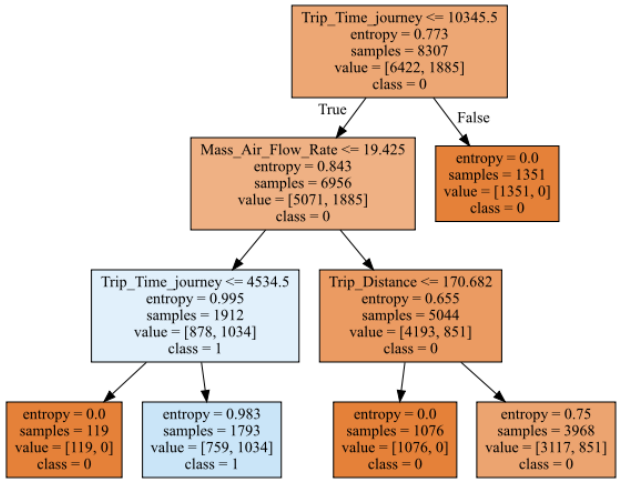
CS 484: Introduction to Machine Learning

Spring 2021 Assignment 3

Question 1 (20 points)

An observation is misclassified if the predicted target category is different from the observed target category. The misclassification rate is the proportion of observations that are misclassified. The following diagram shows the classification tree for a binary target variable. The target categories are 0 and 1. Based on the diagram, please calculate the misclassification rate.



Ans:

There are five leaf nodes in the tree and in which two of the leaf nodes have misclassified samples of number 759 and 581.

$Total\ number\ of\ samples\ =\ 8307$

$$Misclassification\ rate\ =\ \frac{Total\ Number\ of\ Samples\ Misclassified}{Total\ number\ of\ Samples}$$

$$=\ \frac{759+851}{8307}=\ 0.19381244733357408 = 19.3812\%\ (approx.)$$

Question 2 (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:

Target 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 well, you should write custom Python codes to find the optimal split for a string feature. Also, do not encode the nominal features into dummy columns. It is because your classification tree is not deep enough to let 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?

Ans:

$Number\ of\ Commercials\ Cars\ =\ 3789$

$Number\ of\ Private\ Cars\ =\ 6513$

$Total\ number\ of\ cars\ =\ 10302$

$Entropy\ of\ the\ root\ node\ =\ -\left(\left(\frac{3789}{10302}\right)*\ \log_2\left(\frac{3789}{10302}\right)+\left(\frac{6513}{10302}\right)*\ \log_2\left(\frac{6513}{10302}\right)\right)=$
0.9489621493401781

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.

Ans:

We can find the following entropies in the first layer:

Split Entropy of Education: 0.9356142508258437

Split Entropy of Car Type: 0.7684152303050842

Split Entropy of Occupation: 0.712583253573726

- Optimal Split will be the minimum of it (i.e., Occupation) with the entropy value of 0.712583253573726.
- Left branch →('Blue Collar', 'Student', 'Unknown')
- Right branch → ('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional')

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

Ans:

- We have selected feature "Occupation" for splitting in the first layer, because it has the minimum entropy value of 0.712583253573726.
- The values in the branches of the first layer are:
  - ✓ ('Blue Collar', 'Student', 'Unknown')
  - ✓ ('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?

Ans:

- Splitting for second layer is as follows:

➢ **Left:**

Split Entropy of Education in the next layer (left): 0.6670194998377932

Split Entropy of Car Type in the next layer (left): 0.7725782837913743

Split Entropy of Occupation in the next layer (left): 0.8042192219461467

- Minimum entropy is at Education at 0.6670194998377932, so the split will look like (left):

  Left Branch (left): (Below High School)

  Right Brank (left): (High School, Bachelors, Masters, Doctors)

➢ **Right:**

Split Entropy of Education in the next layer (right): 0.6175650406874581

Split Entropy of Car Type in the next layer (right): 0.3274450052616845

Split Entropy of Occupation in the next layer (right): 0.5664540067183996

- Minimum entropy is at Car Type at 0.3274450052616845, so the split will look like (right):

  Left Branch (right): ('Minivan', 'SUV', 'Sports Car')

  Right Branch (right): ('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.

Ans:

The decision rules are given below:

Leaf	Decision Rule	#Total	#Commercial	#Private	Class
1	('Blue Collar', 'Student', 'Unknown') → (Below High School)	823	216	607	Private
2	('Blue Collar', 'Student', 'Unknown') → (High School, Bachelors, Masters, Doctors)	3029	2559	470	Commercial
3	('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') → ('Minivan', 'SUV', 'Sports Car')	4594	30	4564	Private
4	('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') → ('Panel Truck', 'Pickup', 'Van')	1856	984	872	Commercial

Question 3 (40 points)

We provide you the sample_v10.csv that contains 10,000 observations. This data contains a categorical 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 Backward 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.

Ans:

The frequency table of the categorical target field is shown below:

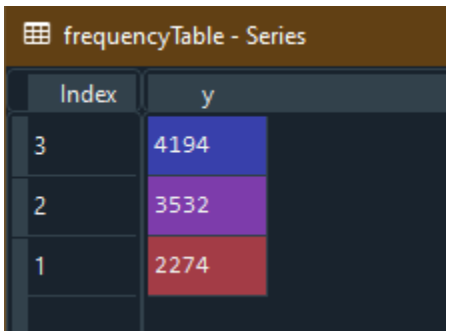


Figure 1

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

Ans:

- The initial model in the Backward Selection method is the one that includes all the parameters. The R style formula for this model can be shown as:

$$y \sim x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10$$

OR

Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

- From the python code we can get the following information:

- ✓ Model Log-Likelihood Value = -1956.055139748098
- ✓ Number of Free Parameters = 22

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

Ans:

Table 1

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
0	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10	22	-1956.05514					
	Step 1							
1.1	Intercept + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10	20	-2225.27135	538.432419	2	1.20E-117	4490.542699	4634.749507
1.2	Intercept + x1 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10	20	-1956.331061	0.5518417 81	2	0.758872963	3952.662121	4096.868929
1.3	Intercept + x1 + x2 + x4 + x5 + x6 + x7 + x8 + x9 + x10	20	-1956.280363	0.450445991	2	0.798338173	3952.560725	4096.767533
1.4	Intercept + x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9 + x10	20	-5780.494365	7648.878455	2	0	11600.98873	11745.19554
1.5	Intercept + x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9 + x10	20	-1956.470194	0.830107891	2	0.660304659	3952.940387	4097.147195
1.6	Intercept + x1 + x2 + x3 + x4 + x5 + x7 + x8 + x9 + x10	20	-1958.083555	4.056831062	2	0.131543783	3956.167111	4100.373918

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
1.7	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9 + x10	20	-1956.074428	0.0385771 67	2	0.980896251	3952.148857	4096.355664
1.8	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x9 + x10	20	-1958.114765	4.1192497 43	2	0.12750179	3956.229529	4100.436337
1.9	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x10	20	-1956.953231	1.7961819 06	2	0.407346562	3953.906461	4098.113269
1.10	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9	20	-8111.136792	12310.163 3	2	0	16262.27358	16406.48039
	Step 2							
2.1	Intercept + x2 + x3 + x4 + x5 + x6 + x8 + x9 + x10	18	-2225.335417	538.52197 7	2	1.15E-117	4486.670834	4616.45696
2.2	Intercept + x1 + x3 + x4 + x5 + x6 + x8 + x9 + x10	18	-1956.349997	0.5511377 74	2	0.759140136	3948.699994	4078.486121
2.3	Intercept + x1 + x2 + x4 + x5 + x6 + x8 + x9 + x10	18	-1956.302332	0.4558033 22	2	0.796202554	3948.604666	4078.390787
2.4	Intercept + x1 + x2 + x3 + x5 + x6 + x8 + x9 + x10	18	-5780.849096	7649.5493 36	2	0	11597.69819	11727.48432
2.5	Intercept + x1 + x2 + x3 + x4 + x6 + x8 + x9 + x10	18	-1956.492425	0.8359936 24	2	0.658364327	3948.98485	4078.770977
2.6	Intercept + x1 + x2 + x3 + x4 + x5 + x8 + x9 + x10	18	-1958.099584	4.0503114 11	2	0.131973293	3952.199168	4081.985295
2.7	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x9 + x10	18	-1958.	4.1121215 2	2	0.127957032	3952.260978	4082.047105

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
			130489					
2.8	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x8 + x10	18	-1956.975343	1.80182922	2	0.406197975	3949.950686	4079.736813
2.9	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9	18	-8111.623048	12311.09724	2	0	16259.2461	16389.03222
	Step 3							
3.1	Intercept + x2 + x4 + x5 + x6 + x8 + x9 + x10	16	-2225.390183	538.1757067	2	1.37E-117	4482.780367	4598.145813
3.2	Intercept + x1 + x4 + x5 + x6 + x8 + x9 + x10	16	-1956.587077	0.569494603	2	0.752204311	3945.174155	4060.539601
3.3	Intercept + x1 + x2 + x5 + x6 + x8 + x9 + x10	16	-5780.969902	7649.335144	2	0	11593.93998	11709.30525
3.4	Intercept + x1 + x2 + x4 + x6 + x8 + x9 + x10	16	-1956.723857	0.843053905	2	0.656044306	3945.447714	4060.813166
3.5	Intercept + x1 + x2 + x4 + x5 + x8 + x9 + x10	16	-1958.306072	4.007484051	2	0.134829801	3948.612144	4063.977599
3.6	Intercept + x1 + x2 + x4 + x5 + x6 + x9 + x10	16	-1958.323514	4.042367742	2	0.132498511	3948.647028	4064.012474
3.7	Intercept + x1 + x2 + x4 + x5 + x6 + x8 + x10	16	-1957.229652	1.854639122	2	0.395612707	3946.459299	4061.824745
3.8	Intercept + x1 + x2 + x4 + x5 + x6 + x8 + x9	16	-8112.240223	12311.87579	2	0	16256.48045	16371.84589

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
	Step 4							
4.1	Intercept + x4 + x5 + x6 + x8 + x9 + x10	14	-2225.544229	537.914303	2	1.56E-117	4479.088458	4580.033223
4.2	Intercept + x1 + x5 + x6 + x8 + x9 + x10	14	-5781.387503	7649.600852	2	0	11590.77501	11691.71977
4.3	Intercept + x1 + x4 + x6 + x8 + x9 + x10	14	-1956.999404	0.824653299	2	0.662107964	3941.998808	4042.943573
4.4	Intercept + x1 + x4 + x5 + x8 + x9 + x10	14	-1958.538668	3.903181915	2	0.142047	3945.077337	4046.022102
4.5	Intercept + x1 + x4 + x5 + x6 + x9 + x10	14	-1958.669897	4.165638 69	2	0.124578487	3945.339793	4046.284558
4.6	Intercept + x1 + x4 + x5 + x6 + x8 + x10	14	-1957.512961	1.851767787	2	0.396181082	3943.025922	4043.970688
4.7	Intercept + x1 + x4 + x5 + x6 + x8 + x9	14	-8112.644438	12312.11472	2	0	16253.28888	16354.23364
	Step 5							
5.1	Intercept + x4 + x6 + x8 + x9 + x10	12	-2226.028231	538.057653 8	2	1.45E-117	4476.056462	4562.580546
5.2	Intercept + x1 + x6 + x8 + x9 + x10	12	-5782.152223	7650.305638	2	0	11588.30445	11674.82853
5.3	Intercept + x1 + x4 + x8 + x9 + x10	12	-1958.999982	4.001155 46	2	0.135257119	3941.999963	4028.524048

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
5.4	Intercept + x1 + x4 + x6 + x9 + x10	12	-1959.087193	4.175578001	2	0.12396091	3942.174386	4028.698447
5.5	Intercept + x1 + x4 + x6 + x8 + x10	12	-1957.947365	1.895922972	2	0.387530204	3939.894731	4026.418815
5.6	Intercept + x1 + x4 + x6 + x8 + x9	12	-8112.948484	12311.89816	2	0	16249.89697	16336.42105
	Step 6							
6.1	Intercept + x4 + x6 + x8 + x10	10	-2226.621494	537.3482568	2	2.07E-117	4473.242988	4545.346391
6.2	Intercept + x1 + x6 + x8 + x10	10	-5782.357348	7648.819965	2	0	11584.7147	11656.81811
6.3	Intercept + x1 + x4 + x8 + x10	10	-1959.957215	4.019698609	2	0.134008868	3939.914429	4012.017833
6.4	Intercept + x1 + x4 + x6 + x10	10	-1959.988862	4.082989227	2	0.129834513	3939.97772	4012.081124
6.5	Intercept + x1 + x4 + x6 + x8	10	-8113.071189	12310.24765	2	0	16246.14238	16318.24578
	Step 7							
7.1	Intercept + x4 + x8 + x10	8	-2228.836796	537.7591624	2	1.69E-117	4473.673592	4531.356315
7.2	Intercept + x1 + x8 + x10	8	-5786.178187	7652.441945	2	0	11588.35637	11646.03911

Index	Model Form	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
7.3	Intercept + x1 + x4 + x10	8	-1961.902611	3.890792233	2	0.142930595	3939.805222	3997.487945
7.4	Intercept + x1 + x4 + x8	8	-8118.796141	12317.67785	2	0	16253.59228	16311.27501
	Step 8							
8.1	Intercept + x4 + x10	6	-2230.408314	537.0114006	2	2.45E-117	4472.816628	4516.078667
8.2	Intercept + x1 + x10	6	-5787.366389	7650.927556	2	0	11586.73278	11629.99482
8.3	Intercept + x1 + x4	6	-8119.883592	12315.96196	2	0	16251.76718	16295.02923

1. **Step 1:** Since index 1.7 (removing x7) has the largest Chi-Square Significance value. X7 will be removed from the model.
2. **Step 2:** Since index 2.3 (removing x3) has the largest Chi-Square Significance value. X3 will be removed from the model.
3. **Step 3:** Since index 3.2 (removing x2) has the largest Chi-Square Significance value. X2 will be removed from the model.
4. **Step 4:** Since index 4.3 (removing x5) has the largest Chi-Square Significance value. X5 will be removed from the model.
5. **Step 5:** Since index 5.5 (removing x9) has the largest Chi-Square Significance value. X9 will be removed from the model.
6. **Step 6:** Since index 6.3 (removing x6) has the largest Chi-Square Significance value. X6 will be removed from the model.
7. **Step 7:** Since index 7.3 (removing x8) has the largest Chi-Square Significance value. X8 will be removed from the model.
8. **Step 8:** Since all Chi-Square Significance values are not greater than 0.05, removing any of the rest will reduce model goodness-of-fit.
9. Therefore, no predictors can be removed.
10. The Backward Selection stops at Model: **Intercept + x1 + x4 + x10.**

- Step summary of the Backward Selection method:

Table 2

Index	Model Form	Removed Predictor	Number of Free Parameters	Log-Likelihood	Deviance	Degrees of Freedom	Chi-Square Significance	AIC	BIC
0	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10	x7	22	-1956.05514					
1.7	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9 + x10	x3	20	-1956.074428	0.038577167	2	0.980896251	3952.148857	4096.355664
2.3	Intercept + x1 + x2 + x4 + x5 + x6 + x8 + x9 + x10	x2	18	-1956.30233	0.455803322	2	0.796202554	3948.60466	4078.390787
3.2	Intercept + x1 + x4 + x5 + x6 + x8 + x9 + x10	x5	16	-1956.587077	0.569494603	2	0.752204311	3945.174155	4060.539601
4.3	Intercept + x1 + x4 + x6 + x8 + x9 + x10	x9	14	-1956.999404	0.824653299	2	0.662107964	3941.998808	4042.943573
5.5	Intercept + x1 + x4 + x6 + x8 + x10	x6	12	-1957.947365	1.895922972	2	0.387530204	3939.894731	4026.418815
6.3	Intercept + x1 + x4 + x8 + x10	x8	10	-1959.957215	4.019698609	2	0.134008868	3939.914429	4012.017833
7.3	Intercept + x1 + x4 + x10		8	-1961.902611	3.890792233	2	0.142930595	3939.805222	3997.487945

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

Ans:

The final model suggested by the Backward Selection method is: **Intercept + x1 + x4 + x10.**

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

Ans:

- The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for all the models that are listed in (C) are already given in *Table 1* above. Only for the removed predictor AIC and BIC are given below:

Table 3

Index	Model Form	Suggestion w. r. t. AIC and BIC	AIC	BIC
0	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10	x7		
1.7	Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9 + x10	x3	3952.148857	4096.355664
2.3	Intercept + x1 + x2 + x4 + x5 + x6 + x8 + x9 + x10	x2	3948.60466	4078.390787
3.2	Intercept + x1 + x4 + x5 + x6 + x8 + x9 + x10	x5	3945.174155	4060.539601
4.3	Intercept + x1 + x4 + x6 + x8 + x9 + x10	x9	3941.998808	4042.943573
5.5	Intercept + x1 + x4 + x6 + x8 + x10	x6	3939.894731	4026.418815
6.3	Intercept + x1 + x4 + x8 + x10	x8	3939.914429	4012.017833
7.3	Intercept + x1 + x4 + x10		3939.805222	3997.487945

- In each of the models, **Lowest AIC** and **Lowest BIC** suggested removing the predictor to be removed as shown above.