Price Prediction of D.C. Residential Properties

指導老師：李家岩 老師

工資系 黃瀚陞H34056110
統計系 林佳霈H24061040
製造所 陳建安P96081067
工科專 吳紹琪N97071094

一、背景與動機

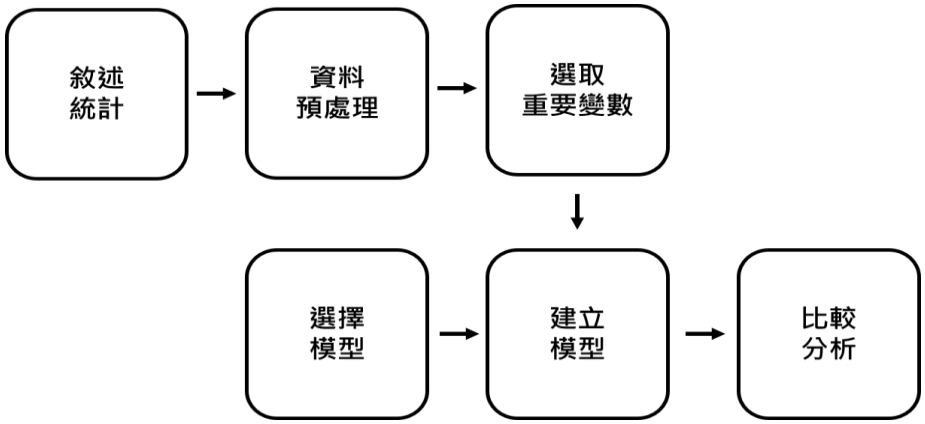
房地產市場預測是指利用已獲取的各種資料，運用資料科學等手段，對房地產市場加以分析和判斷，為房地產開發和經營決策提供依據，透過分析房地產市場，我們可以得到(1)影響房價之重要因素(2)居民對於房地產特徵之喜好(3)新售出房地產之價格。

做好房地產的市場分析與預測，將有利於企業制定良善的房地產開發經營計畫、平衡社會對房地產的需要，透過了解居民之需求，企業可以選擇居民最喜愛之特徵進行開發，增加購獲利，也可以幫助企業在投資時降低風險。經研究發現，房價之趨勢是重要的經濟指標，每個人都受到經濟增長或衰退以及房價下跌或上漲的影響，因此房地產之預測不管是對於企業還是個人，都極為重要。

二、方法論與研究架構

1. 架構

本次專案是要針對華盛頓特區(D.C.)的房價進行預測。針對此次專案，我們收先針對了我們的資料集進行敘述統計的分析，接著透過敘述統計的分析結果進行遺漏值填補、離群值處理等資料預處理，再來則經由統計方法選出會影響預測結果的重要變數，再將重要變數放入我們選好的模型當中，最後以 RMSE 來計算誤差並且提出本次專案所得的洞見。



2. 資料預處理

(1) 人工刪除部分欄位(Column)

在此次選用的資料當中，部分變數對於預測是完全沒有任何幫助的，因此在資料預處理步驟即予以刪除。舉例來說，資料中的 City 以及 State 這兩欄，因為每一筆資料此欄位都相同（都是 WASHINGTON 和 DC），因此直接刪除這兩欄。此外，資料中部分欄位之間的內容與特性是重複的，因此將重複欄位刪除，只留下其中一欄，像是資料中的 X 變數與 LATITUDE 變數、Y 變數與

LONGTITUDE 皆是代表經度與緯度，因此只留下 X 變數與 Y 變數作為接下來的預測資料。

(2) 剔除遺失值過多的欄位

資料中某些欄位存在著大量的遺失值，這些欄位若直接使用將影響預測準確率，但若是要用數理方法填補遺失值又會因為遺失值過多而造成填補效果不佳，因此在此次專案當中，將遺失值超過 20%的欄位刪除。

(3) 剔除變數不完整的資料

部分筆數的資料在蒐集時並沒有記錄到所有的變數值，而一旦此筆資料存在許多空白將會造成最後預測結果不佳，因此一旦資料（每一個 Row）若有超過 10 個變數為遺失值或空白就直接剔除。

(4) 填補遺漏值

部分欄位存在著少數的遺漏值，因此需要將遺漏值補齊後才可將資料放入模型做預測。在資料當中，存在遺漏值的欄位皆屬於類別型的資料（如房子的樓層數、房子的地理方位等），因此選擇以填入眾數的處裡遺漏值。

(5) 移除離群值

在我們希望預測的價格欄位當中存在著少數的離群值，這些房子的價格特別的高或非常的低，這些房子的價格與市場的行情有著非常大的差異（可能是天價豪宅或是沒人想住的凶宅等等），為了避免影響符合市場行情的價格預測，因此予以刪除。

(6) 價格標準化

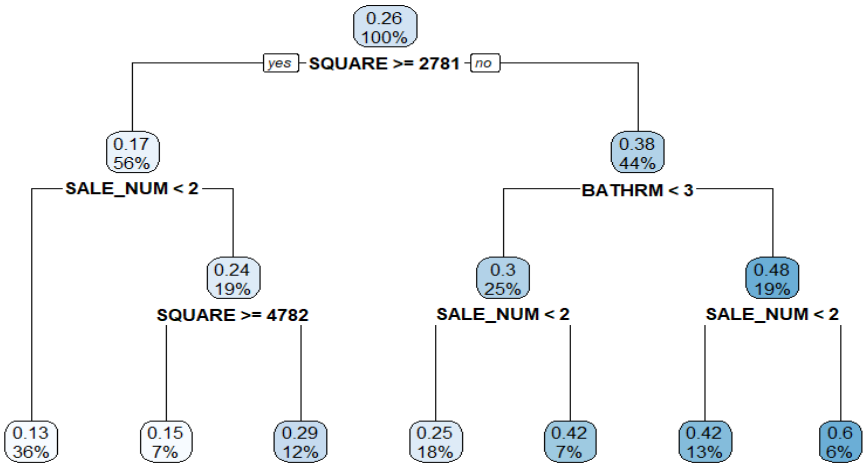
由於欲預測的房價範圍橫跨了數萬元到數百萬元之間，為了不要使模型在收斂時的速度落差過大，因此此次專案將價格標準化，使變數之間的量級調整至相似尺度上，改善預測的速度與準確率。

(7) 類別轉為 Dummy Variable

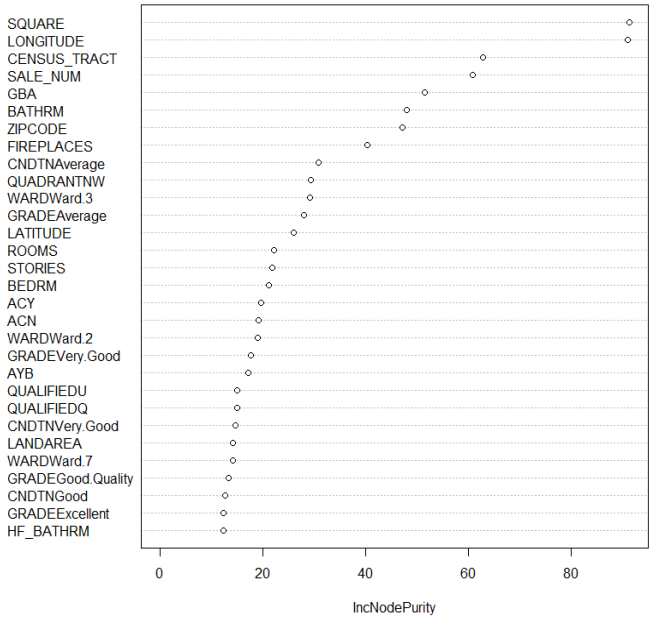
由於資料當中有部分資料屬於類別資料，無法被直接放入模型當中做預測，因此使用 One-hot Encoding 將類別型的資料轉成 dummy variable 的形態。

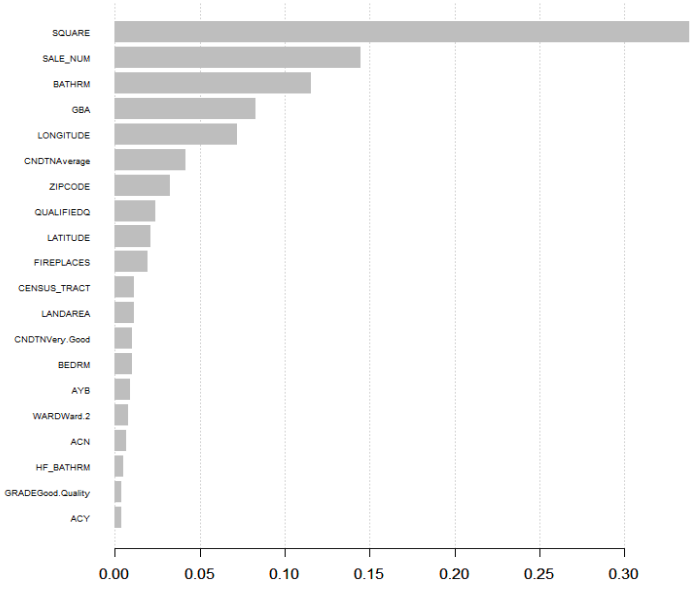
3. 重要特徵篩選

由於此份資料的變數眾多，即使經過了資料預處理步驟之後，仍舊有超過三十個變數，倘若每次都將這麼多變數丟入模型預測與調參數，將會造成預測速度緩慢以及模型難以解釋的問題，因此我們針對此次所使用到的 Decision Tree、Random Forest 以及 xgboost 三個模型列出其重要特徵，以下為各自的重要特徵。


0.26
100%
yes
SQUARE >= 2781
no
0.17
56%
0.38
44%
SALE_NUM < 2
BATHRM < 3
0.24
19%
0.3
25%
0.48
19%
SQUARE >= 4782
SALE_NUM < 2
SALE_NUM < 2
0.13
36%
0.15
7%
0.29
12%
0.25
18%
0.42
7%
0.42
13%
0.6
6%

rf


SQUARE
LONGITUDE
CENSUS_TRACT
SALE_NUM
GBA
BATHRM
ZIPCODE
FIREPLACES
CNDTNAverage
QUADRANTNW
WARDWard.3
GRADEAverage
LATITUDE
ROOMS
STORIES
BEDRM
ACY
ACN
WARDWard.2
GRADEVery.Good
AYB
QUALIFIEDU
QUALIFIEDQ
CNDTNVery.Good
LANDAREA
WARDWard.7
GRADEGood.Quality
CNDTNGood
GRADEExcellent
HF_BATHRM
0
20
40
60
80
IncNodePurity


SQUARE
SALE_NUM
BATHRM
GBA
LONGITUDE
CNDTNAverage
ZIPCODE
QUALIFIEDQ
LATITUDE
FIREPLACES
CENSUS_TRACT
LANDAREA
CNDTNVery.Good
BEDRM
AYB
WARDWard.2
ACN
HF_BATHRM
GRADEGood.Quality
ACY
0.00
0.05
0.10
0.15
0.20
0.25
0.30

	XGB	DecisionTree	RandomForest
SQUARE	1	V	1
SALE_NUM	2	V	4
BATHRM	3	V	6
GBA	4		5
LONGTITUDE	5		2
CNDTNAverage	6		9
ZIPCODE	7		7
QUALIFIEDQ	8		X
LATITUDE	9		X
FIREPLACES	10		8
CENSUS_TRACT	X		3
QUADRANTNW	X		10

以上為三種方法所篩選出的重要變數，從表格中我們可以發現重要變數在不同方法間的重複性很高，因此，我們接著使用簡單線性迴歸驗證選出的變數是否確實顯著及觀察變數與價格之間的相關性。

4. 選擇模型

此次專案中，我們希望可以透過資料中所提供的各種變數，來預測一間房子的價格，屬於迴歸預測的問題。模型選用上，使用了 Linear Regression、Decision Tree、Random Forest 以及 XGBoost (eXtreme Gradient Boosting)四種。

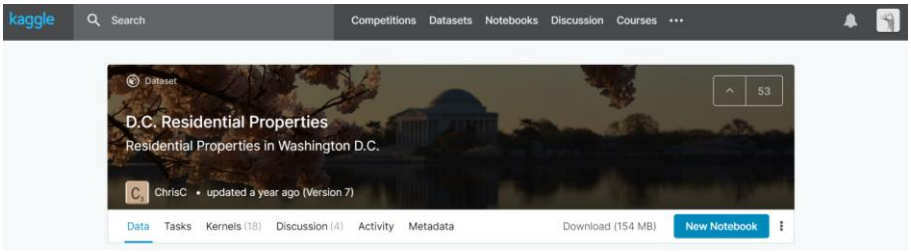
首先我們對重要變數建立簡單線性回歸，看看變數之間的相關性及模型預測力。接著再以解釋性佳的 Decision Tree 作為我們的第二個模型，Decision Tree 主要以樹狀分枝來進行預測，其中一項優點是有著非常好的解釋性，由於我們希望此專案的預測不僅僅只追求準確率，而是可以解釋結果進而提出影響房價高低的因素與判斷準則，因此選用了 Decision Tree 作為模型。而 Random Forest 則是包含了許多 Decision Tree，將 Decision Tree 輸出結果進行投票或是加總，因此可以提升預測準確率。最後選用的 XGBoost 則是加入多個弱學習器進行最終預測，預測的速度較快且準確率高，故選用其做為本次專案的第四個模型嘗試。

三、資料蒐集與分析結果

1. 資料蒐集

本研究之資料集取自 Kaggle (www.kaggle.com) Datasets, 為取自美國華盛頓特區(D.C.)的政府開放資料集，關於地址資料則由 D.C.地理信息系統提供(D.C. Geographic Information System)。記錄自 1982 年 6 月至 2018 年 7 月的 D.C.房地產銷售紀錄的各種真實資料。華盛頓特區是美國的

首都，人口接近 70 萬人，並且自 2000 年以來一直在增長。這個城市高度隔離，生活成本高。 2017 年，該區單戶住宅的平均價格為 649,000 美元。關於 DC open data 的應用條款，詳見美國華盛頓 D.C.的政府開放資料官方網站(https://dc.gov/)。此資料集的原始資料共有 158,957 筆資料與 49 個變數。



2. 資料分析

(1) 簡單線性迴歸(Linear Regression)

我們對重要變數建立簡單線性迴歸，從檢定中可發現篩選出來的變數皆為顯著，且從迴歸估計值中我們可得知房間個數與房價成弱正相關，除此之外，房價也因位置而影響。接著我們使用迴歸模型進行預測，交叉驗證後的 rmse 為 0.2299，在此迴歸模型的預測力表現的並不好，因此，我們使用其他模型來做預測。

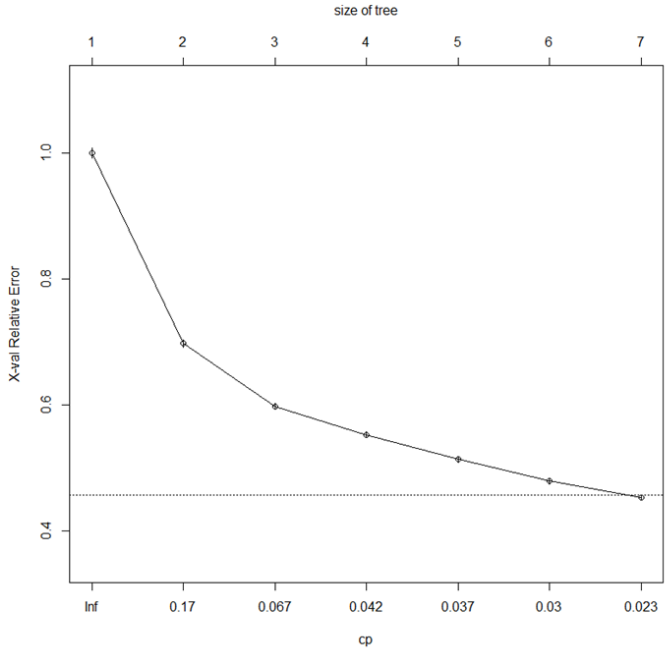
```
Call:
lm(formula = PRICE ~ SQUARE + SALE_NUM + BATHRM + GBA + LONGITUDE +
    CNDTNAverage + ZIPCODE + QUALIFIEDQ + LATITUDE + FIREPLACES +
    CENSUS_TRACT + QUADRANTNW, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61570 -0.06232  0.00255  0.05941  0.82545

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.233e+01  2.233e+00 -27.914   <2e-16 ***
SQUARE       -2.731e-05  4.912e-07 -55.594   <2e-16 ***
SALE_NUM      3.305e-02  3.372e-04  98.000   <2e-16 ***
BATHRM        2.596e-02  6.610e-04  39.278   <2e-16 ***
GBA           5.040e-05  9.801e-07  51.421   <2e-16 ***
LONGITUDE    -8.975e-01  2.367e-02 -37.922   <2e-16 ***
CNDTNAverage -5.952e-02  1.038e-03 -57.321   <2e-16 ***
ZIPCODE       2.031e-04  8.647e-05   2.349   0.0188 *
QUALIFIEDQ    5.914e-02  1.206e-03  49.044   <2e-16 ***
LATITUDE     -2.746e-01  1.860e-02 -14.758   <2e-16 ***
FIREPLACES    3.333e-02  6.833e-04  48.785   <2e-16 ***
CENSUS_TRACT -7.951e-06  3.734e-07 -21.292   <2e-16 ***
QUADRANTNW   -4.273e-02  2.660e-03 -16.062   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.113 on 56592 degrees of freedom
Multiple R-squared:  0.6356,    Adjusted R-squared:  0.6355
F-statistic:  8226 on 12 and 56592 DF,  p-value: < 2.2e-16
```

(2) 決策樹模型(Decision Tree Model)

本研究將資料集裡 33 個變數以及 56,606 筆資料建立決策樹迴歸模型，並以 K 折交叉驗證(K-Fold Cross Validation)，此處 K 值取 5，交叉驗證後決定模型最佳引數複雜度(CP, complexity parameter)，CP 作為控制樹規模的懲罰因子，簡而言之，CP 越大，樹分裂規模

（nsplit）越小。進行 5 折交叉驗證之後，取得 CP＝0.0447991，以此最佳參數建立決策樹迴歸模型。由 CP 值與交叉驗證誤差(X-val relative error)關係圖得知，下 x 軸為 cp complexity parameter（α），上 x 軸為末梢節點數(number of terminal nodes, |T|)，y 軸為交叉驗證誤差(X-val relative error)，隨著模型的複雜度（size of tree）增加，所能改善的模型適合度的空間降低(X-val relative error)。

交叉驗證前、後的 RMSE 誤差值，分別為 0.1397 與 0.1419，可得知模型預測結果，在交叉驗證後有較佳之表現。決策樹建模參數設定與結果如下圖所示。




```
tree <- rpart(data = train, PRICE~. ,control=rpart.control( cp=0.044))
```

```
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 36226, 36228, 36229, 36226, 36227
Resampling results across tuning parameters:

  cp         RMSE       Rsquared   MAE
  0.0447991  0.1418471  0.4243693  0.1075547
  0.1009015  0.1498625  0.3570760  0.1139282
  0.2982195  0.1752151  0.2910307  0.1358199

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.0447991.
```

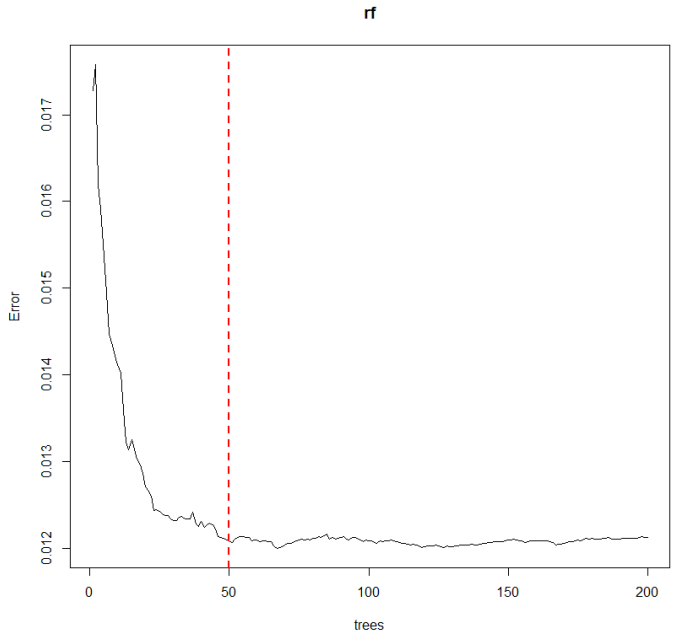
(3) 隨機森林模型(Random Forest Model)

隨機森林由數個決策樹所組成，是一個集成學習法(ensemble learning)，將幾個建立好的模型結果整合在一起，以提升預測準確率。由集成學習法建立的模型較不容易發生過度配適的問題，雖然

能提供較好的預測結果，但在推論和解釋度方面就會有所限制。

本研究將資料集裡 33 個變數以及 56,606 筆資料建立隨機森林迴歸模型，以 K 折交叉驗證(K-Fold Cross Validation)，此處 K 值取 10，並以 OOB(out-of-bag)的 RMSE 值來評估增加每一顆樹的整體誤差改變量，藉此以輔助決策隨機森林模型需要多少顆樹，整體誤差才會趨於穩定。從 OOB(out-of-bag)的 RMSE 值的結果，可得知 OOB(out-of-bag)的 RMSE 值隨著樹的數量增加而遞減，但在模型選用 50 棵樹後則漸趨穩定，故此處選用 50 顆樹作為最佳樹之樹木建立模型，原因在於約 50 棵樹即足夠使誤差趨於穩定，故不需使用更多樹，以避免造成過度擬合的問題。

交叉驗證前、後結果，分別為 0.0890 與 0.0885，可得知模型預測結果，在交叉驗證後有較佳之表現。OOB(out-of-bag)的 RMSE 值與樹的數量趨勢圖、10 折交叉驗證結果與模型參數如下圖所示。




```
rf <- randomForest(data = train ,PRICE~.,mtry = 19,  ntree = 20,nodesize = 50)
```

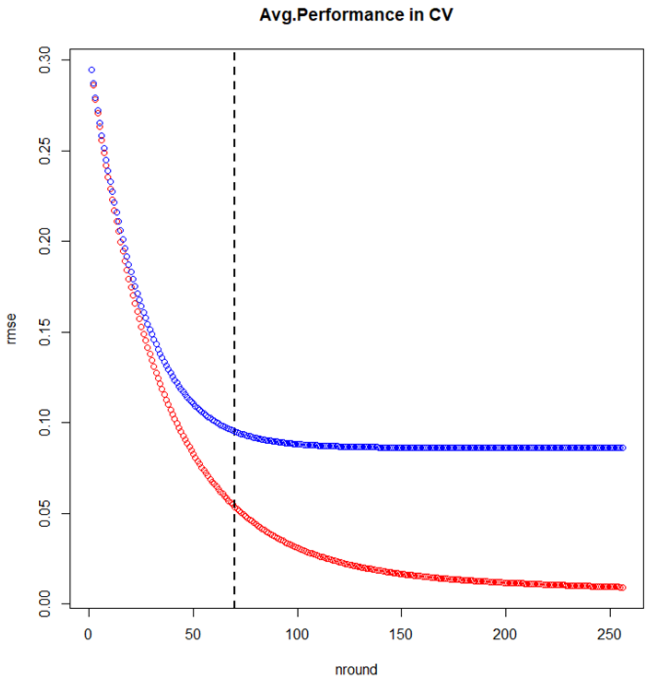
	mtry	node_size	OOB_RMSE
1	19	20	0.08768514
2	19	10	0.08783467
3	19	30	0.08824225
4	17	20	0.08830775
5	19	40	0.08830845
6	17	10	0.08849979
7	19	50	0.08850475
8	17	30	0.08860758
9	15	10	0.08864410
10	17	40	0.08887995

(4) 極限梯度提升模型(XGBoost, Extreme Gradient Boosting Model)

極限梯度提升模型(XGBoost, Extreme Gradient Boosting) 是基於梯度提升決策樹(GBDT,Gradient Boosted Decision Tree)的改良與延伸，主要被應用於解決監督式學習的問題。考慮多棵樹的參數優化，透過增量訓練 (additive training) 的方式，每一次保留原來的模型，加入一個新的函數至模型中，亦即每一步皆會在前一步的基礎之上增加一顆樹，修復前顆樹的不足，以提升目標函數。

本研究將資料集裡 33 個變數以及 56,606 筆資料建立極限梯度提升迴歸模型，以 K 折交叉驗證(K-Fold Cross Validation)，此處 K 值取 10，並以交叉驗證的 RMSE 平均值來決定最佳的決策樹數量(nround)，實驗結果得知 RMSE 值隨著樹的增加而遞減，當樹的數量到達 70 顆時，下降趨勢呈現穩定，故取 70 顆樹最為最佳模型參數，以維持模型最佳效率。

交叉驗證前、後的 RMSE 值，分別為 0.0869 與 0.0950，可得知模型預測結果，在交叉驗證後有較佳之表現。交叉驗證的 RMSE 平均值趨勢與模型參數如下圖所示。




```
#跑有測試資料的rmse，並且選擇70那筆
cv.model1 = xgb.cv(
  params = xgb.params,
  data = ddata_set,
  nfold = 5,
  nrounds=70,
  early_stopping_rounds = 30,
  print_every_n = 23
)
```

下表為四個模型 rmse 的比較，以預測力來說 random forest 與 XGBOOST 的預測力相對較好。

Rmse of different model after CV	
Linear Regression	0.2299
Decision Tree	0.1419
Random Forest	0.0885
Xgboost	0.0950

3. 決策應用

根據此次專案，我們認為分析結果可以應用在兩個層面，第一個是站在投資者的角度，可以透過此次分析使得「投資房產獲利」；第二個則是站在屋主販售的角度，使屋主「販售房屋獲利」，以下分述之。

(1) 投資房產獲利

本研究結果提供美國 D.C.區域房產的投資者有用的訊息，舉例來說，投資美國 DC 房產前，應注意該房產是否有本研究分析的重要特性(重要特徵)，可能是影響 DC 區域房價的關鍵，也是未來是否保值的重要因素，能藉此作整體投資評估。

(2) 販售房屋獲利

本研究結果提供欲販售美國 D.C.區域房產的屋主評估販售房產價格的參考指標。在販售房屋前，應先針對房屋的重要特性(重要特徵)作評估，了解潛在的房屋價值，決定販售價格區間。

四、結論

對於此次專案而言，隨機森林與 XGBoost 之預測結果相似，而我們認為 XGBoost 應能有更好的預測力，可能需要對 XGBoost 重新進行參數調校才能有更好的結果。除此之外，XGBoost 之學習速度較為快速，而決策樹誤差較大但是解釋相對較容易，透過決策數之變數篩選，我們可以看出坪數、房間數、位置等變數為影響房價之最大因素，與我們預期結果符合。由於資料橫跨之時間範圍較長，我們將在未來加入時間與物價指數等變數進行預測，並進行趨勢分析，使預測表現更為理想。