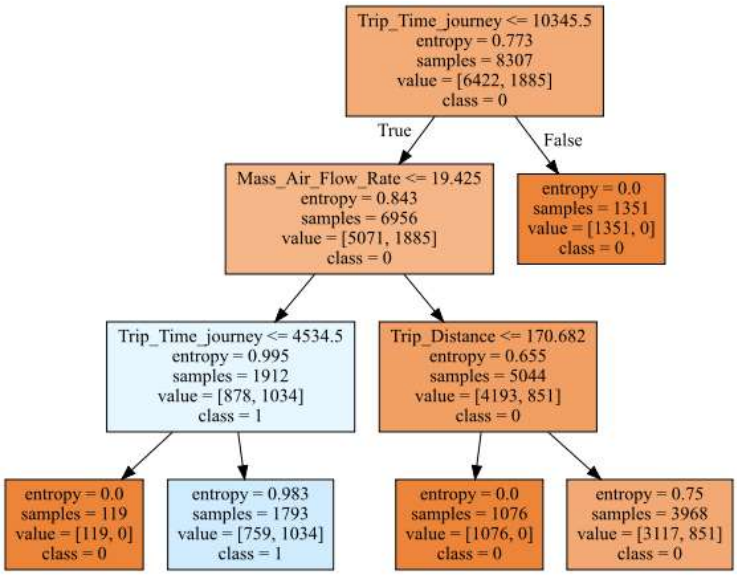
CS 484: Introduction to Machine Learning

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.



false 0 : 851. false 1: 759 . total samples: 8307

(851 + 759)/8307 = 0.1938

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.

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

Entropy for root node is given as 0.9489621493401781

- (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.

  Occupation

  (0.7148805225259208, ('Blue Collar', 'Unknown', 'Student'))

  (4543 2668)

  Car Type

  (0.7573352263531922, ('Minivan', 'SUV', 'Sports Car'))

  (2481 4730)

  Education

  (0.9343298080392602, ('Below High School',))

  (6154 1057)

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

  ('Blue Collar', 'Unknown', 'Student')

  Entropy Value 0.71488

  (4543 2668)

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

  left:

  occupation:

  (0.650587823999444, ('Below High School',))

  Car Type:

  (0.7689481386570244, ('Minivan', 'SUV', 'Sports Car'))

  Education:

[('Blue Collar',), ('Unknown',), ('Student',)]

right:
(0.3212873372854656, ('Minivan', 'SUV', 'Sports Car'))
(0.623532570928089, ('Below High School', 'High School', 'Bachelors'))
(0.5740628685071641, ('Doctor', 'Lawyer'))

- (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.

Predicted probability for CAR_USE: 0.8291814946619217

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.

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

  3 2 1

  [4194, 3532, 2274]

- (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.

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                10000
Model:                        MNLogit   Df Residuals:                     9978
Method:                           MLE   Df Model:                           20
Date:                Sun, 07 Mar 2021   Pseudo R-squ.:                  0.8170
Time:                        19:32:17   Log-Likelihood:                -1956.1
converged:                       True   LL-Null:                       -10688.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
      y=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.0165      0.087     11.636      0.000       0.845       1.188
x1            -1.1172      0.058    -19.343      0.000      -1.230      -1.004
x2            -0.0175      0.026     -0.669      0.503      -0.069       0.034
x3             0.0103      0.018      0.586      0.558      -0.024       0.045
x4            -1.5573      0.041    -38.103      0.000      -1.637      -1.477
x5             0.0030      0.010      0.287      0.774      -0.018       0.024
x6             0.0163      0.009      1.822      0.068      -0.001       0.034
x7         -1.268e-07      0.007   -1.7e-05      1.000      -0.015       0.015
x8            -0.0134      0.007     -2.028      0.043      -0.026      -0.000
x9             0.0076      0.006      1.315      0.189      -0.004       0.019
x10            0.0072      0.009      0.804      0.421      -0.010       0.025
------------------------------------------------------------------------------
      y=3       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4041      0.106      3.817      0.000       0.197       0.612
x1            -1.1685      0.071    -16.354      0.000      -1.309      -1.028
x2             0.0002      0.033      0.005      0.996      -0.064       0.064
x3            -0.0009      0.022     -0.041      0.968      -0.045       0.043
x4            -0.0218      0.027     -0.794      0.427      -0.075       0.032
x5            -0.0088      0.013     -0.671      0.503      -0.034       0.017
x6             0.0004      0.011      0.038      0.970      -0.021       0.022
x7            -0.0017      0.010     -0.179      0.858      -0.021       0.017
x8            -0.0072      0.008     -0.867      0.386      -0.024       0.009
x9             0.0024      0.007      0.324      0.746      -0.012       0.017
x10            1.3464      0.038     35.838      0.000       1.273       1.420
==============================================================================
Model Log-Likelihood Value = -1956.055139748098
Number of Free Parameters = 22
```

Model 0 : y = Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

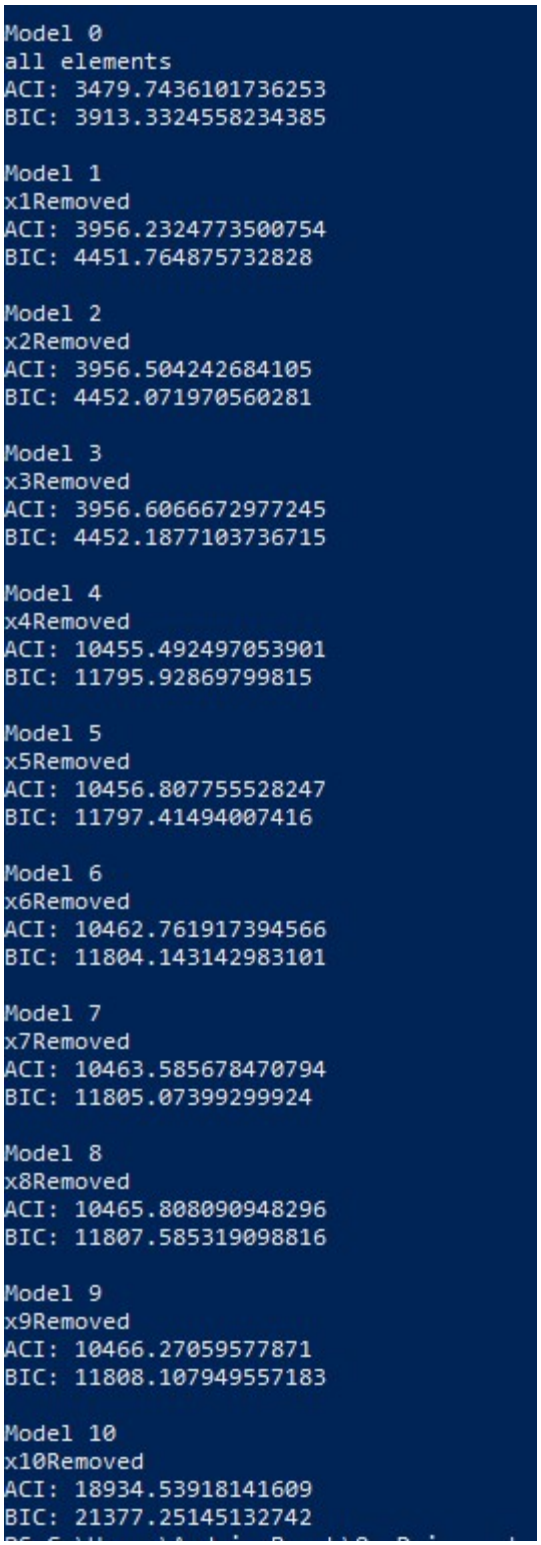
- (20 points). Please show the step summary of the Backward Selection method. The step summary should include the name of the removed feature, the log-likelihood value of the reduced model, the number of free parameters of the reduced model, the Deviance test statistic, the Deviance degree of freedom,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

In the code Q3

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

  Final Model --> y = Intercept + x1 + x4 + x10

- (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?


Model 0
all elements
ACI: 3479.7436101736253
BIC: 3913.3324558234385
Model 1
x1Removed
ACI: 3956.2324773500754
BIC: 4451.764875732828
Model 2
x2Removed
ACI: 3956.504242684105
BIC: 4452.071970560281
Model 3
x3Removed
ACI: 3956.6066672977245
BIC: 4452.1877103736715
Model 4
x4Removed
ACI: 10455.492497053901
BIC: 11795.92869799815
Model 5
x5Removed
ACI: 10456.807755528247
BIC: 11797.41494007416
Model 6
x6Removed
ACI: 10462.761917394566
BIC: 11804.143142983101
Model 7
x7Removed
ACI: 10463.585678470794
BIC: 11805.07399299924
Model 8
x8Removed
ACI: 10465.808090948296
BIC: 11807.585319098816
Model 9
x9Removed
ACI: 10466.27059577871
BIC: 11808.107949557183
Model 10
x10Removed
ACI: 18934.53918141609
BIC: 21377.25145132742

model 7