

Name: Deep Pawar(A20545137)
Professor: Ming Long Lam
Institute: Illinois Institute of Technology

CS 484: Introduction to Machine Learning

Fall Semester 2023 Assignment 3

We provide you with the **claim_history.xlsx** which contains 10,302 observations on various vehicles. You will use the observations in this Excel file to train models that predict the usage of a vehicle. Your models will use the following variables.

Label Field

- **CAR_USE.** Vehicle Usage. It has two categories, namely, *Commercial* and *Private*.

Nominal Predictor

- **CAR_TYPE.** Vehicle Type. It has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
- **OCCUPATION.** Occupation of Vehicle Owner. It has nine categories, namely, *Clerical*, *Home Maker*, *Medical*, *Lawyer*, *Management*, *Skilled Worker*, *STEM*, *Student*, and *Not Reported*.

Ordinal Predictor

- **EDUCATION.** Highest Education Level of Vehicle Owner. It has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

Although a decision tree can accommodate missing values in the predictors, we will use only observations where there are no missing values in all the above four variables. After dropping the missing values, we will use all the 100% complete observations for training both models.

For each observation, you will calculate the predicted probabilities for $CAR_USE = Commercial$ and $CAR_USE = Private$. You will classify the observation in the CAR_USE category that has the highest predicted probability. In case of ties, choose the *Private* category.

Question 1 (50 points)

You will train a classification tree model with the following specifications:

- The maximum depth is two.
- The split criterion is the Entropy metric.
- An observation in the parent node will be assigned to the left child node if the splitting criterion is evaluated to be True. Otherwise, it will be assigned to the right child node.

Since the *sklearn tree* module does not handle string features, you must write Python codes to find the optimal split for a string feature. You must use values of a nominal string AS IS. Do not encode the nominal features into dummy columns. To find all the possible splits of a nominal predictor, we suggest the *itertools.combinations()* function to you.

- a) (20 points) 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.

Ans:

Calculations of Question 1 - a) :

Total Count: 10302
Root Node Entropy: 0.9489621493401781

Prediction probabilities of left observations are :

```
['EDUCATION', 0.6670194998377932, [[0], [1, 2, 3, 4]], [[823, CAR_USE
Commercial    216
Private       607
Name: LEFT, dtype: int64, 0.8304276080710689], [3029, CAR_USE
Commercial    2559
Private       470
Name: RIGHT, dtype: int64, 0.6226204001098349]], 'Entropy', 3029]
```

Prediction probabilities of right observations are :

```
['CAR_TYPE', 0.3274450052616845, [['Minivan', 'SUV', 'Sports Car'], ['Pickup', 'Panel Truck', 'Van']], [[4594, CAR_USE
Commercial    30
Private      4564
Name: LEFT, dtype: int64, 0.056791153992247115], [1856, CAR_USE
Commercial    984
Private      872
Name: RIGHT, dtype: int64, 0.9973716177249364]], 'Entropy', 1856]
```

Total Number of Observations and Probabilities :

CAR_USE	Commercial	Private
0	216	607
1	2559	470
2	30	4564
3	984	872

CAR_USE	Commercial	Private
0	0.2624544350	0.7375455650
1	0.8448332783	0.1551667217
2	0.0065302569	0.9934697431
3	0.5301724138	0.4698275862

- b) (10 points) Let us study a fictitious person. The person works in a *STEM* occupation, has an education level of *Masters*, and owns a *Minivan*. What are the Car Usage probabilities?

Ans:

```
Calculations of Question 1 - b) :

CAR Uses Probability

Commercial, Private
0.006530256856769699  0.9934697431432303
Predicted CAR_USE: Private
```

So, the Car Usage probabilities are Commercial is 0.006530256856769699 and Private is 0.9934697431432303

- c) (10 points) Let us study another fictitious person. The person is a *Student*, has a *High School* level of education, and owns a *Pickup*. What are the Car Usage probabilities?

Ans:

```
Calculations of Question 1 - c) :

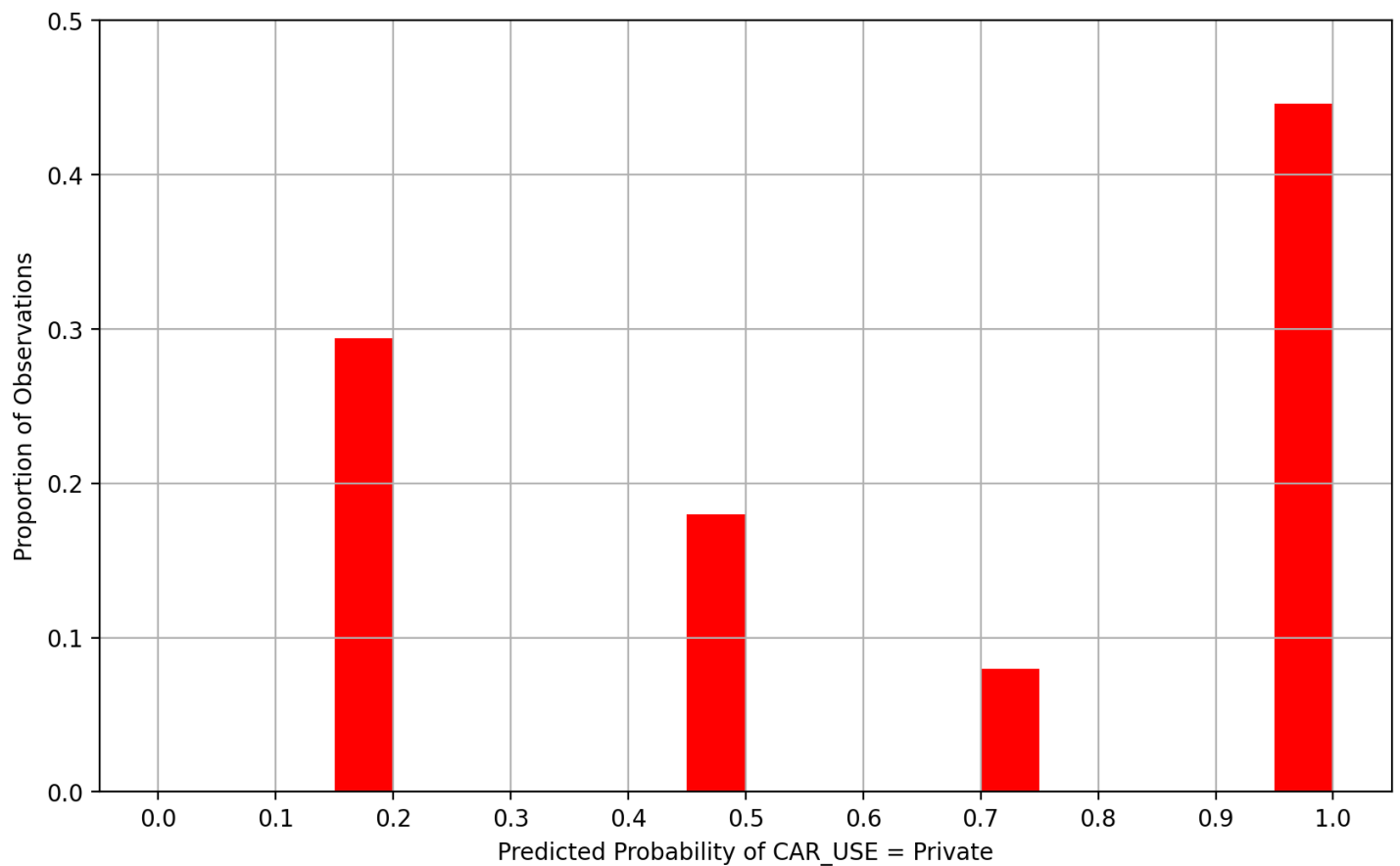
CAR Uses Probability

Commercial, Private
0.8448332783096731  0.15516672169032683
Predicted CAR_USE: Commercial
```

So, the Car Usage probabilities are Commercial is 0.8448332783096731 and Private is 0.15516672169032683

- d) (5 points) 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.

Ans:



- e) (5 points) Finally, what is the misclassification rate of the Classification Tree model?

Ans:

Calculations of Question 1 - e) :

The Misclassification rate of the Classification Tree model :
15.414482624733061

Question 2 (50 points)

You will train a Naïve Bayes model with a Laplace/Lidstone value of 0.01.

a) (10 points) What are the Class Probabilities?

Ans:

```
Calculations of Question 2 - a) :
```

```
The Class Probabilities are :
```

```
Commercial Private  
[0.36779266 0.63220734]
```

So, the class probabilities are Commercial: 0.36779266 and Private: 0.63220734

b) (10 points) 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.

Ans:

Calculations of Question 2 - b) :

The frequency counts and the row probabilities of Car Type :

CAR_TYPE	Minivan	Panel Truck	Pickup	SUV	Sports Car	Van
CAR_USE						
Commercial	553	853	1068	555	200	560
Private	2141	0	704	2328	979	361

CAR_TYPE	Minivan	Panel Truck	Pickup	SUV	Sports Car	Van
CAR_USE						
Commercial	0.145949	0.225125	0.281869	0.146477	0.052784	0.147796
Private	0.328727	0.000000	0.108092	0.357439	0.150315	0.055428

The frequency counts and the row probabilities of Occupation :

OCCUPATION	Clerical	Home Maker	Lawyer	Management	Medical	Not Reported	\
CAR_USE							
Commercial	285	57	0	308	0	593	
Private	1305	786	1031	949	321	72	

OCCUPATION	STEM	Skilled Worker	Student
CAR_USE			
Commercial	364	1735	447
Private	1044	553	452

OCCUPATION	Clerical	Home Maker	Lawyer	Management	Medical	\
CAR_USE						
Commercial	0.075218	0.015044	0.000000	0.081288	0.000000	
Private	0.200368	0.120682	0.158299	0.145709	0.049286	

OCCUPATION	Not Reported	STEM	Skilled Worker	Student
CAR_USE				
Commercial	0.156506	0.096068	0.457904	0.117973
Private	0.011055	0.160295	0.084907	0.069400

The frequency counts and the row probabilities of Education :

EDUCATION	Bachelors	Below High School	Doctors	High School	Masters
CAR_USE					
Commercial	1191	326	302	1438	532
Private	1632	1189	632	1514	1546
EDUCATION	Bachelors	Below High School	Doctors	High School	Masters
CAR_USE					
Commercial	0.314331	0.086039	0.079704	0.379520	0.140406
Private	0.250576	0.182558	0.097037	0.232458	0.237371

- **Tabular Calculations:**

Count	Car Type					
CAR_USE	Minivan	Panel Truck	Pickup	SUV	Sports Car	Van
Commercial	553	853	1068	555	200	560
Private	2141	0	704	2328	979	361

Row Prob	Car Type					
CAR_USE	Minivan	Panel Truck	Pickup	SUV	Sports Car	Van
Commercial	0.1459488	0.2251254	0.2818686	0.1464766	0.0527844	0.1477963
Private	0.3287272	0	0.1080915	0.3574390	0.1503148	0.0554276

Count	Occupation								
CAR_USE	Clerical	Home Maker	Lawyer	Management	Medical	Skilled Worker	STEM	Student	Not Reported
Commercial	285	57	0	308	0	1735	364	447	593
Private	1305	786	1031	949	321	553	1044	452	72

Row Prob	Occupation								
CAR_USE	Clerical	Home Maker	Lawyer	Management	Medical	Skilled Worker	STEM	Student	Not Reported
Commercial	0.0752177	0.0150435	0	0.0812879	0	0.4579045	0.0960676	0.1179731	0.1565057
Private	0.2003685	0.1206817	0.1582988	0.1457086	0.0492860	0.0849071	0.1602948	0.0693997	0.0110548

Count	Education				
CAR_USE	Below High School	High School	Bachelors	Masters	Doctors
Commercial	326	1438	1191	532	302
Private	1189	1514	1632	1546	632

Row Prob	Education				
CAR_USE	Below High School	High School	Bachelors	Masters	Doctors
Commercial	0.0860385	0.3795197	0.3143310	0.1404064	0.0797044
Private	0.1825580	0.2324582	0.2505758	0.2373714	0.0970367

c) (10 points) Let us study a fictitious person. The person works in a *Skilled Worker* occupation, has an education level of *Doctors*, and owns an SUV. What are the Car Usage probabilities?

Ans:


```
Calculations of Question 2 - c) :
```

```
The Car Usage probabilities are :
```

```
Commercial Private  
[[0.5136312 0.4863688]]
```

So, the Car Usage probabilities are Commercial: 0.5136312 and Private: 0.4863688

- d) (10 points) Let us study another fictitious person. The person works in a *Management* occupation, has a *Below High School* level of education, and owns a *Sports Car*. What are the Car Usage probabilities?

Ans:

```
Calculations of Question 2 - d) :
```

```
The Car Usage probabilities are :
```

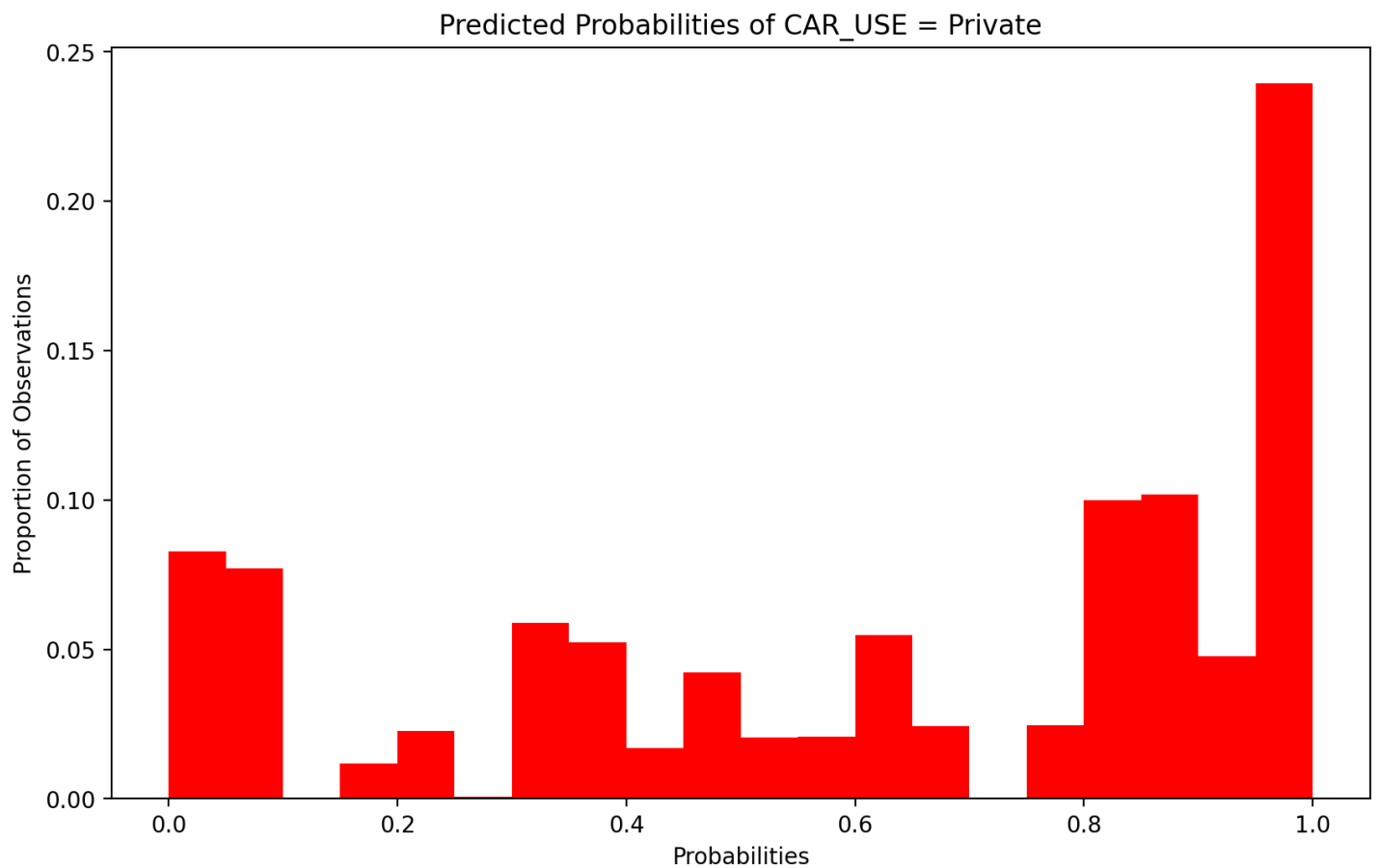
```
Commercial Private  
[[0.0509781 0.9490219]]
```

So, the Car Usage probabilities are Commercial: 0.0509781 and Private: 0.9490219

- e) (5 points) 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.

Ans:

- Histogram of the predicted probabilities of $CAR_USE = Private$



f) (5 points) Finally, what is the misclassification rate of the Naïve Bayes model?

Ans:

Calculations of Question 2 - f) :

The misclassification rate is : 0.1280333915744516

So, the misclassification rate of this Naïve Bayes model is $0.1280334 = 12.80334\%$