CS 584-04: Machine Learning

Autumn 2019 Assignment 3

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?

Answer:

Frequency tables are a basic tool you can use to explore data and get an idea of the relationships between variables. A frequency table is just a data table that shows the counts of one or more categorical variables. One of the most useful aspects of frequency tables is that they allow you to extract the proportion of the data that belongs to each category. With a one-way table, you can do this by dividing each table value by the total number of records in the table:

CAR_USE			
Commercial	2652	2652	36.777146
Private	4559	4559	63.222854
Total	7211	7211	100.000000

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

Answer:

CAR_USE			
Commercial	2652	2652	36.777146
Private	4559	4559	63.222854
Total	7211	7211	100.000000

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

Answer:

P_Train = 0.7; P_Test = 0.3
P_Comm_train = y_train[y_train=='Commercial'].count()/ y_train.shape[0] = 0.3677714602690334
P_Comm_test = y_test[y_test=='Commercial'].count()/y_test.shape[0] = 0.3678421222905209

Prob_Training_Commercial = P_Comm_train*P_Train / ((P_Comm_train*P_Train) + (P_Comm_test*P_Test))

Probability (Training|Commercial): 0.6999596538317057

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

Answer:

P_Pri_train = y_train[y_train=='Private'].count()/ y_train.shape[0] = 0.6322285397309666
P_Pri_test = y_test[y_test=='Private'].count()/y_test.shape[0] = 0.6321578777094792

Prob_Test_Pri = P_Pri_test*P_Test / ((P_Pri_train*P_Train) + (P_Pri_test*P_Test))
Probability (Test|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?

Answer: As we learnt in class, entropy is a measure of a split. Higher the entropy means the distribution is impure.

Entropy for root node is given as 0.9489621493401781

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

Answer: The split criteria for Level 0 is on variable Occupation: with categories ('Blue Collar', 'Unknown', 'Student') on left node and remaining categories on right node.
We have chosen to split on variable Occupation because it has lowest entropy.

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

Answer: Entropy of split of the first layer is 0.7148805225259208

d) (5 points). How many leaves?

Answer: There are 4 leaf nodes on Level 2 of the tree.

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

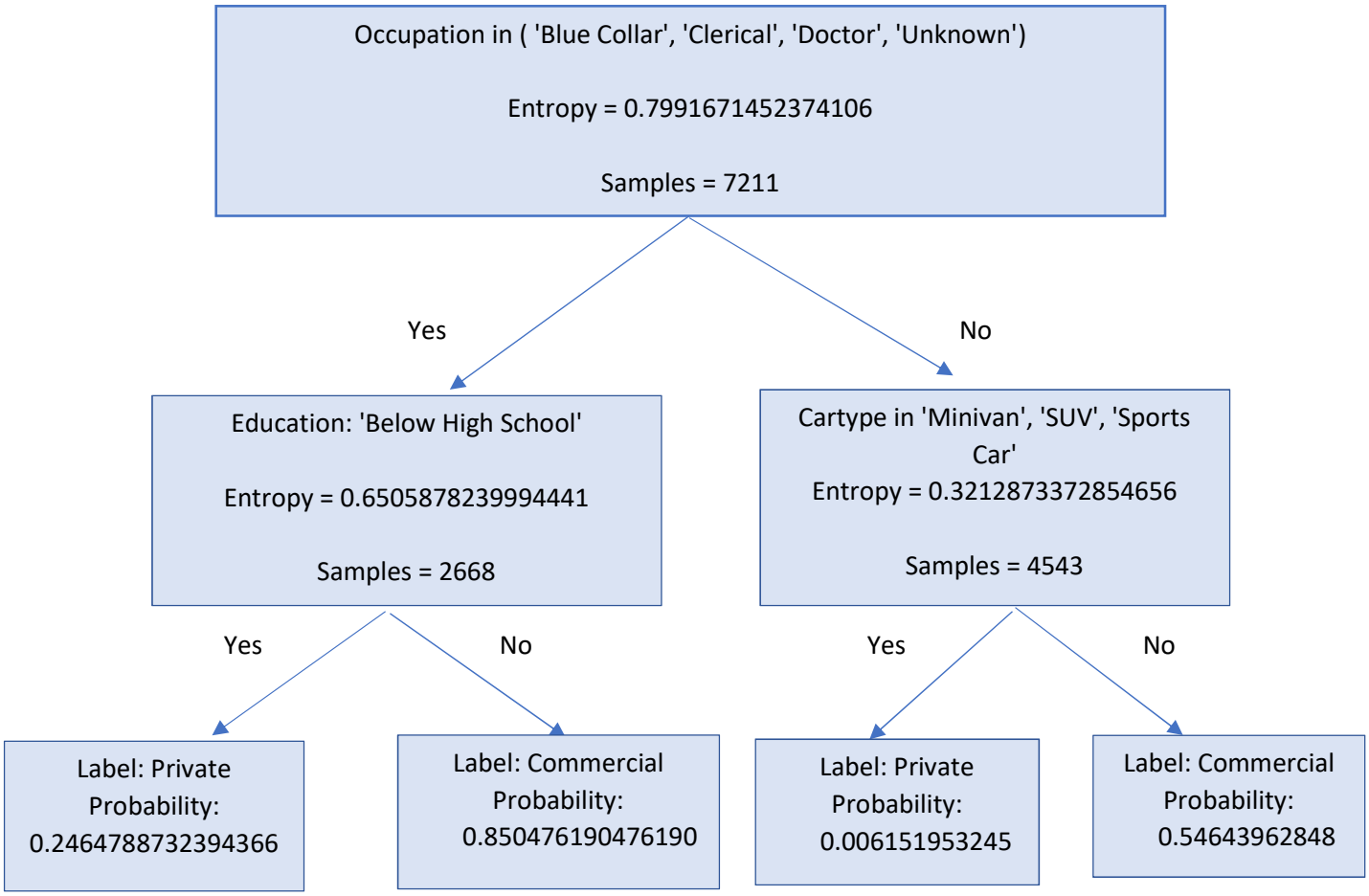
Answer:

Decision rules are as follows:
Confidence of occurance of Event:'Commercial'at leaf nodes
0.24647887323943662
0.8504761904761905
0.006151953245155337
0.5464396284829721

Count of Target values on all the leaf nodes are as follows:
140
1786
20
706

Threshold: 0.3677714602690334

Decision trees being formed is as follows:



Working of the tree:
Tree has 3 levels: Level 0, Level 1, and Level2
The Root node is split on Variable occupation.
On Level 1 there are two splits on variables: Education and Cartype.
Finally, the we calculate the decision rules and apply it on testing dataset. After this we get suitable predicted probabilities as per the rules. This dataset is then assigned suitable labels once the predicted probabilities are compared with the threshold.

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?

Answer:
An observation is misclassified if the observed target value is not equal to the predicted target value.
The misclassification rate formula is: (FP+FN)/(TP+TN+FP+FN)
In our case misclassification rate is: 0.1708185053380783

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

Answer:

RASE in our case is: 0.3300251195213117

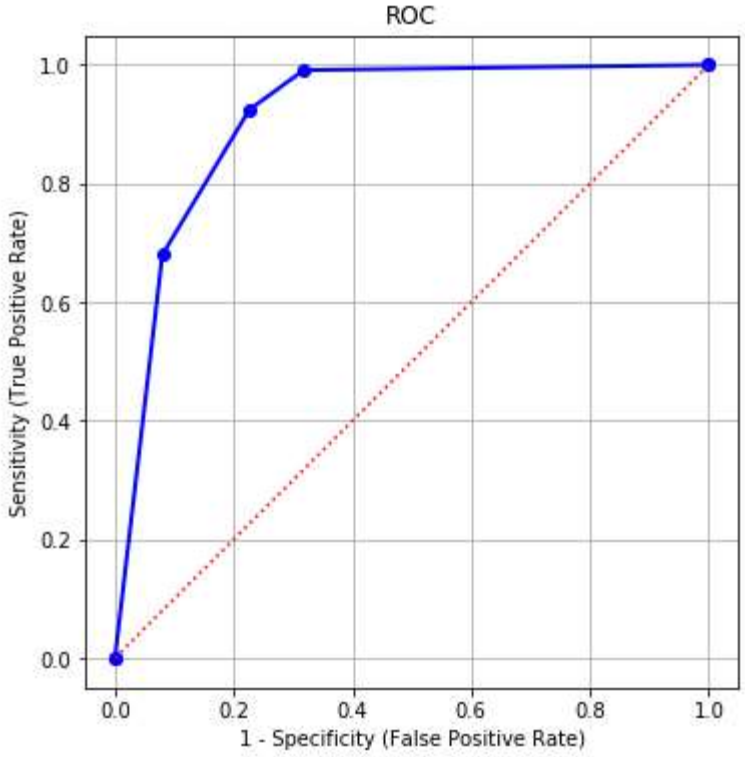
c) (10 points). What is the Area Under Curve in the Test partition?

Answer:

Area Under Curve is: Area Under the curve 0.9114708659772841

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.

The curve is as follows:


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