Data Science 110 Homework 3

Problem 1: One-by-one Feature Selection

1. Describe your feature selection method

    ▼ 計算 feature 和 label 的 chi-squared stats，根據結果選出前 10 高的 feature

2. Show your result and code of the feature selection. Which features are selected?

    a. result

        i. ['Hsa.4689']

        ii. ['Hsa.1130']

        iii. ['Hsa.692']

        iv. ['Hsa.8147']

        v. ['Hsa.692']

        vi. ['Hsa.1221']

        vii. ['Hsa.692']

        viii. ['Hsa.1131']

        ix. ['Hsa.140']

        x. ['Hsa.1832']

    ▼ code 請見附檔

Problem 2: Subset-Based Feature Selection

1. Describe the following details:

    a. your algorithm, the metaheuristic you choose (PSO, SA or GA)

        ▼ GA

    b. your objective function

        ▼ Logistic Regression

c. the tunable parameters and tunable algorithm components (besides the objective function/cost function module) in your metaheuristic.
What are the specific values/methods you use for your tunable parameters and algorithm component(s), if any

- GA 的一些參數設定
    - ▼ 最多的 feature 數（max feature）：10
    - ▼ population 數：50
    - ▼ 進行 crossover 的機率：0.5
    - ▼ 進行 mutation 的機率：0.2
    - ▼ 幾個 generation：40
    - ▼ tournament_size：3

2. Show your result and code of the feature selection.
(How many features are selected? Which features are selected? Etc.)

a. result

i. ['Hsa.467']

ii. ['Hsa.749']

iii. ['Hsa.1272']

iv. ['Hsa.6617']

v. ['Hsa.166']

vi. ['Hsa.2904']

vii. ['Hsa.42826']

viii. ['Hsa.3024']

ix. ['Hsa.2918']

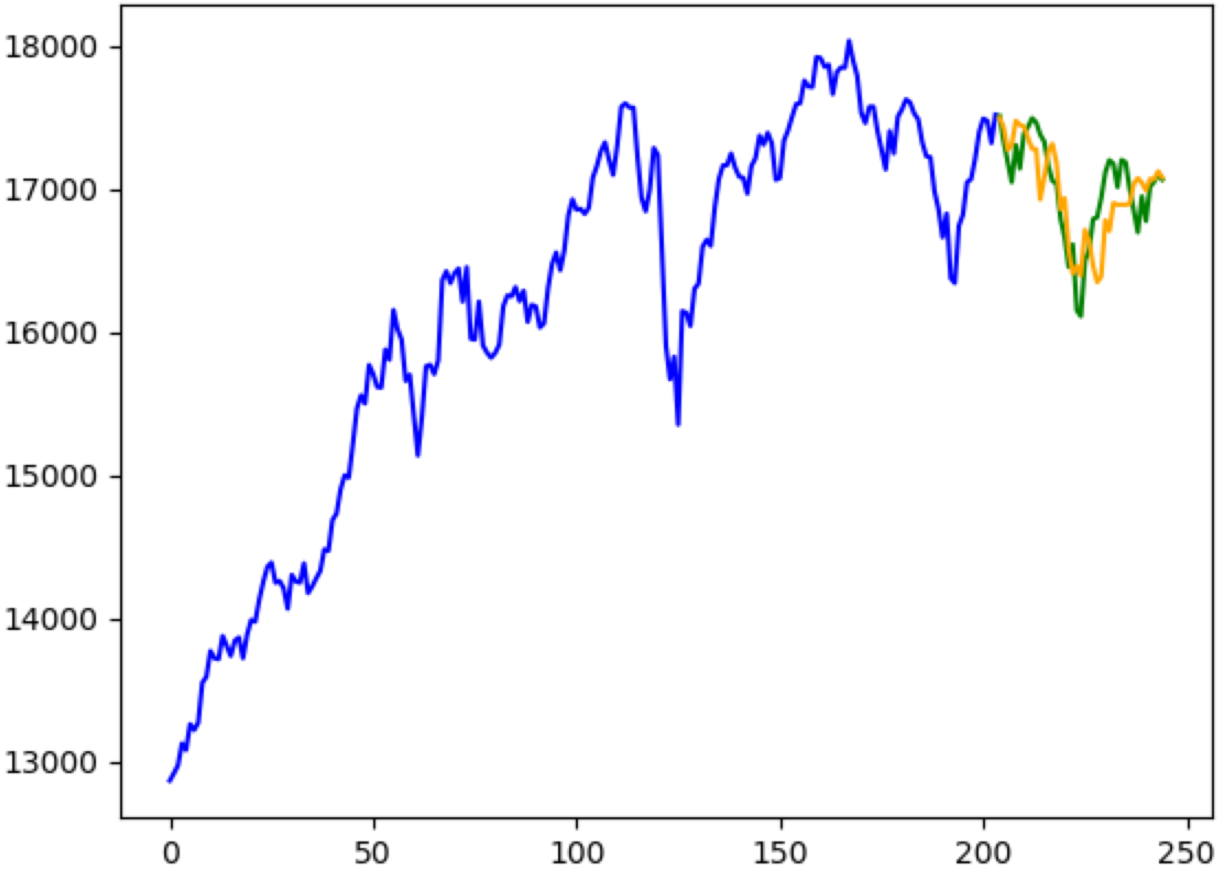
▼ code 請見附檔

Problem 3 ARIMA Forecast

1. What are the ARIMA parameters (p, d, q, P, D, Q, s) that you use?
   And what is the mean square error (MSE) of your forecast?

   - ▼ (p, d, q, P, D, Q, s) —> ((0, 1, 0), (0, 1, 0, 31))
   - ▼ MSE: 61162.46526997664

2. Plot the whole stock (11/04/2020-11/04/2021) and your forecast data (09/06/2021~/11/04/2021) on the same figure. (x-axis: date, y-axis: close value)


18000
17000
16000
15000
14000
13000
0
50
100
150
200
250