* implement **3 classifiers** to predict the stock movement

1. Logistic Regression

2. Neural Network

3. Random Forest Classifier

* How did you preprocess this dataset ?

csv檔內：將每日收盤價減去前一日收盤價，創建收盤價漲跌的欄位（漲：1，跌：-1，不漲不跌：0），作為預測的目標。

程式內：以眾數填補收盤價漲跌欄位的一個缺失值，再創建兩個欄位，分別是開盤價減收盤價的值和最高價減最低價的值。

* Which classifier reaches the highest classification accuracy in this dataset ?

|  |  |  |
| --- | --- | --- |
|  | train accuracy | test accuracy |
| Logistic Regression | 0.546378 | 0.5238 |
| Neural Network | 0.9986749 | 0.996 |
| Random Forest Classifier | 0.947438 | 0.829365 |

因此是Neural Network。

◎Why ?

類神經網路有較良好的推廣性和適應性，較能應對股市的動盪。

隨機森林在訓練模型時可能有overfitting的現象，導致測試時accuracy下降。

Logistic Regression是迴歸分析，可能較不適合用於分類的問題。

◎Can this result remain if the dataset is different ?

為此，將原先的訓練集與測試集對調，用於F74066365\_hw3-1\_for\_another\_dataset .ipynb，出現有趣的結果：

|  |  |  |
| --- | --- | --- |
|  | train accuracy | test accuracy |
| Logistic Regression | 0.5238 | 0.546378 |
| Neural Network | 0.003968254 | 0.001325 |
| Random Forest Classifier | 0.88095238 | 0.881183745583 |

Logistic Regression的兩個accuracy為原先的對調。

類神經網路的兩個accuracy從原本趨近1變成趨近0。

隨機森林沒了overfitting的現象，兩個accuracy相近。

因此變成是Random Forest Classifier有最高的accuracy。

因此若資料集不同，有最高accuracy的模型無法維持相同。

* How did you improve your classifiers ?

改變輸入的特徵與調整模型的超參數。