11010COM525200 Financial Technology

Program HW3

Deadline: 1/3(Monday) 23:59

Grading Policy:

1. In the programing assignment, the code, data and report should be compressed into a **ZIP** file and upload to eeclass website. Also, please write a Readme file to explain how to run your code and discuss characteristics in your report. The report format is not limited.

2. The programming language that can be used on this assignment is Python. Built-in libraries or functions are allowed to use.

3. 30% off for late submission within one week, not accepted after one week.

4. Discussions are encouraged, **but plagiarism is strictly prohibited.**

Problem:

1.
   A. Collect the TAIEX from 2012/12/01 to 2021/12/02 (Day Bar). The data should include open, close, high, low, volume.

   B. Apply the triple-barrier method to label the collected data. The upper bound is set as 4% and labeled as 1 once it is touched. The lower bound is set as 2% and labeled as 2 once it is touched. The vertical barrier is set as 20 days and labeled as 0 once it is touched.

   C. Use “close price” to calculate 8 technical indicators of collected data. The technical indicators are:
      i. Bios of moving average: 5-days, 10-days, 20-days, 60-day.

      ((價格-MA)/MA)

ii. RSI: 14

iii. MACD(快線 DIF), MACD signal(慢線), MACD histogram(柱狀)

iv. Save problem 1, 2, 3 to a csv.

2. Use the collected and analyzed data in problem 1 to train a random forest model. And apply grid search and cross validation method to find the best parameter of trained model.
   A. The test data is the last 30% of the original data.
   B. The parameter are:
      i. 'bootstrap': [True],
      ii. 'max_depth': [80, 90, 100, 110],
      iii. 'max_features': [2, 3],
      iv. 'min_samples_leaf': [3, 4, 5],
      v. 'min_samples_split': [8, 10, 12],
      vi. 'n_estimators': [100, 200, 300, 1000]
   C. Print the distribution of all data, training data and testing data.
   D. Apply 3-fold cross validation and grid search to tune the hyperparameter.
   E. Print the best training score, best parameter and testing score of best estimator found from problem 2-D. The score is defined as accuracy.
   F. Plot the ROC curve of the best estimator found from problem 2-D.