

# 永豐AI GO競賽 攻房戰

---

隊伍名稱：台南GOGO  
Private leaderboard：6.116781 (2<sup>nd</sup>)

吳宇翔，蔡岳霖，王維綱，林芊

# 本次實作之關鍵

---

- 納入關鍵外部資料(實價登錄)
- 對實價登錄資料進行清理
- 加入有效之衍生特徵
- 使用DART演算法

- 可再加入的方法
  - Mapping 資料之方式可增加附屬建物等面積資訊(by 星之卡比)
  - Pseudo label (by Turing team)

# 對實價登錄資料進行清理

---

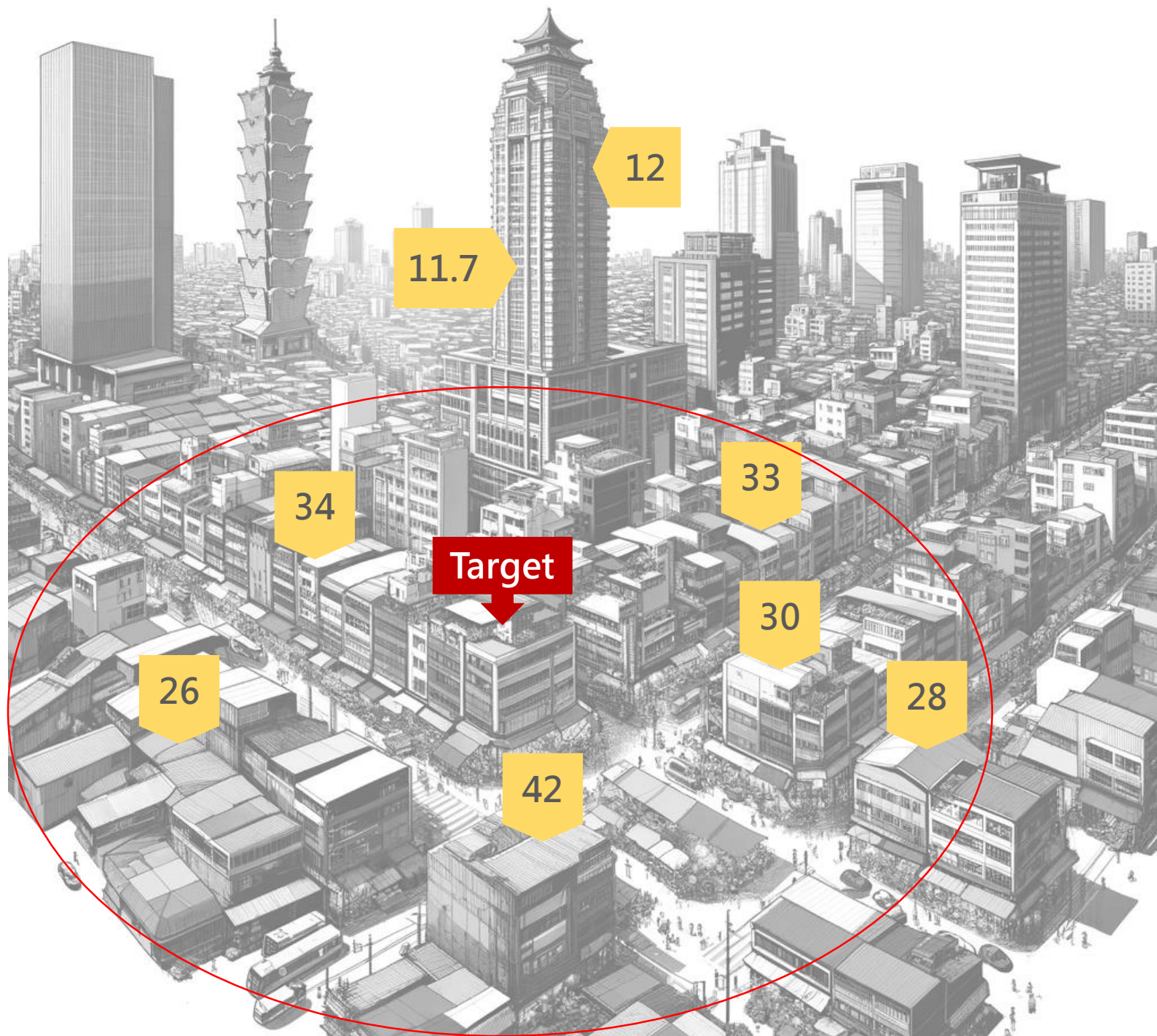
- 以實價登錄2021-22為例
  - 刪除條件
    - 預售屋買賣
    - 非都市土地
    - 交易日期非2021-2022年
    - 非「大樓、華廈、公寓」類別之建物
    - 移轉樓層為1樓或以下
    - 備註中註明“ 特殊交易”
    - 建物面積標準化後z-score>10

# 加入有效之衍生特徵

---

- 訓練與測試資料：
  - 計算每筆資料之間的距離，並且依據不同閾值(500/1000/5000)計算某個範圍內其他資料的不同特徵+統計量

單價(y)與實價登錄亦作相似處理



Age mean  
27.0875

# 使用DART演算法

- 本次使用lightgbm套件，並以相同特徵組合訓練3x2共六個模型後進行averaging得到最後預測結果

**GBDT (2) vs. DART (1) :**  
**差距約在 0.0011 (0.11%)**

編號	超參數設定
1	n_estimators=10000, learning_rate=5e-2, reg_alpha = 1e-2, reg_lambda = 5e-1, max_depth=12, min_child_samples=3, subsample = 0.5, colsample_bytree=0.5, boosting_type = 'dart'
2	n_estimators=10000, learning_rate=1e-2, reg_alpha=3e-1, reg_lambda=3e-1, num_leaves=31, min_child_samples=5, subsample=0.5, colsample_bytree=0.5, subsample_freq=4, boosting_type='gbdt'
3	n_estimators=10000, learning_rate=1e-1, reg_alpha=1e-1, reg_lambda=5e-1, max_depth=12, subsample = 0.5, colsample_bytree=0.5, boosting_type = 'dart', drop_rate = 0.1, skip_drop = 0.8, max_drop = 50,

# 討論

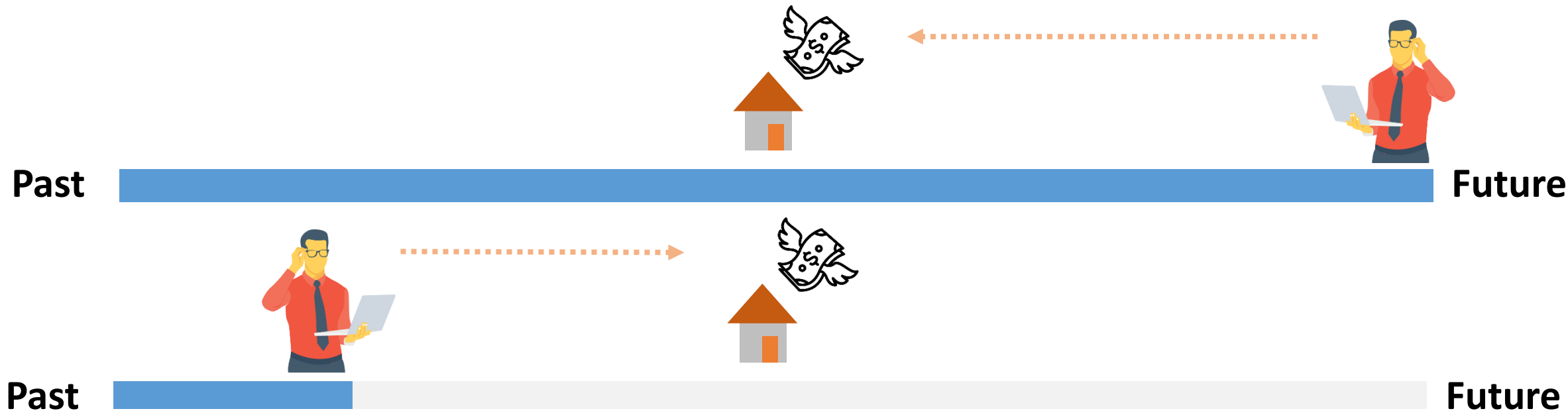
- Q：實價登錄資料的內容比起原本資料產生的衍生特徵還重要？
- A：Yes, 但資料量也是關鍵之一

測試名稱	實價登錄資料量	MAPE(valid)
原始特徵(編號III)	-	0.07637
III + 10%實價登錄資料(2021-22)	23,430	0.07330
III + 50%實價登錄資料(2021-22)	117,148	0.06692
III + 100%實價登錄資料(2021-22)	234,295	0.06254

←約為訓練+測試資料量

# 討論

- Q : data leakage? 實際上線是否會效果會大幅衰退
- A : 衰退幅度取決於實際應用情境





# Thanks for listening

---

Q & A

# 附錄

---

# 外部資料列表

註1：資料為國道標示與其經緯度，  
手動挑選代表各交流道出入口的標誌

註2：此份資料會分別做兩種不同的  
處理產生兩類特徵

## 資料名稱

國小、國中、高中、  
火車站、公車站、  
捷運站、ATM、金  
融機構、便利商店、  
郵局、醫療機構

## 資料名稱

高鐵站點  
機場位置  
快速公路匝道  
國道匝道<sup>註1</sup>

## 資料名稱

監獄位置  
汙水處理廠  
垃圾掩埋場  
焚化爐

## 資料名稱

麥當勞位置  
購物中心位置  
工業區位置

## 資料名稱

實價登錄 2020  
實價登錄 2021-22<sup>註2</sup>  
實價登錄 2023 Q1-Q3

## 其餘曾測試之外部資料：

- 市政府
- 全聯門市
- 火化場

# 資料前處理

---

- 訓練資料
  - 刪除單價較極端的兩筆資料(TR-5660, TR-8800)
- 實價登錄2020 & 實價登錄2023
  - 刪除條件
    - 非都市土地
    - 交易日期非2020年
    - 非「大樓、華廈、公寓」
    - 移轉樓層為1樓或以下
    - 備註中註明“ 特殊交易”
    - 建物面積標準化後z-score > 10

# 資料前處理

---

- 實價登錄2021-22

- 刪除條件

- 預售屋買賣
    - 非都市土地
    - 交易日期非2021-2022年
    - 非「大樓、華廈、公寓」
    - 移轉樓層為1樓或以下
    - 備註中註明“ 特殊交易”
    - 建物面積標準化後z-score>10

- 實價登錄2021-22\_rev

- 刪除條件

- 不動產買賣
    - 非都市土地
    - 交易日期非2021-2022年
    - 移轉樓層為50樓以上

# 特徵工程

---

- 訓練與測試資料：

- Frequency Encoding

- 縣市
    - 縣市+鄉鎮市區

- One-hot Encoding

- 縣市
    - 鄉鎮市區
    - 建物型態
    - 使用分區
    - 主要建材
    - 主要用途

- 衍生特徵

- 主要用途\_relabel (住/商)
    - 移轉層次/總樓層數
    - 每筆資料最鄰近10筆交易的平均距離
    - (id\_2\_count)猜測哪些資料屬於同一棟建築物，加入此建物的交易總筆數、屋齡分布區間、每筆交易在此建物中的交易次序

# 特徵工程

- 訓練與測試資料：
  - 衍生特徵(cont.)
    - 計算每筆資料之間的距離，並且依據不同閾值(500/1000/5000)計算某個範圍內其他資料的不同特徵+統計量

距離閾值	過濾特徵	特徵	統計量
500、1000、5000公尺	無	屋齡	平均、最小值、最大值、數量、平均與此筆資料的差異、最小值與此筆資料的差異、此筆資料的百分位數
		總樓層數	平均、最小值、最大值、平均與此筆資料的差異、最小值與此筆資料的差異
		土地面積	平均、最小值、最大值
		建物面積	平均、最小值、最大值
		主建物面積	平均、最小值、最大值
		陽台面積	平均、最小值、最大值
		附屬建物面積	平均、最小值、最大值
	建物型態	屋齡	平均、平均與此筆資料的差異、最小值與此筆資料的差異
1000公尺	無	路名	資料相同的比例
		<u>房價平均_精準</u>	平均、遺漏值比例
		<u>房價平均_相同屋齡</u>	平均

# 特徵工程

- 訓練與測試資料：

- 衍生特徵(cont.)

- 每筆資料周邊範圍的平均單價(y)，並以特定方式過濾選取的資料。

距離閾值	過濾特徵與標準	計算特徵
1000公尺	建物型態、總樓層數、路名相同，且屋齡差異在+-2以內	平均單價
	建物型態、總樓層數、路名相同	平均單價
	建物型態、路名相同、且屋齡差異在+-5以內	平均單價
	建物型態、路名相同	平均單價
200公尺	建物型態、總樓層數、路名相同，且屋齡差異在+-2以內	平均單價
	建物型態、總樓層數、路名相同	平均單價



# 特徵工程

- 實價登錄：
  - 根據不同篩選特徵組合過濾對應的實價登錄資料，計算部分特徵的統計量。為避免overfitting，計算完後會將此特徵進行quantization ( $q=128$ )

資料名稱	[代稱] / 篩選特徵與標準	未配對比例	計算特徵與統計量
實價登錄資料 (2021Q1-2022Q4)	[屋齡] 縣市、鄉鎮市區相同，且 屋齡差距在3以內	0.043%	平均：單價、屋齡、建物面積、總價、總樓層數、移轉層次、車位總價、車位個數、房間數量  標準差：單價 偏態：單價  其他：資料筆數、訓練/測試資料建物面積高於實價登錄資料建物面積的比例

除此之外，亦對特徵之間進行四則運算以產生其他衍生特徵

# 特徵工程

---

- 周邊重要設施：
  - 設定距離閾值並且計算每筆資料在此範圍內每種重要設施的數量
  - 計算每筆資料距離每種重要設施的最短距離
  - 針對學校類型，除最短距離外會計算此學校的學生總數

# 實價登錄篩選條件與特徵列表

---

資料名稱	[代稱] / 篩選特徵與標準	未配對比例	計算特徵與統計量
實價登錄資料 (2021Q1-2022Q4)	[精準] 縣市、鄉鎮市區、建物型態、移轉層次、總樓層數、路名相同，且屋齡差距0.5以內	55.1%	平均：單價、屋齡、建物面積、總價、總樓層數、移轉層次、車位總價、車位個數、房間數量  標準差：單價  偏態：單價  其他：資料筆數、訓練/測試資料建物面積高於實價登錄資料建物面積的比例
	[同建物] 縣市、鄉鎮市區、建物型態、總樓層數、路名相同，且屋齡差距在0.5以內	26.7%	
	[屋齡_型態_移轉層次] 縣市、鄉鎮市區、建物型態、移轉層次組別(0-4, 5-9, 10以上)相同，且屋齡差距3以內	0.88%	
	[移轉] 縣市、鄉鎮市區、建物型態、移轉層次組別相同	0.009%	
	[屋齡_型態_路名] 縣市、鄉鎮市區、建物型態、路名相同，且屋齡差距在3以內	8.55%	
	[屋齡_型態] 縣市、鄉鎮市區、建物型態相同，且屋齡差距3以內	0.83%	
	[路名_型態] 縣市、鄉鎮市區、建物型態、路名相同	3.73%	
	[路名] 縣市、鄉鎮市區、路名相同	2.14%	
	[型態] 縣市、鄉鎮市區、建物型態相同 (註1)	0.085%	
	[屋齡] 縣市、鄉鎮市區相同，且屋齡差距在3以內	0.043%	

註1：此篩選條件下的單價平均後續未放入模型中

資料名稱	[代稱] / 篩選特徵與標準	未配對比例	計算特徵與統計量
實價登錄資料 _rev(2021Q1- 2022Q4)	[路名_rev] 縣市、鄉鎮市區、路名相同	65.78%	平均：單價、建物面積、總價、 車位總價、車位個數、房間數量  其他：資料筆數
	[型態_rev] 縣市、鄉鎮市區、建物型態相同	18.93%	
實價登錄資料 (2020Q1- 2020Q4)	[路名_型態2020] 縣市、鄉鎮市區、建物型態、路名相同	14.95%	平均：單價、建物面積、總價  其他：資料筆數
	[同建物2020] 縣市、鄉鎮市區、建物型態、總樓層數、路名相同	23.47%	
	[精準2020] 縣市、鄉鎮市區、建物型態、總樓層數、路名相同，且實價登錄 與訓練/測試資料的屋齡差距介於-3到0之間	33.88%	
實價登錄資料 (2023Q1- 2023Q3)	[路名_型態2023] 縣市、鄉鎮市區、建物型態、路名相同	16.38%	平均：單價、建物面積、總價  其他：資料筆數
	[同建物2023] 縣市、鄉鎮市區、建物型態、總樓層數、路名相同	27%	
	[精準2023] 縣市、鄉鎮市區、建物型態、總樓層數、路名相同，且實價登錄 與訓練/測試資料的屋齡差距介於0到3之間	37.95%	

# 實價登錄衍生特徵列表

- 建物面積差距
  - 路名\_型態、屋齡、精準、屋齡\_型態\_路名 篩選條件下的建物面積平均 - 訓練/測試資料的建物面積
- 屋齡差距
  - 路名\_型態、精準、路名、屋齡\_型態\_路名 篩選條件下的屋齡平均 - 訓練/測試資料的屋齡
- 不同篩選條件下價格平均之差距
  - 精準 - (路名\_型態、屋齡、屋齡\_型態\_路名)
  - 屋齡 - (型態、屋齡\_型態、屋齡\_型態\_移轉層次)
  - 路名 - (路名\_型態、移轉、屋齡)
  - 路名\_型態 - (型態、移轉、屋齡\_型態\_路名、路名\_型態2020、路名\_型態2023)
- 比率
  - 資料筆數： $\text{路名\_rev} / (\text{路名} + \text{路名\_rev})$
  - 資料筆數： $\text{型態\_rev} / (\text{型態} + \text{型態\_rev})$
  - 價格： $\text{路名\_rev} / (\text{路名} + \text{路名\_rev})$

# 演算法與超參數


- 演算法

- 本次使用lightgbm套件，並且以相同特徵組合分別訓練3x2共六個模型後進行averaging得到最後預測結果
- 在驗證階段，則以8:2方式切割驗證資料，並未做K-Fold CV
- 超參數請見右表

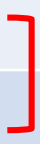
編號	超參數設定
1	n_estimators=10000, learning_rate=5e-2, reg_alpha = 1e-2, reg_lambda = 5e-1, max_depth=12, min_child_samples=3, subsample = 0.5, colsample_bytree=0.5, boosting_type = 'dart'
2	n_estimators=10000, learning_rate=1e-2, reg_alpha=3e-1, reg_lambda=3e-1, num_leaves=31, min_child_samples=5, subsample=0.5, colsample_bytree=0.5, subsample_freq=4, boosting_type='gbdt'
3	n_estimators=10000, learning_rate=1e-1, reg_alpha=1e-1, reg_lambda=5e-1, max_depth=12, subsample = 0.5, colsample_bytree=0.5, boosting_type = 'dart', drop_rate = 0.1, skip_drop = 0.8, max_drop = 50,

# 不同來源特徵重要程度

編號	使用特徵	特徵數量	MAPE(valid)
I	訓練資料之特徵與其衍生特徵 (不含使用預測目標製作之特徵)	281	0.07705
II	編號I加上使用預測目標所製作之衍生特徵	286	0.07753
III	編號II加上周邊設施特徵 (包含主辦單位提供之資料與自行取得之資料)	344	0.07637
IV	編號III加上實價登錄資料(2020年)	356	0.07436
V	編號IV加上實價登錄資料(2021-2022年)	503	0.06267
VI	編號V加上實價登錄資料(2023Q1-Q3) 即為完整特徵組合(leaderboard上之提交)	516	0.06266



0.00201



0.01169



# 不同篩選條件對MAPE之影響

代稱	MAPE(valid)	MAPE差異(較編號III之結果)
同建物	0.06776	-0.00861
精準	0.06908	-0.00728
屋齡	0.06908	-0.00728
屋齡_型態_路名	0.06985	-0.00652
路名	0.07160	-0.00477
路名	0.07205	-0.00432
屋齡_型態	0.07350	-0.00287
屋齡_型態_移轉層次	0.07356	-0.00281
型態	0.07452	-0.00185
移轉	0.07462	-0.00175