CS 484: Introduction to Machine Learning

Spring 2022 Assignment 3

Question 1 (40 points)

You will train a classification tree to predict the usage of a car. The data is the claim_history.csv that contains 10,302 observations. The analysis specifications are:

Label Field

- **CAR_USE**. The car's usage. This field has two categories, namely, *Commercial* and *Private*.

Nominal Feature

- **CAR_TYPE**. The car's type. This feature has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
- **OCCUPATION**. The occupation of the car owner. This feature has nine categories, namely, *Blue Collar*, *Clerical*, *Doctor*, *Home Maker*, *Lawyer*, *Manager*, *Professional*, *Student*, and *Unknown*.

Ordinal Feature

- **EDUCATION**. The education level of the car owner. This feature has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

Decision Tree Specifications

- Use only the complete records.
- The maximum number of branches is two.
- The maximum depth is two.
- The split criterion is the Entropy metric.

Since the sklearn tree module does not handle string features, you have to write your own 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. It is because your classification tree will not have the enough depth to allow all the dummy columns be used for splitting.

Please answer the following questions.

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

The entropy value of the root node is 0.94896

b) (10 points). Please list the optimal split (i.e., feature name, values in the two branches, and the split entropy) for all three features in the first layer.

Predictor	Left Branch	Right Branch	Split Entropy
CAR_TYPE	['Minivan', 'SUV', 'Sports Car']	['Panel Truck', 'Pickup', 'Van']	0.7684152
OCCUPATION	['Blue Collar', 'Student', 'Unknown']	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']	0.7125833
EDUCATION	['Below', 'High', 'School']	['High School', 'Bachelors', 'Masters', 'Doctors']	0.9356143

c) (5 points). Which feature is selected for splitting in the first layer? What are the values in the branches of the first layer?

OCCUPATION has the lowest split entropy. Therefore, the split criterion of the first layer is **OCCUPATION** = ['Blue Collar', 'Student', 'Unknown'] and OCCUPATION = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'].

d) (10 points). Which features are selected for splitting in the second layer? What are the values in the branches of the second layer?

EDUCATION ['Below High School'] ['HIghschool', 'Bachelors', 'Masters', 'Doctors']

CAR_TYPE ['Minivan', 'SUV'], ['Panel Truck', 'Pickup', 'Van']

e) (10 points). Describe the leaf (i.e., terminal) nodes in a table. Please include the decision rules, the counts of the target categories, and the predicted probabilities for CAR_USE.

Terminal nodes (Leaf) are listed below with their predicted probabilities.

Leaf	Predicator			CAR_USE (Count, Probability)	
	OCCUPATION	EDUCATION	CAR_TYPE	Commercial	Private
0	['Blue Collar', 'Student', 'Unknown']	['Below High School']		216 (0.2625)	607 (0.7375)

1	['Blue Collar', 'Student', 'Unknown']	['High School', 'Bachelors', 'Masters', 'Doctors']		2559 (0.8448)	470 (0.1552)
2	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['Minivan', 'SUV', 'Sports Car']	30 (0.0065)	4564 (0.9935)
3	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['Panel Truck', 'Pickup', 'Van']	984 (0.5302)	872 (0.4698)

Question 2 (40 points)

We provide you the sample_v10.csv that contains 10,000 observations. This data contains a categorical label variable **y** and ten continuous features are **x1**, **x2**, **x3**, **x4**, **x5**, **x6**, **x7**, **x8**, **x9**, and **x10**. You will then use this data to train a multinomial logistic regression model that always includes the Intercept term. To include only significant continuous features in the model, you will use the Forward Selection method to determine the list of significant continuous features. The threshold for test significance is 0.05.

a) (5 points). Show the frequency table of the categorical target field.

Target (y): 1 2 3

Frequency: 2274 3532 4194

b) (5 points). What is the initial model in the Forward Selection method? Please also show the log-likelihood value and the number of free parameters.

Initial model in the Backward Selection model must include all the ten continuous features. The model form is logit(y) = Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10.

The **log- likelihood value is -2030.5289** and **number of free parameters is 20**.

c) (20 points). Please show the step summary of the Forward Selection method. The step summary should include the name of the entered feature, the log-likelihood value of the expanded model, the number of free parameters of the expanded model, the Deviance test statistic, the Deviance degree of freedom, and the Deviance significance value.

Step	Removed Predictor	No of Free parameters	Log-likelihood	Deviance		
				Chi-square	Degree of Freedom	Significance
0		20	-2030.5289			
1	X7	18	-2030.5583	0.0588	2	0.9710
2	X3	16	-2030.8236	0.5306	2	0.7670
3	X2	14	-2031.1159	0.5846	2	0.7465
4	X5	12	-2031.5230	0.8141	2	0.6656
5	X9	10	-2032.3195	1.5931	2	0.4509
6	X6	8	-2034.5668	4.4947	2	0.1057
7	X8	6	-2036.8352	4.5367	2	0.1035

d) (5 points). What is the final model suggested by the Forward Selection method?

The final model is **logit(y) = x1 + x4 + x10 + Intercept**

e) (5 points). Please calculate the Akaike Information Criterion and the Bayesian Information Criterion for all the models that you listed in Part (c). What model will each criterion suggest?

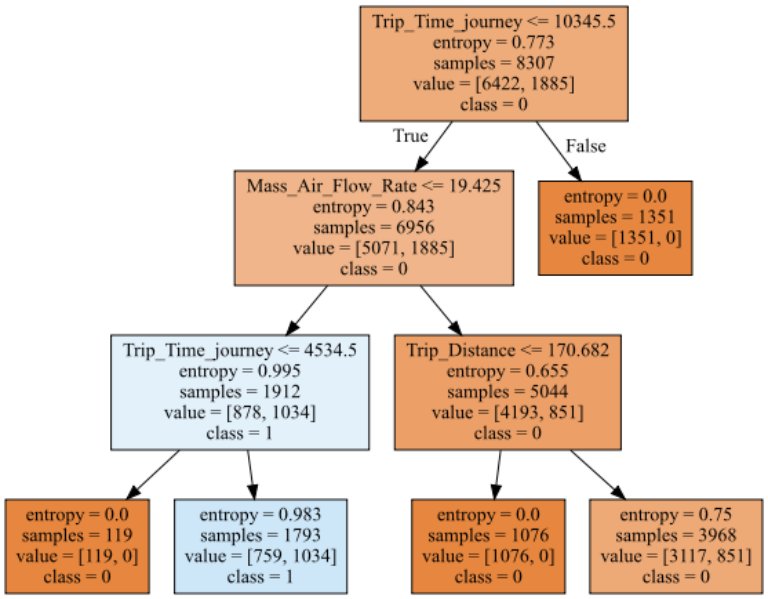
Step	Removed Predictor	No of Free parameters	Log-likelihood	Akaike Information Criterion	Bayesian Information
0		20	-2030.5289	4101.0578	4245.2646
1	X7	18	-2030.5583	4097.1166	4226.9027
2	X3	16	-2030.8236	4093.6472	4209.0126
3	X2	14	-2031.1159	4090.2318	4191.1766
4	X5	12	-2031.5230	4087.0459	4173.5700
5	X9	10	-2032.3195	4084.6390	4156.7424
6	X6	8	-2034.5668	4085.1337	4142.8164
7	X8	6	-2036.8352	4085.6704	4128.9325

The Akaike Information Criterion suggests this model: **x1 + x4 +x6 + x8 +20 + Intercept**

The Bayesian Information Criterion suggest this model: **x1 + x4 + x10 + Intercept**

Question 3 (20 points)

An observation is misclassified if the predicted target category is not the same as the observed target category. The misclassification rate is the proportion of observations that have been misclassified. The following diagram shows the classification tree for a binary target variable. The target categories are 0 and 1. Based on the diagram, what is the misclassification rate?



The tree provided above has five leaf (terminal) nodes. Thus, the classification tree will classify all the observations in a leaf node to the majority target category in the leaf node, the observations that belong to other target categories will be misclassified. We can calculate the misclassification rate by adding up the numbers of misclassified observations in the leaf nodes.

Subtotal sum: 8308 Class 0 sum: 6422 Class 1 sum: 1885 Misclassified sum: 1610

Thus, the misclassification rate is 1610 / 8308 = 0.1938