CS 484: Introduction to Machine Learning

Fall Semester 2023 Assignment 5

SAI RAM ODURI

A20522183

Question 1 (100 points)

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). We will use two of the datasets in the repository for analyses, namely, the **WineQuality_Train.csv** for training and the **WineQuality_Test.csv** for testing.

The categorical target variable is *quality_grp*. It has two categories, namely, 0 and 1. The Event category is 1. The input features are *alcohol*, *citric_acid*, *free_sulfur_dioxide*, *residual_sugar*, and *sulphates*. These five input features are considered interval variables.

We will train two models. One is a classification tree, and another is a binary logistic regression.

The classification tree has the following specifications.

- The Splitting Criterion is Entropy.
- The maximum tree depth is five.
- The initial random state value is 20230101 for the classification tree.

The binary logistic regression has the following specifications.

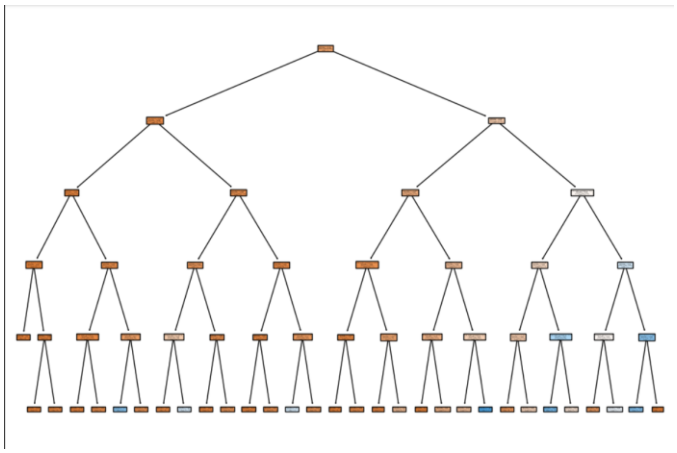
- The model must include the Intercept term.
- Use the All-Possible Subset method to determine the model with the lowest Akaike Information Criterion.

After we train these two models, we will compare them using a suite of model performance metrics and charts.

(a) (20 points) What are the Root Average Squared Error values of both models for both training and testing partitions?

For Classification tree

```
Accuracy: 0.8158974358974359
Confusion Matrix:
[[1466   99]
 [ 260  125]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.94      0.89      1565
           1       0.56      0.32      0.41       385

    accuracy                           0.82      1950
   macro avg       0.70      0.63      0.65      1950
weighted avg       0.79      0.82      0.80      1950
```



```
alcohol: 0.6874894053881022
citric_acid: 0.1384235200085932
free_sulfur_dioxide: 0.057978765712633634
residual_sugar: 0.06314460414629469
sulphates: 0.05296370474437615
```

For logistic regression

```
Optimization terminated successfully.
         Current function value: 0.413258
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.412059
         Iterations 7
Optimization terminated successfully.
         Current function value: 0.413851
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.412167
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.489349
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.500117
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            quality_grp   No. Observations:                 4547
Model:                          Logit   Df Residuals:                     4542
Method:                           MLE   Df Model:                            4
Date:                Tue, 28 Nov 2023   Pseudo R-squ.:                 0.1676
Time:                        23:23:49   Log-Likelihood:                -1873.6
converged:                       True   LL-Null:                       -2251.0
Covariance Type:            nonrobust   LLR p-value:                4.888e-162
```

If we look the logit regression table this the complete report

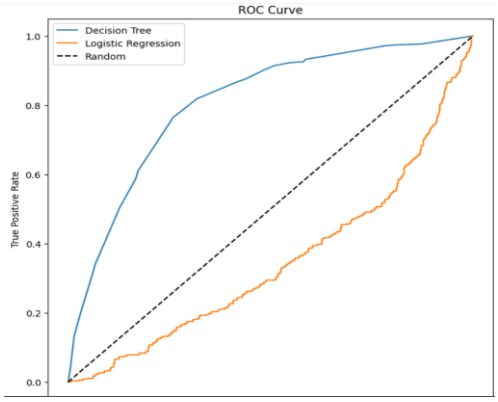
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -12.5048      0.484    -25.821      0.000     -13.454     -11.556
alcohol        0.9014      0.037     24.639      0.000       0.830       0.973
citric_acid    0.9580      0.298      3.220      0.001       0.375       1.541
free_sulfur_dioxide     0.0143      0.002      5.793      0.000       0.009       0.019
sulphates      1.0593      0.267      3.966      0.000       0.536       1.583
==============================================================================
Accuracy: 0.8020512820512821
Confusion Matrix:
[[1564    1]
 [ 385    0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89      1565
           1       0.00      0.00      0.00       385

    accuracy                           0.80      1950
   macro avg       0.40      0.50      0.45      1950
weighted avg       0.64      0.80      0.71      1950
```

```
RASE - Custom Classification Tree (Train): 0.33981146313626537
RASE - Custom Classification Tree (Test): 0.35450835495056965
RASE - Custom Logistic Regression (Train): 0.3594095216698445
RASE - Custom Logistic Regression (Test): 0.36119586645907126
```

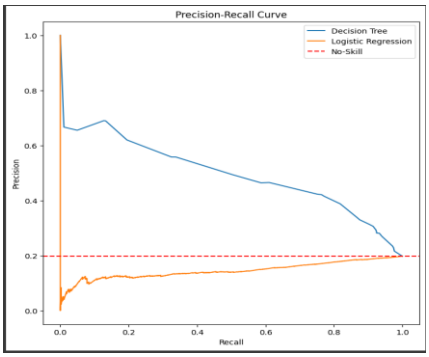
(b) (20 points) What are the Area Under Curve values of both models for both training and testing partitions?

```
AUC (Decision Tree, Training): 0.8444447682086705
AUC (Decision Tree, Testing): 0.8069441102028961
AUC (Logistic Regression, Training): 0.3256025593050861
AUC (Logistic Regression, Testing): 0.3459657275631717
```

(c) (10 points) Generate the Receiver Operating Characteristic curve for both models on the training partition. Please put the two curves in the same chart frame. Don't forget to add the diagonal reference line.



(d) (10 points) Generate the Precision and Recall chart for both models on the training partition. Please put the two curves in the same chart frame. Don't forget to add the No-Skills line to the chart.



(e) (10 points) What is the threshold for the Event probability based on the F1 Score from the training partition? Please calculate the thresholds of both models.

Threshold for logistic regression is generally 0.5, but when the f1 score is take into account,

The threshold for logistic regression is 0.00

The optimal threshold decision tree is 0.16717

```
Optimal Threshold for Logistic Regression: 0.0
Threshold for Decision Tree: 0.16716716716716717
```

(f) (10 points) Using the F1 Score threshold, what are the Misclassification Rates of both models when evaluated only on the testing partition?

The misclassification for decision tree – 0.2513

The misclassification for logistic regression – 0.80

```
Misclassification Rate for Decision Tree: 0.2512820512820513
Misclassification Rate for Logistic Regression: 0.8025641025641026
```

(g) (10 points) Generate the Cumulative Gain and Lift table for both models using the predicted Event probabilities from the testing partition. Which model has the highest Lift value in Decile 1?

```
Cumulative Gain and Lift Table for Decision Tree:
      Cumulative Gains (Decision Tree)  Lift (Decision Tree)
1644                          0.002597              0.002597
1862                          0.005195              0.002597
855                           0.007792              0.002597
1667                          0.007792              0.001948
545                           0.007792              0.001558
...                                ...                   ...
259                           1.000000              0.000514
888                           1.000000              0.000514
1135                          1.000000              0.000513
1536                          1.000000              0.000513
663                           1.000000              0.000513

[1950 rows x 2 columns]

Cumulative Gain and Lift Table for Logistic Regression:
      Cumulative Gains (Logistic Regression)  Lift (Logistic Regression)
1651                                     0.0                    0.000000
490                                      0.0                    0.000000
1853                                     0.0                    0.000000
1093                                     0.0                    0.000000
303                                      0.0                    0.000000
...                                      ...                         ...
64                                       1.0                    0.000514
1287                                     1.0                    0.000514
1897                                     1.0                    0.000513
1137                                     1.0                    0.000513
698                                      1.0                    0.000513

[1950 rows x 2 columns]
```

(h) (10 points) Based on all the above model performance metrics and charts, which model will you pick as the Champion model?

Ans. My choice based on the metrics above considering that classification tree is outperforming logistic regression in most of the metrics, since both have an accuracy close to 81.5 and 80.29 % but the decision classifier does better job in classifying and identifying the features correctly. The misclassification rate of decision tree 0.25 , and logistic regression is 0.19. But in terms of AUC and ROC decision tree performance is better. So, the champion model is decision tree.