Aman Rangapur
A20517739

CS484: Intro to Machine Learning- Assignment 3

Question-1

a) Please describe the leaf nodes of the classification tree. Your description should include these five pieces of information: (1) Splitting Criterion, (2) Number of Observations, (3) Predicted Probabilities of CAR_USE, (4) Predicted CAR_USE category, and (5) Split Entropy Value.

Leaf	OCCUPATION	EDUCATION	CAR_TYPE	Split Entropy	CAR_USE (Count / Probability): Commercial	CAR_USE (Count / Probability): Private
0	['Blue Collar', 'Student', 'Unknown']	['Below High School']		0.8304	216 (0.2625)	607 (0.7375)
1	['Blue Collar', 'Student', 'Unknown']	['High School', 'Bachelors', 'Masters', 'Doctors']		0.6226	2559 (0.8448)	470 (0.1552)
2	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['Minivan', 'SUV', 'Sports Car']	0.0568	30 (0.0065)	4564 (0.9935)
3	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['Panel Truck', 'Pickup', 'Van']	0.9974	984 (0.5302)	872 (0.4698)

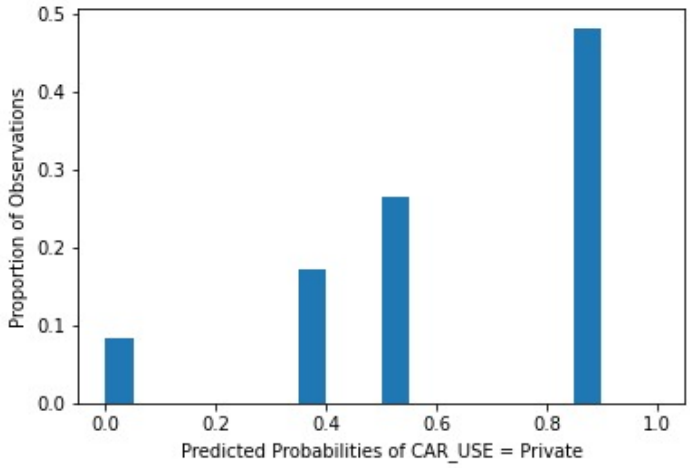
b) Let us study a fictitious person. The person works in a *Professional* occupation, has an education level of *Doctors*, and owns a *Minivan*. What are the Car Usage probabilities?

Commercial	Private
0.1127	0.8873

c) Let us study another fictitious person. The person is a *Student*, has a *Below High School* level of education, and owns a *Sports Car*. What are the Car Usage probabilities?

Commercial	Private
0.3973	0.6027

d) Generate a histogram of the predicted probabilities of CAR_USE = *Private*. The bin width is 0.05. The vertical axis is the proportion of observations.



e) Finally, what is the misclassification rate of the Classification Tree model?

Threshold probability of an Commercial is given as 0.36779266161910307.

Threshold probability of an Private is given as 0.6322073383808969.

Misclassification Rate is 20.3%.

Question- 2

You will train a Naïve Bayes model. You will apply the Laplace/Lidstone value of 0.01 to the cell counts for the purpose of computing the row probabilities. However, you do not change the cell counts.

a) What are the Class Probabilities?

Label	Commercial	Private
Frequency Count	3789	6513
Class Probability	0.3677926616	0.6322073384

b) Cross-tabulate the label variable by each predictor and show the resulting table. The table must contain the frequency counts and the row probabilities in each label class.

Feature: Car type

CAR_TYPE	Commercial(Row Prob./Freq Count)	Private(Row Prob./Freq Count)
Minivan	0.1459487992 553	0.3287271611 2141
Panel Truck	0.2251253629 853	0.0 0

CAR_TYPE	Commercial(Row Prob./Freq Count)	Private(Row Prob./Freq Count)
Pickup	0.2818685669 1068	0.1080915093 704
SUV	0.1464766429 555	0.3574389682 2328
Sports Car	0.05278437582 200	0.1503147551 979
Van	0.1477962523 560	0.05542760633 361

Feature: Occupation

OCCUPATION	Commercial(Row Prob./Freq Count)	Private(Row Prob./Freq Count)
Blue Collar	0.4579044603 1735	0.08490710886 553
Clerical	0.07521773555 285	0.2003684938 1305
Doctor	0.0 0	0.0492860433 321
Home Maker	0.01504354711 57	0.1206817135 786
Lawyer	0.0 0	0.158298787 1031
Manager	0.08128793877 308	0.1457085828 949
Professional	0.096067564 364	0.160294795 1044
Student	0.11797308 447	0.06939966221 452
Unknown	0.1565056743 593	0.01105481345 72

Feature: Education

EDUCATION	Commercial(Row Prob./Freq Count)	Private(Row Prob./Freq Count)
Bachelors	0.314330958 1191	0.2505757715 1632
Below High School	0.08603853259 326	0.182557961 1189

EDUCATION	Commercial(Row Prob./Freq Count)	Private(Row Prob./Freq Count)
High School	0.3795196622 1438	0.2324581606 1514
Masters	0.1404064397 532	0.237371411 1546
PhD	0.0797044075 302	0.09703669584 632

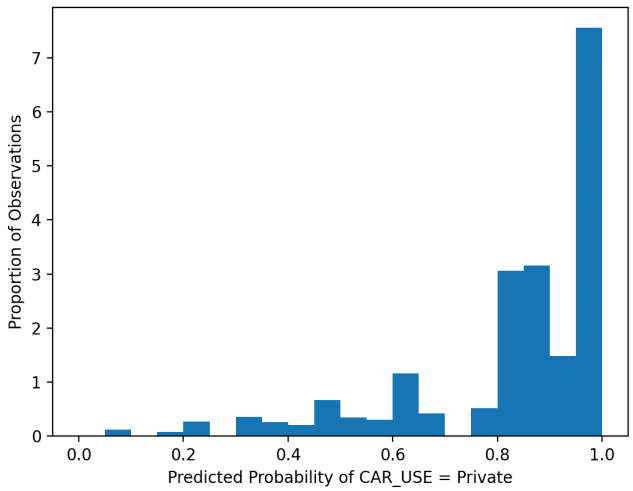
c) Let us study a fictitious person. The person works in a *Professional* occupation, has an education level of *Doctors*, and owns a *Minivan*. What are the Car Usage probabilities?

Commercial	Private
0.112808321249469	0.887191678750532

d) Let us study another fictitious person. The person is a *Student*, has a *Below High School* level of education, and owns a *Sports Car*. What are the Car Usage probabilities?

Commercial	Private
0.140653588914461	0.859346411085538

e) Generate a histogram of the predicted probabilities of CAR_USE = *Private*. The bin width is 0.05. The vertical axis is the proportion of observations.



f) Finally, what is the misclassification rate of the Naïve Bayes model?

Number of mislabeled points out of a total 10302 points : 1319. Misclassification Rate on train dataset: 12.8%.