MFE5310 Assignment 2

Deadline:

- Submit to bb before **23:59:59, Apr.11, 2020**
- Late submission of an assignment would result in a reduced grade for the assignment, unless an extension has been granted by the instructor. A late submission receives an additional 20 points penalty for every 24 hours delay.

Evaluation

- -20 if your program cannot be executed without any running error
    - write a running guideline if you have multiple files
- all the print messages will be counted to the assignment scores

Submission:

Please hand out a zip file named with **your name and Student No.** which includes your source files of your program executions.

For example,

You should hand out .py and compress the them all in a zip file 张三216010000.zip .

Refer to the lr.ipynb in our course, write a program in python. The requirements are shown below

1. **Task: predict any financial assets movement direction (up/down) (classification) or return (regression).** No need to be organized as a trading strategy
2. prepare the dataset for any financial assets in China, Hong Kong or US market. **Print** what kind of dataset you are using
    a. Should check if your dataset has class imbalance issue, and **print** the result in your program
3. Implement the machine learning algorithms (hints: try to use sklearn package)
    a. Random Forrest
    b. GBDT
4. **Fine Tune the parameters at least including max_depth, n_estimators using GridSearch**
    a. Normal cross validation
    b. TimeSeriesSplit cross validation
    c. TimeSeriesSplit cross validation with a fixed training size
5. Find features by yourself. **Print** what kind of features you are using.
    a. Hints: OHLCV
6. **If you want to give different weight for different training samples, how to do that? for example, you expect your model is able to predict more accurate for MOUTAI, such a large cap stock, than some small cap stocks. (Hint: use sample_weight parameter in .fit function in sklearn)**
7. **Try to PRINT the feature importance for Random Forrest and GBDT (Hint: feature_importances_ in xgboost object)**
8. Compare the binary classification performance for all the machine learning models you have implemented. Evaluate your results by the metrics below, and **print** it as a pandas data frame
    a. Accuracy
    b. Precision, recall, f1
    c. ROC/AUC
9. **Print** your conclusion
    a. e.g. "GBDT performs the best for my dataset (daily, CSI300 constituents, from 20160801 to 20180901) with the accuracy 75%, precision **, recall **, f1**, auc **"