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?**

   Ans:

   count of target variable in train data :

   CAR_USE

   Commercial 2652

   Private 4559

   dtype: int64

proportion of target variable in train data :

CAR_USE

Commercial 0.367771

Private 0.632229

dtype: float64

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

Ans:

count of target variable in test data :

CAR_USE

Commercial 1137

Private 1954

dtype: int64

proportion of target variable in test data :

CAR_USE

Commercial 0.367842

Private 0.632158

dtype: float64

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

Ans:

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*?

Ans:

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?**

Ans:

root node entropy : 0.9491621304379432

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: OCCUPAION

Predictor value:

left subset: ('Blue Collar', 'Student', 'Unknown')

right subset: ('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')

entropy: 0.7112852339228054

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

Ans:

entropy of the split of the first layer:

for left node: 0.6141477604154597

for right node: 0.32518571962956416

d) **(5 points). How many leaves?**

Ans:

There are four leaves

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

Ans:

leave 1:

entropy: 0.9008100314320404

total count: 2251

commercial count: 1538

private count: 713

commercial probability: 0.6832518880497557

private probability: 0.3167481119502443

class: Commercial

leave 2:

entropy: 0.49610976358071707

total count: 469

commercial count: 418

private count: 51

commercial probability: 0.8912579957356077

private probability: 0.10874200426439233

class: Commercial

leave 3:

entropy: 0.05901648263570702

total count: 3217

commercial count: 22

private count: 3195

commercial probability: 0.006838669567920423

private probability: 0.9931613304320795

class: Private

leave 4:

entropy: 0.997294381646235

total count: 1274

commercial count: 676

private count: 598

commercial probability: 0.5306122448979592

private probability: 0.46938775510204084

class: Commercial

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?**

Ans:

Accuracy: 0.8075056615981883

Misclassification Rate: 0.19249433840181174

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

Ans:

Root Average Squared Error: 0.3408548724638163

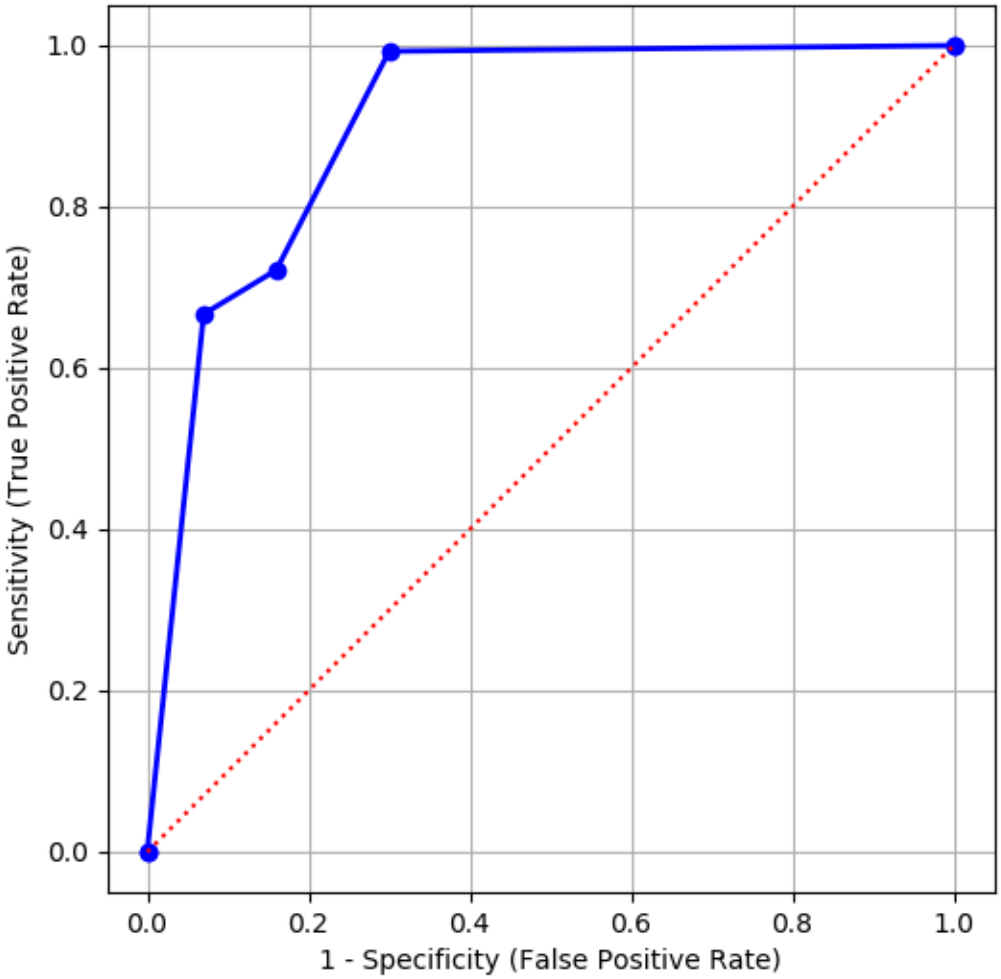
c) **(10 points). What is the Area Under Curve in the Test partition?**

Ans:

Area Under Curve: 0.9033465311748332

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.**

Ans:


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity (True Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity (False Positive Rate)