Applied Statistical Analysis: Problem 3 (Team-Based Project)

Conduct the following subjects by using the data for Census Income Data Set, which is uploaded to Keio.jp. There are two data sets: “train_test.csv” and “submit.csv”. See the document for the detail. This problem is performed by the team-based project.

1. Construct several models to predict the output variable *y* (whether a person makes over 50K a year (binary: 'yes','no')). In order to do it, you can divide the data “train_test.csv” into training sample and test sample data.

2. As for the models, you are requested to use **logistic regression** and **classification tree** models. Furthermore, you can use other advanced models, such as Support Vector Machine, Random Forest, Deep Learning, etc. (these models are not explained in this class.)

3. Show the ROC (Receiver Operating Characteristics) Curve and AUC (Area under the curve) for the several candidate models by using the test data set. Select the best model in terms of the AUC. For the selected model, examine the important input variable which could have an effect on the output variable.

4. As for the selected model, examine the classification limit (or threshold to determine the output variable ('yes','no')). Show the confusion matrix (recall, precision) and F-measure for the determined threshold of the selected model, from test data set.

- Deadline for the submitting file for each team: January 5 (Sunday) at 20:00, 2020.

  Please submit the “submit” file via keio.jp.

  To hsuzuki@ae.keio.ac.jp

  The “submit” file is used to submit the **predicted output variable** (whether a person makes over 50K a year (binary: 'yes','no') and the **probability of ‘yes’** (or score). There are 13 input variables. The output variable is not provided. The number of instances is 2,000.

- Deadline for the presentation file of Power Point for each team: January 5 (Sunday) at 20:00, 2020.

  Please submit your presentation file of Power Point via keio.jp

  Write in English.

  Write the following items in the 1st page.

  - Title of your presentation
  - Tem member name

You are required to make a presentation in the class on January 6 and 20 (Monday), 2020.

The presentation time is 10 minutes.

The presenter is 2-3 persons.

English presentation is required.

The students will be requested to evaluate other team presentations.

The student's evaluation is considered when this problem is graded.

The AUC and F-measure scores on the basis of your submitted data are also considered for the evaluation.