機器學習與 Python 行銷數據分析期末報告

臺中市八大行政區房價預測

及西屯區分析

組員：行銷四乙 D0790192 林崇暟

行銷四乙 D0812006 王偵瀞

行銷三甲 D0932200 栗村響樹

環境三甲 D0961332 蕭淳匀

指導老師：周進華 老師

繳交日期：2023.05.12

目錄

圖目錄

表目錄

一、資料蒐集

本研究報告房屋資料來自內政部不動產交易時價查詢服務，將八大行政區的 url 分別命名，以簡潔的方式利用 def 定義爬蟲流程公式及區域命名，最後利用 for 迴圈將八大行政區的資料及其區名稱彙整到同一個 df，由於一次爬取過多資料可能會造成網頁阻攔，為避免此情況發生，故於 for 迴圈中增設 time.sleep(15)，爬取一行政區資料後便休息 15 秒，最後建立 df_filtered 將 df 資料中過濾出我們要的目標值。

次要的資料為欲新增的額外特徵——捷運，資料來自於臺中市政府資料開放平台，資料並不需要而外做處理(如經緯度轉換等)，下載下來導入即可使用。

二、資料處理

2.1 DDP1

(1) NAN

- 刪掉 NAN 太多欄位

  刪除 NAN 帶多無法使用的欄位，如：community、 build_share1、build_share2 及 note，使用 drop()，由於為直行，故 axis = 1。

- 刪掉有少部分 NAN 資料的行

  由於有些類別中的 NAN 占比不多，故可直接做刪除，利用 dropna 刪除 build_type、main_purpose 以及 layout 類別中的 NAN 的資料列。

(2) 重複值

利用 drop_duplicates()去除資料的重複值。

(3) 屬性轉換

將交易日期中月份 00 者刪除，並轉換成西元年。將 price 與 plain 的逗號利用 replace(',', '')將其抽出，並用 astype 轉為浮點數。

(4) 簡化分類

由於有些欄位資料內容過於冗長，如 main_purpose 及 build_type。將 main_purpose 中的商業用、辦公用及商辦用歸類為商業用，工業用、住工用及工商用皆歸類於工業用，而農業用、其他及見其他登記事項皆歸類於其他，剩於之類別即為簡稱。而 build_type 則將類別減少至 3 字以下。

(5) 擷取

- 路名

  路名位於#之後與數字之間，利用[\u4e00-\u9fff]抓取中文，但由於數字為全形，故須將數字一個一個打出來。其中有少數路名並未依照大多數的路名格式命名以及含有非中文字或數字的路名，皆在路名擷取過程將有問題的資料的索引值收集在一個 list，並將這些有問題的行列從資料庫移除。

- 房衛廳

  利用 for loop 搭配 regex 建立定義函數，再將每一行的房衛廳的數字擷取下來，分別將房、衛、廳的數字回傳至不同的 Series，最後再與原本的 Data Frame 做合併。

- 最高樓層/樓層數=>(清出來要賣的樓層)

  先透過“/”符號過濾掉後面的總樓層，再透過“層”這個關鍵字計算總共販賣幾層樓，如果該物件樓層顯示“全”，則先以“全”做填補。

  其中也透過 unique()觀察最高樓層是第幾層，並依序將一層到四十層建立在一個 list 裡面，最後再搭配 for loop 將資料庫裡每個物件的最高樓層篩選出來。

- 交易年

  新增一個 deal_year 的欄位，並用 apply()及 lambda 將 deal_date 的交易年抽取出來

(6) 新增外部特徵

- 捷運

  將捷運檔案導入編輯器，利用 def 定義計算經位度間的實際距離，利用 for 迴圈及 if 判斷目標物分別與三者的距離是否< 500 m，並各別建立兩種清單，一種為 0 或 1 判斷有無，另一種為回填距離＜ 500 m 之捷運名稱。

同時，也建立另一個欄位，將附近(500m)有捷運站的物件，置入捷運站的名稱

- Label Encoding

利用 for 迴圈及 if 將 elevator 及 manager 的有、無置換成 0、1。

2.2 DDP2

首先將 DDP1 後資料匯入，使用住商及住家類別資料，並刪除含 str 型態的欄位，再將 feature 與 label 分開。進行 DDP2 前先將資料進行切割(train：85 %，test：15 %)，後續資料處理都分別於 train 與 test 中進行。

(1) NAN

處理 age 缺值，缺值得處理方式大致分為兩類，刪掉或補值，由於 age 缺值占比過高，刪掉不太妥當，故而進行補值，補值的方法又可分為三種，平均值、中位數及眾數，本研究擇定利用中位數做填補，考量到的數據分布並非常態分佈，大量的填補中位數不會影響數據的分布型態，比起利用均值填補，更不容易出現額外的偏差，從而影響後續模型的建構與分析，因此擇定利用中位數做填補。

(2) 離群值

所使用方法以極端異常值(extreme outlier)為主 quantile 為輔，若將輕度異常(mild outlier)之資料作為邊界會造成會有過多資料被刪除，故直接使用有些不妥，式(1)~式(5)為輕度與極端異常值之公式，如下所示：

$$IQR = Q3 - Q1 \quad (1)$$
$$IF1 = Q1 - 1.5 \times IQR \quad (2)$$
$$IF2 = Q3 + 1.5 \times IQR \quad (3)$$
$$OF1 = Q1 - 3 \times IQR \quad (4)$$
$$OF2 = Q3 + 3 \times IQR \quad (5)$$

其中 Q1 為第一四分位數，Q3 為第三四分位數，IF1 為輕度離群值之下限，IF2 輕度離群值上限，OF1 為極端離群值之下限，OF2 極端離群值之上限。由上述的公式可以看到，極端離群值比起輕度離群值多了兩倍的 IQR 也因此大幅減少了過多資料被刪的情形。

特徵 livingroom 的經式(4)與式(5)的計算後，得出 Q1 及 Q3 都是 2.0，故 IQR 為 0，所以 livingroom 的特徵處理改用≤quantile(0.99)，即為所刪除的資料大於 99%的數據。

(3) One Hot Encoding

利用.get_dummies()針對以下六個欄位進行編碼：

build_type、main_purpose、district、deal_year、

total_floor 及 MRT(捷運名稱)。

三、視覺化分析

3.1 八大行政區交易量 Treemap

圖一為八大行政區的交易量，如圖所示可知北屯區有最大的交易量，第二大交易量為西屯區，交易量越大可能意味著市場活躍、需求旺盛及具有投資可能性，但並不是絕對的關係，因交易量也深受政策的變化，本研究的視覺分析將著重探討西屯區，除該地區交易量大以外，同時也是本校所在位置。

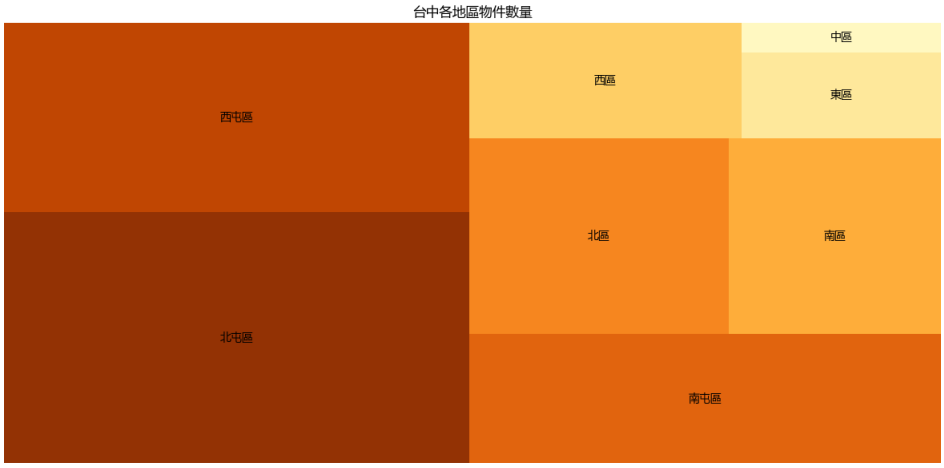

圖 一、八大行政區交易量 treepmap

3.2 物件熱力圖

物件熱力圖是一種以顏色呈現物件分布密度或特定屬性值的地圖，顏色的深淺或濃度反映了該區域內物件的數量或該屬性值的大小。它可以提供一個直觀的視覺化方式，用於顯示物件的集中程度和變化趨勢。

(1) 使用此圖分析物件數量之優勢

- 直觀視覺化：通過顏色映射直觀呈現房價分佈，使得用戶能夠迅速理解不同區域的物件聚集程度。
- 熱點識別：能夠快速識別熱門地區或熱點，即物件數較高或需求較大的區域。

(2) 使用此圖分析房價之劣勢

- 無法提供具體房價數值：僅顯示整體趨勢和集中區域，無法提供每個物件的具體房價數值。
- 標記位置限制：無法詳細標記每個個體物件的屬性和細節。
- 複雜性：當物件數量眾多時，熱力圖可能變得複雜和擁擠，難以清晰地解讀。

(3) 小結

綜合上述分析，此圖較適合用於觀察整體的物件分布情況、熱點識別和空間關聯性及市場趨勢和整體供需情況。

如圖二所示可發現大多的交易量都集中於西屯區的西邊，東南邊的交易量較少，其原因為東邊屬都市計畫七期之範圍，該區房屋價格較昂貴，因此不會有太多的交易量，若

有要買房子之需求，由此圖可知，西屯區西側會有較多的選擇。

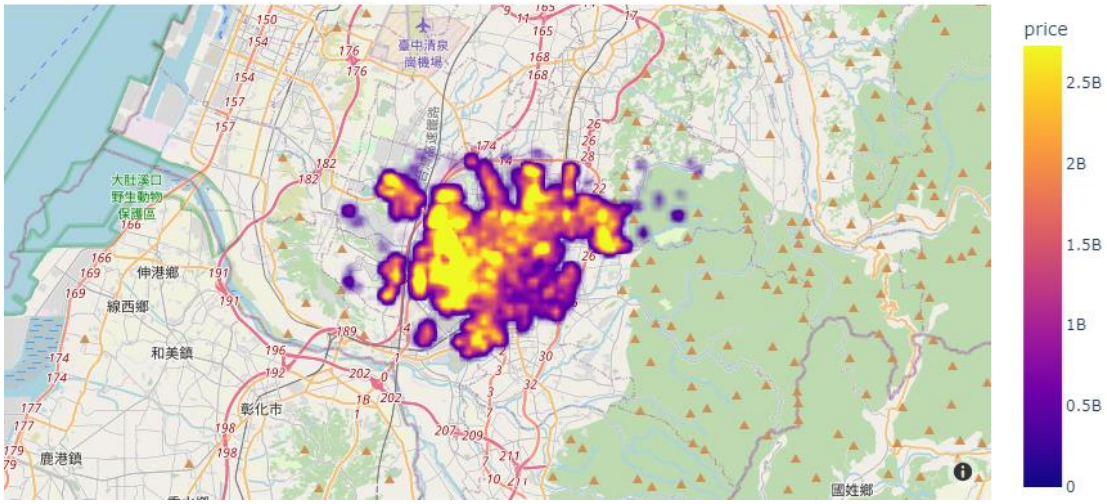

圖 二、西屯區物件熱力圖

3.3 物件分布地圖

(1) 使用此圖分析房價之優勢

- 空間關聯性分析：幫助觀察物件之間的空間關聯性，如與周邊設施的距離、地理環境等。
- 地理信息展示：能夠展示不同區域的地理信息，如道路、公園等，對於環境評估具有重要意義。
- 可個性化定制：可以根據需要自定義地圖標記和顯示內容，提供更靈活的分析選項。

(2) 使用此圖分析房價之劣勢

- 缺乏整體趨勢：難以直觀地呈現整體房價趨勢和集中區域，需要進行更多的分析工作。
- 資料量限制：當物件數量眾多時，地圖可能顯示不完整，無法顯示所有物件。
- 地理限制：僅限於地理位置展示，無法展示其他重要因素如房屋內部特點、建築結構等。
- 資訊過載：當標記物件過多時，地圖上的標記可能變得拥擠，難以清晰解讀。
- 需要專業技能：製作和解讀物件分布地圖可能需要一定的地理信息和數據處理的專業知識。

(3) 小結

綜合上述分析，此圖較適合展示個別物件的屬性和空間關聯性和多樣化的地理特徵，例如：周邊設施分析、地理環境評估等。

圖三為西屯區物件的確切位置，可藉由調整圖的大小來查看物件分布與道路間的關係，並配有數字可明確知道擇定範圍內物件的數量及其周遭環境。

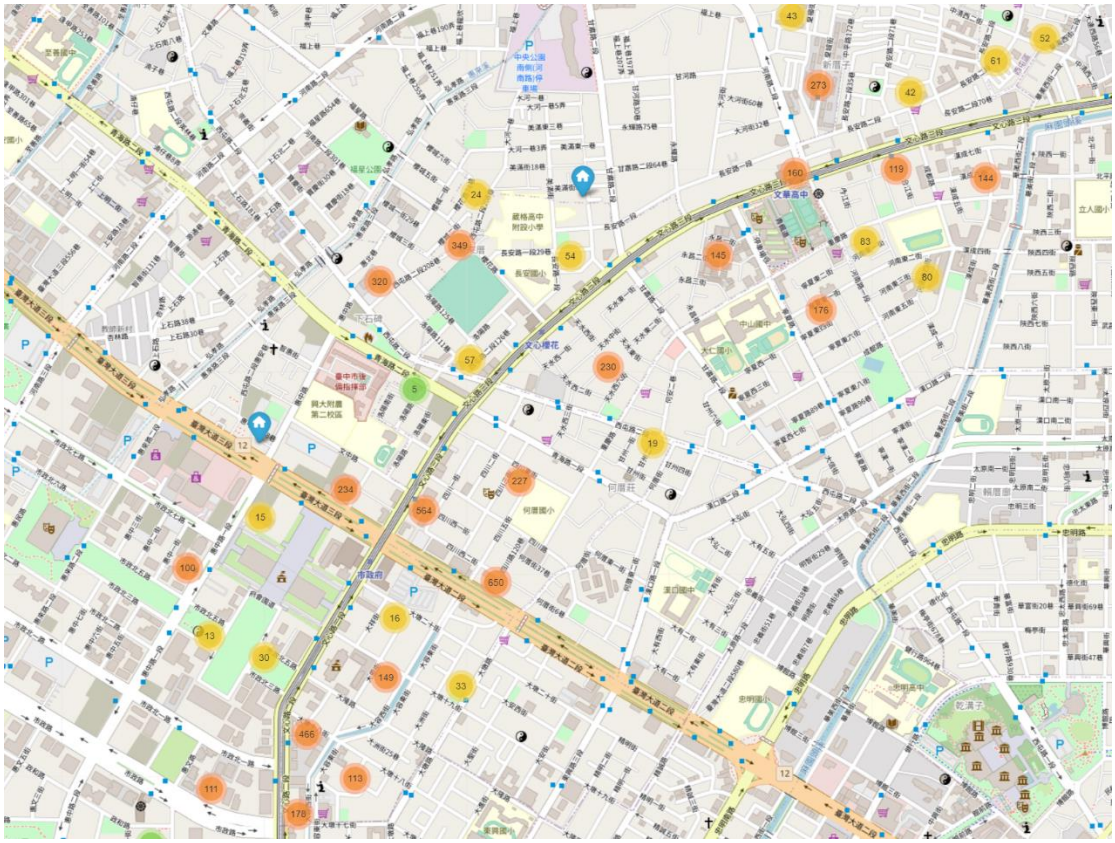

圖 三、物件分布地圖

3.4 交易量與時間趨勢折線圖

西屯區交易量與時間的關係如圖四所示，可以發現在民國102年的時候交易量急遽上升，甚至為這10年以來的高峰，高達6,144筆，造成此原因與政府政策影響，在102年臺中市進行了市地重劃區抵費地標售，這當中就包含了西屯區(https://www.land.taichung.gov.tw/content/?parent_id=123641&type_id=123641)，後續幾年也進行了幾次的都市計畫和疫情等事件，交易量自然也呈現起起落落，單基本上交易量介於3,000~4,000筆。

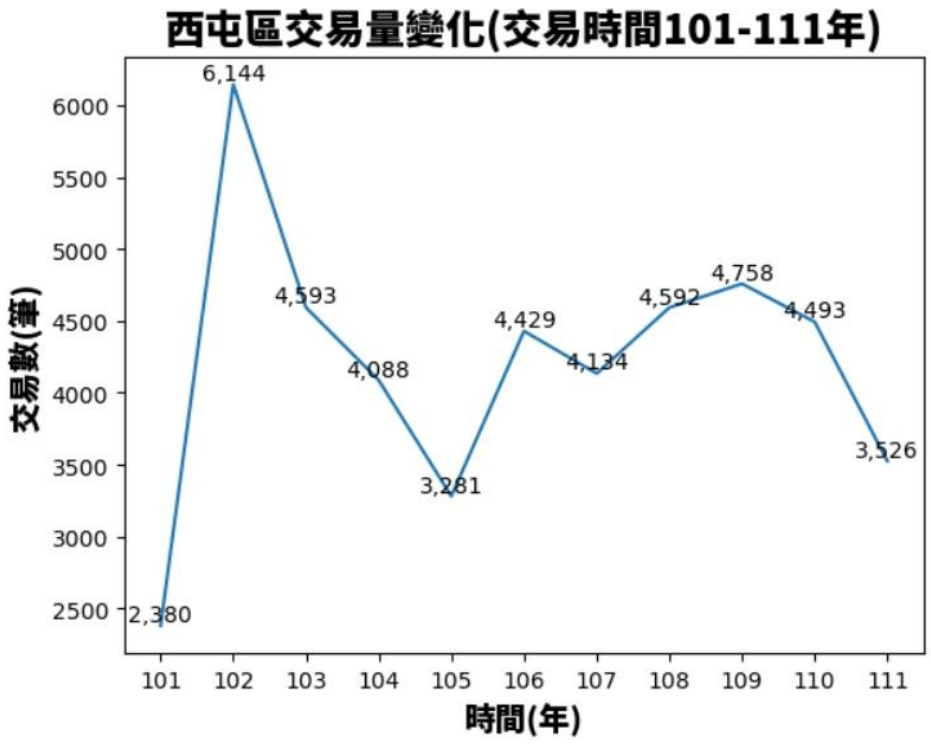

圖 四、西屯區交易量變化(交易時間 101-111 年)

3.5 道路與交易量 Treemap

圖五為前十大交易量的路段，於國安一路與西屯路二段有較大的交易量，藉由此圖更精細的呈現出圖二的的結果，並直覺明確地觀察到各路段交易量的差異。

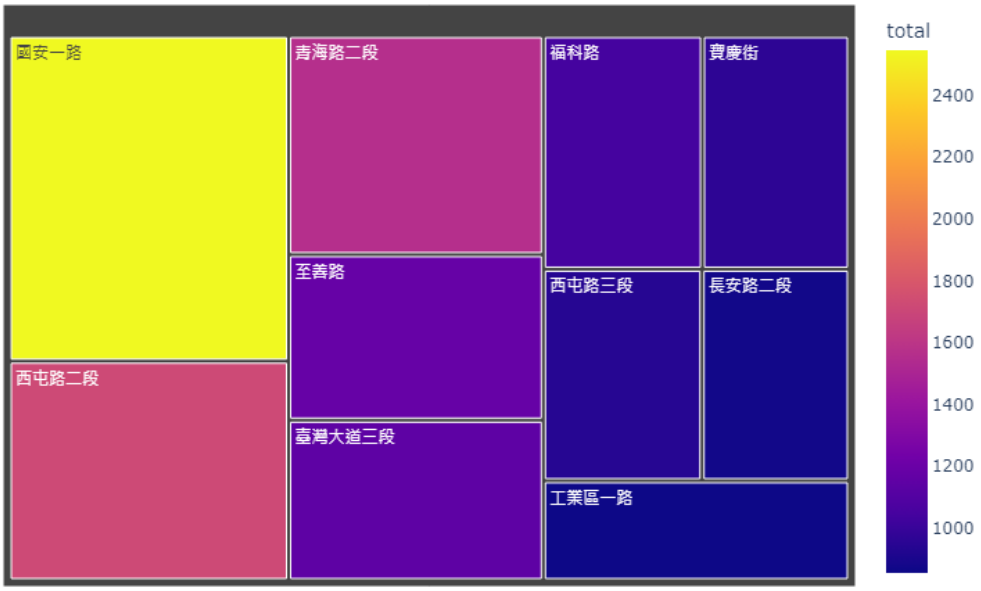

圖 五、西屯區前十筆最大交易量的路段

四、特徵處理

4.1 特徵處理

將所有特徵正歸化，但 Random Forest Regressor 與 XGBoost Regressor 因模型特性的關係，不用使用正歸化數據，因此於導入模型前再進行反正歸化。

4.2 評估最佳模型(特徵挑選&模型最佳化)

截至目前所彙整的特徵挑選一共嘗試了 5 種方式，即利用 forward 挑選 7、8 個特徵，與皮爾遜相關性分析配嵌入法(相關性篩選：0.2/0.3/0.5)，其中利用了三種模型(Linear Regression、RF 及 XGBoost)，將結果依不同模型彙整於表一至表三，結果顯示以模型採用隨機森林，特徵選擇方法使用 Wrapper Method，有最佳的模擬效果，R^2 達 0.9159。

表 一、評估最佳模型– Linear Regression

特徵選擇方法	特徵數量	R^2	模型最佳化方法
Wrapper Method	8	0.7791	N/A
Wrapper Method	7	0.7731	N/A
Pearson's Correlation ($\pm$0.2, $\pm$1) & Embedded Method	4	0.7119	N/A

表 二、評估最佳模型-RF

特徵選擇方法	特徵數量	R^2	模型最佳化方法
Wrapper Method	8	0.9159	N/A
Wrapper Method	7	0.9152	N/A
Pearson's Correlation ($\pm 0.2, \pm 1$) & Embedded Method	8	0.8229	GridSearchCV

表 三、評估最佳模型-XGBoost

特徵選擇方法	特徵數量	R^2	模型最佳化方法
Pearson's Correlation ($\pm 0.2, \pm 1$) & Embedded Method	8	0.7990	GridSearchCV
Pearson's Correlation ($\pm 0.3, \pm 1$) & Embedded Method	4	0.7869	GridSearchCV
Pearson's Correlation ($\pm 0.5, \pm 1$)	2	0.7469	GridSearchCV

4.3 各模型詳細流程

(1) Linear Regression

A. 將 Train Data 中的資料進行 OneHotEncoding, OrdinalEncoding 與 MinMaxScaler。再將 Linear Regression 置入 SequentialFeatureSelector 中，進而利用 Wrapper Method (Forward)的方式選取 8 個特徵。

B. 為了節省時間，再從原先 8 個特徵再做一次 Wrapper Method 選取其中的 7 個特徵作為另一個模型。

C. 為了利用 Embedded Method 選取特徵，先是將正規化後的數據進行 Pearson's Correlation 的分析，藉此篩選掉特徵與特徵間相關係數大於±0.7 的其一特徵，選擇方法則是再將兩特徵與標籤做比較，將相關性較高者留下。同時，特徵與與標籤之間若是介於±0.2 之間，也都將這些特徵去除。再將剩下的 8 個特徵去進行 Linear Regression 的 Embedded Method，留下剩下的 4 個特徵。

(2) Random Forest Regression

A. 將 Train Data 中的資料進行 OneHotEncoding, OrdinalEncoding 與 MinMaxScaler。再將 Linear Regression 置入 SequentialFeatureSelector 中，進而利用 Wrapper Method (Forward)的方式選取特徵。歷經了 12 個小時，將 53 個欄位的資料篩選出了 8 個特徵。

B. 為了節省大量時間，再從原先 8 個特徵再做一次 Wrapper Method(Forward)選取其中的 7 個特徵作為另一個模型。

C. 如同 Linear Regression 的第 3 點，但不同之處再於，在將 train data 正規畫並列出特徵與特徵的相關係數介於±

0.2 的特徵之後，再將 train data 進行反正規劃，好讓後續的 Random Forest Regressor 可以進行 ，但由於在 Embedded Method 的 feature importance 的比較後，發現個個特徵都對模型有貢獻，也因此將全數特徵保留，並進而利用 GridSearchCV 進行模型最佳化，找出適合的超參數以建立模型。

(3) XGBoost

A. 如同 Random Forest Regressor 的第 3 點，只是將 GridSearchCV 內的模型改成 XGBoost Rgressor

B. 將 XGBoost Regressor 第一點的特徵與特徵的相關係數改成介於±0.3，在利用 GridSearchCV 進行模型正規化。

C. 由於經過 Pearson’s Correlation 篩選特徵與特徵的相關係數改成介於±0.5 後，只剩兩個特徵，也就直接進行 GridSearchCV 來建立模型。

五、New instance 結果

利用 3 筆隨機資料進行預測，隨機資料數值如圖六及圖四所示，而預測結果如圖八所示，價格分別各為：6,513,969 元、20,245,969 元及 19,513,540 元。

圖 六、隨機資料(前半部)

build_type	age	longitude	latitude	main_purpose	manager	plain	room
套房	6	120.61	24.145	住家	1	30	2
公寓	15	120.645	24.131	住家	0	50	4
大樓	26	120.72	24.2	住商	1	100	3

圖 七、隨機資料(後半部)

livingroom	bathroom	total_floor	MRT_nearby	elevator	district	deal_year	MRT
2	1	1	1	1	南屯區	109	MRT_南屯站
3	2	2	1	1	西屯區	110	MRT_文心櫻花站
1	1	1	0	1	南區	111	無

圖 八、價格預測結果

```
In [5]: print("the price of three instances are{}".format(y_pred))
   ...: #the price of three instances are:
the price of three instances are[ 6513969.05444444 20245969.16666667 19513540.        ]
```