科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/5/28

Data Analytics – HW3-1

Predict the stock movement

1. 資料欄位 (data from S&P 500)

train.csv (2009.01.02 to 2017.12.29, 共 2264 筆資料)

- Date **[type: object]**

  資料的日期，可以觀察到並沒有週末的資料，格式: dd-mmm-yyyy

  Ex. “06-Jan-2009”
- Open Price **[type: float64]**

  當日的開盤股價，整體資料最小值為 679.28，最大值為 2692.71

  Ex. “931.17”
- Close Price **[type: float64]**

  當日的收盤股價，整體資料最小值為 676.53，最大值為 2690.16

  Ex. “934.70”
- High Price **[type: float64]**

  當日的最高股價，整體資料最小值為 695.27，最大值為 2694.97

  Ex. “943.85”
- Low Price **[type: float64]**

  當日的最低股價，整體資料最小值為 666.79，最大值為 2685.92

  Ex. “927.28”
- Volume **[type: int64]**

  當日股票總交易量

  Ex. “5392620032”

test.csv (2018.01.02 to 2018.12.31，共 252 筆資料)

- Date **[type: object]**

  資料的日期，可以觀察到並沒有週末的資料，格式: dd-mmm-yyyy

Ex. “04-Jan-2018”

- Open Price **[type: float64]**

  當日的開盤股價，整體資料最小值為 2363.12，最大值為 2936.76

  Ex. “2719.31”

- Close Price **[type: float64]**

  當日的收盤股價，整體資料最小值為 2351.10，最大值為 2930.75

  Ex. “2723.99”

- High Price **[type: float64]**

  當日的最高股價，整體資料最小值為 2410.34，最大值為 2940.91

  Ex. “2729.29”

- Low Price **[type: float64]**

  當日的最低股價，整體資料最小值為 2346.58，最大值為 2927.11

  Ex. “2719.07”

- Volume **[type: int64]**

  當日股票總交易量

  Ex. “2100767744”

2. 資料前處理

◆ 預覽資料並進行資料視覺化

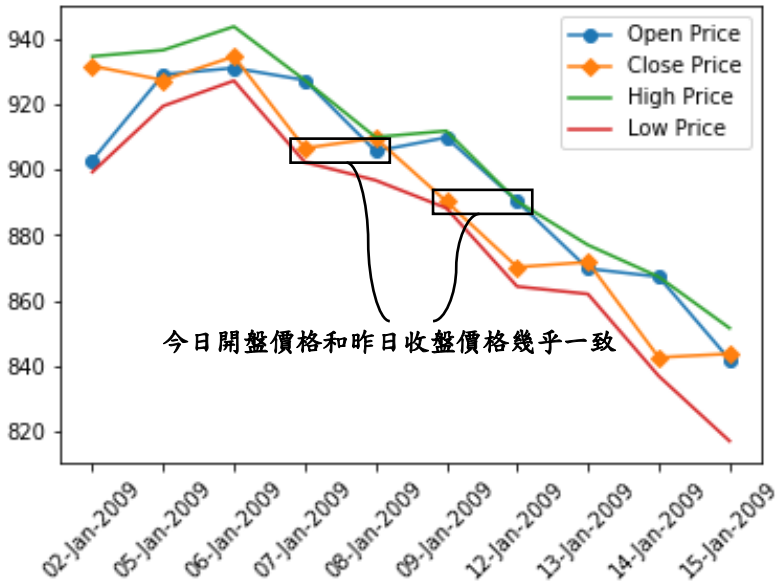
➢ 先利用 head()、info()、describe() 來預覽資料，得知以下訊息:

1. 資料欄位的型別
2. 是否有缺失值
3. 哪些欄位有可能當作特徵值

➢ 接著利用 matplotlib 來將資料視覺化，此時可以更加深入了解資

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/5/28

料欄位間的關聯性，並決定要選擇哪些欄位作為特徵值；從下方折線圖中可以發現今日開盤價格(藍色)和昨日收盤價格(橘色)基本上相去不大，因此可以作為特徵值餵給模型來找出兩者間的關聯性。



◆ **生成驗證測試的答案**

➢ **目標:**

■ 以每日收盤價相對於前一天的收盤價之漲跌作為驗證測試的答案

➢ **RF pred:** 將今日收盤價格與昨日收盤價格相比，若今日大於昨日，即為漲 (1)；今日小於昨日，即為跌 (0)。

◆ **加入特徵值**

➢ 將 Open Price 和 Close Price 作為特徵值。

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/5/28

3. 模型訓練與測試結果

- 三個模型先用 5-fold cross validation 對訓練資料進行測試及驗證，再將全部訓練資料丟進模型進行訓練，分別套用到以下 3 個模型:
  1. Logistic Regression
  2. NN – Multilayer Perceptron Classifier
  3. K Nearest Neighbor Classifier
- 其中 Logistic Regression 在 5-fold cross validation 和丟入全部訓練資料後的表現都是最好的，準確度高的同時確保不會 overfitting。
- 套用至 Testing Set 時，也如預期是 Logistic Regression 表現最好。

4. 結論

- 而之所以 Logistic Regression 在這次的預測中表現較另外兩種亮眼，可能取決於兩個點:
  1. 特徵值選得好
  2. 需要預測的結果比較單純，只有漲或跌
- Logistic Regression 在進行這種二分式的分類問題，實際上都能有不錯的效果，因此它的準確度高也是預料中的結果；但如果需要進行三種以上的分類問題，效果可能就不是那麼顯著了。
- 如何改良 Classifiers?
  - 選用不同的特徵值

科系: 資訊系
姓名: 林啟允
學號: C44056223
最後修改日期: 2020/5/28

剛開始只有 Close Price 作為特徵值時，Logistic Regression 的在 5-fold cross validation 中的準確度只有 54%左右，但加入 Open Price 後，準確度就能提升到 93%。

- 嘗試改變超參數的值

  K Nearest Neighbor Classifier 在 n_neighbors 小於 3 的時候，會有 overfitting 的問題產生，建議使用 default 的 5，雖準確度略為下降，但 overfitting 的問題就不會那麼明顯。