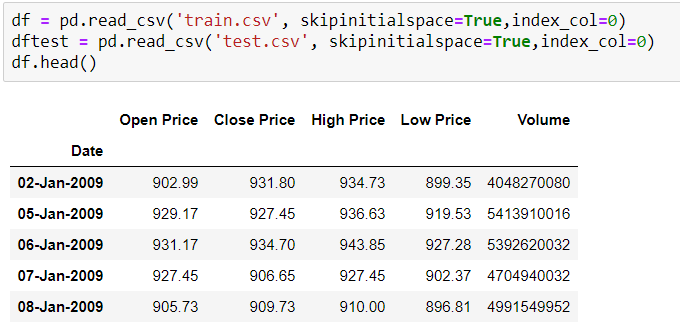
資料分析與學習基石HW3

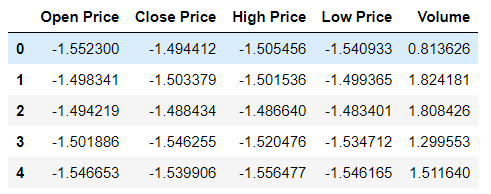
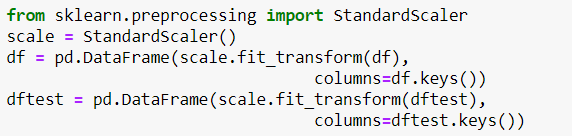
Discussion:

1. **How did you preprocess the data?**

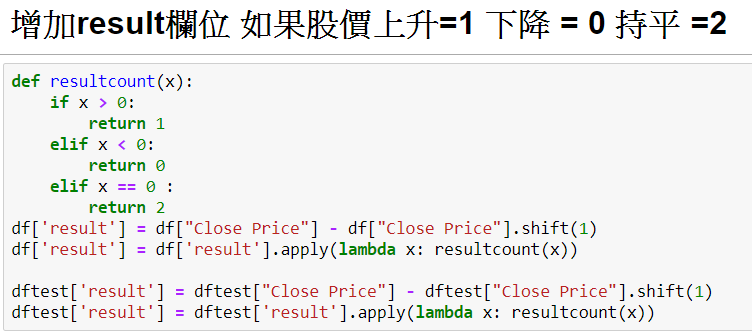
將日期行當成index，以利後續操作



將資料展現出來，之後將各列數據進行標準化



接著利用function 將上漲或下跌加在新欄位Result



為了增加一些變數 我上網搜尋 很多人做的時候都會新增close\_to\_open close\_to\_high close\_to\_low等等數據，能找出比較明確close price與其他價格間的關係



我會再去測試有無這組數據與準確率間的關係

像在Logistic Regression內 不加的準確率比較高

在SVM內加的準確率比較高

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

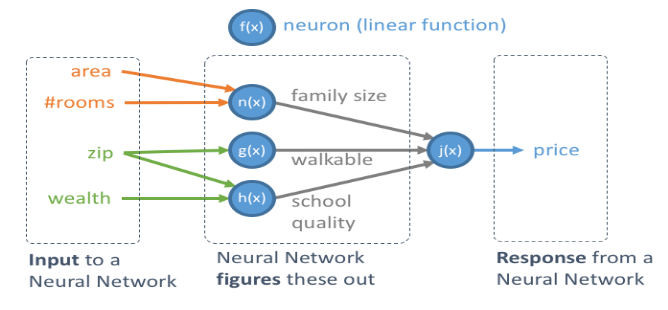
根據我實測結果 使用Logistic Regression準確率比較高

LR(81.7%) VS SVM(80%) VS NN(67%)

原因可能是Logistic Regression 與SVM不同的點是 LR不用像SVM要找最大距離，而SVM對於異常數據比較敏感，因為training時只需要找支持向量，有效樣本本來就不高，一旦被干擾，預測結果就會不如預期，而在本例子中，可能股票的異常數據較小，比較呈線性發展，所以與LR相差不會太大

補充: 我覺得Logistic Regression在加入更多特徵時期準確率變低是在於，LR不擅長處理大的特徵空間，而SVM反之，所以在加入更多特徵時準確率提升，當然在超過某個特徵空間數時準確率下降

而Neural Networks準確率不高是因為這個例子裡資料不夠多，而且這個例子的變數環環相扣，用NN會讓他們分散train，所以我覺得比較適合用在鐵達尼號跟房價預測這種例子。



**Can this result remain if the dataset is different?**

1. 有可能

因為如果去觀察HW3 train 跟 test的數據 ，會發現test的股價相較於train高非常多，train從900開始 但是test高到2000多，如果test的數據能比較與train一樣分布比較平均，結果可能會不同，但是應該不會差太多。

從例子來看看起來像拿一個尺度大的model去跑極端值(有點誇飾)

但如果加入其他變因，非股價類，結果應該會有不同。

1. **How did you improve your classifiers?**

就嘗試更改model裡的參數而已，慢慢試。然後對資料做標準化