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.

1. (20 points) What are the Root Average Squared Error values of both models for both training and testing partitions?

**For Classification tree**

A screenshot of a computer

Description automatically generated

A diagram of a structure

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**For logistic regression**

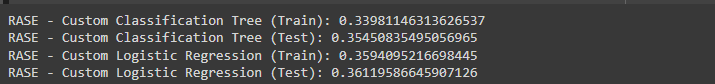
**A screenshot of a computer

Description automatically generated**

**If we look the logit regression table this the complete report**

**A screenshot of a computer

Description automatically generated**

****

1. (20 points) What are the Area Under Curve values of both models for both training and testing partitions?

A screenshot of a computer code

Description automatically generated

1. (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.

A graph of a line

Description automatically generated

1. (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.

A graph of a line graph

Description automatically generated

1. (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



1. (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



1. (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?

A screenshot of a computer

Description automatically generated

1. (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.