Readme
Please put input and output folders into code folder

Table of Files
01_Data_preprocessing.ipynb
02_EDA_plot.ipynb
03_model_01_DNN_Gluon.ipynb
03_model_02_LGB_XGB.ipynb
pm.py
04_1Package_demo for internal data.ipynb
04_2Package_demo for external data.ipynb

FILE NAME: 01_Data_preprocessing.ipynb
Run the 1st cell to load packages
Run the 2nd cell to set data types for category features
Run the 3rd cell to load internal and external data
Run the 4th cell to remove outliers for test data
Run the 5th cell to remove outliers for train data
Run the 6th cell to combine test and train dataset
Run the 7th cell to remove useless features
Run the 8th cell to create feature month
Run the 9th cell to combine feature spine, limb and onset into a new feature and convert it into a value
Run the 10th cell to create new features related with t
Run the 11th cell to change data type
Run the 12th cell to transform OHE for feature race_bl, race_bl and rop
Run the 13th cell to fill in missing values with mode and median values for category and continuous variables separately
Run the 14th cell to output preprocessed data(combined internal and external data)

FILE NAME: 02_EDA_plot.ipynb

Run the 1st cell to load library
Run the 2nd cell to load preprocessed data
Run the 3rd cell to Response variable frequency distribution with month
Run the 4th cell to Onset differences with month
Run the 5th cell to Gender differences with month
Run the 6th cell to Performance of six cases of patients

FILE NAME: 03_model_01_DNN_Gluon.ipynb
Run the 1st cell to load packages
Run the 2nd cell to load preprocessed data
Run the 3rd cell to define loss function and network structure
Run the 4th cell to split patient id into 10 folds
Run the 5th cell to define train model function
Run the 6th cell to define k-fold cross validation function
Run the 7th cell to set hyper-parameters
Run the 8th cell to run k_fold_cross_valid
Run the 9th cell to output predicted train value
Run the 10th cell to output predicted test value
Notice: the training process needs at least 6 hours to finish.

FILE NAME: 03_model_02_LGB_XGB.ipynb
Run the 1st cell to load packages
Run the 2nd cell to load preprocessed data
Run the 3rd cell to split patient id into 10 folds
Run the 4th cell to run LightGBM and XGBoost in 10-folds
Run the 5th cell to obtain predicted train value table from DNN and combine the predicted train result of three models
Run the 6th cell to obtain best ratios for three models
Run the 7th cell to define prediction summary function to compare different predicted result for differnt models
Run the 8th cell to show train dataset prediction summary
Run the 9th cell to load predicted test value
Run the 10th cell to combine predicted test result of three models
Run the 11th cell to show test dataset prediction summary

FILE NAME: pm.py
NAME: package_gu.pm
DESCRIPTION: This is a package for evaluating different models and compare the performances for Python.
PACKAGE CONTENTS:
        function1: descriptive statistics
        function2: histogram of a column
        function3: box plot of predicted value for every model
        function4: plot predicted VS observed; plot residual VS observed
        function5: MPE for filtered feature
        function6 and 7: compare the perform of models with filtered feature
        function 8: get the confidence interval
    @author: eileen
FUNCTIONS
    binned_prediction_summary(data)
        # function 1
    histogram(df, label)
        # function2 histogram for the column
    draw_box(df, category, lower, upper)
        # function 3
    draw_scatter(df, filter_feature=None, lb=0, ub=0)
        # function 4
    MPE_histogram(df, category, lower, upper, model)
        # function 5
    prediction_summary(df, filter_feature=None, lb=0, ub=0)
        # function 6 function 7

FILE NAME: 04_1Package_demo for internal data.ipynb
Run 1st cells to load packages and load the result data.
Run 2nd cells to show the instructions of the package.
Run 3rd cells to call the descriptive statistics from the package we created.
Run 4rd cells to call the histogram from the package we created.
Run 5th cells to call the draw scatters for the whole period from the package we created.
Run 6th cells to call the draw scatters for the first one year from the package we created.
Run 7th cells to call the prediction_summary from the package we created.
Run 8th cells to call the prediction_summary for the first 456 days

from the package we created.
Run 9th cells to call the prediction_summary for the first 10 months from the package we created.
Run 10th cells to call the draw_box for the first 10 months from the package we created.
Run 11th cells to call the MPE for the first 12 months from the package we created.

FILE NAME: 04_2Package_demo for external data.ipynb
Run 1st cells to load packages and load the result data.
Run 2nd cells to call the descriptive statistics from the package we created.
Run 3rd cells to call the histogram from the package we created.
Run 4th cells to call the draw scatters for the whole period from the package we created.
Run 5th cells to call the draw scatters for the first one year from the package we created.
Run 6th cells to call the prediction_summary from the package we created.
Run 7th cells to call the prediction_summary for the first one year from the package we created.
Run 8th cells to call the prediction_summary for the first 40 months from the package we created.
Run 9th cells to call the draw_box for the first 10 months from the package we created.
Run 10th cells to call the MPE for the first 12 months from the package we created.