# CS 584-04: Machine Learning

## Fall 2019: Assignment 3

Amitdeb Prasad Bhattacharya A20402789

## Question 1

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

```
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?

```
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*?

```
probability that an observation is in the Training partition given t
hat 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*?

```
probability that an observation is in the Test partition given that
CAR_USE = Private: 0.29997652823125087
```

## Question 2

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

```
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?

```
layer0-education:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False           2419     3729  6148
True             235      828  1063
All             2654     4557  7211
 entropy: 0.9367954214398647
 split interval: 0.5

layer0-car-type:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False           1736      734  2470
True             918     3823  4741
All             2654     4557  7211
 entropy: 0.7668215614477197
 left subset: ('Minivan', 'SUV', 'Sports Car')
 right subset: ('Panel Truck', 'Pickup', 'Van')

layer0-occupation:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False            698     3793  4491
True            1956      764  2720
All             2654     4557  7211
 entropy: 0.7112852339228054
 left subset: ('Blue Collar', 'Student', 'Unknown')
 right subset: ('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manag
er', 'Professional')

split criterion for first layer
predictor name: OCCUPATION
predictor value:
 left subset: ('Blue Collar', 'Student', 'Unknown')
 right subset: ('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manag
er', 'Professional')
```

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

```
layer1-left-node-education:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False           1802      333  2135
True             154      431   585
All             1956      764  2720
```

```
 entropy: 0.6691104563656328
 split interval: 0.5

layer1-left-node-car-type:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False           1060      136  1196
True             896      628  1524
All             1956      764  2720
 entropy: 0.7724257598476323
 left subset: ('Minivan', 'SUV', 'Sports Car')
 right subset: ('Panel Truck', 'Pickup', 'Van')

layer1-left-node-occupation:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False           1638      442  2080
True             318      322   640
All             1956      764  2720
 entropy: 0.8059372474392577
 left subset: ('Student',)
 right subset: ('Blue Collar', 'Unknown')

layer1-right-node-education:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False            172     1488  1660
True             526     2305  2831
All              698     3793  4491
 entropy: 0.6141477604154597
 split interval: 2.5

layer1-right-node-car-type:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False            676      598  1274
True              22     3195  3217
All              698     3793  4491
 entropy: 0.32518571962956416
 left subset: ('Minivan', 'SUV', 'Sports Car')
 right subset: ('Panel Truck', 'Pickup', 'Van')

layer1-left-node-occupation:
 cross table:
CAR_USE   Commercial  Private   All
LE_Split
False             39     1505  1544
True             659     2288  2947
All              698     3793  4491
 entropy: 0.5615766200308671
 left subset: ('Clerical', 'Manager', 'Professional')
 right subset: ('Doctor', 'Home Maker', 'Lawyer')

entropy of the split of the first layer:
 for left node: 0.6141477604154597
 for right node: 0.32518571962956416
```

d) (5 points). How many leaves?

```
There are four leaves
```

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

```
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

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?

```
threshold is 0.3680488143114686
Accuracy: 0.8075056615981883
Misclassification Rate: 0.19249433840181174
```

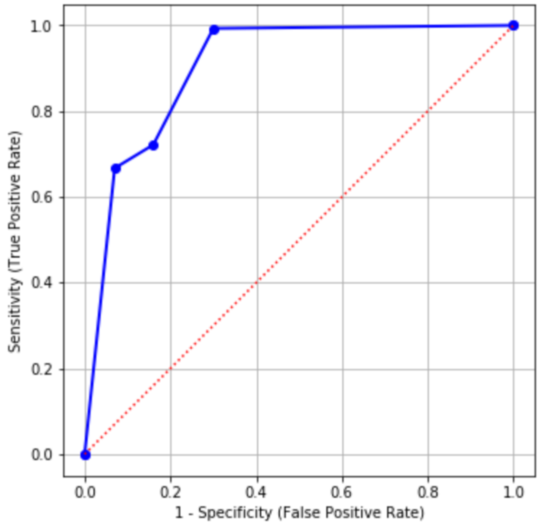
b) (10 points). What is the Root Average Squared Error in the Test partition?

```
Root Average Squared Error: 0.3408548724638163
```

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.


Sensitivity (True Positive Rate)
1 - Specificity (False Positive Rate)