統計學習與深度學習期末專案

房價預測模型

統計學習互助群：李維農、陳薇守、羅元駿、蔡鉎驊、陳沛妤

Agenda

- **資料蒐集與處理**
  1. 內政部實價登錄資料庫
  2. 以 Google Map API 自行建構生活機能資料庫
  3. 以 Google Street API 街景圖片建構 VGG 模型
- **探索性資料分析 (EDA)**
- **預測模型一：實價登錄**
- **預測模型二：實價登錄+生活機能**
- **預測模型三：實價登錄+生活機能+街景/衛星圖**
- **總結**

資料蒐集與處理-內政部實價登錄資料庫

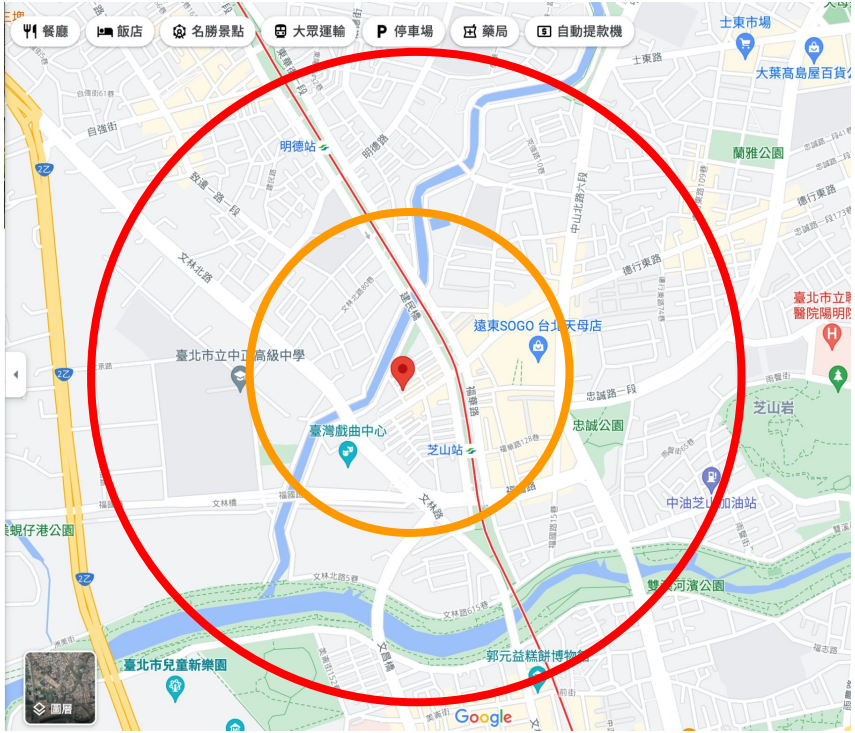
- 台北市 2020 - 2021Q3 房地不動產交易資料，共 16230 筆。
- 32 個項目 → 21 個項目 → 82 個參數。
- 使用「單價元 / 平方公尺」為「房價」預測變數。

總項目：	
鄉鎮市區	主要用途
土地位置建物門牌	主要建材
土地移轉總面積平方公尺	建物移轉總面積平方公尺
都市土地使用分區	建物現況格局-房
土地數	建物現況格局-廳
建物數	建物現況格局-衛
車位數	建物現況格局-隔間
移轉層次	有無管理組織
移轉層次項目	單價元平方公尺
總樓層數	屋齡
建物型態	

資料蒐集與處理-生活機能資料庫

1. 參考永慶房仲看屋檢核表定義生活機能店家(如下表)
2. Geocoding Api 將地址轉成經緯度
3. Google Place Api Textsearch 尋找店家

範圍	指標
鄰里環境評估(800m)	捷運、超商、公園
區域環境評估(5km)	學校(托兒所、國小、國高中職、大學)、金融機構、醫院、大賣場、超市、百貨公司、警察局、消防局



資料蒐集與處理-街景圖片 & 衛星圖片

透過Google API 取得圖片資料

(Streetview & Staticmap)

ex : 台北市士林區德行西路 111巷2弄4號5樓

資料蒐集與處理-街景圖選取適當角度

「 圖片人工篩選 」後剩下13691筆

選取合適的街景圖

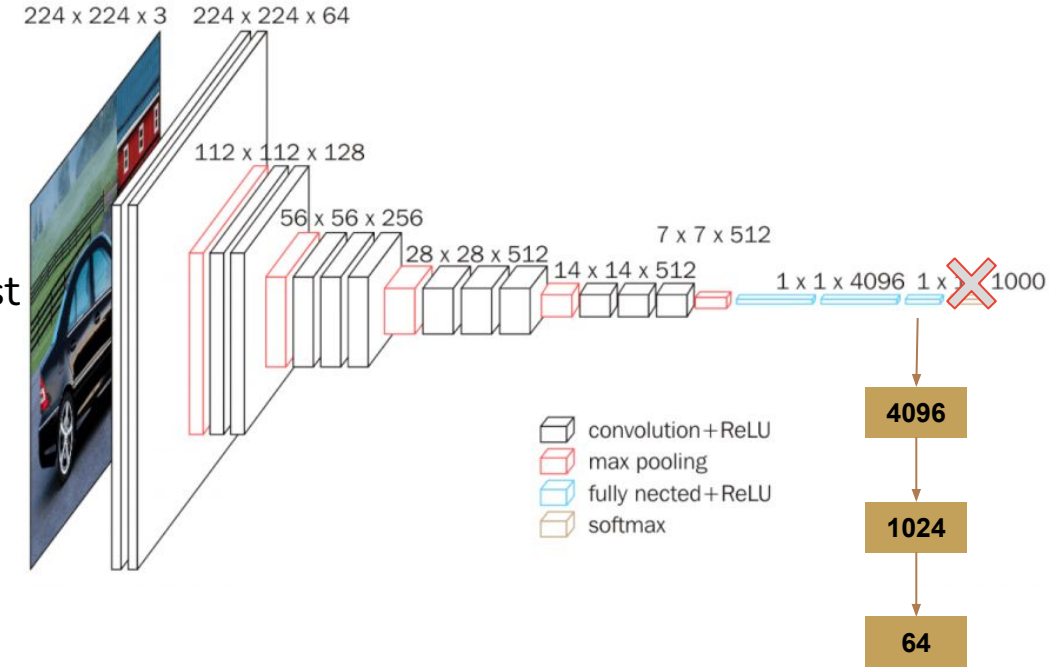
（盡量正面對街道，兩側為建築物）

資料蒐集與處理-VGG模型資料庫

1. VGG16 輸入至 fc2 層

2. 降維方法
   a. 接三層 Dense
      維數選擇:Random Forest
      預測計算 RMSE

   b. PCA (Principal components analysis)



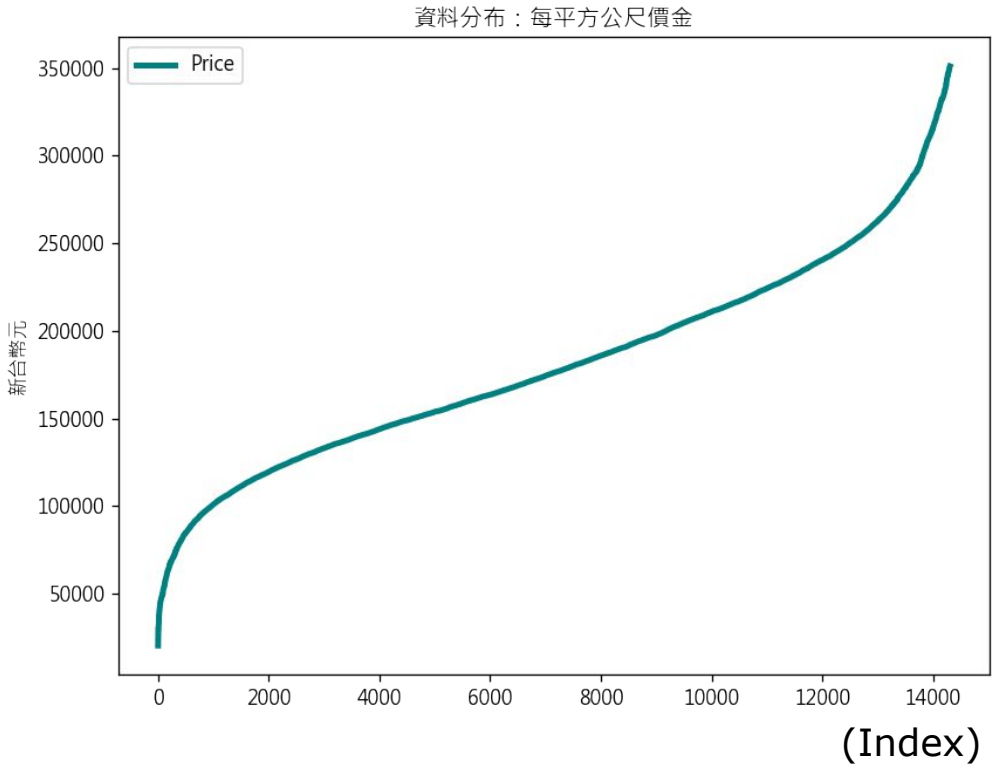
Source: Neurohive

探索性資料分析(EDA)

房價分佈

1. 依照房價由小至大排序

2. 多數房子的每平方公尺價格，分布於 $15000~$25000

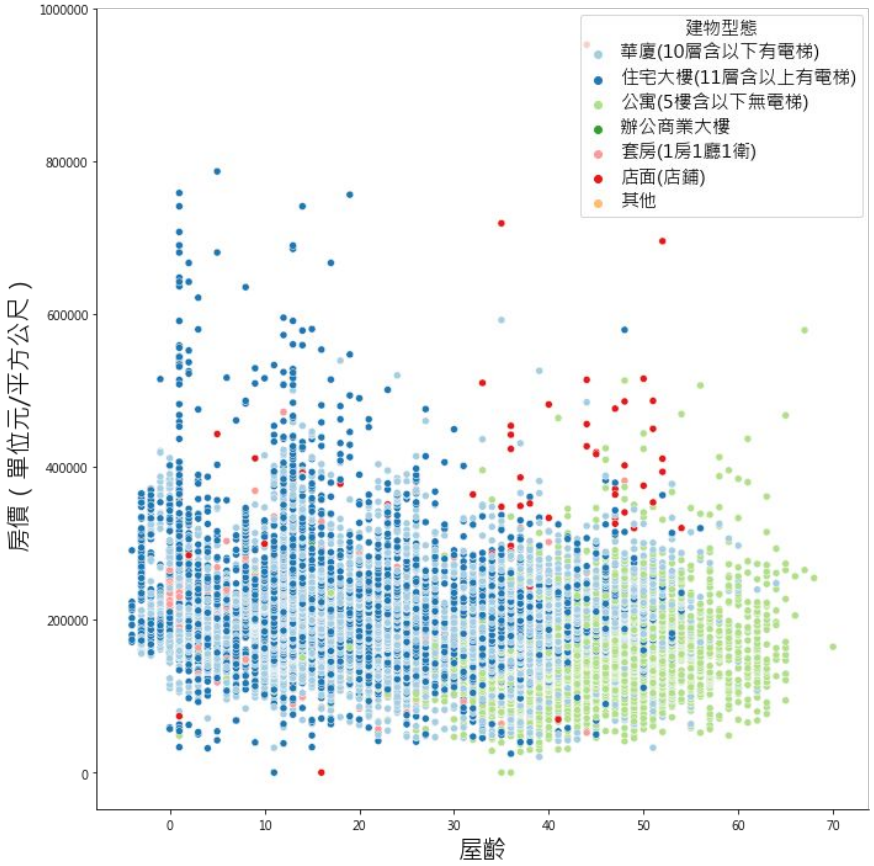


屋齡&建物型態 vs 房價

1. 僅擷取與「住宅」相關區域（包含住商混合區）

2. 5 層樓以下無電梯公寓屋齡明顯偏高、住宅大樓與華夏屋齡較低

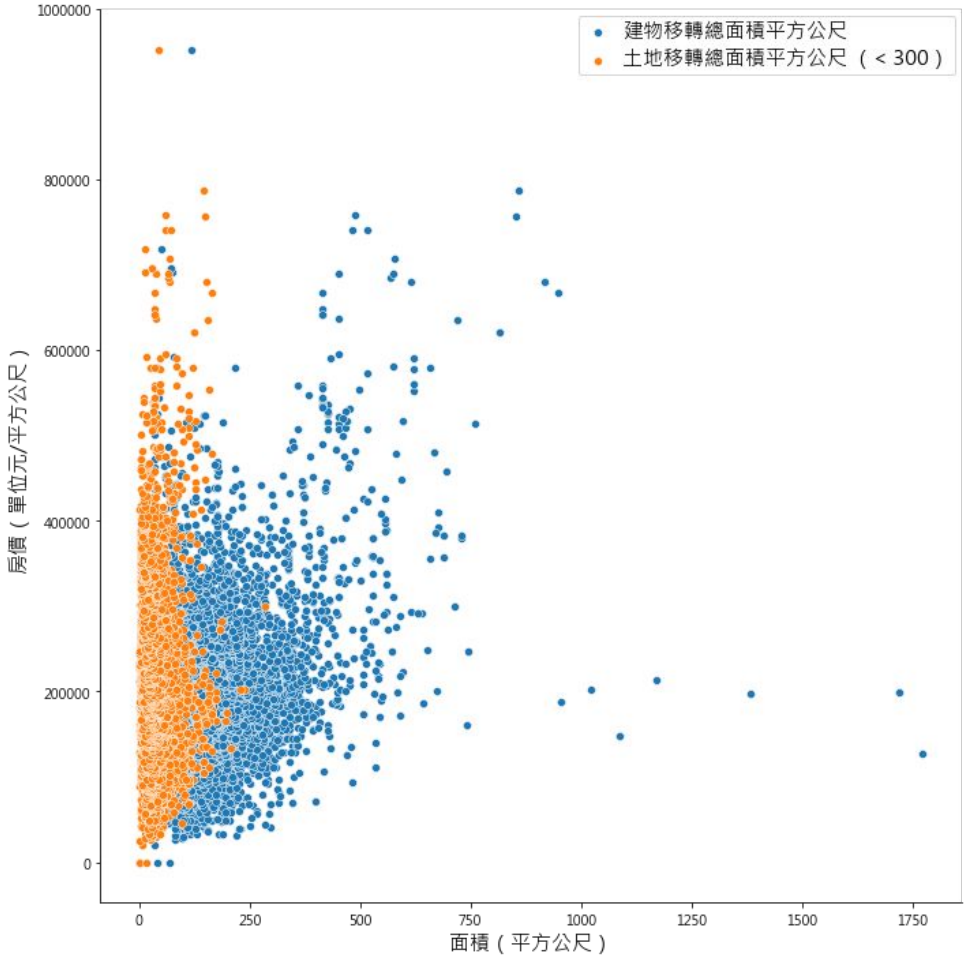
3. 屋齡與房價無明顯相關性



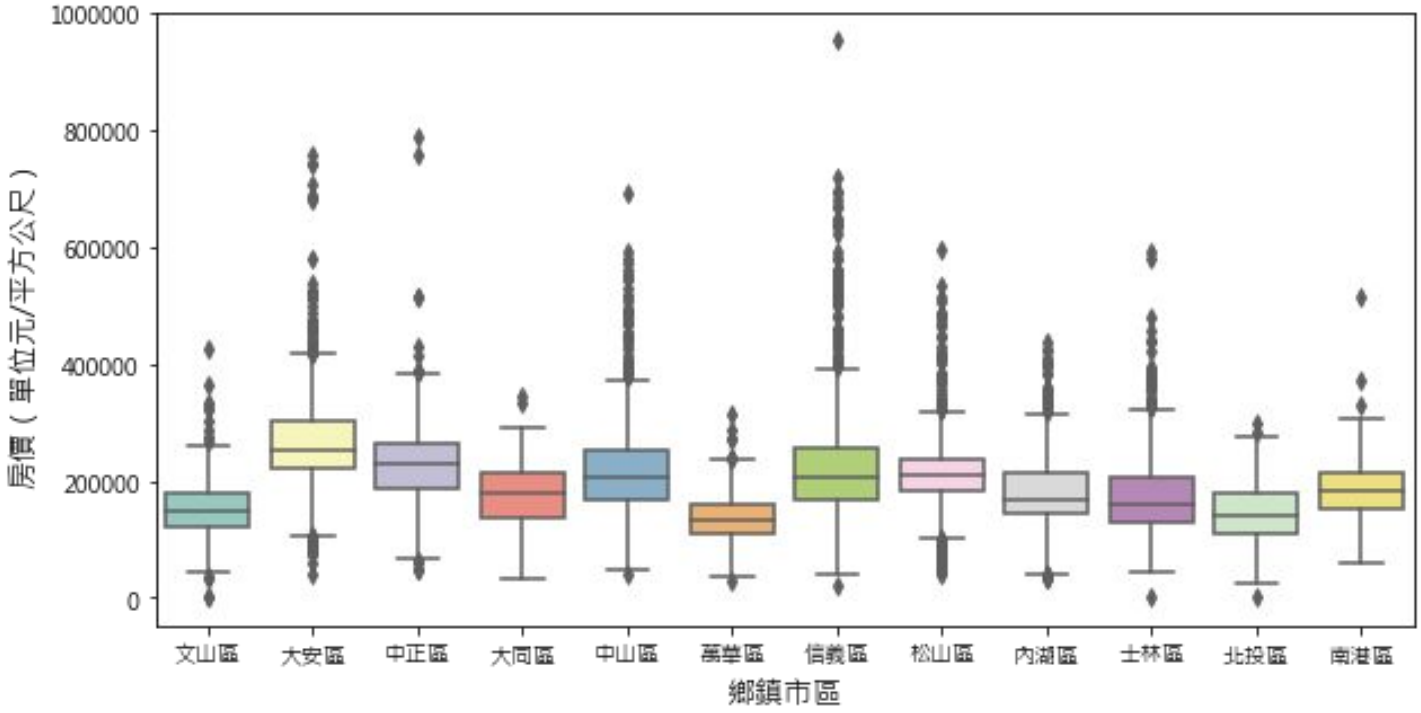
面積 vs 房價

1. 僅擷取與「房地」相關交易，土地面積大的交易會被歸類在「土地」交易

2. 房地面積較大，價格上升，可能為豪宅寬闊；相較之下土地面積無此特徵



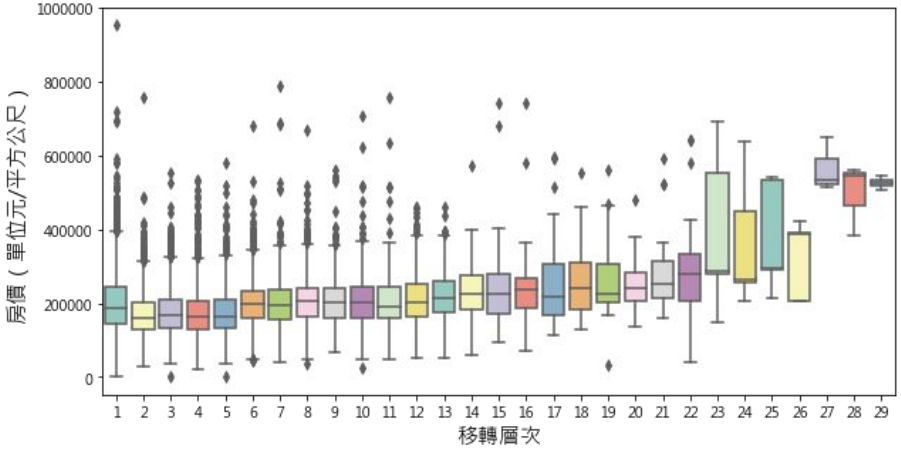
區域 vs 房價

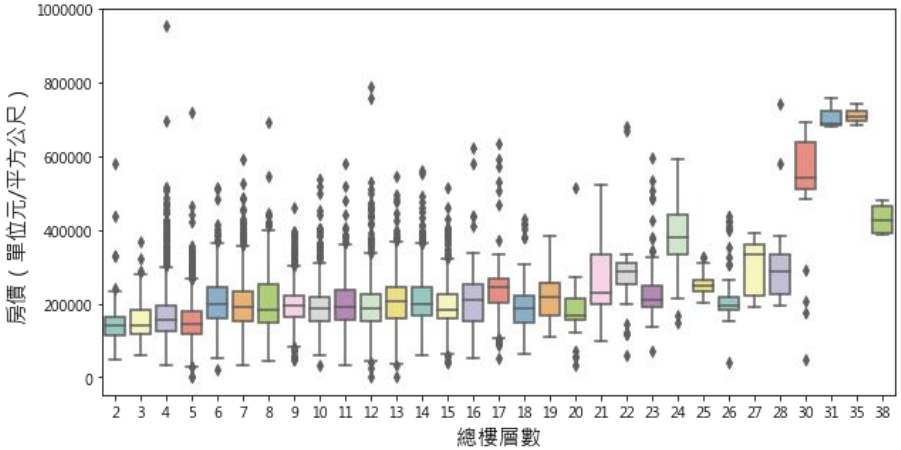

1000000
800000
600000
400000
200000
0
房價（單位元/平方公尺）
文山區
大安區
中正區
大同區
中山區
萬華區
信義區
松山區
內湖區
士林區
北投區
南港區
鄉鎮市區

層數 vs 房價

1. 移轉層次越高越貴，在22樓以上起伏明顯；總樓層數則在30樓以上明顯高價。

2. 總樓層數38層並沒有顯著的高房價，原因為38層其實皆為同一棟建築，資料量小且侷限。



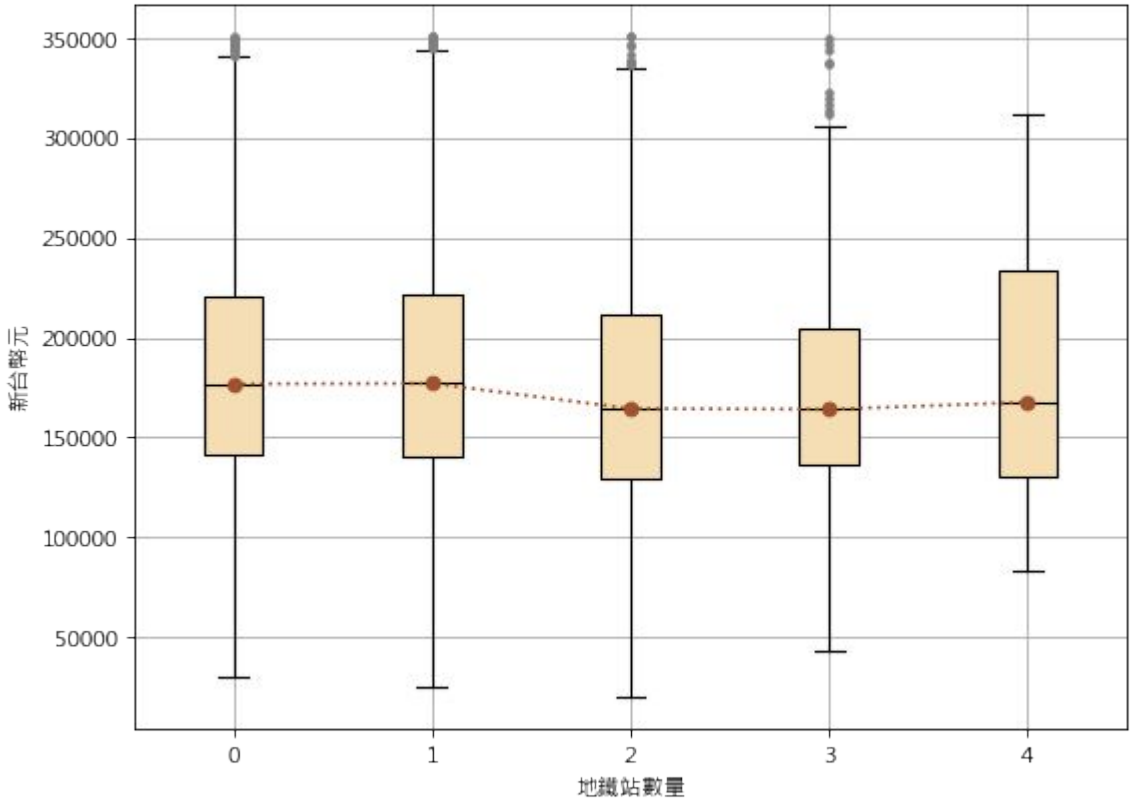


地鐵站數量 vs 房價

1. 半徑800m內的地鐵站數量

2. 並不是交通越便利，房價越高

3. 800m內有4座地鐵站之處範圍侷限



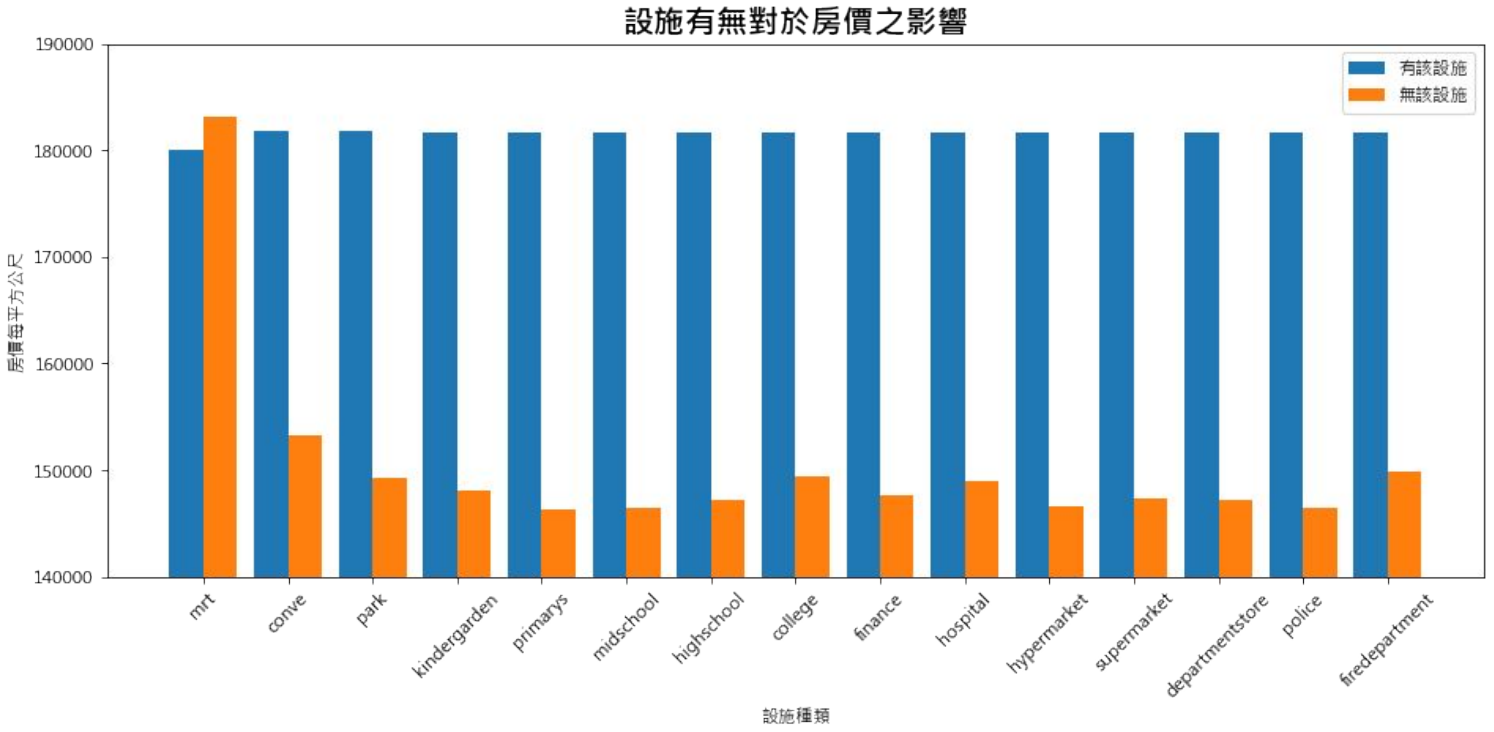
地鐵站

1. 半徑
站數

2. 並不
房價

3. 800m
之處

設施vs 房價


設施有無對於房價之影響
有該設施
無該設施
房價每平方公尺
190000
180000
170000
160000
150000
140000
mrt
conve
park
kindergarden
primarys
midschool
highschool
college
finance
hospital
hypermarket
supermarket
departmentstore
police
firedepartment
設施種類

預測模型

預測模型一：僅實價登錄

使用 13 種模型進行預測，其中表現最佳為 Stacking，RMSE 為 3.5 萬元／單位平方公尺，所有資料點平均為 18.8 萬元單位平方／公尺。

Model	Stacking	Random Forest Method	Bagging Method	XGBoost	Gradient Boosting	KNN Algorithms
Parameters	Random forest, Bagging, Gradient boosting, KNN, SGD, Ridge, XGB	best n_estimators: 280	best n_estimators: 79	best n_estimators: 15	best loss: huber best learning_rate: 0.2 best n_estimators: 80	best n_neighbors: 15 best weights: distance best leaf_size: 1
RMSE	35103	35400	35518	37736	38068	41109
Rank	1	2	3	4	5	6
Model	Lasso Regression	Ridge Regression	SGD	Decision Tree	Adaboost	Supported Vector Machine Regression
Parameters	best alpha: 2.7	best alpha: 2.7	best loss: squared_error best penalty: l2 best alpha: 0.0001	best criterion: squared_error best max_depth: 4 best min_samples_leaf: 5	best loss: linear best learning_rate: 3 best n_estimators: 140	best kernel: linear best C: 2.9
RMSE	41403	41418	41528	45538	46981	53490
Rank	7	8	9	10	11	12

預測模型一：僅實價登錄

1. 以 Random Forest 的 feature_importances_ 挑選前十大重要特徵

2. 較為重要的變數
- 類別：建物型態、鄉鎮市區
- 連續：屋齡、面積

前十大重要特徵：	
建物型態_公寓(5樓含以下無電梯)	0.133
屋齡	0.111
鄉鎮市區_大安區	0.092
土地移轉總面積平方公尺	0.089
建物移轉總面積平方公尺	0.087
移轉層次	0.056
總樓層數	0.038
建材_鋼筋混凝土造	0.038
鄉鎮市區_北投區	0.032
鄉鎮市區_文山區	0.030

預測模型二：實價登錄+生活機能

加上生活機能資料庫後 RMSE 顯著下降，為各種 Source 排列組合之中表現最好的，顯現台北市房價與生活機能有較大相關

Source	僅實價登錄		實價登錄+生活機能	
Model	Stacking	Random Forest Method	Stacking	Random Forest Method
Parameters	Random forest, Bagging, Gradient boosting, KNN, SGD, Ridge, XGB	best n_estimators: 94	Random forest, Bagging, Gradient boosting, KNN, SGD, Ridge, XGB	best n_estimators: 90
RMSE	35103	35340	32934	33387
Comparison			較佳, -2169	較佳, -1953

預測模型二：實價登錄+生活機能

1. 金融機構、超商、高中職、大賣場為新增的重要特徵

2. 猜測：金融機構包括銀行、人壽、證券，基本上會在主要幹道附近，因此房價可能較高。

前十大重要特徵：	
金融機構	0.136
屋齡	0.135
建物型態_公寓(5樓含以下無電梯)	0.082
建物移轉總面積平方公尺	0.049
超商	0.045
高中職	0.045
土地移轉總面積平方公尺	0.043
大賣場	0.043
移轉層次	0.043
建材_鋼筋混凝土造	0.035

預測模型三：實價登錄+生活機能+街景/衛星圖

加上街景圖或衛星圖的 VGG Features 後 RMSE 表現不如預期。

Source	實價登錄+生活機能		實價登錄+機能生活+街景/衛星	
Model	Stacking	Random Forest Method	Stacking	Random Forest Method
Parameters	Random forest, Bagging, Gradient boosting, KNN, SGD, Ridge, XGB	best n_estimators: 90	Random forest, Bagging, Gradient boosting, KNN, SGD, Ridge, XGB	best n_estimators: 280 / 240
RMSE	32934	33387	34466 / 33214	35188 / 34161
Comparison			較差, +1532 /較差, +280	較差, +1801 / 較差, +774

預測模型三：實價登錄+生活機能+街景/衛星圖

兩者的前五大重要特徵相同，而衛星圖「Sat3」有進前十重要特徵

證實衛星圖對房價預測比街景圖更有影響，而街景圖對房價預測幫助不大

街景/ 前十大重要特徵：		衛星/ 前十大重要特徵：	
金融機構	0.130	金融機構	0.130
屋齡	0.099	屋齡	0.093
建物型態_公寓(5樓含以下無電梯)	0.091	建物型態_公寓(5樓含以下無電梯)	0.090
高中職	0.038	高中職	0.034
建材_鋼筋混凝土造	0.035	建材_鋼筋混凝土造	0.034
大賣場	0.029	移轉層次	0.028
移轉層次	0.027	鄉鎮市區_大安區	0.024
超商	0.026	大賣場	0.023
鄉鎮市區_大安區	0.025	超商	0.021
建物移轉總面積平方公尺	0.020	Sat3	0.020

預測模型三：檢討

主觀因素	客觀因素

提及房價時，想到什麼？

預測模型三：檢討

主觀因素	客觀因素

提及房價時，想到什麼？

1. 地段 → 便利性 → 機能

2. 新房 → 屋齡 → 實價登錄

3. 電梯 → 大樓 → 建物型態

沒有街景樣貌、周遭環境

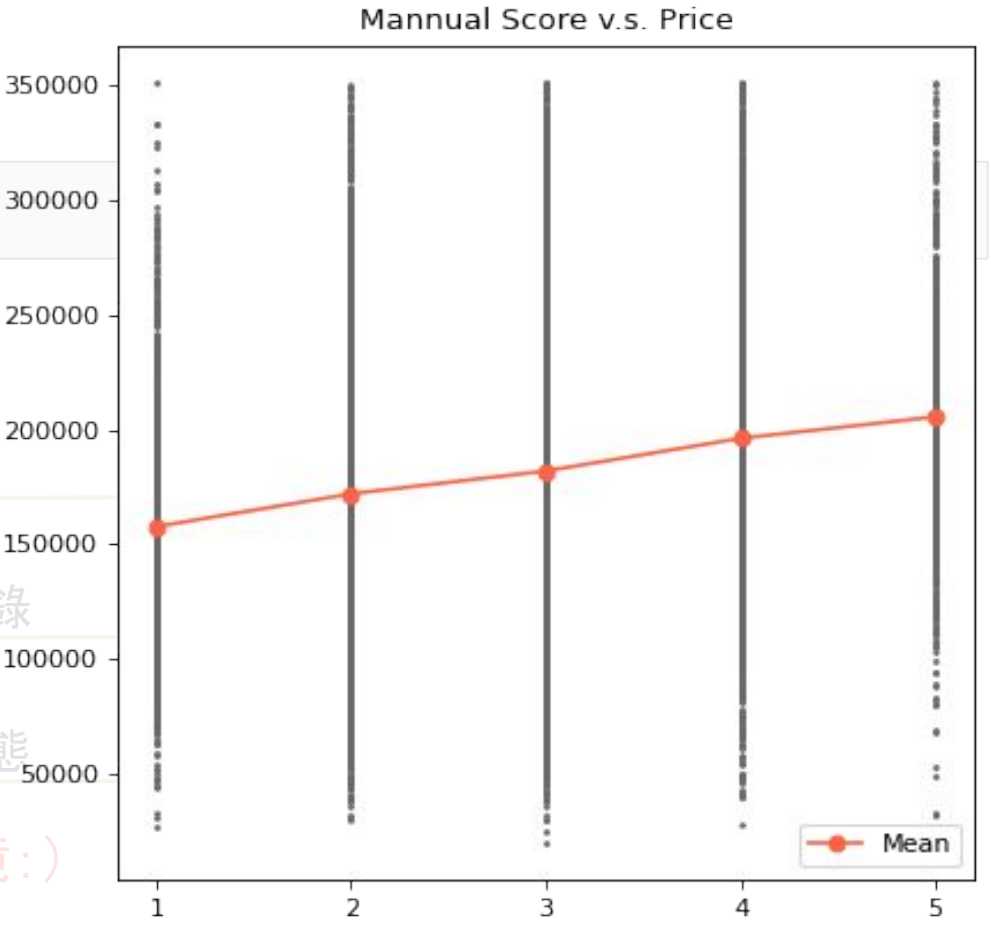
預測模型：檢討

主觀因素

提及房價時，想到什麼？

1. 地段 → 便利性 → 機能
2. 新房 → 屋齡 → 實價登錄
3. 電梯 → 公寓 → 建物型態

沒有街景樣貌、周遭環境：）



預測模型三：檢討

主觀因素	客觀因素
提及房價時，想到什麼？	模型、資料限制
1. 地段 → 便利性 → 機能	1. 分類
2. 新房 → 屋齡 → 實價登錄	
3. 電梯 → 公寓 → 建物型態	

沒有街景樣貌、周遭環境

預測模型三：檢討

主觀因素	客觀因素
提及房價時，想到什麼？	模型、資料限制
1. 地段 → 便利性 → 機能	1. 分類
2. 新房 → 屋齡 → 實價登錄	2. VGG建構
3. 電梯 → 公寓 → 建物型態	

沒有街景樣貌、周遭環境

預測模型三：檢討

主觀因素	客觀因素
提及房價時，想到什麼？	模型、資料限制
1. 地段 → 便利性 → 機能	1. 分類
2. 新房 → 屋齡 → 實價登錄	2. VGG建構
3. 電梯 → 公寓 → 建物型態	3. 區域特性

沒有街景樣貌、周遭環境

預測模型三：檢討

主觀因素	客觀因素
提及房價時，想到什麼？	模型、資料限制
1. 地段 → 便利性 → 機能	1. 分類
2. 新房 → 屋齡 → 實價登錄	2. VGG建構
3. 電梯 → 公寓 → 建物型態	3. 區域特性
沒有街景樣貌、周遭環境	4. 街景更新

總結

- 將不同 Source 排列組合後，比較 RMSE 表現：

  實價+機能 > 實價+機能+衛星 > 實價+機能+街景 > 實價+街景
  > 僅機能 > 僅實價 > 僅衛星 > 僅街景

- 後續改善方向：

  1. Random Forest：Test RMSE >> Train RMSE
     → 原因待探討
  2. 都市房價：機能便利性 >> 街景/衛星圖
     →擴大地區範圍，將全台灣資料納入
  3. VGG model：分類 >> 連續數字
     → 將房價分成五個區隔作為預測變數，取代原先的單位元 / 平方公尺

Reference

- 永慶房屋看屋檢核表
- Fine-Tuning Pre-trained Model VGG-16
- Fine-tuning with Keras and Deep Learning
- Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery
- Take a Look Around: Using Street View and Satellite Images to Estimate House Prices
- Google Places API
- How to Use the Google Places API for Location Analysis and More
- 使用Google Map API (Geocoding API) 得到點位縣市鄉鎮資料
- How to Query Google Street View Static API with Python (UPDATED IN 2020)
- google-street view · PyPI

Thank you