大數據、機器學習與程式交易

Automatic Valuation Systems of Real Estate

─ House Price Prediction using Machine Learning Models

指導老師: 謝明華 教授
組員: 楊宗泰 李郁萱 許雅筑 鄭博仁

目錄

一. 研究介紹

(一) 痛點

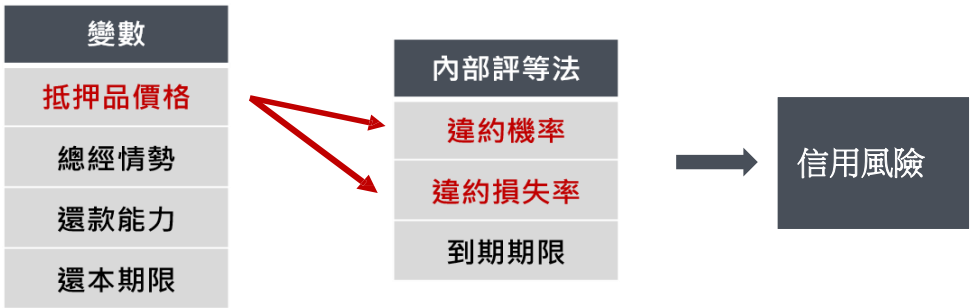
綜觀全球，尤其已開發國家多面臨房價高居不下的挑戰，也因此對一般民眾，購買屬於自己的一間房子是屬於非常重要的花費，且往往需要向金融機構進行借貸，希望透過此專案，利用機器學習模型對波動性大的房價做較精準的預估，協助達成以下目標:

1. 避免銀行過度放貸
2. 銀行即時掌握不動產價格

(二) 目的與應用

以 2008 次貸危機為例，主因為 2000 年以來美國房地產價格持續走揚，使得銀行讓信用評等不佳的借款人也獲得貸款，發生過度放貸或超額貸款的情形。加上當時缺乏即時掌握不動產價格的觀念和系統，造成全球嚴重的金融危機。也在此事件後，銀行和放款機構意識到監控價格的觀念，已進行放貸評估和避險的機制。近年由於大數據和機器學習的技術蓬勃發展，使得具即時性且大量運算的「自動估價模型」應運而生，期望透過本次研究即時預估房價，協助銀行判斷房屋價值是否低於房貸現值，以供銀行房貸部門做運營決策。

由國際清算銀行的巴塞爾銀行監理委員會新修改的的新巴塞爾資本協定（Basel II），強調目標為標準化風險控管制度，以提升金融機構內部風險控管能力。台灣也不例外，金管會預計 2022 年上路新巴塞爾資本協定，台灣的銀行可使用內部評等法來評估信用風險，銀行建立的不動產自動估價系統可作為風險指標之一，輔助銀行自行建構風險評估模型。



(三) 自動估價系統

利用估價標的物的成交案例，配合時間、交易資訊與不動產其他屬性資料，利用大數據建立估價模型，提供不動產物件的價格預測，讓銀行體系能即時掌握不動產價格，以進行核貸評估作業。針對自動估價系統，具有以下優點與價值:

1. 客觀、省時、低成本:
   建立不動產估價共同基礎，減少人為因素與溝通成本，同時降低不動產重估之鑑價成本及時間。

2. 控制放款風險:
   可就現有新承做的不動產擔保放款案件提供客觀的估價參考，進一步控制放款風險，針對未來的房貸調整貸款成數。

3. 可隨時監控不動產擔保適足率:
   模型經訓練後，取得新資訊即能夠自動學習，可即時監控不動產價格，對擔保品市場價值低於未償還本金的個案，可即時監控，並制定對應策略。

二. 資料簡介

(一) 資料內容

1. 資料來源: 內政部「不動產交易實價登錄」提供的不動產買賣資料集

2. 資料筆數: 105~108 年台北市的不動產實價資訊，共 81151 筆，28 個欄位

3. 變數欄位: 包含不動產個別資料與交易資料

   - **不動產個別資料**：該建物所在的鄉鎮市區、土地使用分區及主要用途、建物型態、建物移轉總面積、房間衛浴及廳個數、車位資訊等等
   - **不動產交易資料**：總價、交易年月日、單價平方公尺、車位總價

鄉鎮市區	交易標的	土地移轉總面積	土地使用分區	交易日期	總樓層數	建物型態	主要用途	建築完成年月	建物移轉總面積	格局-房	格局-廳	格局-衛	總價元(萬)	車位類別	.....
文山區	房地(土地+建物)	30.04	住	1061126	十一層	住宅大樓	住家用	841122	112.71	3	2	2	1550	升降機械	
中山區	土地	2	其他	1070201	十二層	公寓			0	0	0	0	20		

不動產實價登錄原始資料範例

(二) 資料篩選

1. 非研究標的，直接刪除
   - 交易標的：刪除僅有土地或僅有車位的標的
   - 都市土地使用分區：僅留下住、商，其餘類別直接刪除
   - 建物型態：刪除倉庫、廠辦、工廠、農舍的類別
   - 主要用途：刪除工業用及商業用

2. 刪除欄位
   - 不相關欄位：交易筆棟數、編號
   - 無法量化：土地區段位置建物區段門牌、主要建材、備註
   - 欄位缺失值過多：非都市土地使用分區、非都市土地使用編定、車位類別

三. 資料前處理

(一) 檢查缺失值

變數缺失值百分比			
鄉鎮市區	0	建物移轉總面積	0
交易標的	0	建物格局.房	0
土地移轉總面積	0	建物格局.廳	0
土地使用分區	1%	建物格局.衛	0
交易年月日	0	建物格局.隔間	0
移轉層次	10.22%	有無管理組織	0
總樓層數	10.36%	總價元	0
建物型態	0	車位移轉總面積	0
主要用途	13%	車位總價元	0
建築完成年月	20.08%		

(二) 缺失值處理

- 土地使用分區 – 直接刪除
- 移轉層次 – 直接刪除
- 總樓層數 – 補均值
- 主要用途 – 補眾數

(三) 特殊資料處理

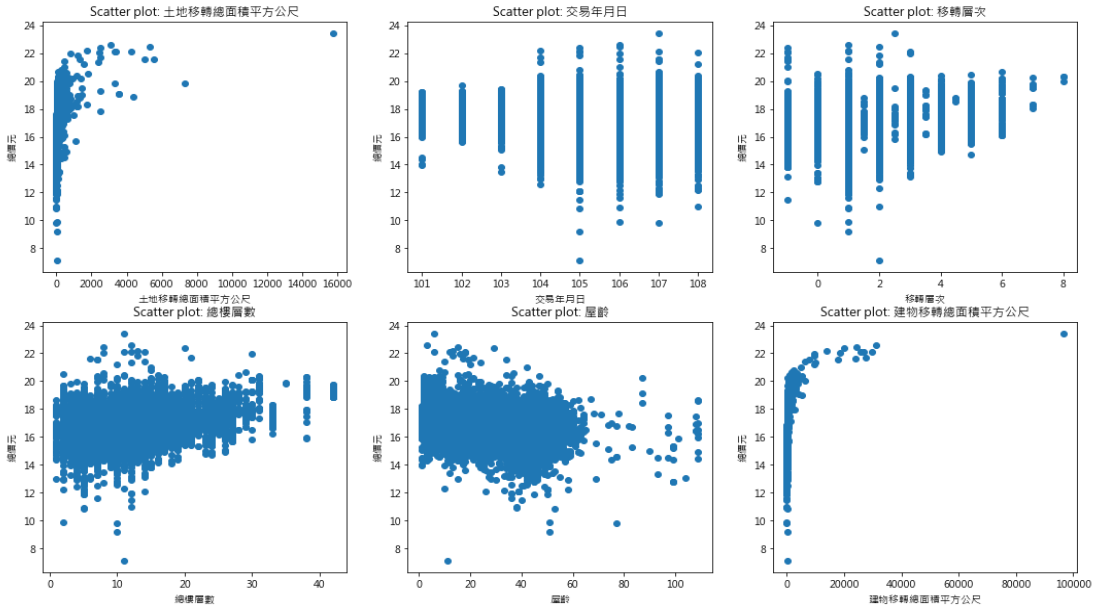
1. 移轉層次:
   - ■ 轉換成樓層區間
   - ■ 根據業內專家建議，將 1~5 層做一區間，6～10 層做一區間，以此類推
   - ■ 將陽台、騎樓、平台、走廊、電梯樓梯間、透天厝、停車場、夾層、露台各增設類別變數
2. 交易年月日:
   認為時間季度具有重要性，拆分成年和季兩欄，共分四季：Q1,Q2,Q3,Q4
3. 建築完成年月:
   與交易日期做計算，轉換成屋齡
4. 對類別資料進行 one-hot encoding: 鄉鎮市區、建物型態、主要用途、建物現況格局.隔間、有無管理組織

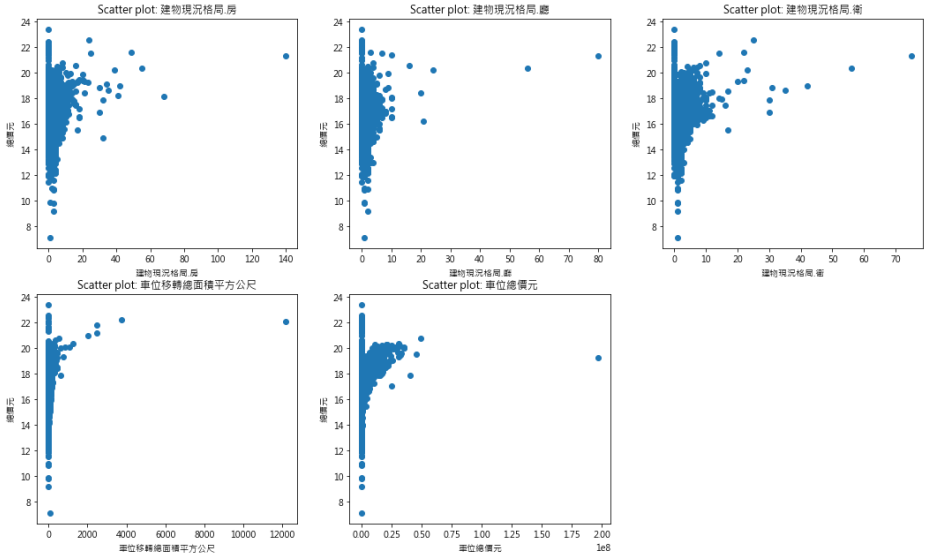
	鄉鎮市區
0	文山區
1	北投區
2	萬華區
3	萬華區
4	萬華區

	中山區	中正區	信義區	內湖區	北投區	南港區	士林區	大同區	大安區	文山區	松山區	萬華區
0	0	0	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	1

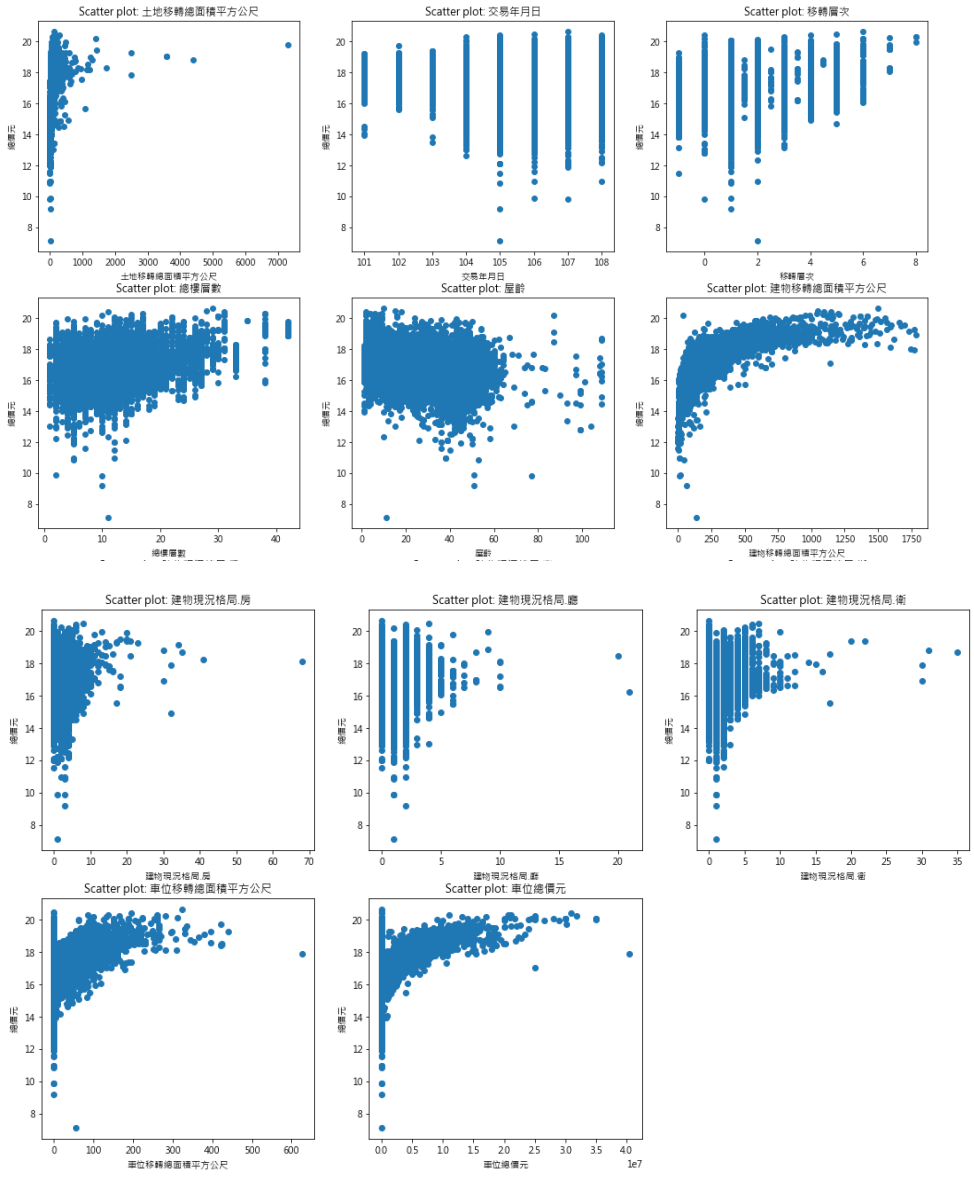
(四) 處理異常值

1. 原始資料





2. 去除三個標準差外的異常值

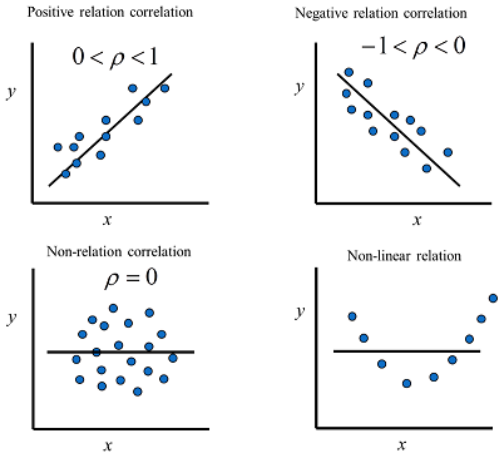


(五) 相關性

1. 相關性: 相關係數是協助了解各欄位與預期目標之間「線性」關係的指標，作為後續是否把一個欄位當成特徵的初步依據。

   - r>0: positively correlated
   - r<0: negatively correlated
   - r=o: no linear correlation

2. 相關係數大小，只是說明哪個欄位與目標更相關一點，且相關係數為零，不代表沒有相關，只是不具備「線性相關」，以下圖為例:
   右下圖為二次函式，為非線性相關，故相關係數為零

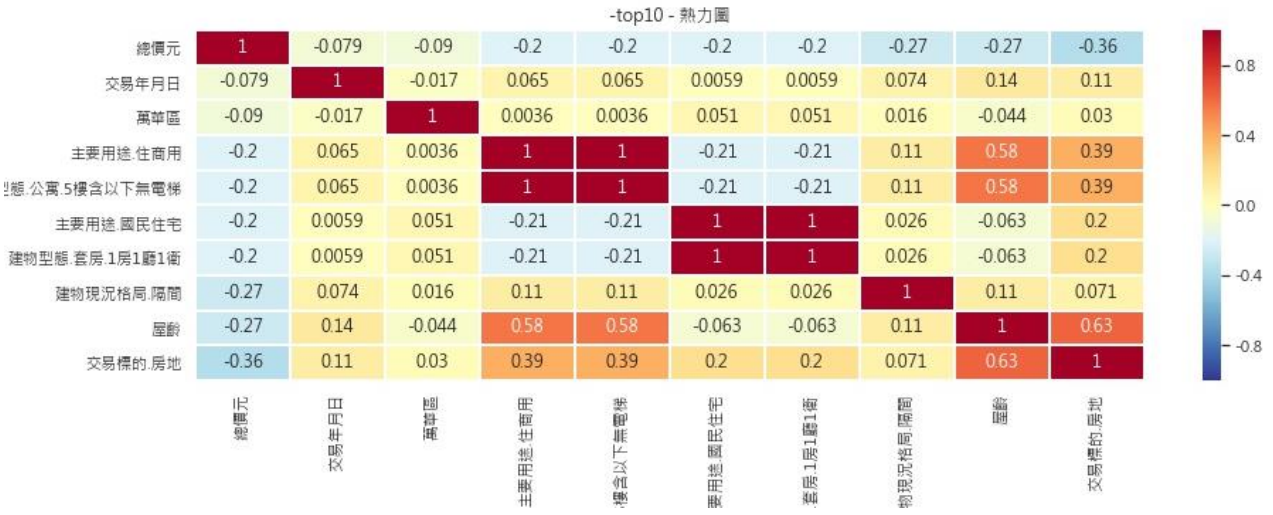


3. 熱力圖

由各變數與預測目標房地產「總價元」的相關係數所繪:

(1) 正相關前十個變數

(2) 負相關前十個變數



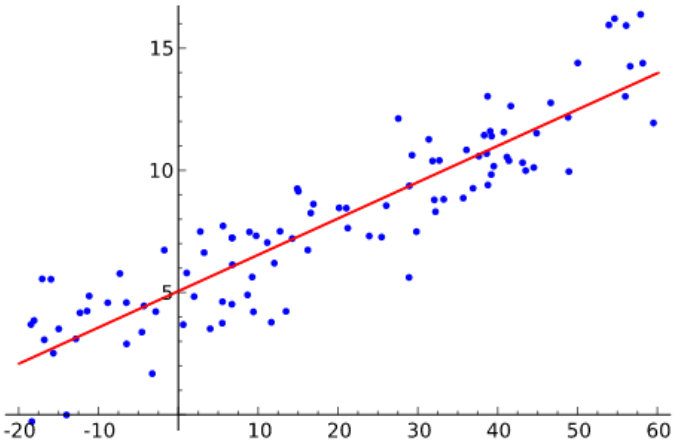
四. 研究方法

(一) 模型介紹

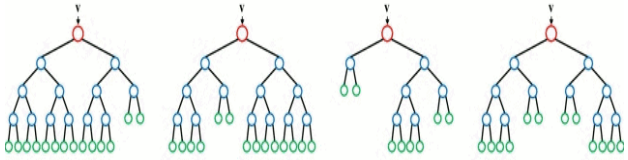
首先，透過確認房價的重要特徵，如:地區所在地、物件本身條件和建物格局等，建立機器學習模型，因為預測的問題屬於監督式學習的回歸問題，將會選擇常見的回歸模型來做嘗試，如：

1. 多變量線性回歸

- 原理: 探討自變數與應變數之間的線性關係
- 優點: 建模速度快，在模型結構不複雜可以很好地在商業決策時提出一個合理的故事。
- 缺點: 對非線性資料來說，模型設計困難，必須相當瞭解資料的結構與特徵關係。



2. 決策樹與隨機森林



- 原理: 決策者透過每一次選擇，從而製造出一個選擇的樹，非常直觀，且隨機森林是決策樹的一種集成，通常能夠表現得更好。
- 優點: 對複雜、非線性資料有良好表現，通常表現得比多變量回歸好。且容易解釋，因每一次選擇相當於人類的一連串決策，解釋力相當高。
- 缺點: 由於決策樹本身特質，容易過度擬合，需適當的修剪才能減輕這些狀況。隨機森林需要的運算量龐大，需要比較高的計算量。
- 結論: 在資料複雜、需要解釋性高，且運算資源允許情況下，是一個相當不錯的選擇。

將會嘗試許多機器學習模型，因為沒有一款機器學習模型適用任何問題。不同的機器學習模型表現取決於資料的大小與結構。所以通常會反覆試驗(trail and error)來嘗試不同的模型，並配合網格搜尋(GridSearch)找出較適合的超參數，或在時間考量下，可能會採用隨機搜尋(RandomSearch)來找尋超參數，若個別模型表現都不錯且模型具有差異，將可考慮用集成學習進一步優化。

(二) 評估標準

國際自動估價平臺測試標準多以命中率（Hit-Rate）作為模型準確度好壞的評估指標:

1. 在某誤差水準之下，可準確被預測到的資料樣本數比例
2. 命中率愈高代表模型精確度越高。一般實務研究上，誤差在 10%的命中率為 50%，而誤差在 20%命中率為 80%才符合自動大量估價命中率標準。
3. Hit-Rate 的計算共分為兩步驟：
   - **首先比對各筆不動產預測($\hat{y}_i$)和真實價格(y_i)的誤差水準是否小於等於 10%，若是則為 1；若否則為 0。**

$$Z_i = 1\ if\ \left|\frac{\hat{y}_i - y_i}{y_i}\right| \leq 10\%\ \ else\ \ 0$$

   - **最終將各數值取平均，得到最終的 Hit-Rate：**

$$\frac{\sum_{i=1}^{n} Z_i}{n}$$

(三) 模型訓練

1. 模型選擇

   - 線性回歸
   - Lasso
   - Bayes
   - 決策樹
   - 隨機森林
   - GDBT
   - MLPR
   - SVM: 資料過大跑不動

2. 年份切割，避免時間窺探
   各年份數據筆數:

年份	筆數
108 年	7864
107 年	14997
106 年	14343
105 年	12932
104 年	5392
103 年	909
102 年	798
101 年	378

3. 切割訓練集和測試集，並將資料依照 6:2:2 分別做為訓練、測試和驗證
4. 模型初始化(使用默認參數)
   用驗證集的 hit-rate 確認沒有過度擬和

(1) 沒有時間切割

	Linear Regression	Decision Tree	Random Forest
10%誤差	22.17%	36%	42.27%
20%誤差	41.37%	57.75%	68.18%

(2) 時間切割: 以 106 年以前資料集做訓練，並以 107 年資料做驗證、測試

	Linear Regression	Decision Tree	Random Forest	
10%誤差	20.99%	36.98%	43.26%	
20%誤差	39.60%	58.02%	68.74%	
	Lasso	貝氏回歸	MLPR	GDBT
10%誤差	20.98%	21.14%	3.34%	32.85%
20%誤差	39.60%	39.88%	6.34%	58.04%

由上表可知，最好的預測模型為**隨機森林**，且本篇研究與一般實務研究相比: 誤差在 10%的命中率為 50%，而誤差在 20%的命中率為 80%，可知本篇研究結果與自動大量估價模型命中率的標準差距不會過大。

(四) 微調模型

1. 調整模型超參數
   - GridSearch:
     隨機森林: {'max_depth': 100, 'max_features': 8, 'n_estimators': 200}
     比較 hit-rate 後發現，使用最佳超參數，模型效果有些微提升。

- RandomSearch
  隨機森林: { bootstrap=True, max_depth=100, max_features=10, min_samples_split=3, n_estimators=200)

2. 嘗試集成學習，組合多個好模型往往比單個模型來得好，但大多數模型太差，所以不考慮
3. 得到最終模型，在測試集上測量性能

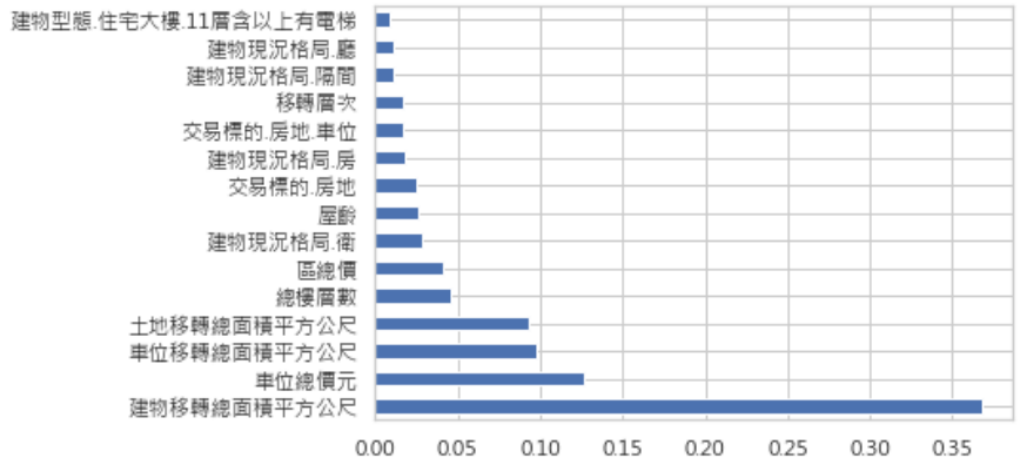
(五) 特徵工程

1. 房價指數
   透過經濟理論，可能是一個重要特徵。
2. 行政區改為區總價
   原本的 0, 1 二元特徵，轉為區總價亦可以保持區別，特徵數也會得到減少
3. 根據特徵重要性，使不必要特徵減少，增加效率(保有 hit-rate 的同時)

Random Forest		
	房價指數	區總價
10%誤差	下跌	下跌
20%誤差	下跌	下跌
放入模型	否	否

(六) 特徵重要性

1. 繪圖



2. 僅擷取重要特徵再跑一次模型，檢驗效果是否差距巨大。如果差距不多，就可以減少計算量, 保有一定的準確度，由 hit-rate 結果下降 4 個百分點，故不考慮使用。

五. 結論

(一) 模型維護

1. 需定期更新資料案件並且讓模型訓練，若模型表現下降則重新檢視模型，以維持模型的品質。
   - ■ 監控系統效能
     透過設定 hit-rate 下降 5%時提醒，查看實際狀況。
   - ■ 評估輸入系統的資料品質
     這部分可以透過, 類似像 box-plot 或者是分布情形查看
   - ■ 定期用新的資料訓練
     保有彈性，可以回溯之前的模型

(二) 未來展望

1. 在新增變數或進行其他演算法組合或優化後，期望達到國際自動估價平臺標準。