CS 584-04: Machine Learning

Autumn 2019 Assignment 3 Answer Key

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 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.

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

|  |  |  |
| --- | --- | --- |
| **CAR\_USE** | **Count** | **Proportion** |
| Commercial | 2,652 | 0.3677715 |
| Private | 4,559 | 0.6322285 |

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

|  |  |  |
| --- | --- | --- |
| **CAR\_USE** | **Count** | **Proportion** |
| Commercial | 1,137 | 0.3678421 |
| Private | 1,954 | 0.6321579 |

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

The Training partition has 7,211 observations and the Test partition has 3,091 observations. Therefore, Prob(Training) = 7211 / 10302 = 0.6999612 and Prob(Test) = 3091 / 10302 = 0.3000388.

Prob(Training | CAR\_USE = *Commercial*)  
= Prob(CAR\_USE = *Commercial* | Training) × Prob(Training) / (Prob(CAR\_USE = *Commercial* | Training) × Prob(Training) + Prob(CAR\_USE = *Commercial* | Test) × Prob(Test))  
= (0.3677715 × 0.6999612) / (0.3677715 × 0.6999612 + 0.3678421 × 0.3000388)  
= 0.6999208

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

Prob(Test | CAR\_USE = *Private*)  
= Prob(CAR\_USE = *Private* | Test) × Prob(Test) / (Prob(CAR\_USE = *Private* | Training) × Prob(Training) + Prob(CAR\_USE = *Private* | Test) × Prob(Test))  
= (0.6321579 × 0.3000388) / (0.6322285 × 0.6999612 + 0.6321579 × 0.3000388)  
= 0.3000154

# Question 2 (40 points)

Please provide information about your decision tree.

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

The distribution of the target variable CAR\_USE in the training partition is

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CAR\_USE** | | |
|  | **Commercial** | **Private** | **Total** |
| **Count** | 2652 | 4559 | 7211 |
| **Proportion** | 0.3677715 | 0.6322285 | 1.0 |

The root entropy is .

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

The split with the lowest entropy for each predictor is

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Left Branch** | **Right Branch** | **Entropy** |
| CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.7573352 |
| OCCUPATION | ['Blue Collar', 'Student', 'Unknown'] | ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] | 0.7148805 |
| EDUCATION | ['Below High School'] | ['High School', 'Bachelors', 'Masters', 'Doctors'] | 0.9343298 |

Since the split of OCCUPATION has the lowest entropy among the three predictors, we will choose it as the split criterion of the first layer.

* The first layer is split by OCCUPATION.
* The left branch consists of ['Blue Collar', 'Student', 'Unknown'].
* The right branch consists of ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'].

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

The distribution of CAR\_USE in the first split is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Count**  **(Proportion)** | **CAR\_USE** | | | |
| **OCCUPATION** | **Commercial** | **Private** | **Total** | **Entropy** |
| ['Blue Collar', 'Student', 'Unknown'] | 1926  (0.7218891) | 742  (0.2781109) | 2668  (1.0) | 0.8528641 |
| ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] | 726  (0.1598063) | 3817  (0.8401937) | 4543  (1.0) | 0.6338460 |

The entropy of the split of the first layer is (0.8528641×2668 + 0.6338460×4543)/(2668+4543) = 0.7148805.

1. (5 points). How many leaves?

In the second layer, the split with the lowest entropy for each predictor is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **First Layer** | **Predictor** | **Left Branch** | **Right Branch** | **Entropy** |
| OCCUPATION: ['Blue Collar', 'Student', 'Unknown'] | CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.7689481 |
| OCCUPATION | ['Student'] | ['Blue Collar', 'Unknown'] | 0.7991671 |
| EDUCATION | ['Below High School'] | ['High School', 'Bachelors', 'Masters', 'Doctors'] | 0.6505878 |
| OCCUPATION: ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] | CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.3212873 |
| OCCUPATION | ['Clerical', 'Manager', 'Professional'] | ['Doctor', 'Home Maker', 'Lawyer'] | 0.5731246 |
| EDUCATION | ['Below High School', 'High School', 'Bachelors'] | ['Masters', 'Doctors'] | 0.6235326 |

We will choose the split which has the lowest entropy among the three predictors in each branch of the first later. Therefore, we will choose the split criteria of the second layer as:

* When OCCUPATION is one of these values: ['Blue Collar', 'Student', 'Unknown'],
  + The second layer is split by EDUCATION.
  + The left branch consists of ['Below High School']
  + The right branch consists of ['High School', 'Bachelors', 'Masters', 'Doctors']
* When OCCUPATION is one of these values: ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'],
  + The second layer is split by CAR\_TYPE
  + The left branch consists of ['Minivan', 'SUV', 'Sports Car']
  + The right branch consists of ['Panel Truck', 'Pickup', 'Van']

There are four leaves in this depth 2 tree.

1. (15 points). Describe all your leaves. Please include the decision rules and the counts of the target values.

The decision rules and the distributions of CAR\_USE in the four leaves are as follows.

|  |  | **CAR\_USE** | |
| --- | --- | --- | --- |
| **Leaf** | **Decision Rule** | **Commercial** | **Private** |
| 0 | OCCUPATION in ['Blue Collar', 'Student', 'Unknown'] and EDUCATION in ['Below High School'] | 140  (0.2464789) | 428  (0.7535211) |
| 1 | OCCUPATION in ['Blue Collar', 'Student', 'Unknown'] and EDUCATION in ['High School', 'Bachelors', 'Masters', 'Doctors'] | 1786  (0.8504762) | 314  (0.1495238) |
| 2 | OCCUPATION in ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] and 'CAR\_TYPE in ['Minivan', 'SUV', 'Sports Car'] | 20  (0.0061520) | 3231  (0.9938480) |
| 3 | OCCUPATION in ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] and 'CAR\_TYPE in ['Panel Truck', 'Pickup', 'Van'] | 706  (0.5464396) | 586  (0.4535604) |
|  | **Overall** | **2652**  **(0.3677715)** | **4559**  **(0.6322285)** |

# Question 3 (40 points)

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

1. (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?

The target event value is Commercial. The proportion of Commercial in the training partition is 0.3677714602690334. An observation will be predicted Commercial if the predicted probability for Commercial is greater than or equal to the threshold. Based on this information, the confusion matrix for the Test partition is

|  |  |  |
| --- | --- | --- |
| Observed CAR\_USE | Predicted CAR\_USE | |
| Commercial | Private |
| Commercial | 1051 | 86 |
| Private | 442 | 1512 |

The Misclassification Rate in the Test partition is (442 + 86) / 3091 = 0.1708185.

1. (10 points). What is the Root Average Squared Error in the Test partition?

The Root Average Squared Error in the Test partition is 0.3300251.

1. (10 points). What is the Area Under Curve in the Test partition?

The Area Under Curve in the Test partition is 0.9114709.

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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | **1-Specificity** | **Sensitivity** | | 0 | 0 | | 0.0798362 | 0.6798593 | | 0.2262027 | 0.9243624 | | 0.3178096 | 0.9912049 | | 1 | 1 | |