

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?

⇒

| | Counts | Proportions |
|------------|--------|-------------|
| Private | 4559 | 0.632229 |
| Commercial | 2652 | 0.367771 |

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

⇒

| | Counts | Proportions |
|------------|--------|-------------|
| Private | 1954 | 0.632158 |
| Commercial | 1137 | 0.367842 |

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

⇒ The 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 $CAR_USE = Private$?

⇒ The 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?

⇒ entropy value of the root node = 0.9489455789827704

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

⇒ The split criterion of the first layer is OCCUPATION.

⇒ Left subtree: OCCUPATION = {'Blue Collar', 'Student', 'Unknown'}

⇒ Right subtree: OCCUPATION = {'Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'}

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

⇒ The entropy of the split of the first layer is 0.7148805225259208

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

⇒ 4 leaves

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

Total | count | Commercial | Private | Class

Leaf 1

2195 | 1494 | 701 | Commercial

Leaf 2

473 | 432 | 41 | Commercial

Leaf 3

3251 | 20 | 3231 | Private

Leaf 4

1292 | 706 | 586 | 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?
⇒ Misclassification Rate: 0.19249433840181174
- b) (10 points). What is the Root Average Squared Error in the Test partition?
⇒ Root Average Squared Error: 0.34172528579286543
- c) (10 points). What is the Area Under Curve in the Test partition?
⇒ 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.

