**房價預測**

臺大國企所 陳帝文 r10724050

**摘要**

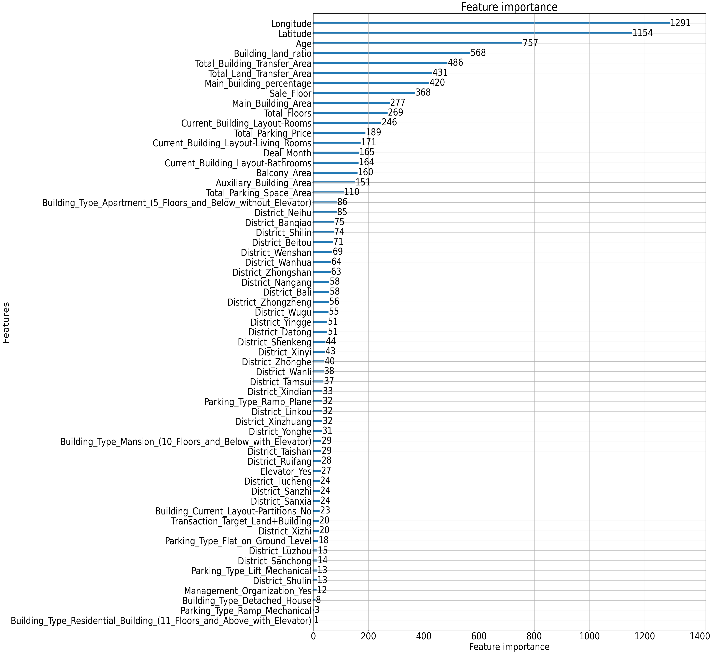
本次房價資料含有臺北市和新北市，直觀上可以感覺到兩縣市的房價和房屋特徵差異較大，屬於不同分布，需要分開處理，因此我參考 DiCiurcio et al. [1] 使用的兩階段機器學習方法。首先對價格使用Kmeans分聚後得到兩群並且標註，接著使用資料的特徵去做 classification。得到兩組資料後再分別建立兩個房價預測模型，最後合併得到最終結果。

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

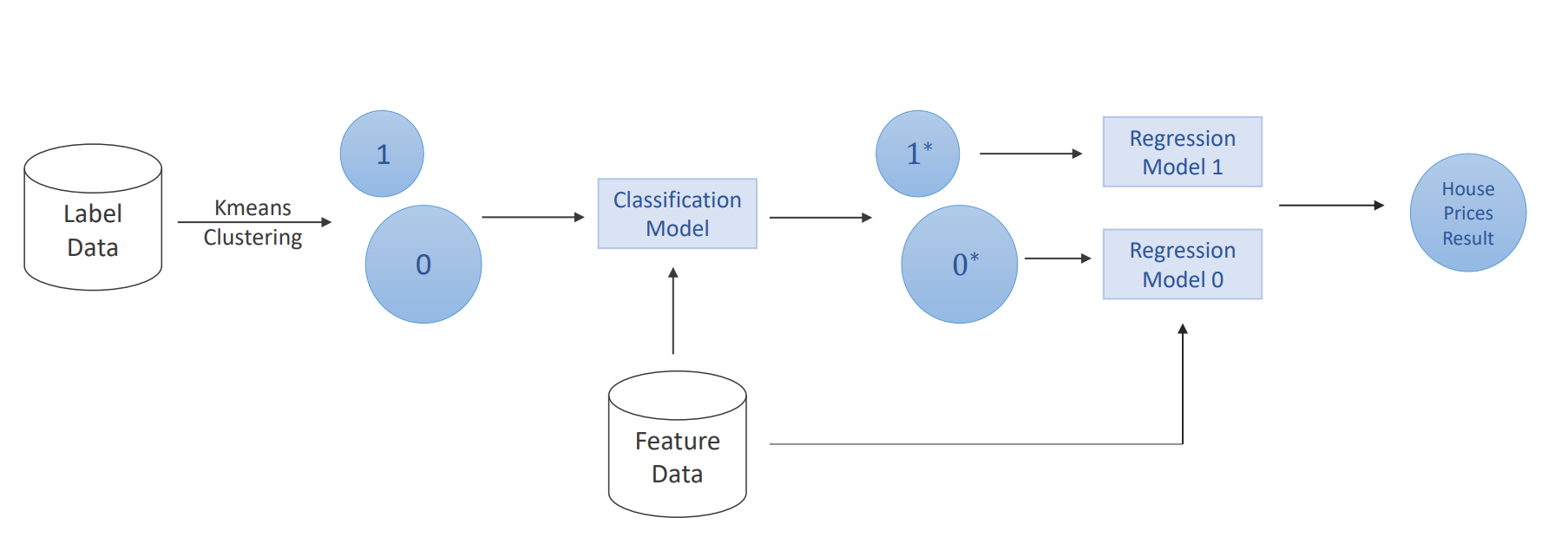
從圖(一)中可以觀察到這些特徵對房價預測的重要性偏低，因此我將各鄉鎮市區和車位類別使用target encoded算出各鄉鎮市區和車位類別的房價平均來代表。至於建物型態中有幾個類別和total floor高度相關，因此我選擇刪除這些行，只保留店面，廠辦，農舍和透天厝的特徵。

* *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。

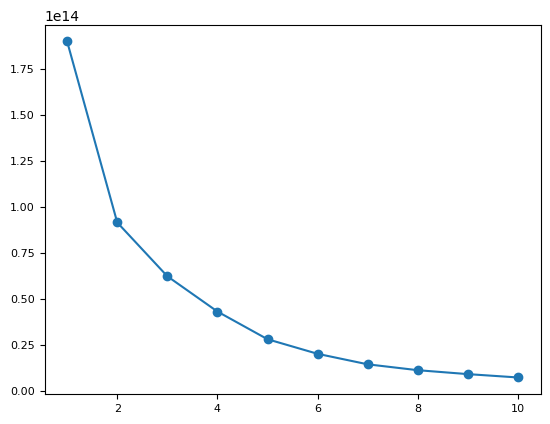
1. **實驗方法**

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(二)所示。



圖(二) 模型流程圖

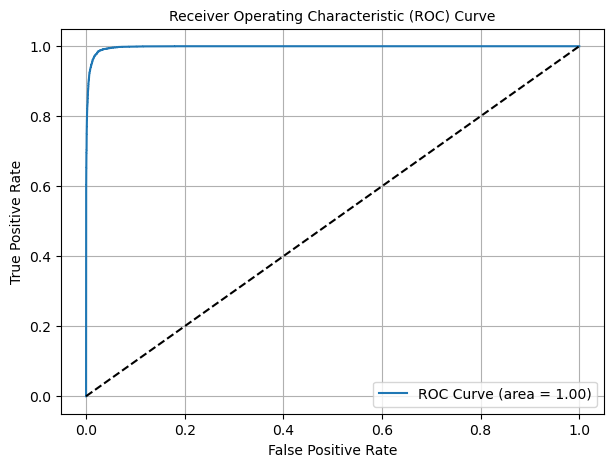
首先，我對label使用Kmeans觀察需要被分成幾聚，結果如圖(三)。可以觀察到將資料分成兩聚時出現較明顯的肘點，因此在此就將其分成兩聚。



圖(三) 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(四)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。

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。



圖(四) 分類結果

1. **實驗結果與分析**