

房屋估價預測

國立政治大學統計碩二許振榆



c o n t e n t

錄

1 資料敘述

視覺化

2 特徵工程

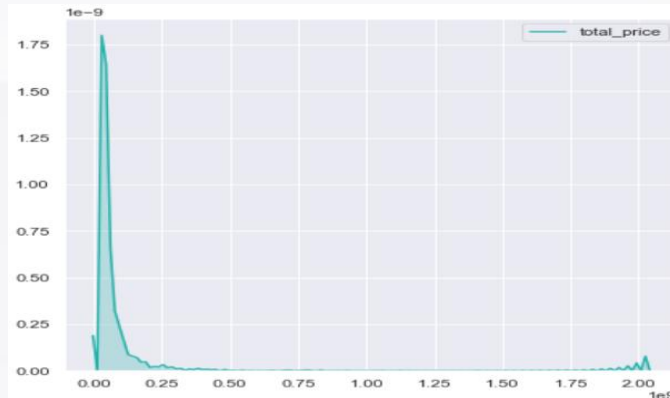
資料預處理、變數篩選

3 模型

建立及衡量結果

資料敘述

應變數



total_price分佈圖

偏度(skewness):24.84

峰度(kurtosis):945.307

資料為明顯的右偏分配
且具有極大的離群值

特徵工程

自變數

物件本身條件

(連續型) total_floor,
parking_area, parking_price,
land_area, building_area

(名目型) building_material,
building_type, building_use,
parking way

物件外部環境

(連續型) town_population, lat,
lon, village_income_median...

(名目型) city, town, village

交易資訊

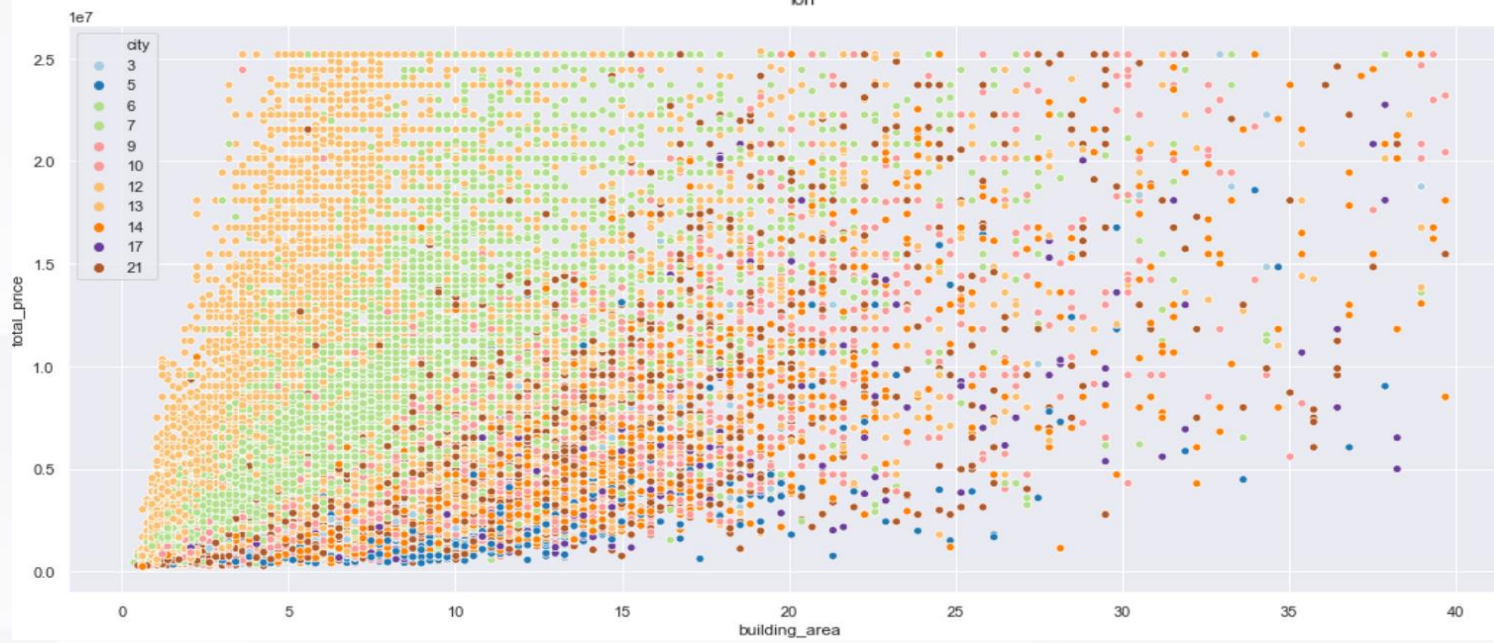
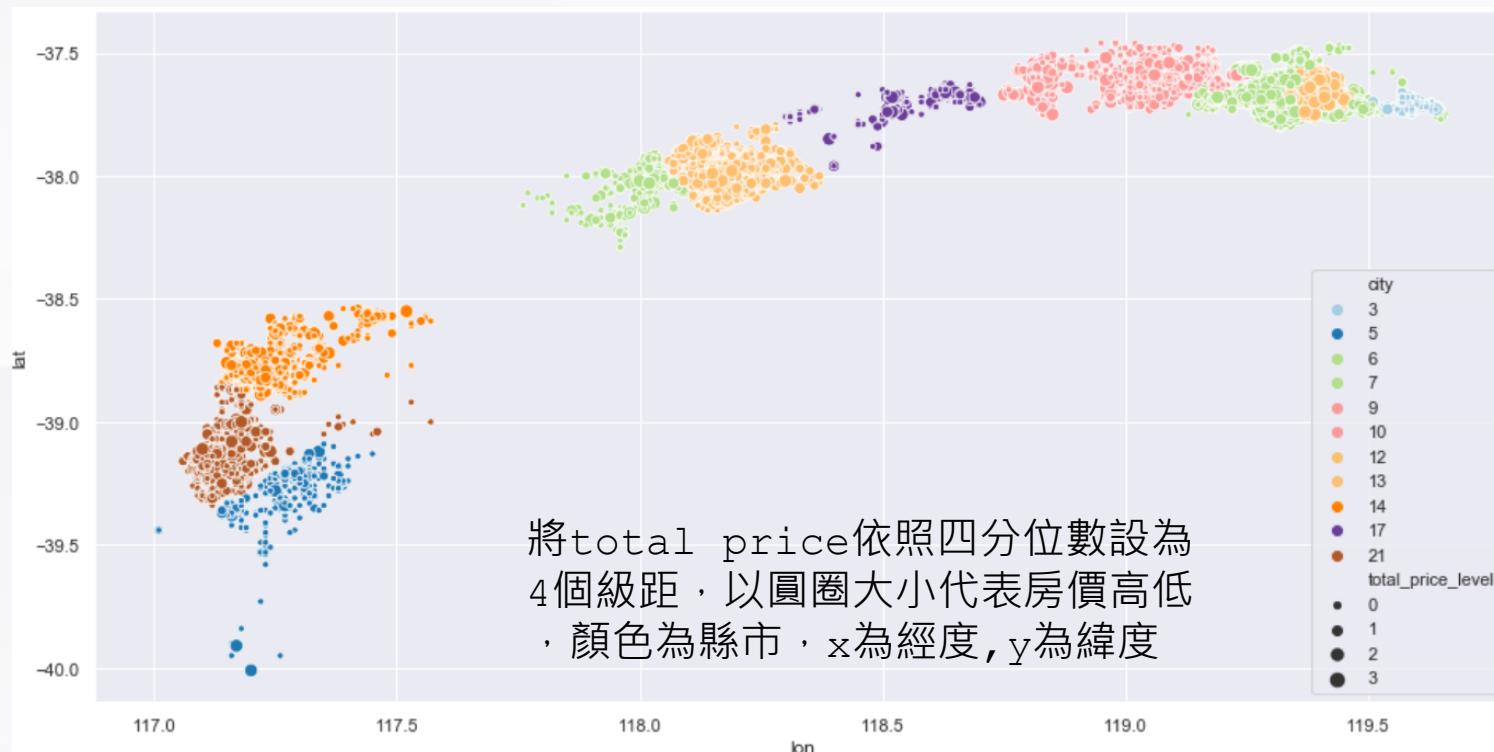
(連續型) txn_dt, txn_floor,
building_complete_dt

模型

資料敘述

特徵工程

模型



畫出不同**城市地理位置及房價關係圖**，可觀察到圖中右上角的城市，有較多高房價物件，左下角以21號棕色城市為附近房價較高的地區

建物面積及房價成正向關係，不同縣市房價對面積斜率明顯不同，推測帶有**地理位置資訊**的變數可能會是重要特徵

資料預處理

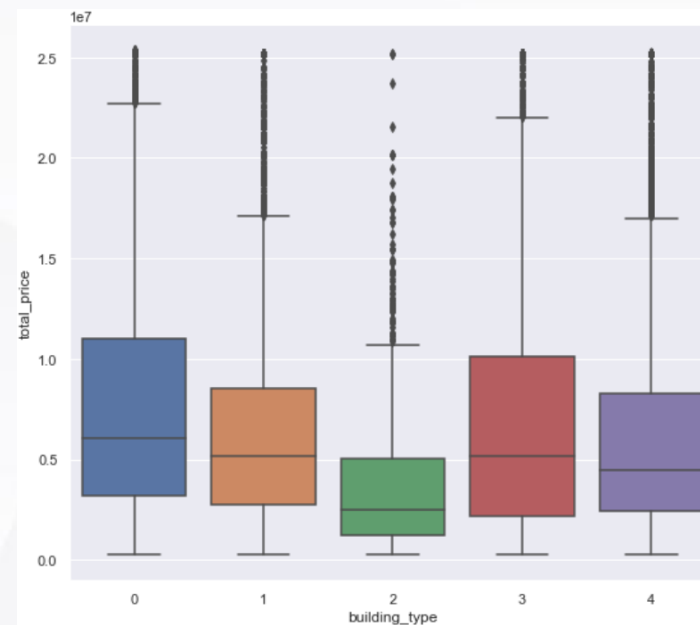
資料敘述

1. 缺失值處理

- **village_income_median**與地緣相關，使用**knn**鄰近法以lat, lon, village做**填補**
- 計算與**txn_floor**相關係數較高的total_floor, building_material, building_type, parking_way, building_complete_df做**MCMC填補**
- **parking_area**缺失值比例達95.16%，**parking_price** 缺失值佔77.84%，兩變數皆有大量缺失值，且資料集中變數parking_way可提供與停車相關的資訊，因此直接**刪除**

2. 名目型變數 (無序) 先刪減類別，再做虛擬變數

對房價具有相似分佈的類別進行合併以減少類別數量。以右圖5種**建物型態**對房價的盒狀圖為例，2號房價分佈較低，0, 3及1, 4分佈較接近，將該名目變數降為3類別，再以OneHot建立3個虛擬變數



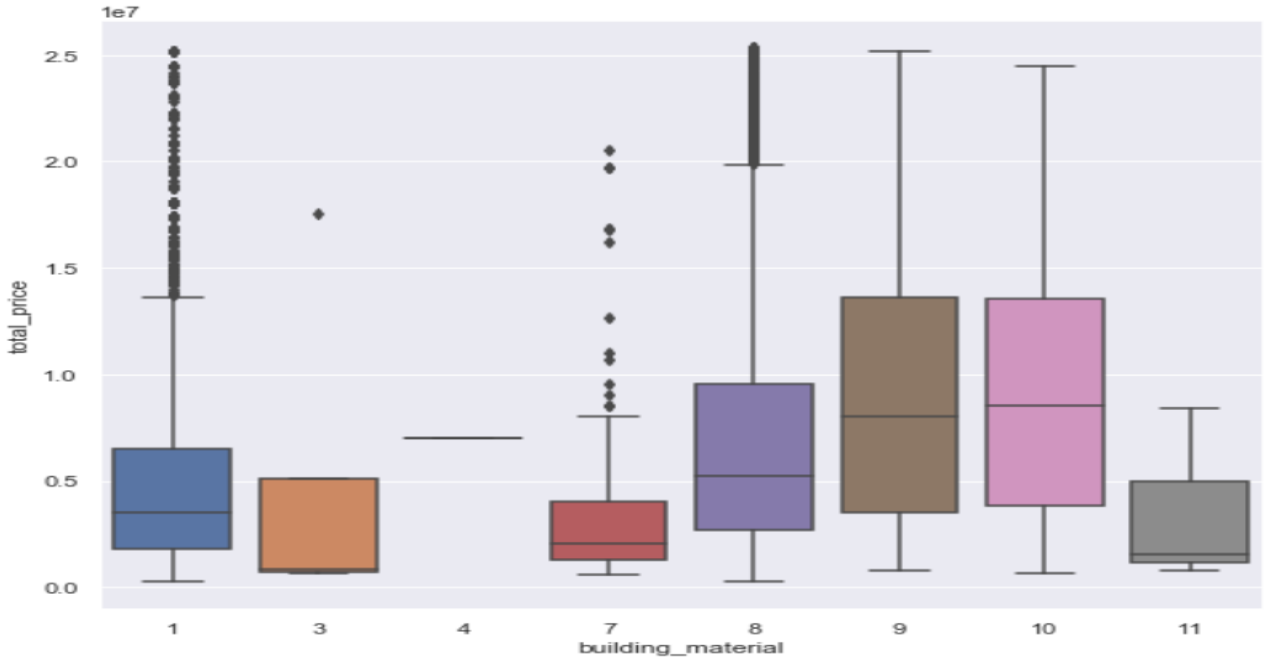
特徵工程

模型

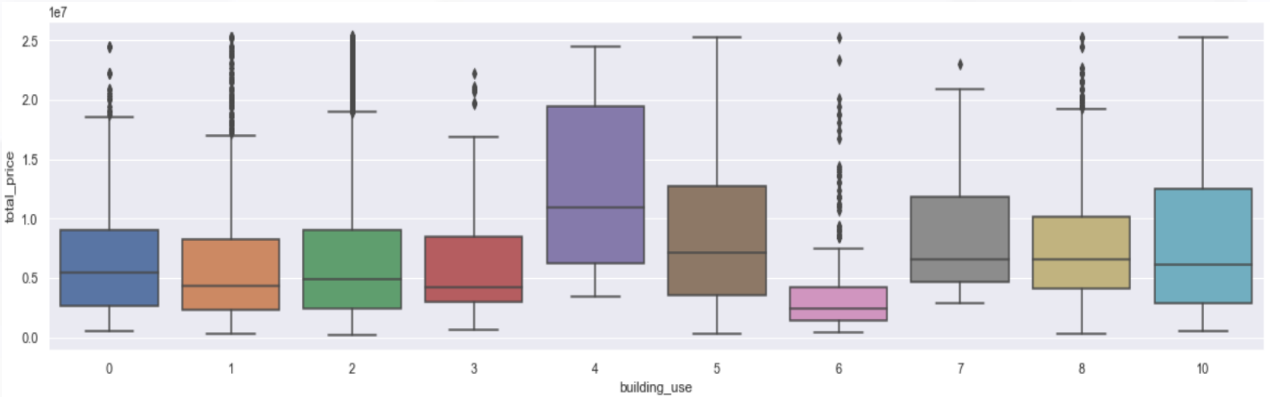
資料敘述

特徵工程

模型



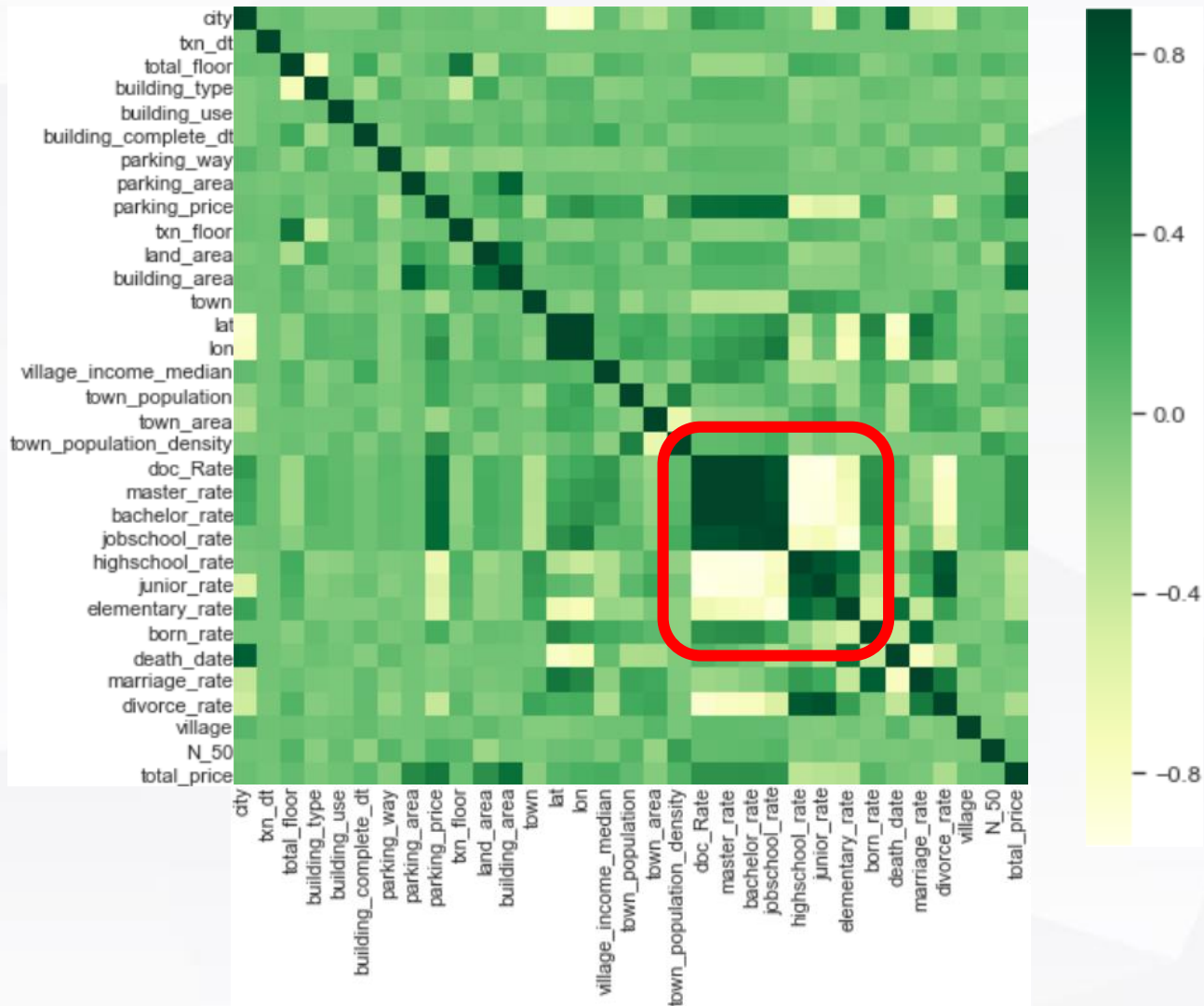
建材及建物用途合併類別如右表，分別從8類別降為5類別及10類別降為4類別



變數	原始類別	合併後類別
Building type	0, 3	0
	1, 3	1
	2	2
Building material	1, 8	1
	3, 11	3
	4	4
	7	7
	9, 10	10
Building use	0, 1, 2, 3, 7, 8	0
	4	4
	5, 10	5
	6	10

3. 篩選變數降維

相關係數矩陣



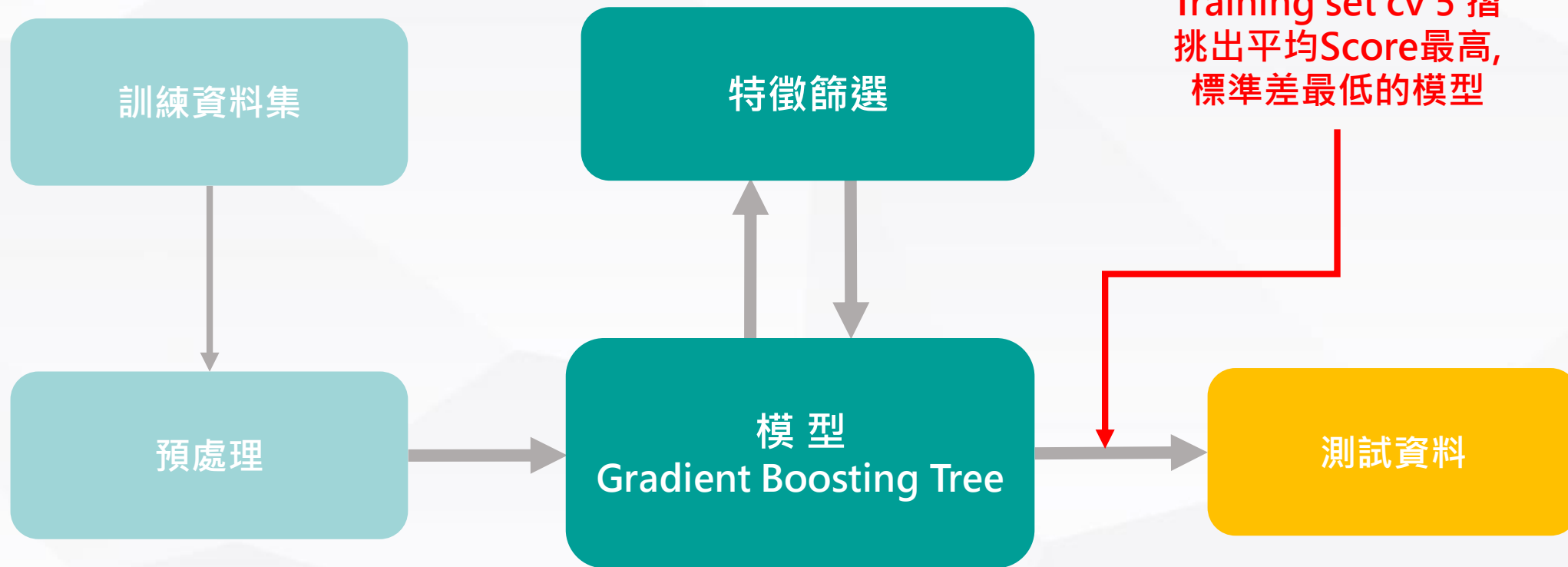
- doc_Rate, master_Rate, bachelor_Rate, jobschool_rate, highschool_rate, junior_rate, elementary_rate 此類11個縣市的各級教育比例具有高度相關性
- born_rate, death_rate, marriage_rate, divorce_rate 為11個縣市描述人口增減婚姻的特徵
- 上述變數用來描述11個縣市的狀況，為避免與city提供重疊資訊，將此11個變數刪除

建模流程

資料敘述

特徵工程

模型



- ✓ `sklearn.GradientBoostingRegressor()`
 - ✓ `xgboost.XGBRegressor()`
 - ✓ LightGBM
- 以5摺cv tune 超參數

特徵篩選

資料敘述

特徵工程

模型



- 舉LightGBM為例，左圖以變數作為節點次數繪圖，擔任節點愈多次的變數視為LightGBM模型的重要特徵
- **建物面積, 交易日期** 為最被頻繁使用的特徵，其他如與**鄰近特定類別的最短距離**也常被拿來當作樹節點
- 最終重要特徵的個數將會以Grid Search篩選挑出


參數: param = {'objective': 'mape', 'feature_fraction': 0.9, 'learning_rate': 0.05, 'max_bin': 250, 'max_depth': 80, 'min_data_in_leaf': 5, 'num_iterations': 1600, 'num_leaves': 480}

模型衡量

資料敘述

特徵工程

模型

模型	5摺平均分數 (標準差)
GradientBoostingRegressor	4581.4 (20.42)
XGBRegressor	4935.4 (49.75)
LightGBM	3207.5 (61.61)
Model_Assembling_1	5022.6 (43.46)
Model_Assembling_2	5023.2 (44.82) 

平均分數最高，且標準差相較其他模型偏低，屬於較為穩健的模型

$\text{Score} = \text{round}(\text{Hit Rate}, 4) * 10^4 + (1 - (1 \text{ if MAPE else MAPE}))$

- ◆ **Model_Assembling_1** : 將GBR, XGBR, LightGBM 預測出的Y值做平均
- ◆ **Model_Assembling_2** : 計算各模型在訓練資料集mape，以mape倒數比例 (0.349, 0.383, 0.268) 作為權重，對三模型預測值權重加總