Instructions

- This homework assignment is worth 160 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: March 31, 2023 by 11:59 pm.**

For this homework assignment and for future one, we will work on the challenge presented in the `data mining cup 2019`. Please read the task and get familiar with the data. For this week homework assignment, answer the following:

Exercise 1

(5 points) Using the bucket, that you create in the last homework assignment, and the pandas library, read the `train.csv` and `test.csv` data files and create two data-frames called `train` and `test`, respectively.

Exercise 2

(85 points) Using the `train` data-frame (including the top 7 features from homework assignment 5), do the following:

(i) Consider a model to predict `fraud`. Then, do the following:

- With the top 5 important features and using the `GridSearchCV` function with `cv = 3`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 6 important features and using the `GridSearchCV` function with `cv = 3`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 7 important features and using the `GridSearchCV` function with `cv = 3`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.

From above three scenarios, identify the best model; that is, the model model (input features and hyper-parameters) that has the best performance.

(ii) Consider a model different from part (i) to predict `fraud`. Then, do the following:

- With the top 5 important features and using the RandomizedSearchCV function with `cv = 3` and `n_iter = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 6 important features and using the RandomizedSearchCV function with `cv = 3` and `n_iter = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 7 important features and using the RandomizedSearchCV function with `cv = 3` and `n_iter = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.

From above three scenarios, identify the best model; that is, the model model (input features and hyper-parameters) that has the best performance.

(iii) Consider a model different from parts (i) & (ii) to predict `fraud`. Then, do the following:

- With the top 5 important features and using the Optuna framework using 3 folds and `N_TRIALS = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 6 important features and using the Optuna framework using 3 folds and `N_TRIALS = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.
- With the top 7 important features and using the Optuna framework using 3 folds and `N_TRIALS = 30`, run a hyper-parameter tuning procedure on the model. Please see page 4 of `DATA-MINING-CUP-2019-task.pdf` file to understand how the model should be evaluated.

Exercise 3

(70 points) Using the `train` data-frame and the models from exercise 2, split the `train` data-frame into two data-frames: `training` (80%) and `validation` (20%) taking into account the proportions of 0s and 1s. Then, do the following:

(i) Consider the best model from exercise 2(i). Build that model on the `training` data-frame. After that, predict the likelihood of `fraud` on the `validation` and `test` data-frames.

(ii) Consider the best model from exercise 2(ii). Build that model on the `training` data-frame. After that, predict the likelihood of `fraud` on the `validation` and `test` data-frames.

(iii) Consider the best model from exercise 2(iii). Build that model on the `training` data-frame. After that, predict the likelihood of `fraud` on the `validation` and `test` data-frames.

Using the prediction on the `validation` data-frame as inputs from parts (i)-(ii)-(iii) and the actual `fraud` values from the `validation` data-frame as the target variable, build a meta-learner to predict `fraud`. Make sure to tune the hyper-parameters of the meta-learner keeping in mind how the results are going to be evaluated. For more info, see page 4 of `DATA-MINING-CUP-2019-task.pdf` file. Finally, use the best meta-learner to predict the likelihood of `fraud` in the `test` data-frame. Submit the likelihoods in a csv file. Also submit the associated cut-off value.