Assignment 2 - ML – Early Prediction of dropouts from a course

Instructions:

a) "Learning is the Goal"… "NOT grades".
b) Students are expected to have good knowledge in feature engineering and classification algorithms. Revise the concepts before solving the problem.
c) It is an individual assignment, not a group activity.
d) You must provide complete solution, with analysis, not just answer. Right approach with appropriate explanation/analysis will be appreciated even though final answer is wrong.
e) Model tuning and evaluation are mandatory.
f) Submit the solution as a softcopy with the file name as "RollNumber_Name_Exercise2_ML.ipynb". No other format is allowed.
g) The solutions will be evaluated automatically using scripts. Strictly follow the instructions.

Problem Statement

The data set of 10,000 students who enrolled in an online course (MOOC) is given. It contains several attributes related to the events/activities on the course portal. It also has a ground truth or label called "dropout". Build the models to predict and analyze which student may continue the course till the end (dropout=0) and which student may discontinue the course (dropout=1).

Details of Attributes: (File: dropout_train.csv & dropout_test.csv)

1) n_events_lst_wk: Number of events in the last week
2) days_course_strt_access1: Number of days between the end of the course and the last day of access of the course material
3) n_access_lst2_wk: Number of accesses in the last two weeks
4) n_events: Total number of events
5) unique_days_accessed: The number of unique days accessed
6) n_access: Total number of accesses till the prediction time
7) n_access_lst_wk: Total number of accesses in the last week
8) n_navigate: Total number of page navigations
9) n_page_close: Total number of page closes

10) n_problem: Total number of problems solved
11) n_videos: Total number of videos watched
12) days_course_end_access_lst: From the start date of the course, after how many days a student accessed course content
13) n_discussion: Total number of discussions on forum
14) n_wiki: Total number of wiki views

a) Analyze the data

- Find out if there are any attributes with correlation more than 0.50
- Perform data exploration (Visual Analysis) and write your observations

b) Curate the data

- Identify the missing values and fill them with an appropriate method.

c) Build dropout prediction models using Random Forest, GBM, XGBOOST and Multilayer Perceptron (MLP).

- Build the model & Perform 5-fold cross validation
- Evaluate the model using accuracy, confusion matrix, precision, recall and F1
- Compare the performance of all the models against test data (choose an appropriate train and test ratio)
- Perform hyper-parameter tuning

d) Identify the top 5 important predictors and visualize the importance scores

e) Save the model and load it for predictions

f) Analyze the results

NOTE: Use less number of samples if your machine does not have enough resources.