資科系 110916015 林翡

機器學習HW1\_Regression Analysis

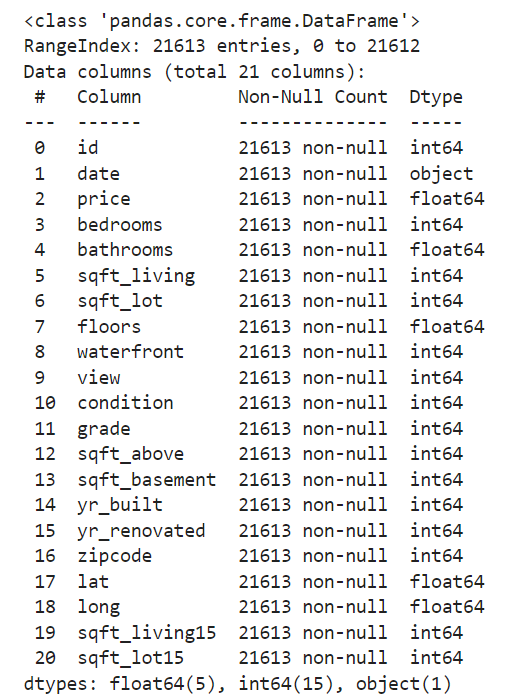
程式碼語言:Python(在Jyputer Lab上寫)

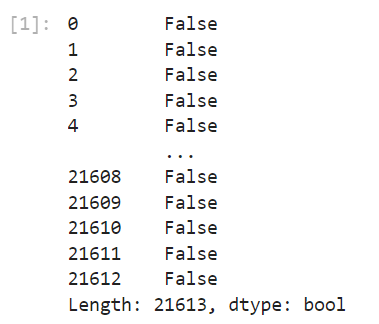
內容:

Part I:

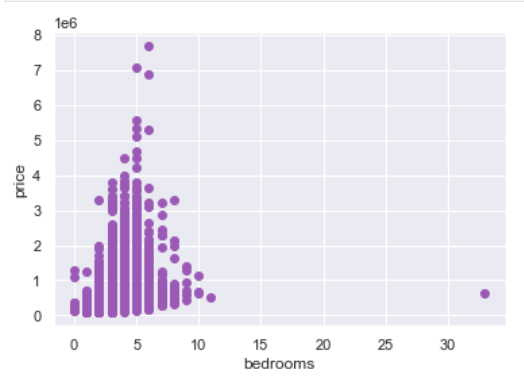
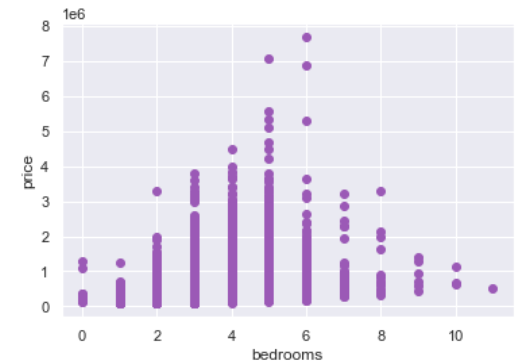
1. 資料清理與視覺化圖表

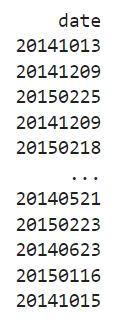
* 資料清理

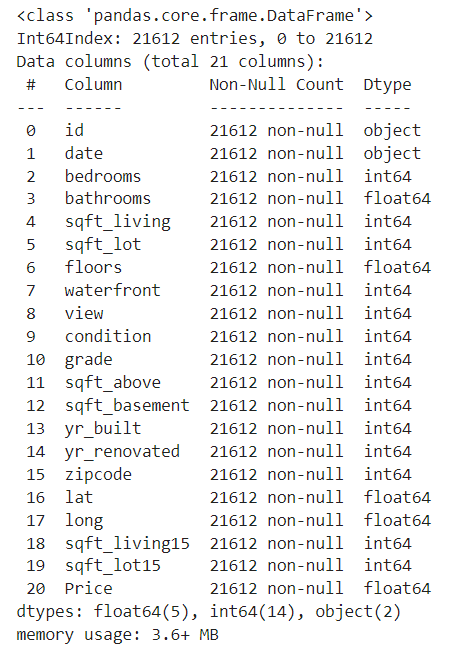
1. 先檢查資料集有無遺漏值，由圖中可看出沒有遺漏值。 (程式碼[1])
2. 檢查資料有無重複值，圖中可看出此資料集無重複值。 (程式碼[1])



1. 用bedrooms-price的散點圖發現有一個離群值，房間數有33間，但房價卻異常低，因此，我要把這筆離群值刪除。 (程式碼[5]~程式碼[6])

 刪除前的散點圖 刪除後的散點圖

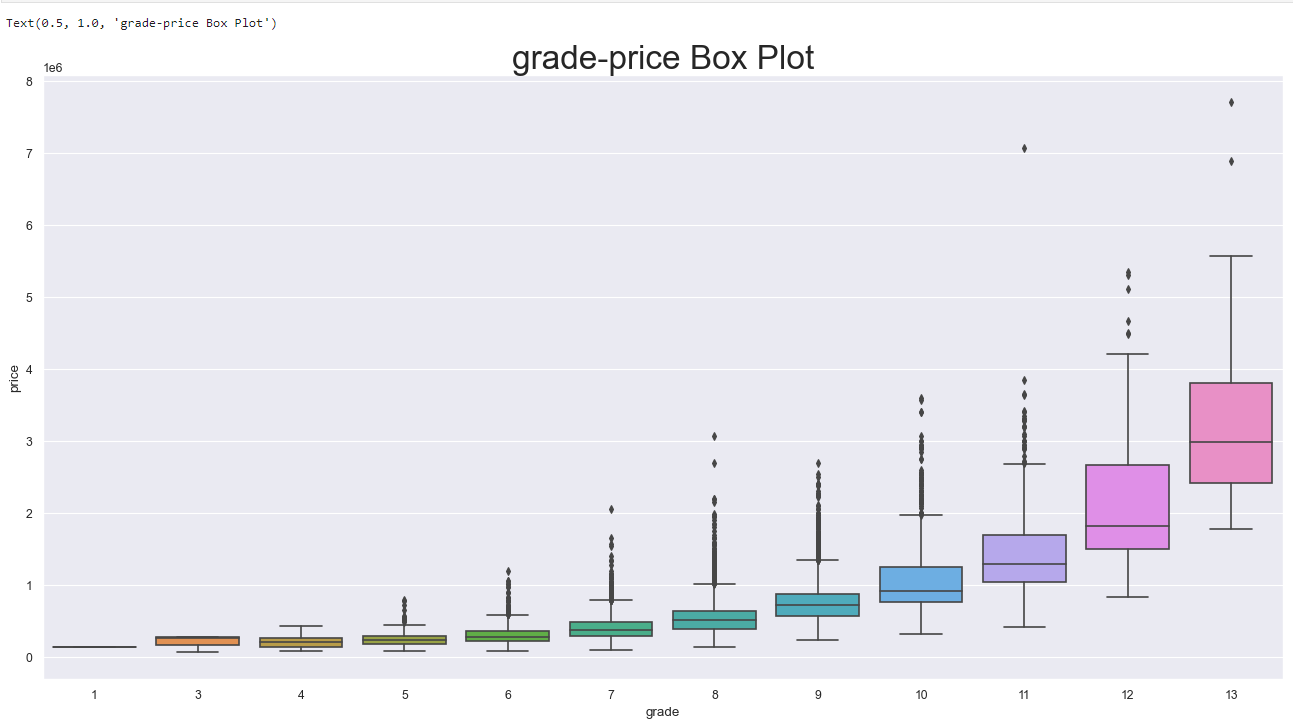
1. 刪除每筆”date”後面的T000000，感覺跟日期無關。(程式碼[7])
2. 把price移到最後，改名為Price。(程式碼[9]~程式碼[13])



* 視覺化圖表 (程式碼[3]~程式碼[4])

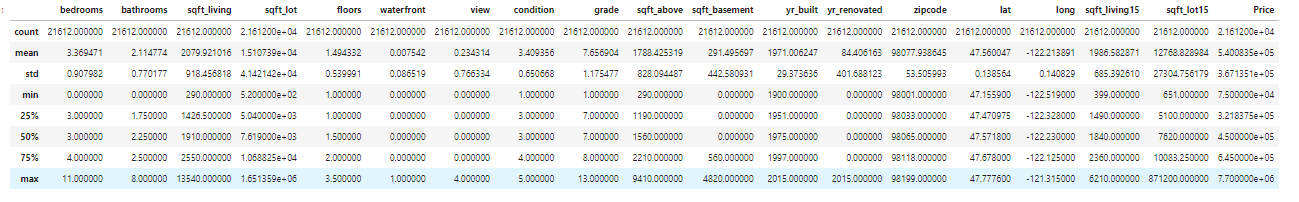
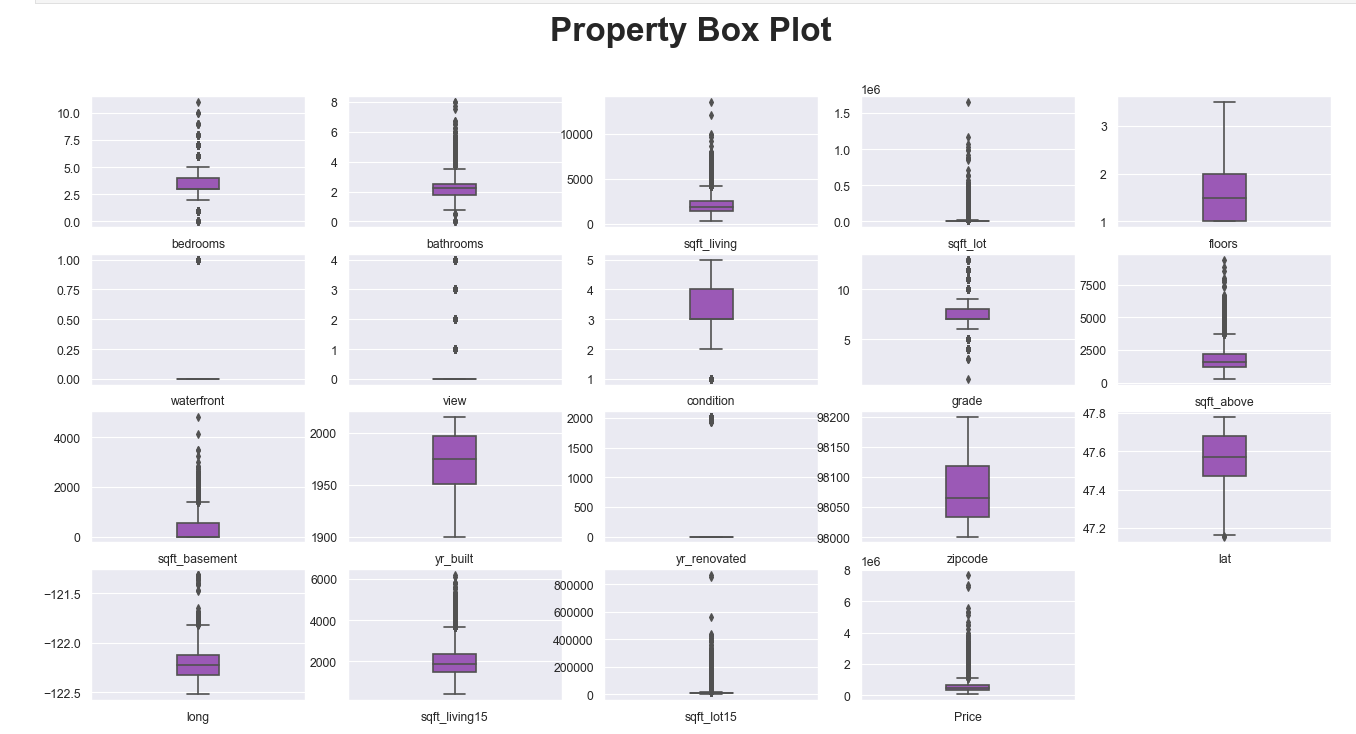
 Price的直方圖

* 說明:從上方的直方圖可以發現房價大約在300000~600000這個區間較多

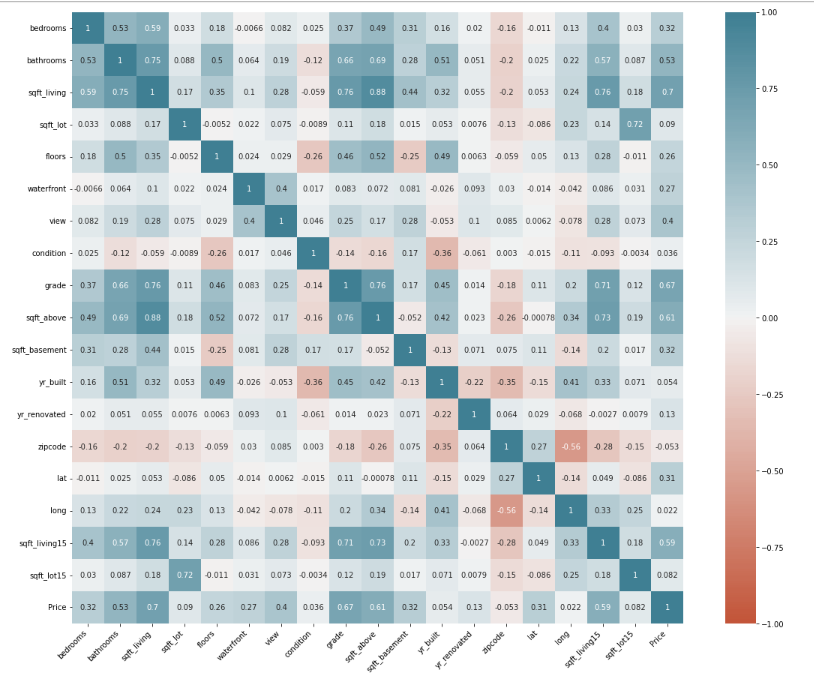
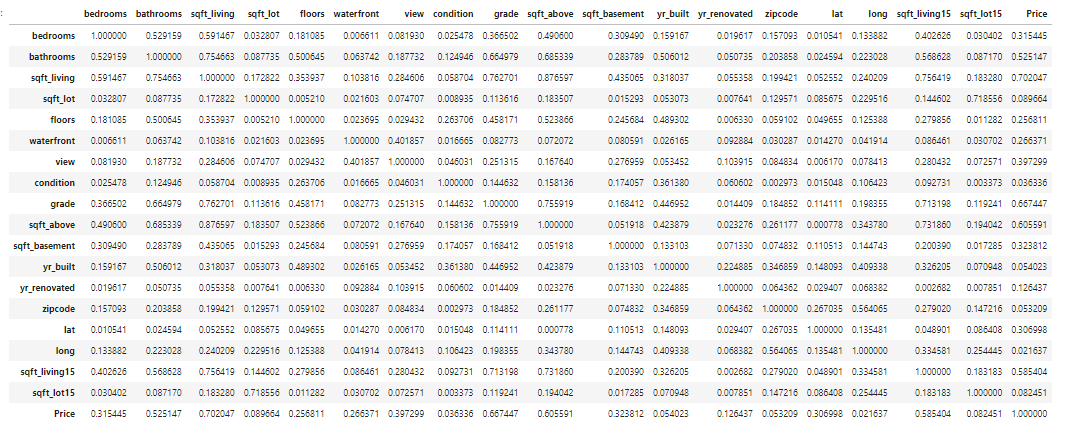
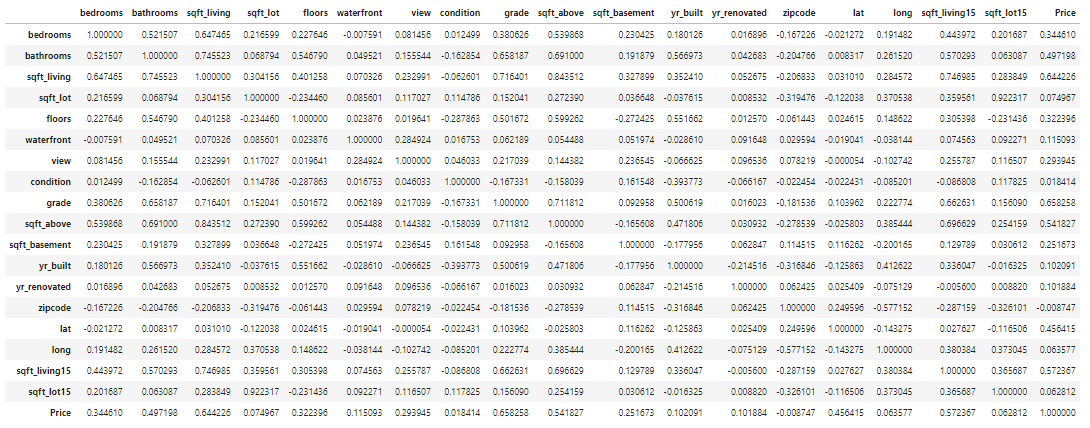
 grade-price盒鬚圖

* 說明:從上方的盒鬚圖可以發現房價會隨著grade的增加而增加

1. 敘述性統計分析

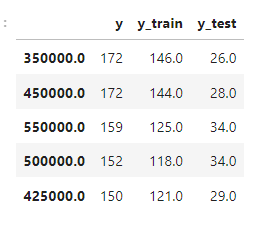
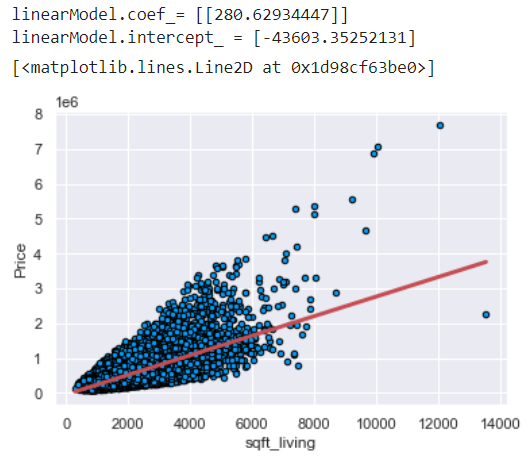
* 敘述性統計分析表 (程式碼[15]):計算各個欄位的平均值、標準差、最小值、四分位數、中位數、最大值。
* 敘述性統計分析圖(程式碼[16])

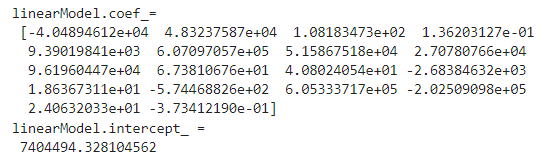
三、特徵相關性分析

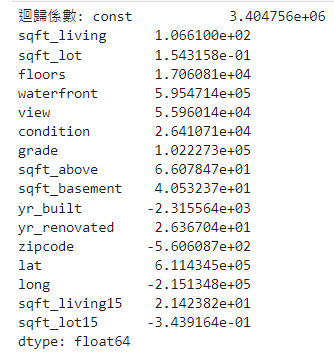
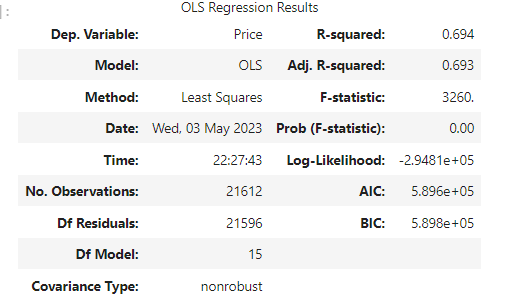
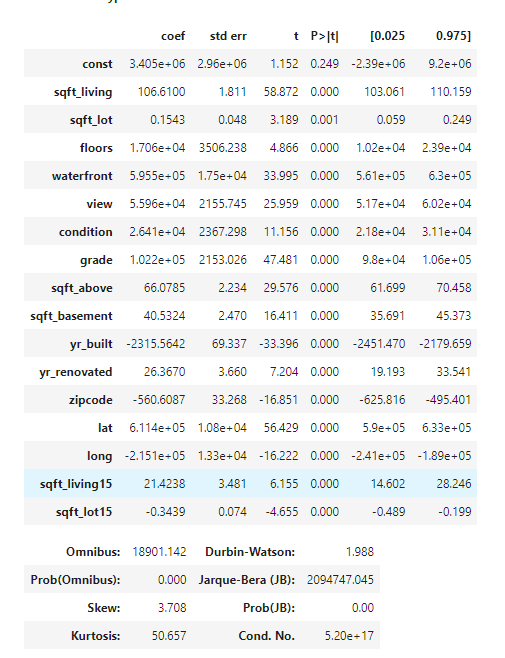
* 相關係數熱度圖 (程式碼[17])
* 皮爾森相關係數 (程式碼[18]):計算各個欄位之間是否有線性相關。
* 斯皮爾曼等級相關係數 (程式碼[19]):計算時會先排序原始數據，計算過程用觀察值的等級取代真實數值。
* 肯德爾等級相關係數 (程式碼[20]) :計算一致配對的比例。

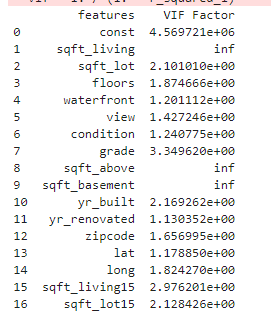
Part II:

1. 資料分割與建置迴歸模型 (線性 or 非線性)

* 資料分割 (程式碼[21]):將原始資料的80%分為訓練集(X\_train、y\_train)，20%分為測試集(X\_test、y\_test)，特徵選擇bedrooms到sqft\_lot15，目標類別Price，最後印出資料集、訓練和測試集的類別比例。
* 簡單線性迴歸 (程式碼[22]):這邊我選擇sqft\_living為自變數、Price為應變數建立迴歸模型。
* 多元線性迴歸 (程式碼[23]~程式碼[25]):因為房子資料集的特徵超過一個，跟房價有關的變數也不是只有sqft\_living而已，所以需要多個自變數放到迴歸模型中。

 以原始資料進行線性擬合 (程式碼[23])

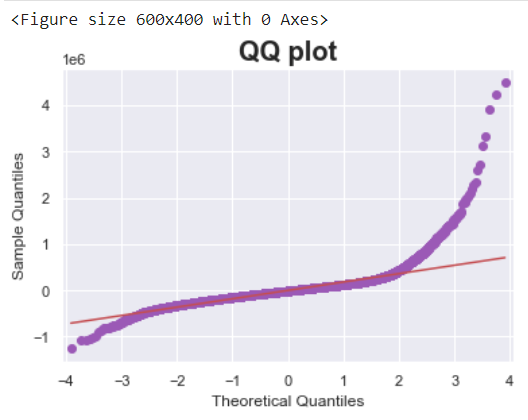
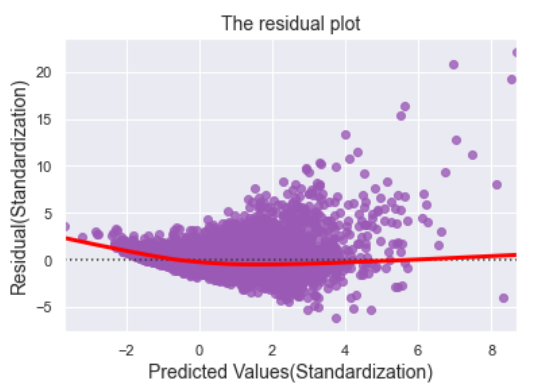
 OLS迴歸分析 (程式碼[24]):對資料集的多個自變數進行迴歸分析，輸出迴歸模型的結果。

當增加多個自變數時要注意多元共線性問題，這邊我使用VIF偵測共線性。由結果可發現sqft\_living、sqft\_above、sqft\_basement有極大的VIF，代表它們可能有嚴重的多元共線性問題。(程式碼[25])

1. 模型效能評估

* 評估指標:MSE、R2、MAE (程式碼[26]~程式碼[28])
* 殘差分析

SW test (程式碼[29]) KS test (程式碼[30])

 QQ圖 (程式碼[31]) 殘差圖(程式碼[33])

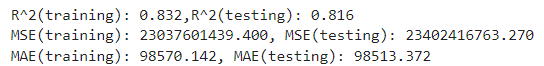
 Durbin-Waston檢測獨立性(程式碼[32])

1. 預測結果分析:從訓練和測試的MSE來看，模型的預測能力不是很理想，MSE過大；但R2都沒有太低，表示模型的擬合程度尚可，可能數據的有些特徵影響了模型的預測能力，畢竟我還沒做標準化、正規化那些。然後可能也有跟上面提到的共線性問題有關，sqft\_living、sqft\_above、sqft\_basement的VIF都極大，表示它們之間有嚴重的共線性，導致模型預測能力不佳。

Part III:

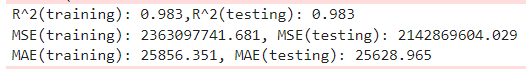
1. 非線性迴歸模型 or 特徵變數變換

* 二階多項式迴歸 (程式碼[34])

 R2、MSE、MAE值

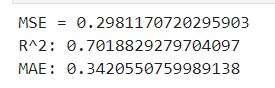
 殘差圖

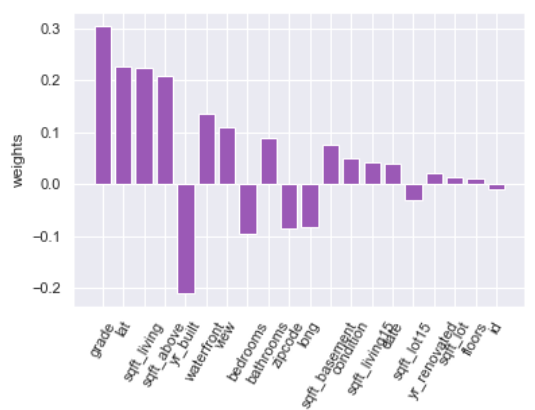
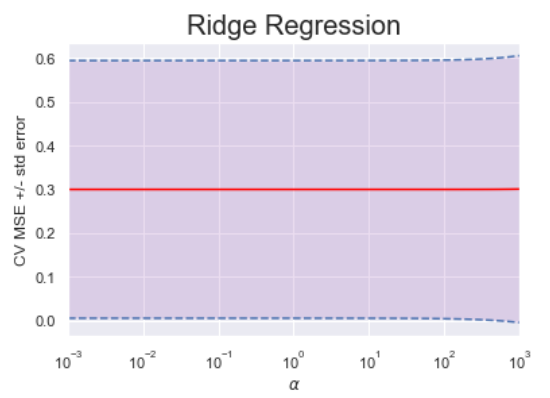
* 隨機森林 (程式碼[35])

 R2、MSE、MAE值

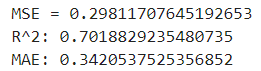
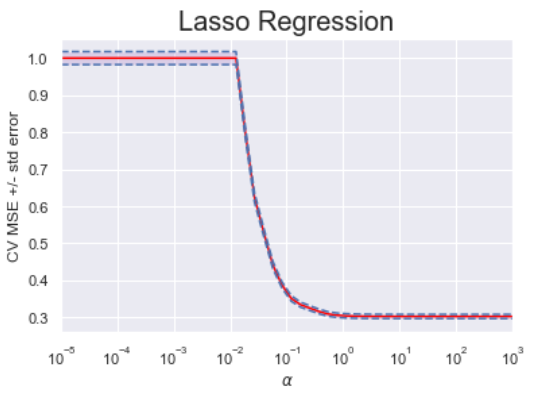
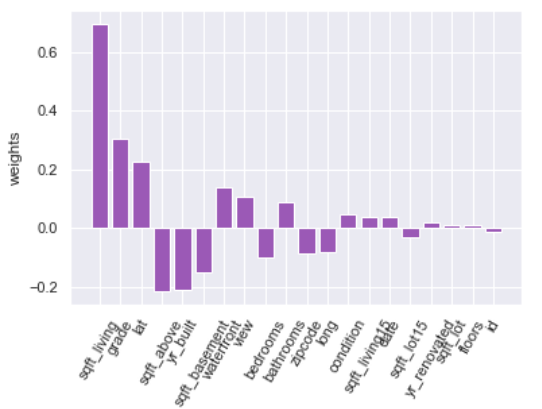
 殘差圖

1. 迴歸模型之正規化:特徵選擇bedrooms到sqft\_lot15。(這邊我是套用課本的範例)

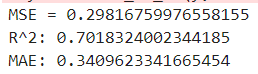
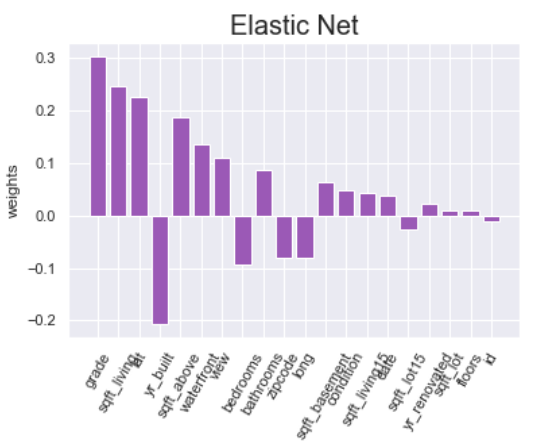
* 脊迴歸 (程式碼[37]~程式碼[41])

 最佳 α值及MSE、R2、MAE

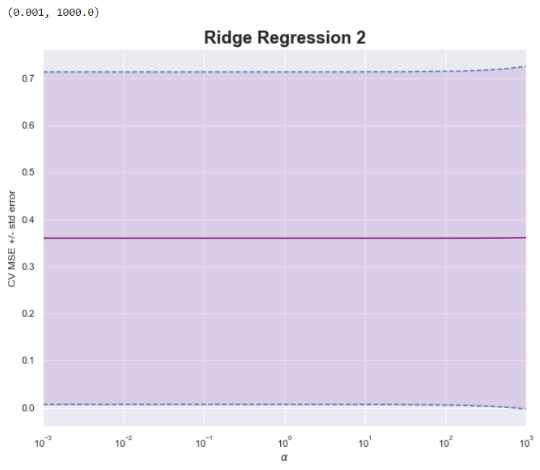
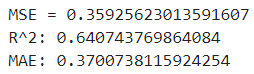
* 套索迴歸 (程式碼[42]~程式碼[44])

 最佳α值及MSE、R2、MAE

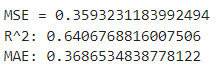
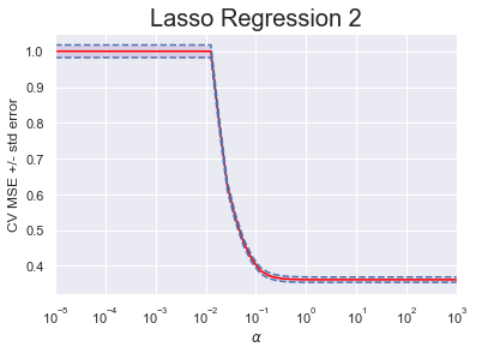
* 彈性網 (程式碼[47]~程式碼[48])

 MSE、R2、MAE

* 選前五個相關特徵重新做脊回歸及套索迴歸:由於前面在做脊迴歸時是選擇bedrooms到sqft\_lot15的特徵，可能有包含了很多不重要的特徵造成alpha值有點大，所以接下來我是是看選前五個重要特徵:grade、sqft\_living、yr\_built、lat、sqft\_above。
* 脊迴歸 (程式碼[49]~程式碼[52])

 最佳α值及MSE、R2、MAE

* 套所迴歸 (程式碼[53]~程式碼[55])

 最佳α值及MSE、R2、MAE

1. 相關模型結果比較分析:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 多元線性迴歸 | 二階多項式迴歸 | 隨機森林 |
| R2(training) | 0.706 | 0.832 | 0.983 |
| R2(testing) | 0.677 | 0.816 | 0.983 |
| MSE(training) | 39086196390.909 | 23037601439.400 | 2363097741.681 |
| MSE(testing) | 45756309217.032 | 23402416763.270 | 2142869604.029 |
| MAE(training) | 125041.012 | 98570.142 | 25856.351 |
| MAE(testing) | 128672.272 | 98513.372 | 25628.965 |
| Shapiro-Wilk | 0.793 |
| Kolmogorov-Smirnov | 0.530 |
| 是否為常態分布 | 否 |
| Durbin-Watson | 1.988 |
| 誤差項是否獨立 | 否 |

|  |  |  |  |
| --- | --- | --- | --- |
| 正規化迴歸 | 脊迴歸1 | 套索迴歸1 | 彈性網1 |
| R2 | 0.7018791140967994 | 0.7018829235480735 | 0.7018324002344185 |
| MSE | 0.2981208859032006 | 0.29811707645192653 | 0.29816759976558155 |
| MAE | 0.34175255716131436 | 0.3420537525356852 | 0.3409623341665454 |

|  |  |  |
| --- | --- | --- |
| 選前五個重要特徵 | 脊迴歸2 | 套索迴歸2 |
| R2 | 0.640743769864084 | 0.6406768816007506 |
| MSE | 0.35925623013591607 | 0.3593231183992494 |
| MAE | 0.3700738115924254 | 0.3686534838778122 |

分析:由上述表格可以發現在我還沒做標準化之前MSE都過大，在正規化迴歸時有先做了標準化，都有的到較小的MSE，但三種正規化迴歸得到的R2、MSE、MAE都很相近，讓我蠻疑惑這樣是否正常，然後隨機森林的R2到達0.9多可能有過擬合的情況，再來是我在做選取除Price以外特徵的脊迴歸時，它的alpha值對應的交叉驗證MSE和MSE標準誤差圖看起來超怪，有詢問老師後選取前五個重要特徵試試看，alpha值有變小，但圖看起來還是跟之前一樣，這邊也挺讓我疑惑的。但大致上看起來正規化的迴歸能降低模型的複雜度。