房價預測

臺大國企所 陳帝文 r10724050

摘要

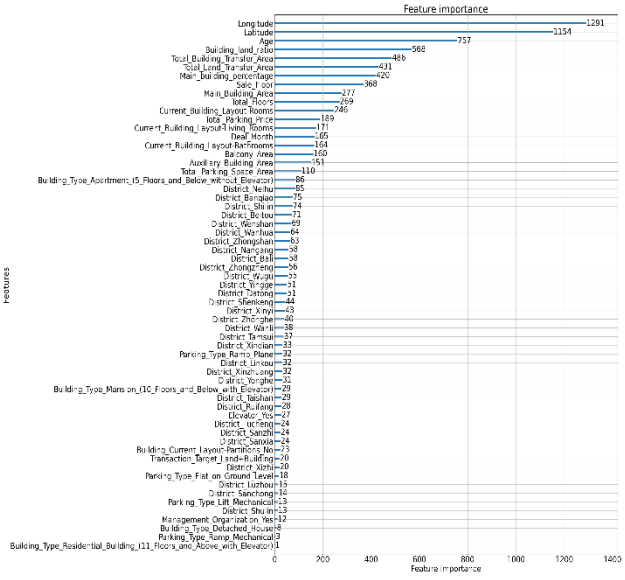
本次房價資料含有臺北市和新北市，直觀上可以感覺到兩縣市的房價和房屋特徵差異較大，屬於不同分布，需要分開處理，因此我參考 DiCiurcio et al. [1] 使用的兩階段機器學習方法。首先對價格使用 Kmeans 分聚後得到兩群並且標註，接著使用資料的特徵去做 classification。得到兩組資料後再分別建立兩個房價預測模型，最後合併得到最終結果。

1. 數據觀察

- *特徵觀察*

首先，我們可以觀察到資料中的特徵有些直覺上與房價沒有關係的特徵，例如 **deal day**, **built month**, **built day** 等，這些特徵我優先排除。另外，也有一些特徵和其他特徵完全線性相依，例如**電梯_有**，**電梯_無**等，這些特徵我會保留「有」的特徵。

接著可以觀察到有大量 one hot encoded 的特徵，例如臺北和新北各**鄉鎮市區**，**建物型態**，**車位類別**等。為了確認這些特徵是否含有「不可調整」成其他代表方式的資訊，我先將資料丟入 lightGBM 做初步迴歸觀察各特徵的重要程度，結果如圖(一)。



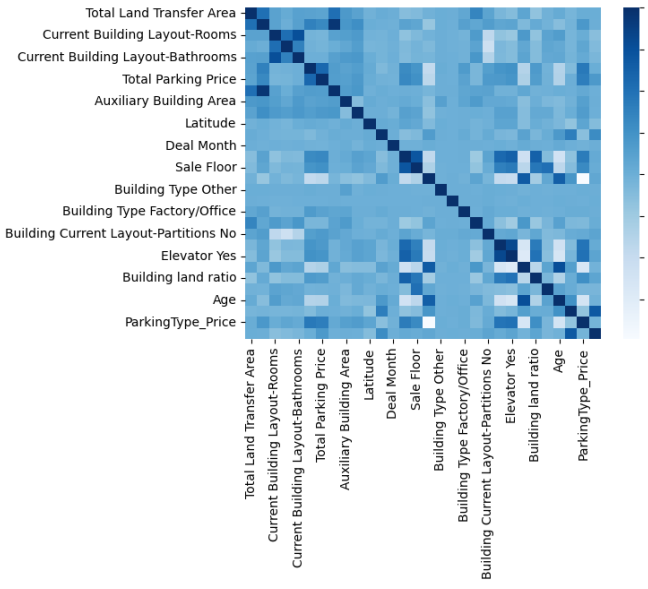
圖(一) 各特徵重要程度

從圖(一)中可以觀察到這些特徵對房價預測的重要性偏低，因此我將各鄉鎮市區和車位類別使用 target encoded 算出各鄉鎮市區和車位類別的房價平均來代表。至於建物型態中有幾個類別和 total floor 高度相關，因此我選擇刪除這些行，只保留店面，廠辦和透天厝的特徵。

最後，由於我使用的 lightGBM 模型屬於 tree model，並不會自動建立特徵間的交叉項，因此我自行加入了隱含公設比資訊的 **Main building percentage** 和 **Building land ratio**，還有該物件所在樓層和整棟樓高度的比例，**Living floor ratio**。我也加入 deal year 扣掉 built year 形成的 **Age**，方便更直觀的觀察房屋年齡和房價的關係，並且刪除前兩特徵。

- *異常值處理*

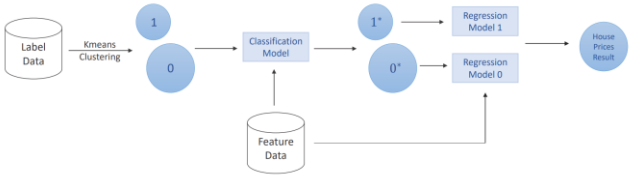
首先，可以觀察到有大量房屋的 **built year** 是 0，我認為這是錯誤的建照年分，因此我選擇將這些資料以中位數代替。接著可以看到有大量資料的 **total floor** 和 **sale floor** 為 0，這些資料我選擇以後面的建物型態特徵代替，以住宅大樓(11 層含以上有電梯)為例，我將屬於這類的 total floor 和 sale floor 設成 11 以貼近真實。最後，我也選擇把少量**土地移轉總面積**為零的訓練資料刪除。



圖(二) 各特徵與價格的相關性

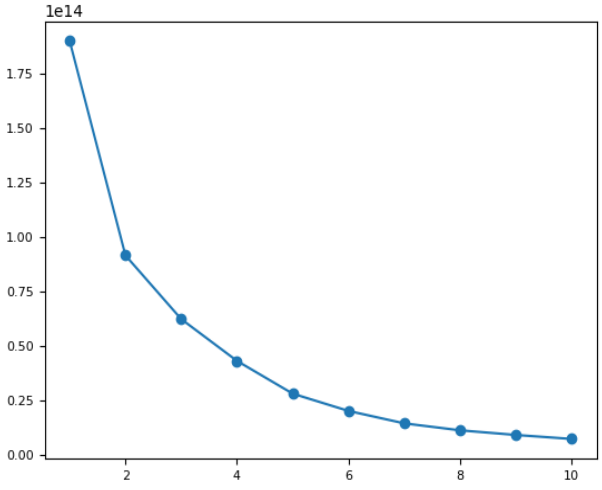
2. 實驗方法

前面有提到我先使用 lightGBM 初步實驗。觀察結果發現新北市一些偏遠鄉鎮市區價格嚴重被高估，例如石門區等，而臺北市大安區等則是嚴重被低估，因此我參考 DiCiurcio et al. [1]，使用兩階段機器學習方法來做此次預測，整個流程如圖(三)所示。



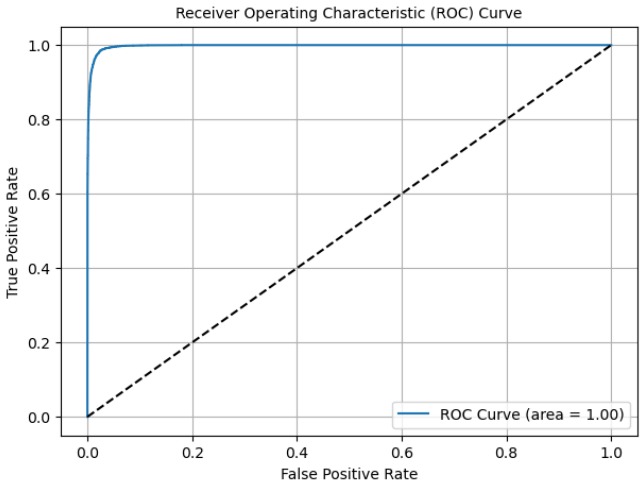
圖(三) 模型流程圖

首先，我對 label 使用 Kmeans 觀察需要被分成幾聚，結果如圖(四)。可以觀察到將資料分成兩聚時出現較明顯的肘點，因此在此就將其分成兩聚。



圖(四) Kmeans 分聚結果

接著，我使用特徵資料並且建立一 lightGBM 分類模型對此兩聚類進行預測，得到一預測準確度高達 98.1%的模型，圖(五)顯示了此分類模型的 ROC curve。深度的對此分類模型預測的結果進行觀察，可以觀察到被預測成 class 0 的資料大多是新北市的房子，而預測為 class 1 則是臺北市加上板橋等房價相對高的區域，可見我們可以從特徵中挖掘出一些隱狀態，幫助我們對房價進行預測。



圖(五) 分類結果

最後，我對預測出來的兩群分別建立 lightGBM 迴歸模型得到的預測數值後依照 index 重新合併，此為最終結果。

3. 實驗結果與分析

- *不分群預測結果*

表(一)展示了不分群下直接使用模型預測房價，訓練資料中 MAE 前五大的鄉鎮縣市和總訓練資料的 MAE。由表中可以看到只有四筆資料的石門區誤差極大，仔細觀察後發現石門區的預測結果嚴重高估。另外，信義區，中正區等高價區則是被嚴重低估，因此有分開處理的打算。

鄉鎮縣市	MAE
士林區	19152.24
大安區	19315.72
中正區	19728.81
信義區	21251.44
石門區	22977.26
訓練資料 MAE: 12160.59	

表(一) 不分群模型的 MAE 比較

- *分群預測結果*

首先，觀察從分聚和分類預測出的 0 和 1[1]，可以發現大多數新北市的房子都被分類在 0* 群，而大部分臺北市的房子和新莊，板橋等新北高房價區則是被分類在 1* 群。表(二)則呈現

[1] 由於版面限制，欲檢視者可以查看 ipynb 檔。

了先分群再預測的模型下的 MAE 前五大鄉鎮市區和總訓練集的 MAE。和表(一)相比，表(二)中信義，中正，大安的誤差大幅下降且總訓練集的 MAE 也大幅下降。

鄉鎮縣市	MAE
大同區	12292.83
中正區	12473.18
松山區	12534.11
信義區	12811.44
大安區	13928.54
訓練資料 MAE: 8485. 91	

表(二) 分群模型的 MAE 比較

雖然兩階段模型的預測結果大幅進步，但從分類結果中可以看到仍有少數新北市較偏遠的地區被分類到1*群，像是石門區有一個，蘆洲有兩個，金山有一個等。因此我決定使用武斷的方式將訓練集分成臺北市和新北市，再訓練兩個預測模型。

- *武斷分群預測結果*

結果[2]顯示用此方法下，大安區，信義區等臺北市的房價誤差相較於前一個方法高，但新北市的各鄉鎮市區則是下降，比較明顯的有淡水區，林口區和中和區等，如表(三)所示。我猜測由於此訓練集大多數都是新北市的房子，若使用此方法，臺北市的訓練集則減少，因此臺北市房價誤差上升。

區域	模型	武斷	數量
淡水	4920	4823	2322
林口	5015	4981	1104
中和	8297	8136	2036

表(三) 新北市部分區域 MAE 變化

由於武斷的方法可以將新北市的房屋價格預測的不錯，加上資料中大部分屬於新北市，因此我決定結合前一個分群法和武斷分群法，有點類似 ensemble 的概念得出最終結果。

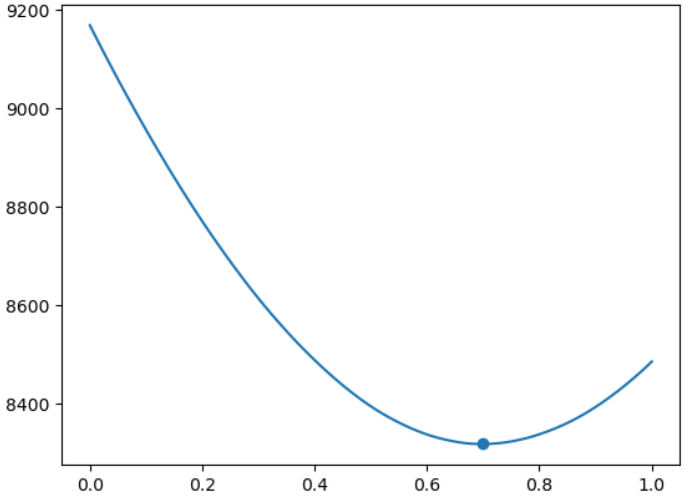
- *集成模型*

此處集成模型為兩個模型建立，分別標示為 $f_1(x)$ 和 $f_2(x)$。而我所需解的如下方程式表示，

$$\min w_1 f_1(x) + w_2 f_2(x)$$
$$\text{s.t } w_1 + w_2 = 1$$

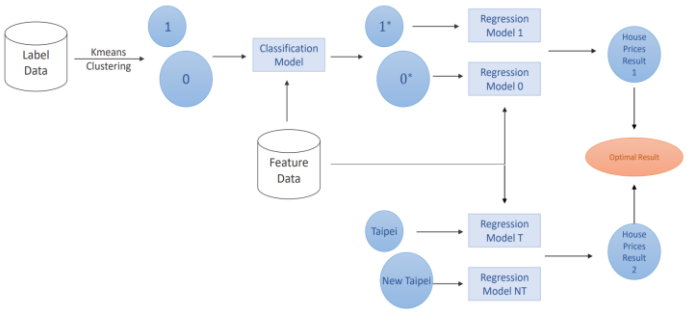
此處為了方便起見，我使用 brute force 法算出 $w_1 = 0, 0.01, 0.02, \ldots, 1$ 下，所有訓練集中的 MAE，得到圖(六)。最後，將得到的最小訓練 MAE 下的權重當最佳權重得到最終結果。

從結果中可以看到，大安，信義，松山等房價誤差雖然仍比模型分群的模型大，但比武斷模型預測的進步。更重要的是，不知道什麼緣故，新北市各區的訓練資料房價在集成模型中預測誤差比武斷模型來的小很多，結果在 ipynb 檔中可以觀察到。



圖(六) 各權重下的訓練集 MAE

最後，圖(七)為最終預測的流程圖。



圖(七) 最終模型流程圖

4. 總結與討論

此次作業中，我認識到使用兩階段模型是

[2] 由於版面限制，欲檢視者可以查看 ipynb 檔。

如何找出資料中的隱狀態並且藉由此優化預測結果。但是，不可避免地，本次實驗仍然存在一些問題。首先，由於訓練資料太少，尤其是在分群後1*只會剩下 6000 筆資料，若是我再事先將資料分成 training data 和 validation data 則會造成誤差太大，自認為用此調參數也失去意義。因此，我選擇使用全訓練集訓練，並且使用 public testing dataset 來觀察是否 overfitting。需注意的是，我使用 public testing dataset 並不是用來調整參數，而是用來觀察上述不同模型想法是否有更佳。表(四)是各種模型在 public testing dataset 的表現。

另外，集成模型使用的是訓練集資料下的最優 MAE 所得到的權重，有 overfitting 的風險。但由於我將此最佳權重丟入 public testing dataset 後得到的結果仍比模型分群，武斷分群和權重各半優異，因此我選擇相信此罪加權重。

模型	testing dataset MAE
不分群	62338.63
模型分群	62210.34
武斷分群	61665.12
最佳權重	61101.50

表(四) 各模型在 public testing dataset 之 MAE

5. 參考文獻

[1] DiCiurcio, K. J., Wu, B., Xu, F., Rodemer, S., & Wang, Q. (2024). Equity Factor Timing: A Two-Stage Machine Learning Approach. *The Journal of Portfolio Management*, *50*(3), 132-148.