?



1. (5 points). What is the probability that an observation is in the Training partition given that CAR\_USE = *Commercial*?



1. (5 points). What is the probability that an observation is in the Test partition given that CAR\_USE = *Private*?



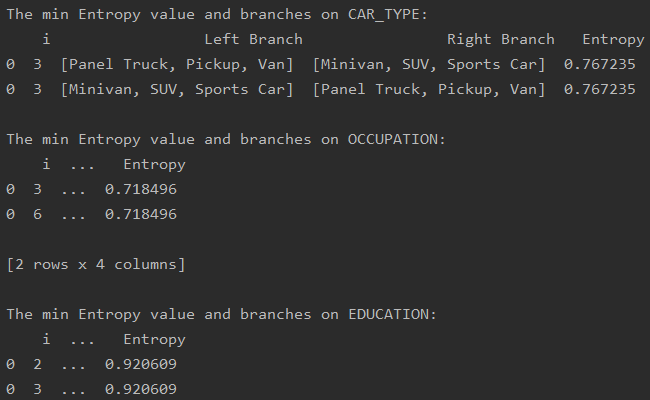
# Question 2 (40 points)

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

1. (5 points). What is the entropy value of the root node?



1. (5 points). What is the split criterion (i.e., predictor name and values in the two branches) of the first layer?





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

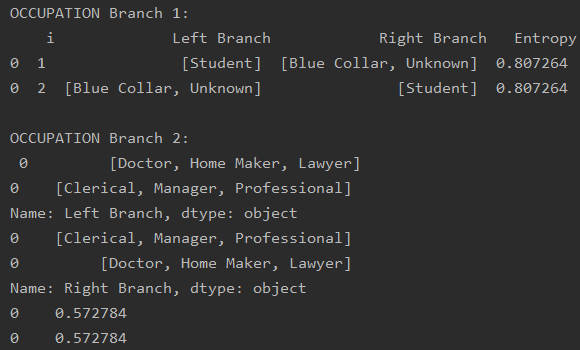
1. (10 points). What is the entropy of the split of the first layer?

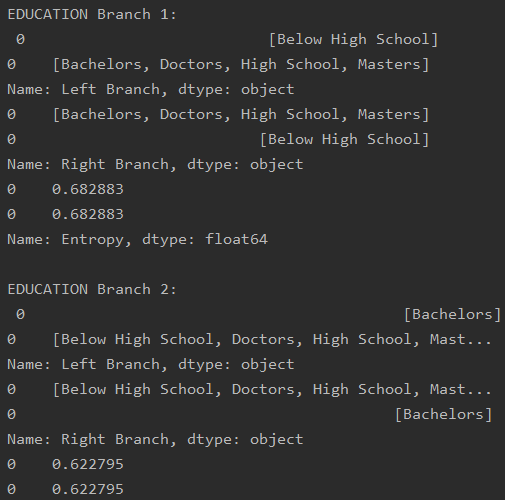
The entropy of the split of the first layer is 0.7184955941364275.

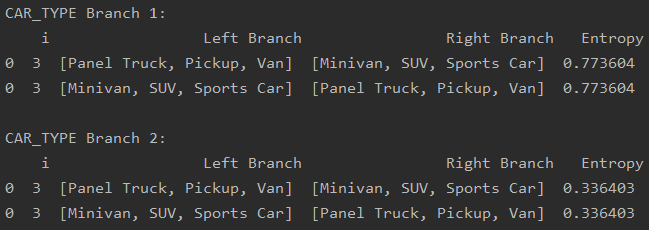
1. (5 points). How many leaves?

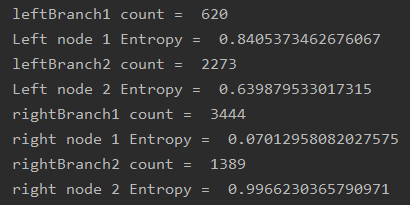
Because the maximum number of branches is two, and the maximum depth is two, so there are 2^2 = 4 leaves.

1. (10 points). Describe all your leaves. Please include the decision rules and the counts of the target values.









|  |  |
| --- | --- |
| 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 |

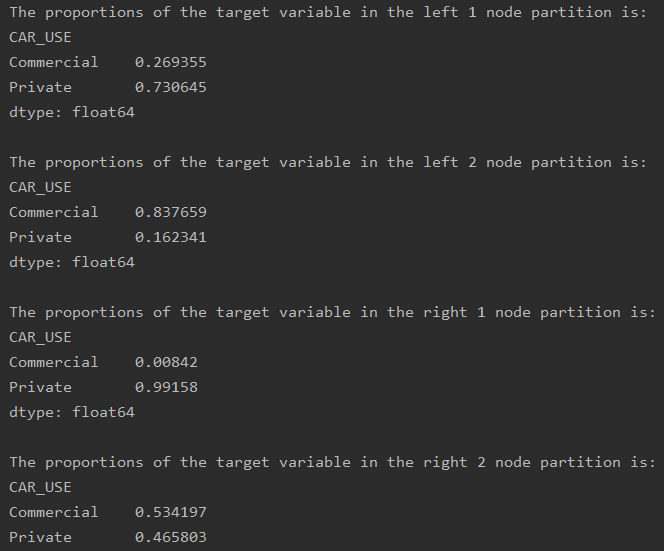
1. (5 points). What are the Kolmogorov-Smirnov statistic and the event probability cutoff value?



# Question 3 (40 points)

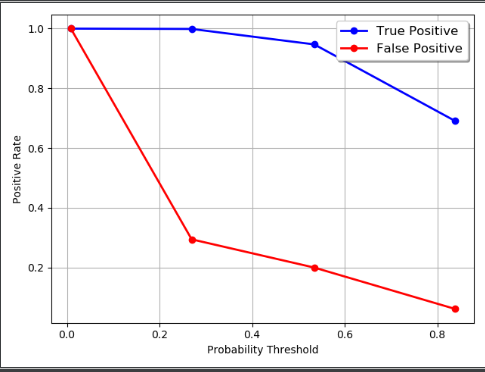
Please apply your decision tree to the Test partition and then provide the following information. You will choose whether to call sklearn functions or write your own Python program to find the answers.

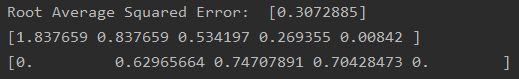
1. (5 points). Use the proportion of target Event value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?



Misclassification Rate: 0.1459627329193

1. (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?





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%.

1. (5 points). What is the Root Average Squared Error in the Test partition?



1. (5 points). What is the Area Under Curve in the Test partition?



1. (5 points). What is the Gini Coefficient in the Test partition?



1. (5 points). What is the Goodman-Kruskal Gamma statistic in the Test partition?



1. (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.

