CS 584-04: Machine Learning

Autumn 2019 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 70% of the records into the Training partition, and the remaining 30% of the records into the Test partition. The random state is 27513.
- **Decision Tree**. The maximum number of branches is two. The maximum depth is two. The split criterion is the Entropy metric.

You need to write a few Python programs to assist you in answering the questions.

Question 1 (20 points)

Please provide information about your Data Partition step.

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

SOLUTION:-

	COMMERCIAL	PRIVATE
COUNT	**2652**	**4559**
PROPORTION	**0.3677714602690334**	**0.6322285397309666**

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

SOLUTION:-

	COMMERCIAL	PRIVATE
COUNT	**1137**	**1954**
PROPORTION	**0.367842122290520**	**0.6321578777094792**

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

SOLUTION:- probability that an observation is in the Training partition given that 'CAR_USE = Commercial':- **0.6999596538317057**

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

SOLUTION:- probability that an observation is in the Test partition given that 'CAR_USE = Private':- **0.29997652823125087**

Question 2 (40 points)

Please provide information about your decision tree.

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

SOLUTION:- Entropy of the root node is **0.9489455789827704**

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

SOLUTION:- SPLIT CRITERION:-

PREDICTOR NAME:- **Occupation**
VALUE IN FIRST BRANCH:- **['Blue Collar', 'Student', 'Unknown']**
VALUE IN SECOND BRANCH:- **['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']**

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

SOLUTION:-

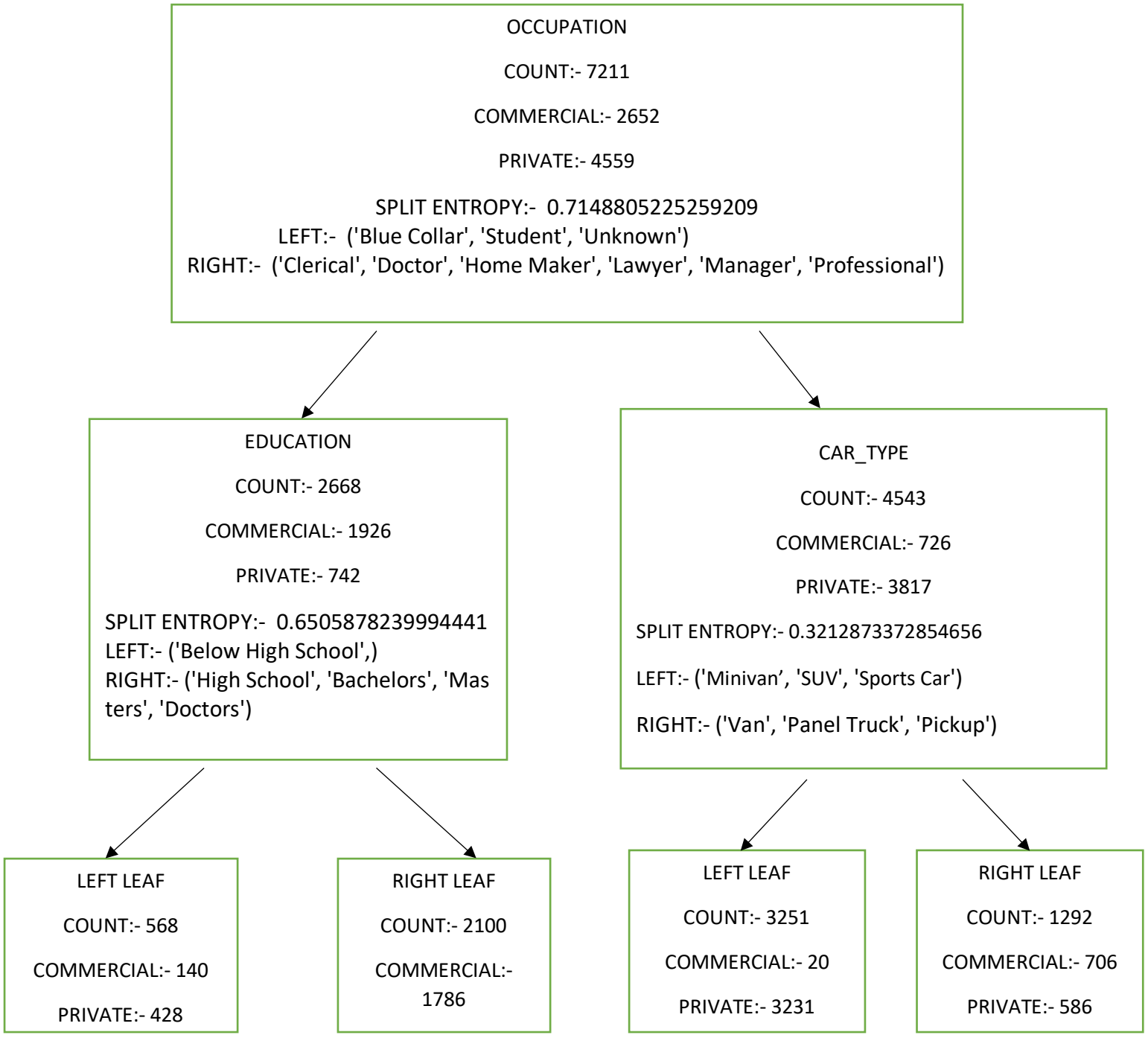
Entropy of the split of the first layer is : **0.7148805225259209**

d) (5 points). How many leaves?

SOLUTION:- Number of leaves = **4**

e) (15 points). Describe all your leaves. Please include the decision rules and the counts of the target values.

SOLUTION:-



Question 3 (40 points)

Please apply your decision tree to the Test partition and then provide the following information.

a) (10 points). Use the proportion of target Event value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?

SOLUTION: – Misclassification Rate:- **0.1708185053381**

b) (10 points). What is the Root Average Squared Error in the Test partition?

SOLUTION: – Root Average Squared Error:- **0.3287429274503**

c) (10 points). What is the Area Under Curve in the Test partition?

SOLUTION: – Area Under Curve:- **0.9114708659773**

d) (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.

SOLUTION:-