大數據分析與資料探勘-第三次作業

繳交日期: 2023/4/26

使用Python程式進行分析並回答以下所有問題。請直接在word檔每個問題的下面，貼上python程式與執行結果，最後將檔案上傳到MOODLE，包含此word檔及python程式檔(佐證用)。兩個檔案麻煩請壓縮成一個檔在上傳，檔名請命名為Homework3\_姓名.zip，另程式碼中請詳細標明對應的題數，若看不懂的話，一律不給分；此外，若檔案有缺少的話，亦會扣分。上傳期限為4/26晚上11點59分前，超過繳交期限就無法上傳。

1. 一張含有 文字 的圖片

   自動產生的描述請匯入Taipei\_house資料檔(台北市房價資料集)至DataFrame，並加以顯示。
2. 一張含有 資料表 的圖片

   自動產生的描述請確認此資料集中是否有空缺值(亦即計算DataFrame中每個欄位的空缺值數目)，並加以顯示。
3. 一張含有 文字 的圖片

   自動產生的描述由於在此資料檔中，「行政區」欄位為類別型資料(包含：信義區、大安區、文山區、松山區)，請使用One-hot Encoding方法將其擴展為四個columns (亦即對應於該列的column為1，其餘columns為0)，將這四個新columns增加到DataFrame的最後，column名稱分別為「行政區\_信義區」、「行政區\_大安區」、「行政區\_文山區」和「行政區\_松山區」，再將原始資料中的第一欄行政區刪除，並加以顯示。
4. 一張含有 文字 的圖片

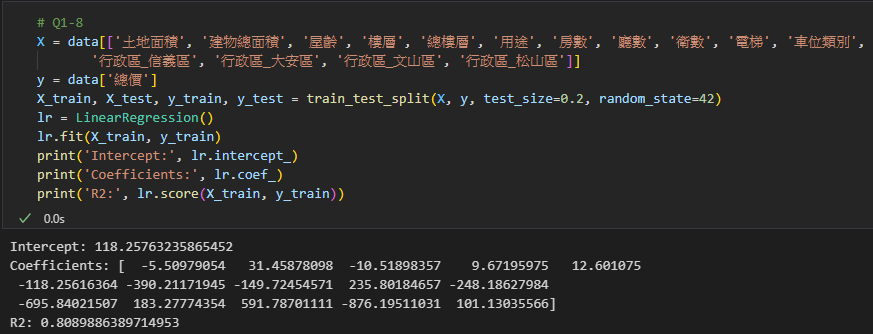
   自動產生的描述承上題，在此資料檔中「車位類別」欄位亦為類別型資料(包含：無(無車位)、坡道平面、坡道機械、升降機械、升降平面等)，請使用編碼的方式將此欄位中「無」置換為0、其他類型的車位全部置換為1，並加以顯示。
5. 一張含有 文字 的圖片

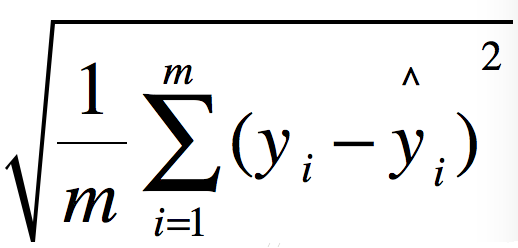
   自動產生的描述透過相關係數，可輔助迴歸模型之建立，目前資料集除了行政區為使用One-hot Encoding方法擴充為四個欄位，車位類別為編碼方式後所得到的欄位，交易日期為日期變數外，其他欄位皆為數值型資料，請把其他欄位(共12個欄位)與總價欄位跑相關係數，並加以呈現；其中，相關係數請顯示至小數點第二位。
6. 一張含有 文字 的圖片

   自動產生的描述請挑選出與總價相關係數最高的欄位，將其設定為X變數，Y變數則設定為總價。接著使用sklearn中的函式把X、Y變數分為training及test datasets。其中，training的資料占80 %、test 20%，並設定Random\_state = 42，請秀出 X\_train的前五筆資料。
7. 承上題，請使用sklearn中的線性迴歸模式套件，以training datasets的資料建立簡單迴歸模型，並分別秀出截距項、X變數之係數及R2。

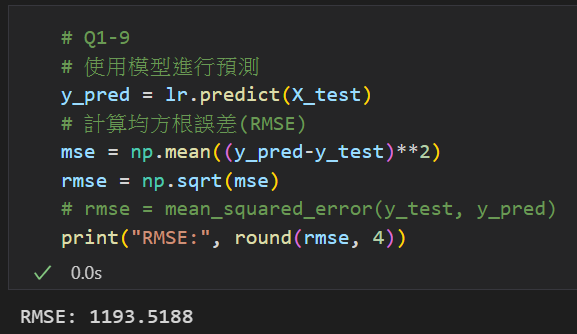
一張含有 文字 的圖片

自動產生的描述

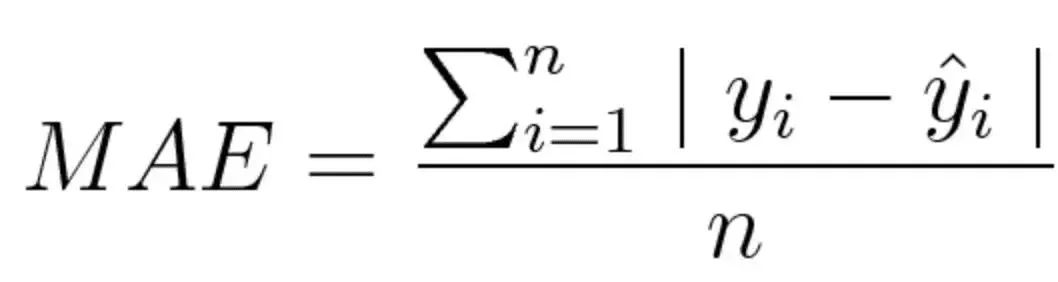
1. 為建立複迴歸模型，請將土地面積、建物總面積、屋齡、樓層、總樓層、用途、房數、廳數、衛數、電梯、車位類別、行政區\_信義區、行政區\_大安區、行政區\_文山區、行政區\_松山區同時放入X變數，而Y變數為總價，接著使用sklearn中的函式把X、Y變數分為training及test datasets；其中，training的資料占80 %、test 20%，並設定Random\_state =42，再使用sklearn中的線性迴歸模式套件，以training datasets的資料建立線性複迴歸模型，並分別秀出截距項、X變數之係數及R2。
2. 承上題，衡量迴歸方程式好壞，常用的指標為均方根誤差(RMSE)，其公式如下：



請用test datasets計算複迴歸模型的均方根誤差(RMSE)，請顯示至小數點第四位，並秀出結果。



1. 另一個常用於衡量迴歸方程式好壞指標為平均絕對誤差(MAE)，其公式如下：



請用test datasets計算複迴歸模型的平均絕對誤差(MAE)，請顯示至小數點第四位，並秀出結果。

一張含有 文字 的圖片

自動產生的描述

1. 承上題，假如土地面積、建物總面積、屋齡、樓層、總樓層、用途、房數、廳數、衛數、電梯、車位類別、行政區\_信義區、行政區\_大安區、行政區\_文山區、行政區\_松山區變數的值分別為36、99、32、4、4、0、3、2、1、0、0、0、0、0、1請依據第(8)小題所建出的複迴歸模型，預測房屋的總價，請顯示至小數一張含有 文字 的圖片

   自動產生的描述點第四位，並秀出結果。
2. 如上課所提，自變數的數目增加，亦會造成迴歸模式的R2增加，故有時會使用調整後判定係數Ra2作為迴歸模式評估的基準，其公式如下。請使用自定函數的功能，建構調整後判定係數的公式，重新計算第8小題所建構之複迴歸模型的調整後判定係數，並加以顯示。

一張含有 文字 的圖片

自動產生的描述

1. 請使用sklearn中的多項式(Polynomial)非線性迴歸套件，以第(8)小題所分割出的training datasets建立模型，並分別印出截距項、X變數之係數及R2（請設定degree = 2）
2. 一張含有 文字 的圖片

   自動產生的描述