CS 584-04: Machine Learning

Fall 2018 Final Examination

# Question 1 (100 points)

You are a data scientist in a trucking company in Europe. The truck company operates their trucks in three regions. The regulations require the company to maintain the trucks in good operating conditions. At the same time, the company wants to minimize the downtime due to taking the trucks offline for maintenance. Therefore, you are asked to develop business rules that help you determine if a truck is due for maintenance.

You have access to historical data about the trucks in your company. The data contain 9,398 observations which you have divided into the training and the testing partitions. The training partition is made available in the **fleet\_train.csv** file and it has 7,504 observations. The testing partition is made available in the **fleet\_monitor\_notscored\_2.csv** file and it has 1,894 observations.

Both partitions contain the following variables for modeling.

**Target Variable**

1. Maintenance\_flag: 1 = Offline for maintenance, 0 = Otherwise. The event value is 1.

**Nominal Predictor**

1. Region: 1, 2, and 3

**Interval Predictors**

1. Accel\_Pedal\_Pos\_D
2. Accel\_Ssor\_Total
3. Ambient\_air\_temp
4. CO2\_in\_g\_per\_km\_Inst
5. Engine\_Coolant\_Temp
6. Engine\_Load
7. Engine\_Oil\_Temp
8. Engine\_RPM
9. GPS\_Altitude
10. GPS\_Bearing
11. GPS\_Latitude
12. GPS\_Longitude
13. Intake\_Air\_Temp
14. Intake\_Manifold\_Pressure
15. Litres\_Per\_100km\_Inst
16. Mass\_Air\_Flow\_Rate
17. Speed\_GPS
18. Speed\_OBD
19. Throttle\_Pos\_Manifold
20. Trip\_Distance
21. Trip\_Time\_journey
22. Turbo\_Boost\_And\_Vcm\_Gauge
23. Vehicle\_speed\_sensor
24. Vibration
25. Voltage\_Control\_Module

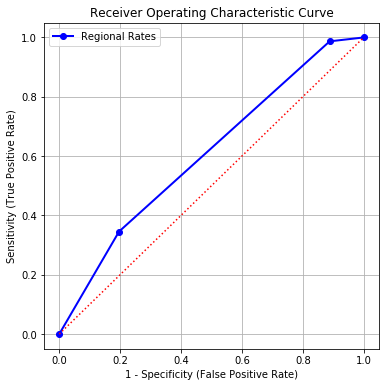
Currently, a truck is taken offline for maintenance according to the region it belongs to. A truck in Region 1 is maintained once every five days. A truck in Region 2 is maintained once every three days. A truck in Region 3 is maintained once every fifty days. This maintenance schedule is equivalent to this business rule.

|  |  |
| --- | --- |
| Region | Probability that Maintenance\_flag = 1 |
| 1 | 1/5 |
| 2 | 1/3 |
| 3 | 1/50 |

When applying this model to the testing partition, we obtained the following model metrics.

1. Area Under Curve = 0.60587637
2. Root Average Squared Error = 0.39477599
3. Misclassification Rate = 0.28880676 (the threshold value is 0.20469083)

The Receiver Operating Characteristic curve for the testing partition is



You are challenged to develop **a new model for each region** which collectively reduces the misclassification rate for the testing partition further. In addition, you will try to neither decrease the Area Under Curve nor increase the Root Average Squared Error.

You can consider any classification model, except the nearest neighbor model, for each region. For example, you can try a classification tree for Region 1, a logistic model for Region 2, and a gradient boosting model for Region 3. You, of course, can use the logistic model for all three regions. After you are satisfied with your models, you will pool the predicted probabilities from these three regions and compare with the observed maintenance flag in the testing partition.

You will prepare a report for your new models. Your report must contain the following five sections.

1. (20 points) Describe your strategies for developing the new models.
2. (20 points) Show how you selected the predictors into each new model. Your objective is to exclude variables that do not contribute to the goodness-of-fit of the new model.
3. (10 points) List the primary model specifications and the key model results (e.g., decision tree diagram, logistic regression, parameter estimates, support vector machine hyperplane equation, etc.)
4. (30 points) Show the model comparison results and list all supporting tables and charts (e.g., Area Under Curve, Root Average Squared Error, Misclassification Rate, ROC curve, Lift or Accumulated Lift curves)
5. (20 points) Argue that you have actually found a better model than the current model.

You must limit your report to no more than four pages including all texts, tables, and charts. Don’t use any overly small font sizes. If I cannot read your report, then I cannot give you points.