Jason Yeoh
A20457826

Question 1

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

	Count	**Proportion**
Private	4884	0.6321511778410561
Commercial	2842	0.3678488221589438

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

	Count	**Proportion**
Private	1629	0.6323757763975155
Commercial	947	0.3676242236024845

c) (5 points). What is the probability that an observation is in the Training partition given that CAR_USE = Commercial?
P('Commercial' | train) = 2842/(4884+2842) = **0.3678** **P(train)** = **0.75**
P('Commercial' | test) = 947/(1629+947) = **0.3676** **P(test)** = **0.25**

P(train | 'Commercial') = $\frac{P('Commercial' \mid train) \bullet P(train)}{P('Commercial')}$ = $\frac{0.3678 \bullet 0.75}{0.3678 \bullet 0.75 + 0.3676 \bullet 0.25}$ = **0.75011449992**

d) (5 points). What is the probability that an observation is in the Test partition given that CAR_USE = Private?
P('Private' | train) = 4884/(4884+2842) = **0.6322** **P(train)** = **0.75**
P('Private' | test) = 1629/(1629+947) = **0.6324** **P(test)** = **0.25**

P(test | CAR_USE = 'Private') = $\frac{P('Private' \mid test) \bullet P(test)}{P('Commercial')}$ = $\frac{0.6324 \bullet 0.25}{0.6324 \bullet 0.25 + 0.6322 \bullet 0.75}$ = **0.2500666114**

Question 2

a) (5 points). What is the entropy value of the root node?
Entropy = 0.9490060293033189

b) (5 points). What is the split criterion (i.e., predictor name and values in the two branches) of the first layer?
Predictor name: Occupation
Branches:
LEFT: {Blue Collar, Student, Unknown}
RIGHT: {Home Maker, Lawyer, Doctor, Professional, Clerical, Manager}
Entropy: 0.7184955941364275

c) (10 points). What is the entropy of the split of the first layer?
After splitting the root layer on the criterion of Occupation, the entropy is 0.7184955941364275. The following split criterions on the left and right branches of root layer are as follows:

LEFT BRANCH:

i) **Predictor name:** Occupation
Branches: ['Student'], ['Blue Collar', 'Unknown']
Entropy: 0.8072641585823174

ii) **Predictor name:** Car Types
Branches: ['Minivan', 'SUV', 'Sports Car'], ['Panel Truck', 'Van', 'Pickup']
Entropy: 0.7736038505678898

iii) **Predictor name:** Education
Branches: [‘Below High School’], [‘High School’, ‘College’, ‘Masters’, ‘Doctors’]
Entropy: 0.6828825901259153 ***OPTIMAL***

RIGHT BRANCH:

iv) **Predictor name:** Occupation
Branches: ['Lawyer', ‘Home Maker’, 'Doctor'], ['Clerical', 'Professional', 'Manager']
Entropy: 0.5727843462818438

v) **Predictor name:** Car Types
Branches: ['Minivan', 'SUV', 'Sports Car'], ['Panel Truck', 'Van', 'Pickup']
Entropy: 0.3364029948589687 ***OPTIMAL***

vi) **Predictor name:** Education
Branches: [‘Below High School’, ‘High School’, ‘College’], [‘Masters’, ‘Doctors’]
Entropy: 0.6241205457081782

OPTIMAL SPLIT ENTROPY ON THE FIRST LAYER:
Left Branch on **Education:** 0.6828825901259153 (see iii)
Right Branch on **Car Types:** 0.3364029948589687 (see v)

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

e) (10 points). Describe all your leaves. Please include the decision rules and the counts of the target values.
LEFT-LEFT BRANCH:
- Decision Rules: **Occupation in {Blue Collar, Student, Unknown}, Education ≤ 'Below High School'**
- Private Count: 460
- Commercial Count: 173
- Total Count: 633
- Entropy: 0.8461626265285531
- Class: **Private**

LEFT-RIGHT BRANCH:
- Decision Rules: **Occupation in {Blue Collar, Student, Unknown}, Education > 'Below High School'**
- Private Count: 356
- Commercial Count: 1920
- Total Count: 2276
- Entropy: 0.6256631177932281
- Class: **Commercial**

RIGHT-LEFT BRANCH:
- Decision Rules: **Occupation in {Home Maker, Lawyer, Doctor, Professional, Clerical, Manager}, Car Type = ['Minivan', 'SUV', 'Sports Car']**
- Private Count: 3409
- Commercial Count: 23
- Total Count: 3432
- Entropy: 0.05803024570980552
- Class: **Private**

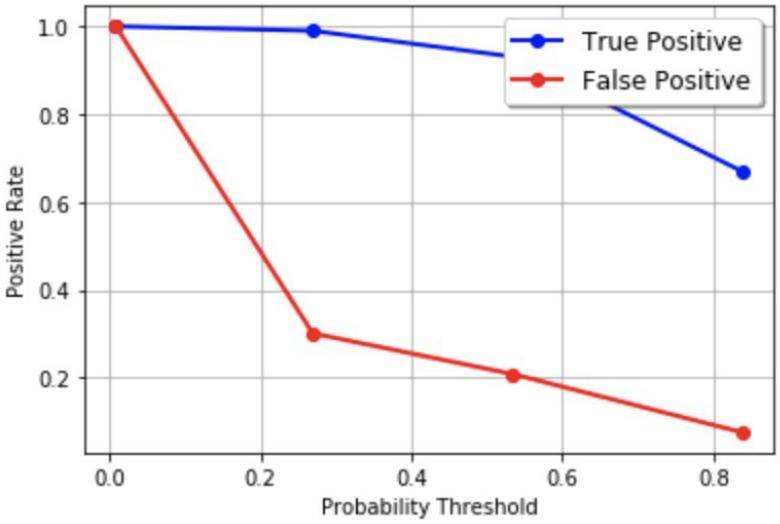
RIGHT-RIGHT BRANCH

- Decision Rules: **Occupation in {Home Maker, Lawyer, Doctor, Professional, Clerical, Manager}, Car Type = ['Panel Truck', 'Van', 'Pickup']**
- Private Count: 650
- Commercial Count: 735
- Total Count: 1385
- Entropy: 0.9972813343356697
- Class: **Commercial**

f) (5 points). What are the Kolmogorov-Smirnov statistic and the event probability cutoff value?

KS statistic = 0.72300827

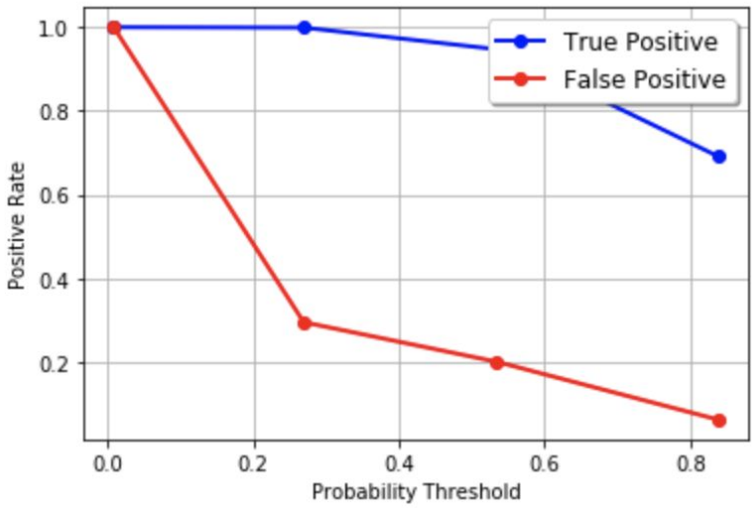
Probability cutoff value = 0.53419726



Question 3

a) (5 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 = P(Commercial | Train) = 0.3678488221589438
Accuracy = 0.8540372670807
Misclassification Rate = 0.1459627329193

b) (5 points). Use the Kolmogorov-Smirnov event probability cutoff value in the training partition as the threshold, what is the Misclassification Rate in the Test partition?



KS Statistic = 0.62965664
KS Threshold = 0.53419726
Misclassification Rate = 0.14596273291925466

c) (5 points). What is the Root Average Squared Error in the Test partition?
Root Average Squared Error (RASE) = 0.3072884960164

d) (5 points). What is the Area Under Curve in the Test partition?
Area Under Curve (AUC) = 0.9315819462838

e) (5 points). What is the Gini Coefficient in the Test partition?
C (# of concordant pairs) = 1372467
D (# of discordant pairs) = 40896
T (# of tie pairs) = 129300

Gini = 2*AUC – 1 = 0.8631638925676
Or
Gini = (C-D)/(C+D+T) = 0.8631638925676

f) (5 points). What is the Goodman-Kruskal Gamma statistic in the Test partition?
C (# of concordant pairs) = 1372467
D (# of discordant pairs) = 40896
T (# of tie pairs) = 129300

Gamma = (C-D)/(C+D) = 0.9421295166209954

g) (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.