FDA HW 3-1

S&P 500 股市預測

系級:統計四
姓名:陳宥任

資料介紹

資料來源:https://www.sharecast.com/index/SP_500/prices/download

資料區間:

Training data : 02-Jan-2009 to 29-Dec-2017，共2264筆

Testing data : 02-Jan-2018 to 31-Dec- 2018，共252筆

資料變數:

Date(日期) Open Price(開盤價) Close Price(收盤價)
High Price(最高價) Low Price(最低價) Volume(成交量)

預測方法及結果

建立預測目標y: 以四天前後的收盤價高低為準則，若四天後收盤價高於四天前收盤價則為1，反之則為0

切割資料: 建立完y後會把資料內的最後四筆資料去除，原因為在train資料內最後四筆是沒有對象可以比較的，因此結果都為0，而在test資料則是不知道預測出來的結果是否正確，因此也去除

建立訓練資及測試資料: 在train_x及test_x中，會把預測目標y以及Date變數去除。 在train_y及test_y中，只留有預測目標y

預測方法及結果

Logistic Regression

準確率為 0.5645161290322581

這邊沒有進行模型的調整讓準確率更佳。原因為此預測法為一個凸優化模型，要調的參數不多。上網檢視過許多人使用logistic regression，參數都不會更改。

預測方法及結果

Neural Network

準確率為(後者) `test accuracy: [0.6862675855236668, 0.5645161271095276]`

很奇怪的事情是這邊NN與Logistic Regression的準確率完全一樣，這邊我找不出來原因

在NN模型中參數我也沒有進行調整，因為對於NN的概念及操作我不太熟悉，不知道該使用甚麼方式才能找到最好的調整方法

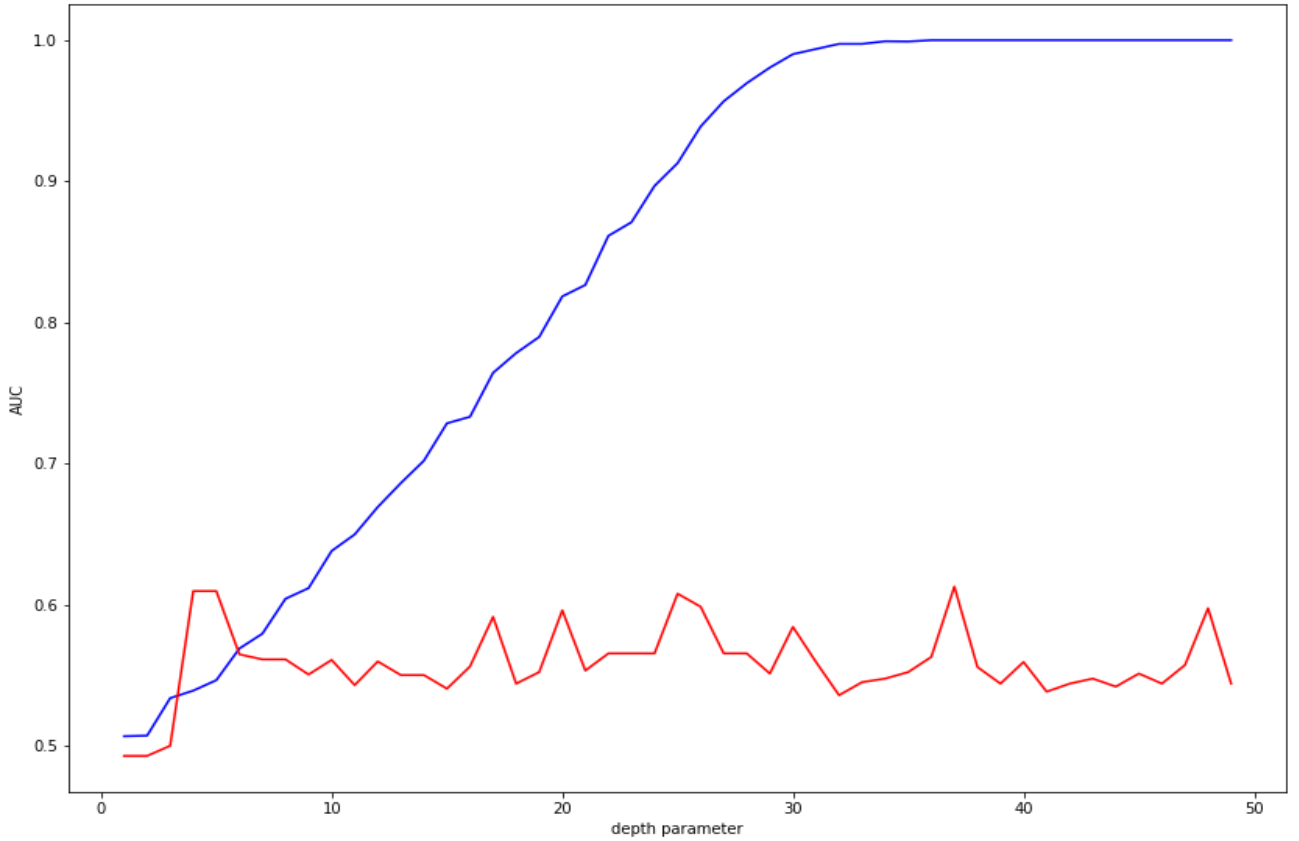
預測方法及結果

Decision Tree Classifier

準確率為 0.4959677419354839

使用ROC Curve, AUC
調整max_depth參數

將max_depth以1-50畫出ROC, AUC，可以看到當test data(紅線)的depth parameter約為4-6時，AUC為最高，這邊選擇使用4來調整參數。



預測方法及結果

Decision Tree Classifier

調整參數後準確率為
(max_depth = 4)

```
0.592741935483871
```

從以上結果可以看到，將max_depth調整到AUC為最高的結果，
得出來的準確率較沒有調整參數的準確率高了約10%左右

結論與QA

Q: How did you preprocess this dataset?

A:從資料探索中可以得知資料並不需要進行預處理，主要有幾個原因:

1.資料並為缺失值
2.資料並無類別數據需要處理
3.資料單位相同不需要進行標準化

Q: Which classifier reaches the highest classification accuracy in this dataset?
A: Why? Can this result remain if the dataset is different?

這邊使用了Logistic Regression、Neural Network和Decision Tree，accuracy分別為0.5645、0.5645及0.5927，以Decision Tree的分類稍微準確一些。主要是因為原先的Decision Tree只有0.5的accuracy，我稍微調整了參數中的max_depth使他的accuracy提高一些

如果是一樣從S&P500爬下其他時間的資料，我認為預測結果會是差不多的，因為並不需要進行太多的預處理，目標的定義差別也不會太大，得出來的結果應該類似

結論與QA

Q: How did you improve your classifiers?

A: 這邊我只針對Decision Tree進行調整參數讓他的預測效果更準確。主要是因為在Logistic Regression中，要調整的參數不多，這邊就沒有動任何參數

再來是Neural Network，因為是第一次接觸及使用，對NN也沒有很熟悉，不太曉得該怎麼動參數，這邊就只使用網路上一般人最常使用的層數以及一些參數設定，沒有做更改使模型更加，所以這裡屬於能力上的不足

最後Decision Tree是根據了ROC curve和AUC來檢視max_depth為多少的時候，他的結果會是最棒的。這邊讓他從1-50去畫出AUC最後的結果，可以看到當在Test data中AUC差不多為4~6的時候，他的AUC會是最高的，所以最後選擇4作為我們max_depth參數的調整