Python DAPT 622

Homework #2 (Group Assignment)

Due on 9-APR-2019

This homework assignment is an extension of the class exercises from the Python session on 22-Feb-2019, and the Data Mining session on Mar-9-2019. You will use the same dataset that was used in those sessions: credit_default_model_data.csv. You can directly read this file from the following (Dropbox) link: 'https://www.dropbox.com/s/qac84v449rjytv5/credit_default_model_data.csv?dl=1'.

Please note that this data is sourced from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. This link includes a brief description of most columns available in these files. Alternatively, you can also download these two files from BlackBoard under Python→Data folder to your local computer and read those files into Python from your local folder.

Using this data set, perform a random-search to optimize the following hyper-parameters for a classification model using **Gradient Boosting**: `n_estimators`, `learning_rate`, `max_depth`, `min_samples_leaf`, and `subsample`. Use `'default payment next month'` column as the target variable. Use a couple of iterations (of random-search) if needed. Once the optimal values for all hyper-parameters are finalized, plot the ROC curve and calculate AUC for the final model. Provide the list of optimal values for all hyper-parameters.

Submit the Python program (either a Jupyter Notebook of a `.py` file) that performs the above task. You can post your submission via BlackBoard or email (patelvj2@vcu.edu). This program should be fully executable. I should be able to run your code after changing the location of the input files. No other changes should be required.