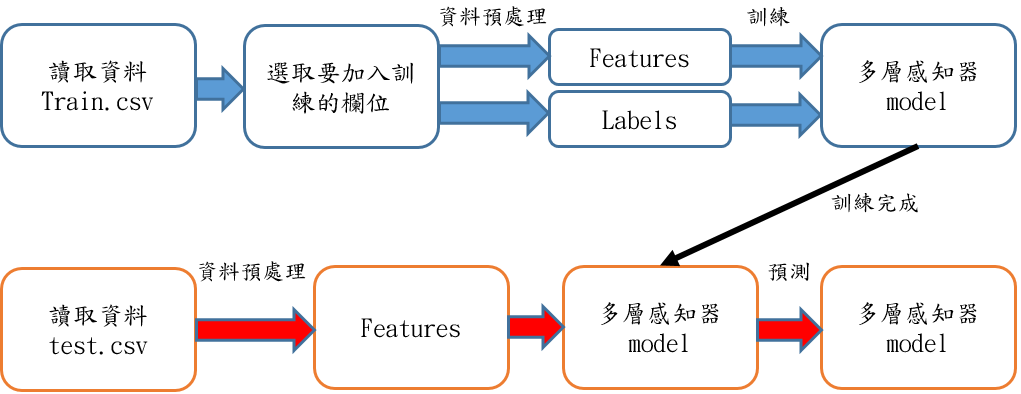
Machine-learning HW1

regression

109368085 臧英宏

電子碩一甲

1. **程式方塊圖**

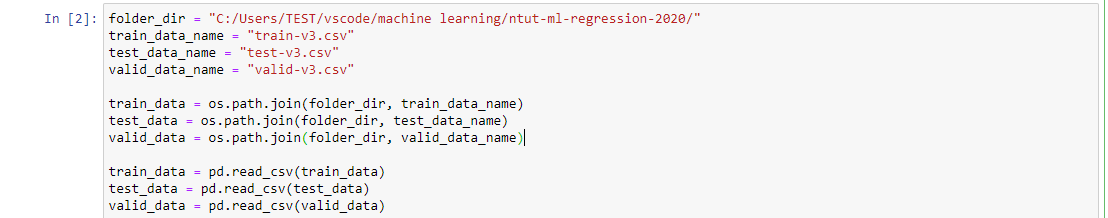


1. **程式碼寫法、作法說明**

匯入所需模組、sklearn的資料預處理模組、關係矩陣判斷模組



輸入訓練測試驗證三筆資料

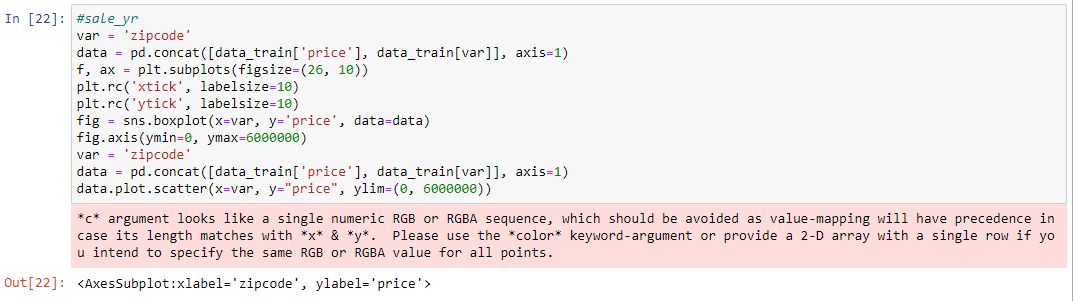


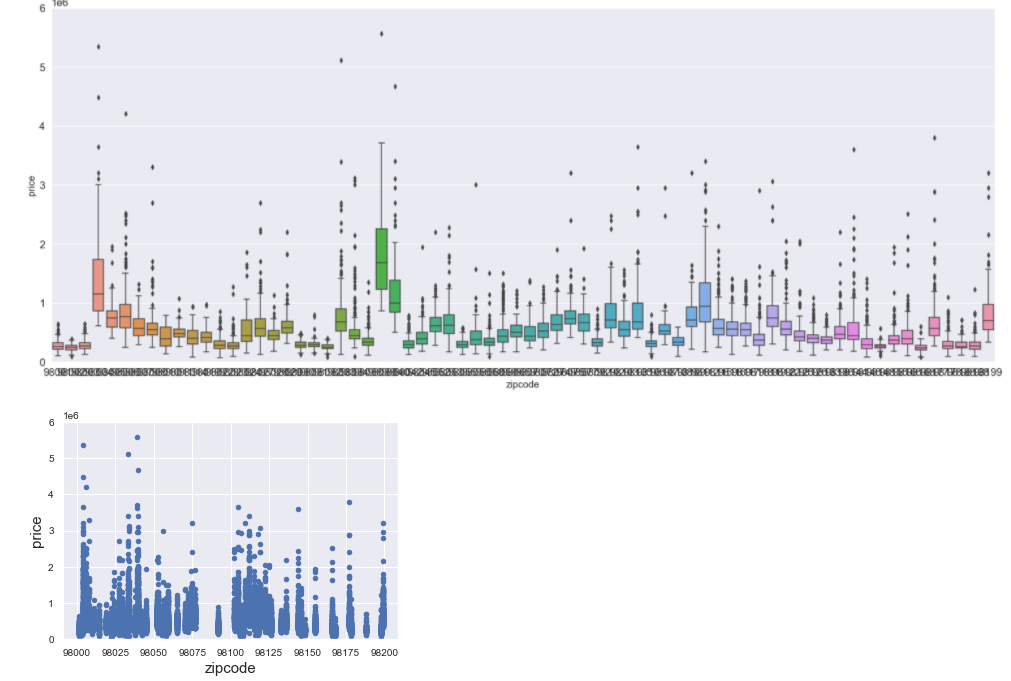
為了有更多的資料源，將驗證資料加入訓練資料內

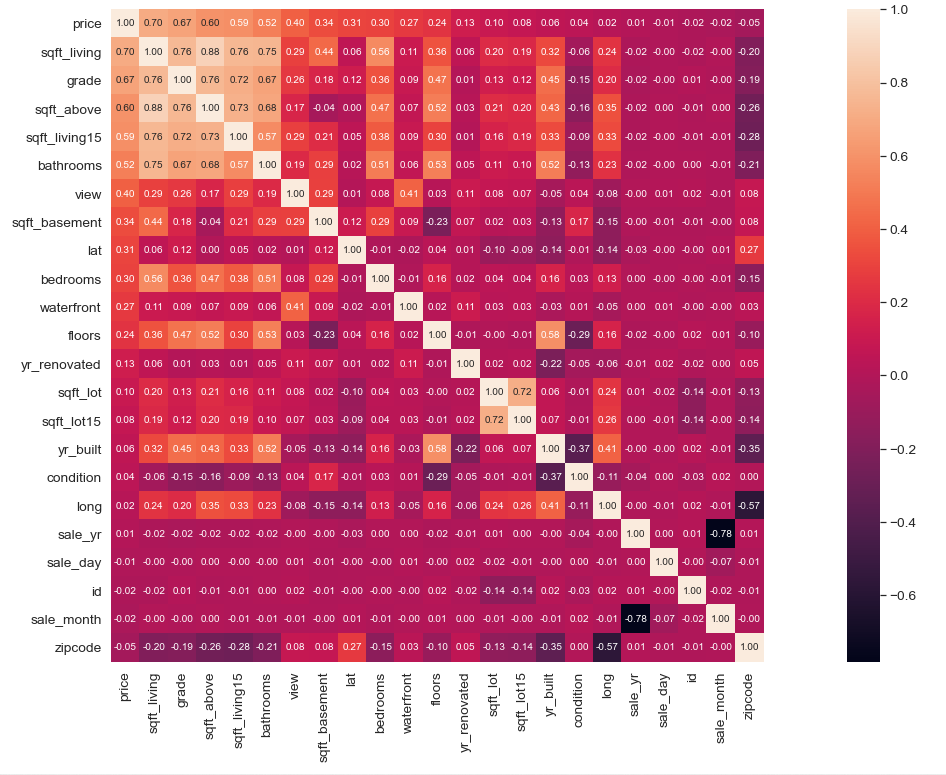
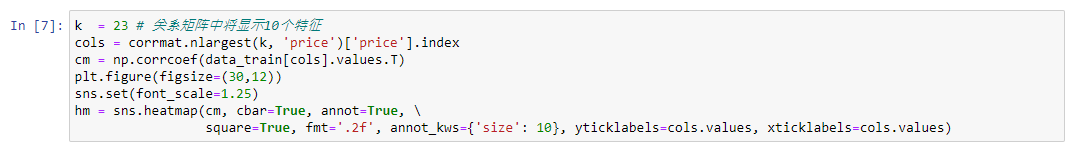


選取要加入訓練的資料欄位至dataframe

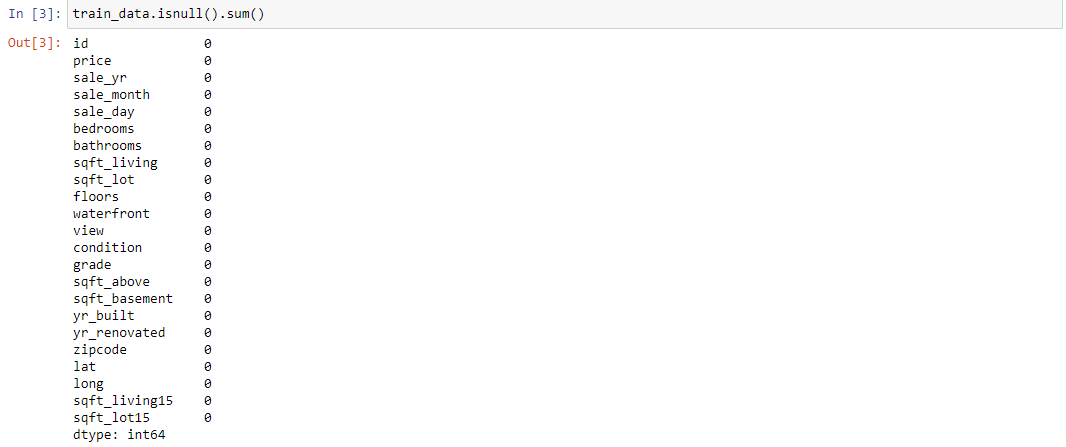
分析資料之分部狀態，透過盒狀圖與點狀圖分析各元素與價格的關係



盒狀圖與點狀圖

產生關係矩陣進一步驗證

檢測缺漏項，發現沒有。



將train\_data存成陣列的形式，並分出Features和Label兩列。

將Features作正規化



建立keras Sequential模型，為一線性堆疊模型



使用model.add方法，建立輸出層與隱藏層1

輸出層是「隱藏層1」共1000個神經元

輸入層為22個神經元輸入，即為22個特徵欄位

輸出層是「隱藏層2」共有1000個神經元

輸出層是「隱藏層3」共有1000個神經元

Dropout為每輪疊代時，每10個輸入值就會被隨機拋棄5個。

用於避免overfitting，讓學習效果更好。

建立輸出層，使用normal分布，定義激活函數linear

定義訓練方式

設定損失函數為MAE，使用adam為優化方法，提高準確率。

開始訓練

設定validation\_split = 0.15，將85%做為訓練資料，15%作為驗證資料。

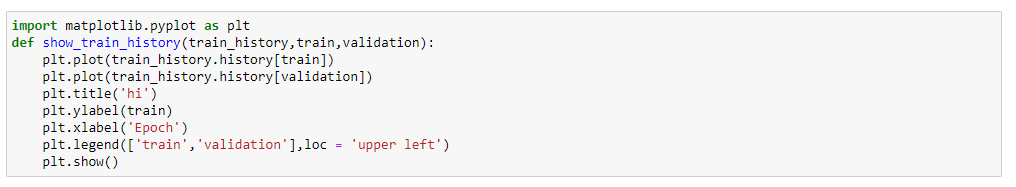
Epochs = 550，執行550次訓練週期

Batch\_size = 20，每一次批20筆資料

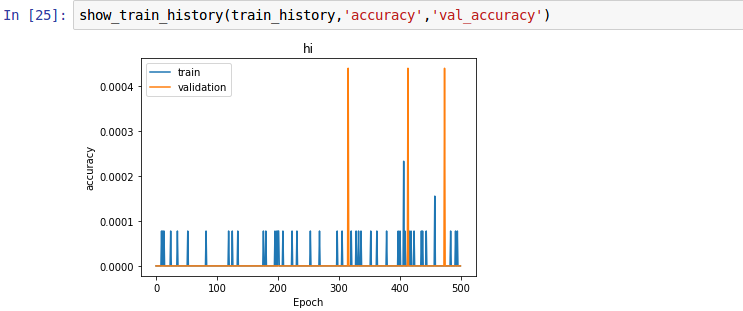


匯入matplotlib.pyplot模組

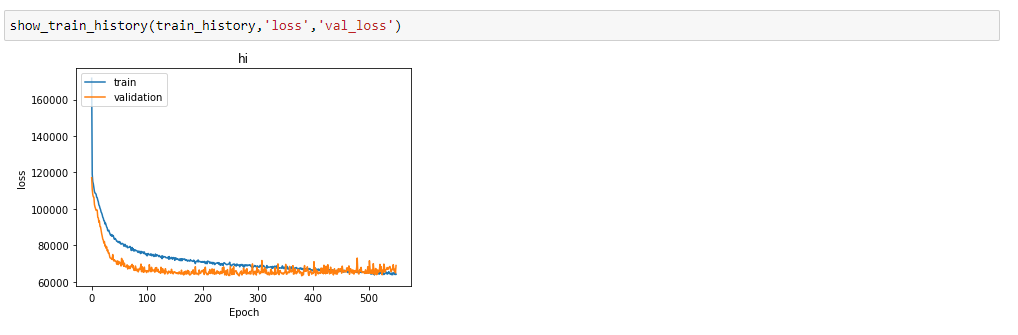
輸入參數(訓練與驗證資料的執行結果)



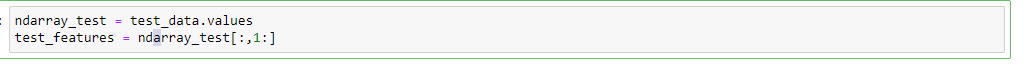
畫出accuracy執行結果



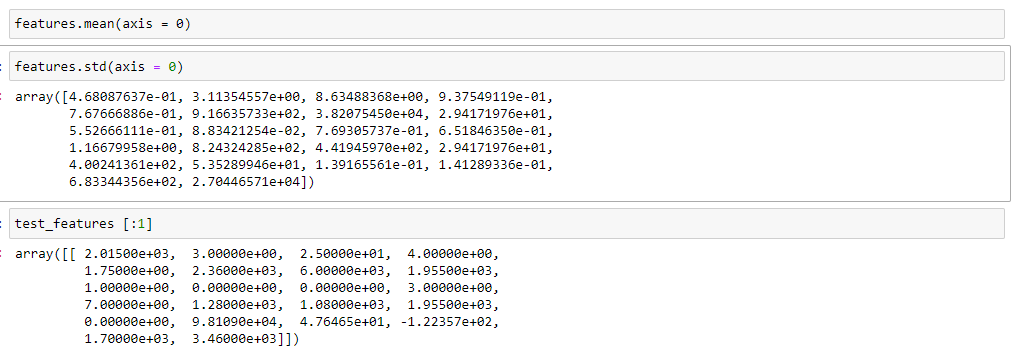
畫出loss誤差執行結果



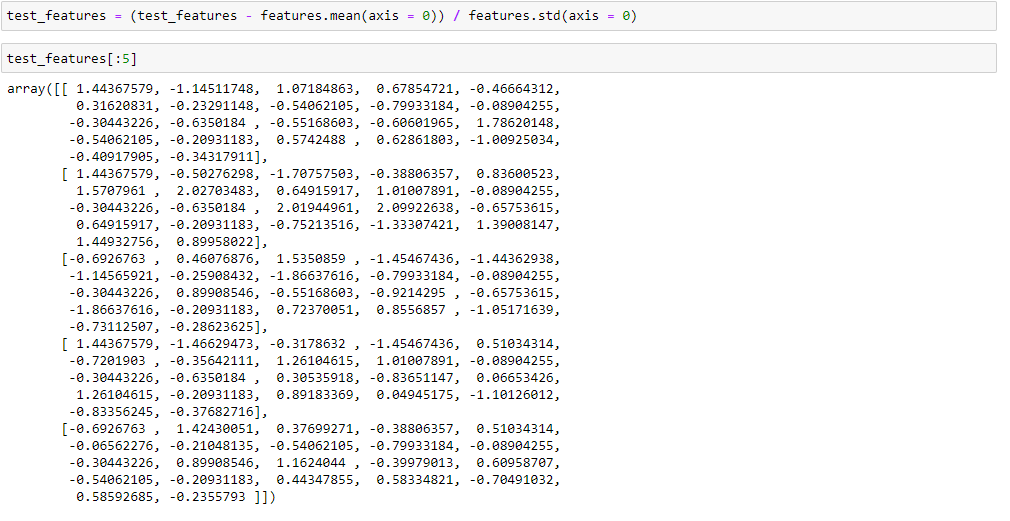
將test\_data存成陣列的形式



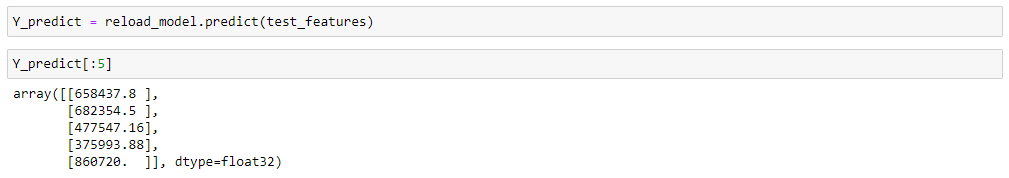
算出Features平均及Features標準差



將test\_Features作正規化 ()



將test\_Features丟到model裡面做預測，輸出為Y\_predict

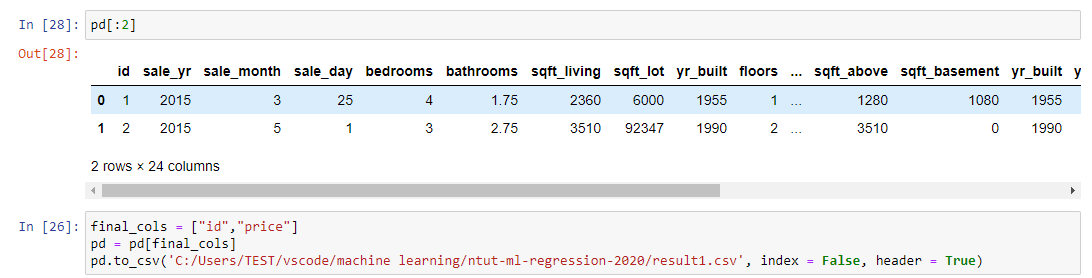


將array資料轉換成DataFrame形式

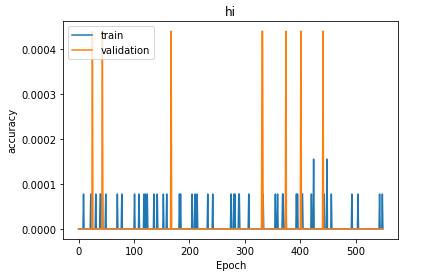
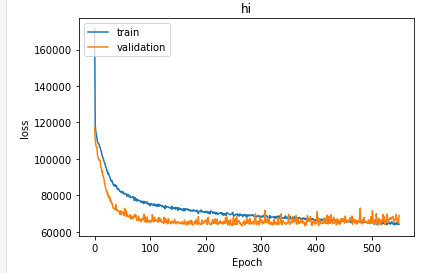


將預測完成之price加入test\_data

再擷取id及price製作成csv檔



1. 畫圖做結果分析



loss執行結果

accuracy執行結果

1. **討論預測值誤差很大的，是怎麼回事？**

模型可能出現的誤差來自兩個主要來源:

1. 因模型無法表示基本資料的複雜度而造成的偏差（bias）

雖然擁有足夠的資料量，但因為複雜度不夠，而出現偏差

故模型一直判斷錯誤，造成資料具大誤差。

1. 因模型對訓練它所用的有限資料過度敏感而造成的方差（variance）

通常出現在模型過於複雜或我們沒有足夠的資料支援它時

1. 資料的屬性分布可能不適合直接做訓練，須對資料做預處理的動作，如對資料取對數或將資料做轉換，若有缺漏項利用平均數補足缺漏項等..
2. **如何改進？**
3. 例如資料參數過多，訓練樣本太少時，Dropout就是每次訓練按概率拿走一部分神經元，只在訓練時使用，解決過度擬合的問題。
4. 使用keras內建函數包含Earlystop()與儲存優良模型加以訓練來減少訓練上的時間浪費。