A20474413
Jorge Gonzalez Lopez

Assignment 3

CS484 Spring 2021
March 05, 2021

Question 1

- **Answer**

Misclassification Rate = 0.1938

- **Explanation**

Rule	Entropy	Cluster 0	Cluster 1	Classification	Misclassification
Trip_Time_journey > 10345.5	0.0	1351	0	Cluster 0	0,00%
Trip_Time_journey <= 10345.5 & Mass_Air_Flow_Rate > 19.425 & Trip_Distance > 170.682	0.75	3117	851	Cluster 0	21,45%
Trip_Time_journey <= 10345.5 & Mass_Air_Flow_Rate > 19.425 & Trip_Distance <= 170.682	0.0	1076	0	Cluster 0	0,00%
Trip_Time_journey <= 10345.5 & Mass_Air_Flow_Rate <= 19.425 & Trip_Time_journey <= 4534.5	0.0	119	0	Cluster 0	0,00%
Trip_Time_journey <= 10345.5 & Mass_Air_Flow_Rate <= 19.425 & Trip_Time_journey > 4534.5	0.983	759	1034	Cluster 1	42,33%

Total number of samples	8307
Total miscalssifications	1610
Misclassification rate	19,38%

Question 2

a)

- **Answer**

  0.948962149

- **Explanation**

  The entropy of the root value is calculated as the entropy of the target (‘CAR_USE’) as follows:

  $$E = -p_{private} log_2(p_{private}) - p_{commercial} log_2(p_{commercial})$$

  To do so, the pandas’ function df['CAR_USE'].value_counts(normalize=True) gives back both probabilities.

b)

- **Answer**

	Left Split	Right Split	Entropy	Gain
CAR_TYPE	Minivan, SUV, Sports Car	Van, Panel Truck, Pickup	0.768415	0.18055
OCCUPATION	Blue Collar, Unknown, Student	Professional, Manager, Clerical, Doctor, Lawyer, Home Maker	0.712583	0.23638
EDUCATION	Below High School	High School, Bachelors, Masters, Doctors	0.935614	0.01335

- **Explanation**

To get the optimal splits, all the possible combinations of values in CAR_TYPE, OCCUPATION and EDUCATION have been calculated. Then, for all of them, the split Entropy has been estimated by getting the probabilities of ‘CAR_USE’ being private or commercial for the number of samples that met the left and right split requirements (df[df[column].isin(split_xx)).

Finally, the split entropy is both split entropies escalated by their probability of occurrence.

c)

- **Answer**

OCCUPATION (the decision rule: value.isin([‘Blue collar’, ‘Unknown’, ‘Student’]))

- **Explanation**

The feature selected for the first split is OCCUPATION because it has the lower entropy split out of the three features (or the maximum gain).

The values in the left branch are: Blue collar, Unknown or Student.

The values in the right branch are: Professional, Manager, Clerical, Doctor, Lawyer or Home Maker

d)

- **Answer**

Left Branch: EDUCATION (the decision rule: value.isin([‘Below High School’]))

Right Branch: CAR_TYPE (the decision rule: value.isin([‘Minivan’, ‘SUV’, ‘Sports Car’]))

- **Explanation**

To calculate the splits for the second layer, the same procedure as the first layer has to be followed. However, in this case, the dataset used to analyze the features is different in each branch:

Left branch -> the subset from the original dataset that meets the first layer decision rule.

Right branch -> the subset from the original dataset that does not meet the first layer decision rule.

The features selected for the second layer split (the ones with the lower entropy split out of all the features (or the maximum gain)) are:

Left Branch: EDUCATION (left branch: Below High School and right branch: High School, Bachelors, Masters, Doctors).

Right branch: CAR_TYPE (left branch: Minivan, SUV, Sports Car and right branch: Van, Pickup, Panel Truck.

e)

- **Answer**

Rule	Entropy	Private	Commercial	Total	Pobability of Private	Pobability of Commercial
OCCUPATION C {Blue collar, Unknown, Student} & EDUCATION C {Below High School}	0.8304	607	216	823	73,75%	26,25%
OCCUPATION C {Blue collar, Unknown, Student} & EDUCATION C {High School, Bachelors, Masters, Doctors}	0.6226	2559	470	3029	84,48%	15,52%
OCCUPATION C {Professional, Manager, Clerical, Doctor, Lawyer, Home Maker} & CAR_TYPE C {Minivan, SUV, Sports Car}	0.0568	4564	30	4594	99,35%	0,65%
OCCUPATION C {Professional, Manager, Clerical, Doctor, Lawyer, Home Maker} & CAR_TYPE C {Van, Pickup, Panel Truck}	0.9974	984	872	1856	53,02%	46,98%

Question 3

a)

- **Command Prompt Output**

```
a) The frequency table of the categorical target field is:
3    4194
2    3532
1    2274
Name: y, dtype: int64
```

- **Explanation**

Just by doing a value_counts() of the target column.

b)

- **Command Prompt Output**

```
Optimization terminated successfully.
         Current function value: 0.195606
         Iterations 10
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                10000
Model:                        MNLogit   Df Residuals:                     9978
Method:                           MLE   Df Model:                           20
Date:                Tue, 02 Mar 2021   Pseudo R-squ.:                  0.8170
Time:                        18:45:18   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 = -1956.0551397480979
Number of Free Parameters = 22
```

- **Explanation**

Get the target values (column ‘y’) and the input dataset (all other columns). Then, the intercept value is added to the input dataset as a constant column (all values equal to 1).

Now, it is possible to calculate the number of free parameters with the rank of the input matrix multiply by the total number of output categories minus 1.

Finally, the Multinomial Logistic Regression model is created with the input and target values and trained. Then, applying the function logit.loglike() to the values of the parameters of the model, the Log-Likelihood value is computed.

c)

- **Command Prompt Output**

```
Model 1  --> Removed feature:  ['x7']
Model Log-Likelihood Value = -1956.0744283318356
Number of Free Parameters = 20
Deviance (Statistic, DF, Significance) 0.03857716747552331 2 0.9808962506876956
Akaike Information Criterion = 3952.1488566636713
Bayesian Information Criterion = 4096.355664103195

Model 2  --> Removed feature:  ['x7', 'x3']
Model Log-Likelihood Value = -1956.30232999277
Number of Free Parameters = 18
Deviance (Statistic, DF, Significance) 0.45580332186864325 2 0.7962025538009445
Akaike Information Criterion = 3948.60465998554
Bayesian Information Criterion = 4078.3907866811114

Model 3  --> Removed feature:  ['x7', 'x3', 'x2']
Model Log-Likelihood Value = -1956.5870772945038
Number of Free Parameters = 16
Deviance (Statistic, DF, Significance) 0.5694946034677741 2 0.7522043110300307
Akaike Information Criterion = 3945.1741545890077
Bayesian Information Criterion = 4060.5396005406265

Model 4  --> Removed feature:  ['x7', 'x3', 'x2', 'x5']
Model Log-Likelihood Value = -1956.9994039437684
Number of Free Parameters = 14
Deviance (Statistic, DF, Significance) 0.8246532985290287 2 0.6621079636455474
Akaike Information Criterion = 3941.9988078875367
Bayesian Information Criterion = 4042.943573095203

Model 5  --> Removed feature:  ['x7', 'x3', 'x2', 'x5', 'x9']
Model Log-Likelihood Value = -1957.9473654299018
Number of Free Parameters = 12
Deviance (Statistic, DF, Significance) 1.895922972266817 2 0.38753020449855446
Akaike Information Criterion = 3939.8947308598035
Bayesian Information Criterion = 4026.4188153235177

Model 6  --> Removed feature:  ['x7', 'x3', 'x2', 'x5', 'x9', 'x6']
Model Log-Likelihood Value = -1959.9572147343606
Number of Free Parameters = 10
Deviance (Statistic, DF, Significance) 4.019698608917679 2 0.13400886768610418
Akaike Information Criterion = 3939.914429468721
Bayesian Information Criterion = 4012.017833188483
```

Model 7 --> Removed feature: ['x7', 'x3', 'x2', 'x5', 'x9', 'x6', 'x8']
Model Log-Likelihood Value = -1961.9026108507444
Number of Free Parameters = 8
Deviance (Statistic, DF, Significance) 3.8907922327675806 2 0.1429305949727081
Akaike Information Criterion = 3939.805221701489
Bayesian Information Criterion = 3997.4879446772984

Best model --> Removed feature: ['x7', 'x3', 'x2', 'x5', 'x9', 'x6', 'x8'] if we try to remove another feature: x1
Model Log-Likelihood Value = -2230.4083138546553
Number of Free Parameters = 6
Deviance (Statistic, DF, Significance) 537.0114060078217 2 2.4516294286020626e-117
Akaike Information Criterion = 4472.8166277093105
Bayesian Information Criterion = 4516.07866994116

- **Explanation**

The features x1, x4 and x10 are the most important ones, as if one of them gets removed from the inputs, the Deviance significance values will get lower than 0.05.

The backward method has been used. For that reason, the significance value is checked when each and every feature of the model is removed. Then, if the greater significance value among all the features is over 0.05, the feature gets eliminated from the model. These steps are carried out until the significance value of all the features remaining is less than 0.05.

d)

- **Command Prompt Output**

```
Optimization terminated successfully.
         Current function value: 0.196190
         Iterations 10
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                10000
Model:                        MNLogit   Df Residuals:                     9992
Method:                           MLE   Df Model:                            6
Date:                Tue, 02 Mar 2021   Pseudo R-squ.:                  0.8164
Time:                        18:49:41   Log-Likelihood:                -1961.9
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          1.0168      0.087     11.675      0.000       0.846       1.188
x1            -1.1145      0.058    -19.332      0.000      -1.227      -1.001
x4            -1.5540      0.041    -38.203      0.000      -1.634      -1.474
x10            0.0073      0.009      0.811      0.417      -0.010       0.025
------------------------------------------------------------------------------
       y=3       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4064      0.105      3.857      0.000       0.200       0.613
x1            -1.1664      0.071    -16.350      0.000      -1.306      -1.027
x4            -0.0224      0.027     -0.821      0.412      -0.076       0.031
x10            1.3449      0.037     35.922      0.000       1.272       1.418
==============================================================================
Model Log-Likelihood Value = -1961.9026108507444
Number of Free Parameters = 8
```

- **Explanation**

The final model is computed by subtracting the least relevant columns from the dataset. In this case, those columns are x2, x3, x5, x6, x7, x8 and x9. These columns are found observing the Deviance significance value of the model when some features are removed. If this value is greater than 0.05, subtracting that feature would not greatly impact the model.

e)

- **Answer**

For the models listed in c:

- The Akaike Information suggests the best model would be removing: ['x7', 'x3', 'x2', 'x5', 'x9'].
- The Bayesian Information suggests the best model would be removing: ['x7', 'x3', 'x2', 'x5', 'x9', 'x6', 'x8']

- **Explanation**

The Akaike and Bayesian Information are other criteria for choosing the features to get the best model. It is intended to get the lowest possible values in both Criterion.
From the models listed in c), the Bayesian information suggests that the best model is Model 7 -> removing the same columns than while checking the p-value, because every successive model reduces its value. However, for the Akaike information, its value with Model 5 is bigger than with the previous one. Therefore, the model it suggests is Model 5 (even though Model 7 has a lower value).

```
Model 1  --> Removed feature:  ['x7']
Akaike Information Criterion = 3952.1488566636713
Bayesian Information Criterion = 4096.355664103195

Model 2  --> Removed feature:  ['x7', 'x3']
Akaike Information Criterion = 3948.60465998554
Bayesian Information Criterion = 4078.3907866811114

Model 3  --> Removed feature:  ['x7', 'x3', 'x2']
Akaike Information Criterion = 3945.1741545890077
Bayesian Information Criterion = 4060.5396005406265

Model 4  --> Removed feature:  ['x7', 'x3', 'x2', 'x5']
Akaike Information Criterion = 3941.9988078875367
Bayesian Information Criterion = 4042.943573095203

Model 5  --> Removed feature:  ['x7', 'x3', 'x2', 'x5', 'x9']
Akaike Information Criterion = 3939.8947308598035
Bayesian Information Criterion = 4026.4188153235177

Model 6  --> Removed feature:  ['x7', 'x3', 'x2', 'x5', 'x9', 'x6']
Akaike Information Criterion = 3939.914429468721
Bayesian Information Criterion = 4012.017833188483

Model 7  --> Removed feature:  ['x7', 'x3', 'x2', 'x5', 'x9', 'x6', 'x8']
Akaike Information Criterion = 3939.805221701489
Bayesian Information Criterion = 3997.4879446772984
```