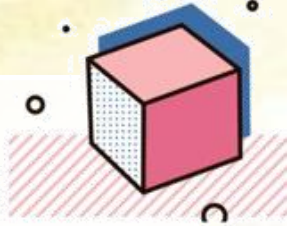


機器學習期末專題報告

筆電價格預測

班　　級：電通四甲

組　　別：第5組

報 告 者：許至佑、林垣志、許楷俊

指導教授：曾士桓

日　　期：112.06.19

背景

- 電腦配置需求日益漸增
  - 擁有一台機體設備良好的電腦非常重要
- 市面上的硬體規格五花八門
  - 消費者對於價格與規格不了解，選 CP 值高就對了?
  - 欲找到最理想且合理價錢的款式便需要花很多時間做功課

動機與目的

- 欲找到最理想且合理價錢的筆電
  - 市場利用平均價格計算硬體規格的價錢
  - 了解市場趨勢，進行價格比較，找到最好的交易

- 利用機器學習的技術推估筆電價格
  - 運用 scikit-learn 套件中的機器學習模型
  - Linear Regression、Random Forest、XGBoost

挑戰

- 資料前處理與特徵選擇
  - 缺失值與異常值的取捨
  - 選擇資料中有用的特徵
- 使用 scikit-learn 套件的模型
  - 評估模型的選擇
  - 超參數設定

資料前處理

&

特徵選取

資料集介紹

- 統計所有廠商販售相關筆電硬體規格的資料紀錄

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model Name	Category	Screen Size	Screen	CPU	RAM	Storage	GPU	Operating System	Operating System Version	Weight	Price
2	Apple	MacBook Pro	Ultrabook	13.3"	IPS Panel F	Intel Core i	8GB	128GB SSI	Intel Iris Pl	macOS		1.37kg	11912523
3	Apple	Macbook Air	Ultrabook	13.3"	1440x900	Intel Core i	8GB	128GB Fla	Intel HD G	macOS		1.34kg	7993374
4	HP	250 G6	Notebook	15.6"	Full HD 19	Intel Core i	8GB	256GB SSI	Intel HD G	No OS		1.86kg	5112900
5	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	512GB SSI	AMD Rade	macOS		1.83kg	22563005
6	Apple	MacBook Pro	Ultrabook	13.3"	IPS Panel F	Intel Core i	8GB	256GB SSI	Intel Iris Pl	macOS		1.37kg	16037611
7	Acer	Aspire 3	Notebook	15.6"	1366x768	AMD A9-9	4GB	500GB HD	AMD Rade	Windows	10	2.1kg	3556800
8	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	256GB Fla	Intel Iris Pr	Mac OS	X	2.04kg	19028613
9	Apple	Macbook Air	Ultrabook	13.3"	1440x900	Intel Core i	8GB	256GB Fla	Intel HD G	macOS		1.34kg	10303160
10	Asus	ZenBook UX43	Ultrabook	14.0"	Full HD 19	Intel Core i	16GB	512GB SSI	Nvidia GeF	Windows	10	1.3kg	13293540
11	Acer	Swift 3	Ultrabook	14.0"	IPS Panel F	Intel Core i	8GB	256GB SSI	Intel UHD	Windows	10	1.6kg	6846840
12	HP	250 G6	Notebook	15.6"	1366x768	Intel Core i	4GB	500GB HD	Intel HD G	No OS		1.86kg	3502559
13	HP	250 G6	Notebook	15.6"	Full HD 19	Intel Core i	4GB	500GB HD	Intel HD G	No OS		1.86kg	3067651
14	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	256GB SSI	AMD Rade	macOS		1.83kg	21696213
15	Dell	Inspiron 3567	Notebook	15.6"	Full HD 19	Intel Core i	4GB	256GB SSI	AMD Rade	Windows	10	2.2kg	4436219
16	Apple	MacBook 12"	Ultrabook	12.0"	IPS Panel F	Intel Core I	8GB	256GB SSI	Intel HD G	macOS		0.92kg	11225261
17	Apple	MacBook Pro	Ultrabook	13.3"	IPS Panel F	Intel Core i	8GB	256GB SSI	Intel Iris Pl	macOS		1.37kg	13502947
18	Dell	Inspiron 3567	Notebook	15.6"	Full HD 19	Intel Core i	8GB	256GB SSI	AMD Rade	Windows	10	2.2kg	6624540
19	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	512GB SSI	AMD Rade	macOS		1.83kg	25413336
20	Lenovo	IdeaPad 320-15I	Notebook	15.6"	Full HD 19	Intel Core i	8GB	1TB HDD	Nvidia GeF	No OS		2.2kg	4437108

Manufacturer：製造商　Model Name：產品名稱　Category：類型　Srceen Size：螢幕尺寸

Screen：螢幕類型　CPU：處理器型號　RAM：記憶體　Storage：硬碟規格

GPU：顯示卡型號　Operating System：作業系統　Operating System Version：版本

Weight：重量　Price：價格 (印尼盾)

資料前處理

- **缺失值與異常值處理**
  - 欄位 [Operating System Version] 的值為NAN
    - 作業系統通常也影響筆電價格，需進一步探討
  - 欄位 [Screen Size, RAM, Weight] 的值認定為物件 (object)
    - 將符號與單位都移除，只留數值
  - 欄位 [Price] 的值是國外幣值，需做幣值轉換成台幣

缺失值　　異常值

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Manufacturer	Model Name	Category	Screen Size	Screen	CPU	RAM	Storage	GPU	Operating System	Operating System Version	Weight	Price
2	Apple	MacBook Pro	Ultrabook	13.3"	IPS Panel F	Intel Core i	8GB	128GB SSI	Intel Iris Pl	macOS		1.37kg	11912523
3	Apple	Macbook Air	Ultrabook	13.3"	1440x900	Intel Core i	8GB	128GB Fla	Intel HD G	macOS		1.34kg	7993374
4	HP	250 G6	Notebook	15.6"	Full HD 19	Intel Core i	8GB	256GB SSI	Intel HD G	No OS		1.86kg	5112900
5	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	512GB SSI	AMD Rade	macOS		1.83kg	22563005
6	Apple	MacBook Pro	Ultrabook	13.3"	IPS Panel F	Intel Core i	8GB	256GB SSI	Intel Iris Pl	macOS		1.37kg	16037611
7	Acer	Aspire 3	Notebook	15.6"	1366x768	AMD A9-9	4GB	500GB HD	AMD Rade	Windows	10	2.1kg	3556800
8	Apple	MacBook Pro	Ultrabook	15.4"	IPS Panel F	Intel Core i	16GB	256GB Fla	Intel Iris Pr	Mac OS	X	2.04kg	19028613
9	Apple	Macbook Air	Ultrabook	13.3"	1440x900	Intel Core i	8GB	256GB Fla	Intel HD G	macOS		1.34kg	10303160
10	Asus	ZenBook UX43	Ultrabook	14.0"	Full HD 19	Intel Core i	16GB	512GB SSI	Nvidia GeF	Windows	10	1.3kg	13293540

轉換為數值

資料前處理

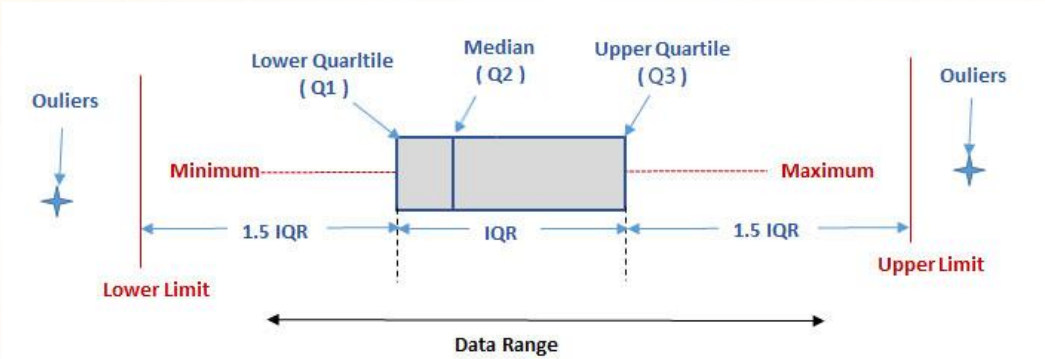
- **缺失值與異常值處理**
  - 欄位 [Operating System Version] 的值為NAN
    - 作業系統通常也影響筆電價格，需進一步探討
  - 欄位 [Screen Size, RAM, Weight] 的值認定為物件 (object)
    - 將符號與單位都移除，只留數值
  - 欄位 [Price] 的值是國外幣值，需做幣值轉換成台幣

資料合併

	A	B	C	D	E	F	G	H	I	J	K	L
1	Manufacturer	Model Name	Category	Screen Size	Screen	CPU	RAM	Storage	GPU	OS	Weight	Price
2	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel	Intel Core	8	128GB SSD	Intel Iris Plus Graphics 6	macOS	1.37	24611.63481
3	Apple	Macbook Air	Ultrabook	13.3	1440x900	Intel Core	8	128GB Flash	Intel HD Graphics 6000	macOS	1.34	16514.55411
4	HP	250 G6	Notebook	15.6	Full HD 1	Intel Core	8	256GB SSD	Intel HD Graphics 620	No OS	1.86	10563.40647
5	Apple	MacBook Pro	Ultrabook	15.4	IPS Panel	Intel Core	16	512GB SSD	AMD Radeon Pro 455	macOS	1.83	46615.85348
6	Apple	MacBook Pro	Ultrabook	13.3	IPS Panel	Intel Core	8	256GB SSD	Intel Iris Plus Graphics 6	macOS	1.37	33134.19115
7	Acer	Aspire 3	Notebook	15.6	1366x768	AMD A9-	4	500GB HDD	AMD Radeon R5	Windows 10	2.1	7348.456675
8	Apple	MacBook Pro	Ultrabook	15.4	IPS Panel	Intel Core	16	256GB Flash	Intel Iris Pro Graphics	Mac OS X	2.04	39313.69208
9	Apple	Macbook Air	Ultrabook	13.3	1440x900	Intel Core	8	256GB Flash	Intel HD Graphics 6000	macOS	1.34	21286.64187
10	Asus	ZenBook UX430UI	Ultrabook	14	Full HD 1	Intel Core	16	512GB SSD	Nvidia GeForce MX150	Windows 10	1.3	27464.85682

資料前處理

- **移除離群值**
  - IQR (四分位距)➔Q3 – Q1，移除 Outliers



  - 針對目標特徵 [Price] 做移除離群值
    - 資料集筆數 1302 ➔ 1267

資料前處理

- **提取資料轉換成新特徵**
  - 將特徵 [Screen, CPU, Storage] 拆成個別獨立的子特徵

- Screen ➔ 可拆分成螢幕解析度、種類，以及是否有觸控功能

Full HD 1920x1080
IPS Panel Retina Display 2880x1800
Full HD 1920x1080
IPS Panel Retina Display 2304x1440
IPS Panel Retina Display 2560x1600
Full HD 1920x1080
IPS Panel Retina Display 2880x1800
Full HD 1920x1080
IPS Panel Full HD / Touchscreen 1920x1080
1366x768
IPS Panel Full HD 1920x1080
1366x768
Full HD / Touchscreen 1920x1080
Full HD 1920x1080
1366x768
1440x900
Full HD 1920x1080
Full HD 1920x1080
Full HD 1920x1080
Full HD 1920x1080
1366x768
Full HD 1920x1080
Touchscreen / Quad HD+ 3200x1800

resolution	screentype	touchscreen
2560x1600	IPSPanelRetinaDispl	0
2304x1440	IPSPanelRetinaDispl	0
2560x1600	IPSPanelRetinaDispl	0
1920x1080		0
1920x1080		0
1920x1080	IPSPanel	1
1366x768		0
1920x1080	IPSPanel	0
1366x768		0
1920x1080		1
1920x1080		0
1366x768		0

資料前處理

- **提取資料轉換成新特徵**
  - 將特徵 [Screen, CPU, Storage] 拆成個別獨立的子特徵

- CPU ➔ 將CPU特徵留下來，且提取處理器速度成新的特徵

CPU
Intel Core i5 2.3GHz
Intel Core i5 1.8GHz
Intel Core i5 7200U 2.5GHz
Intel Core i7 2.7GHz
Intel Core i5 3.1GHz
AMD A9-Series 9420 3GHz
Intel Core i7 2.2GHz
Intel Core i5 1.8GHz
Intel Core i7 8550U 1.8GHz
Intel Core i5 8250U 1.6GHz
Intel Core i5 7200U 2.5GHz
Intel Core i3 6006U 2GHz
Intel Core i7 2.8GHz

freq
2.3
1.2
2.3
2.7
2.4
1.6
1.44
2.5
1.5
1.6
1.8
2

資料前處理

- **提取資料轉換成新特徵**
  - 將特徵 [Screen, CPU, Storage] 拆成個別獨立的子特徵

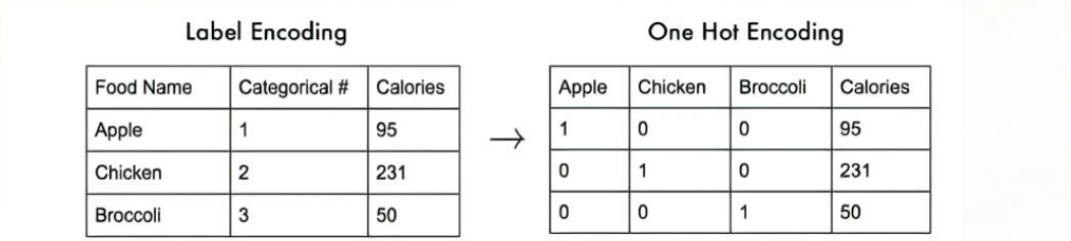
- Storage ➔ 筆電有分單/雙硬碟，因此可拆分統計硬碟大小與種類

32GB Flash Storage
128GB SSD + 1TB HDD
500GB HDD
256GB SSD
256GB SSD
1TB HDD
128GB Flash Storage
256GB SSD
256GB SSD + 256GB SSD
1TB HDD
64GB Flash Storage

primarystorage_size	primarystorage_type	secondarystorage_size	secondarystorage_type
128	SSD	0	
256	SSD	0	
256	SSD	0	
256	SSD	0	
1000	HDD	0	
128	SSD	0	
32	FlashStorage	0	
128	SSD	1000	HDD
500	HDD	0	
256	SSD	0	
256	SSD	0	

資料前處理

- 資料型態轉換
  - One-Hot Encoding



Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

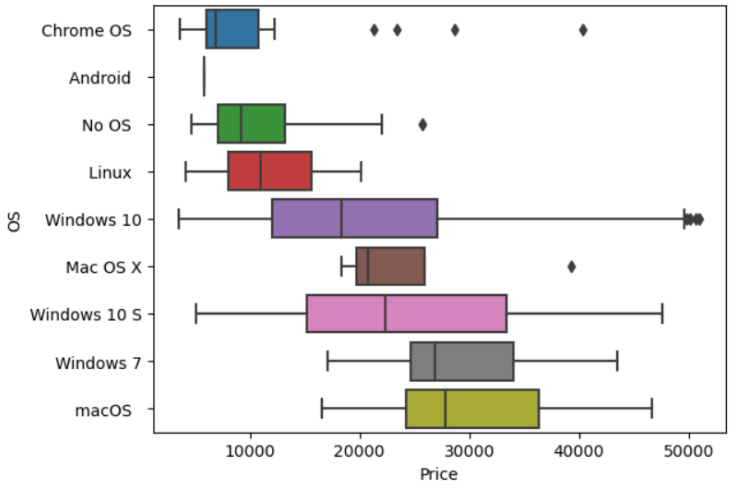
manufactu	model nan	category	screen size	cpu	ram	gpu	os	weight	price	resolution	screentype	touchscree	freq	primarystc	primarystc	secondary
Apple	MacBook	Ultrabook	13.3	Intel Core	8	Intel Iris P	macOS	1.37	24611.63	2560x160(	IPSPanelR	0	2.3	128	SSD	0
Apple	Macbook .	Ultrabook	13.3	Intel Core	8	Intel HD (	macOS	1.34	16514.55	1440x900		0	1.8	128	FlashStora	0
HP	250 G6	Notebook	15.6	Intel Core	8	Intel HD (	No OS	1.86	10563.41	1920x1080		0	2.5	256	SSD	0
Apple	MacBook	Ultrabook	15.4	Intel Core	16	AMD Rac	macOS	1.83	46615.85	2880x180(	IPSPanelR	0	2.7	512	SSD	0
Apple	MacBook	Ultrabook	13.3	Intel Core	8	Intel Iris P	macOS	1.37	33134.19	2560x160(	IPSPanelR	0	3.1	256	SSD	0
Acer	Aspire 3	Notebook	15.6	AMD A9-	4	AMD Rac	Windows	2.1	7348.457	1366x768		0	3	500	HDD	0
Apple	MacBook	Ultrabook	15.4	Intel Core	16	Intel Iris P	Mac OS X	2.04	39313.69	2880x180(	IPSPanelR	0	2.2	256	FlashStora	0
Apple	Macbook .	Ultrabook	13.3	Intel Core	8	Intel HD (	macOS	1.34	21286.64	1440x900		0	1.8	256	FlashStora	0
Asus	ZenBook 1	Ultrabook	14	Intel Core	16	Nvidia Ge	Windows	1.3	27464.86	1920x1080		0	1.8	512	SSD	0
Acer	Swift 3	Ultrabook	14	Intel Core	8	Intel UHL	Windows	1.6	14145.78	1920x108(	IPSPanel	0	1.6	256	SSD	0

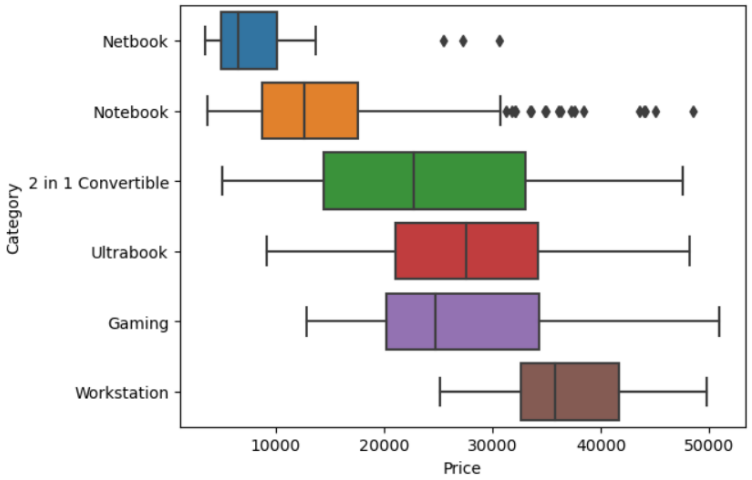
manufactu	model nan	category	screen size	cpu	ram	gpu	os	weight	price	resolution	screentype	touchscree	freq	primarystc	primarystc	secondary
1	234	4	13.3	46	8	50	8	1.37	24611.63	10	1	0	2.3	128	3	0
1	235	4	13.3	46	8	44	8	1.34	16514.55	1	2	0	1.8	128	0	0
7	38	3	15.6	52	8	46	4	1.86	10563.41	3	2	0	2.5	256	3	0
1	234	4	15.4	60	16	7	8	1.83	46615.85	12	1	0	2.7	512	3	0
1	234	4	13.3	46	8	51	8	1.37	33134.19	10	1	0	3.1	256	3	0
0	45	3	15.6	10	4	14	5	2.1	7348.457	0	2	0	3	500	1	0
1	234	4	15.4	60	16	52	3	2.04	39313.69	12	1	0	2.2	256	0	0
1	235	4	13.3	46	8	44	8	1.34	21286.64	1	2	0	1.8	256	0	0
2	466	4	14	75	16	85	5	1.3	27464.86	3	2	0	1.8	512	3	0

圖片來源：初學Python手記#3-資料前處理(Label encoding、 One hot encoding) | by Pat Huang | Medium

資料前處理

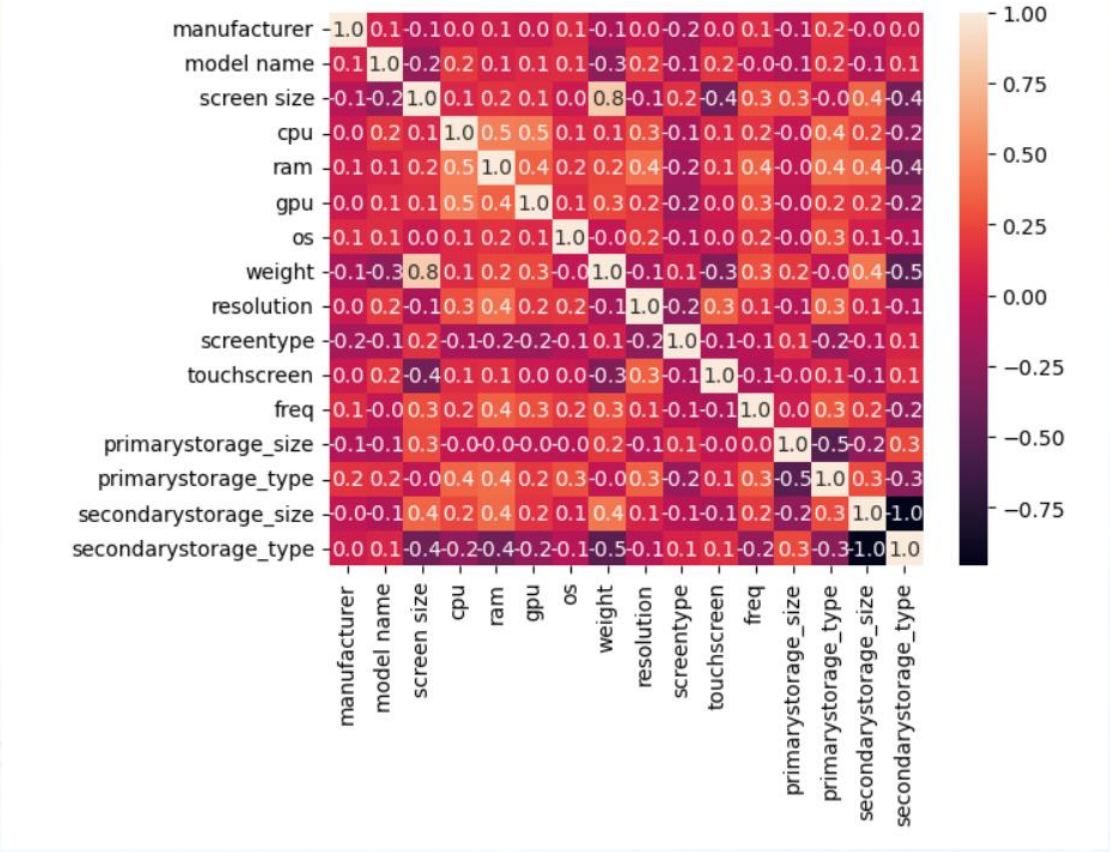
- **觀察Price與其他特徵的關係**
  - 統計所有作業系統的筆電價格
  - 不同種類的筆電也會影響價格





資料前處理

- **觀察Price與其他特徵的關係**
  - 使用相關熱圖分析數值變量與價格之間的關係



特徵選取

- Pearson 相關係數
  - 衡量兩變數間「線性」關聯性的高低程度
  - 兩變數之間的相關係數絕對值較大，則表示彼此之間的關聯性程度較大

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

→ 計算 X, Y 的共變異數

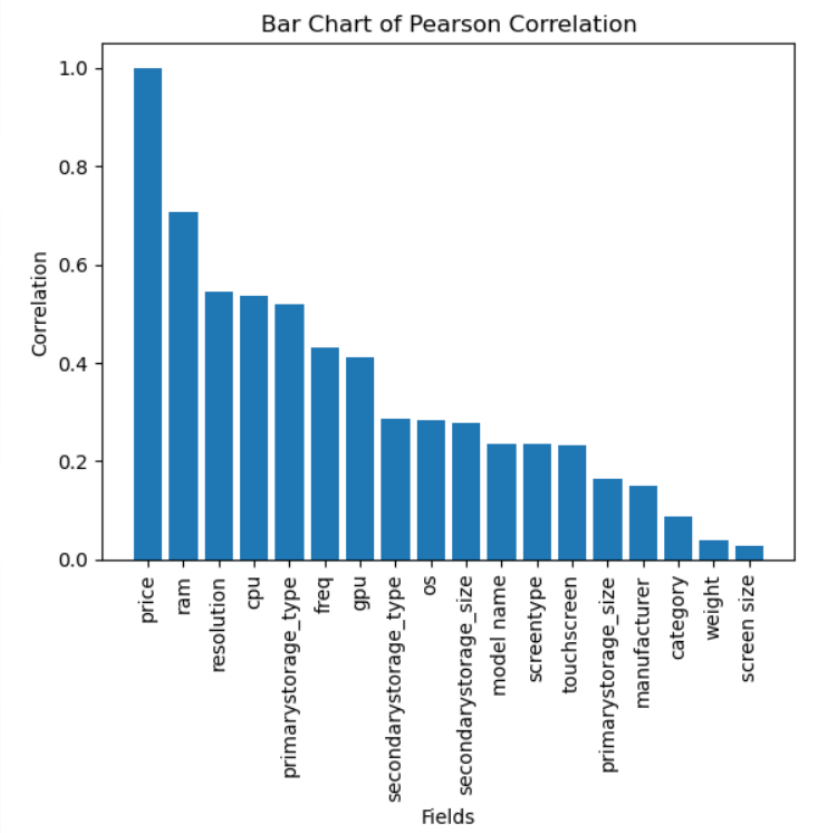
→ X 標準差 * Y 標準差

參考文獻：Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

Feature Selection(特徵選取)

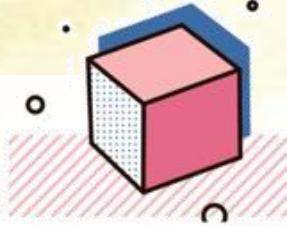
- Pearson 相關係數

```
price                   1.000000
ram                     0.707329
resolution              0.544236
cpu                     0.534894
primarystorage_type     0.520281
freq                    0.430789
gpu                     0.410796
secondarystorage_type   0.286927
os                      0.284218
secondarystorage_size   0.278199
model name              0.235660
screentype              0.234600
touchscreen             0.231391
primarystorage_size     0.162649
manufacturer            0.149137
category                0.087265
weight                  0.038241
screen size             0.028759
Name: price, dtype: float64
```



四種資料 ➔

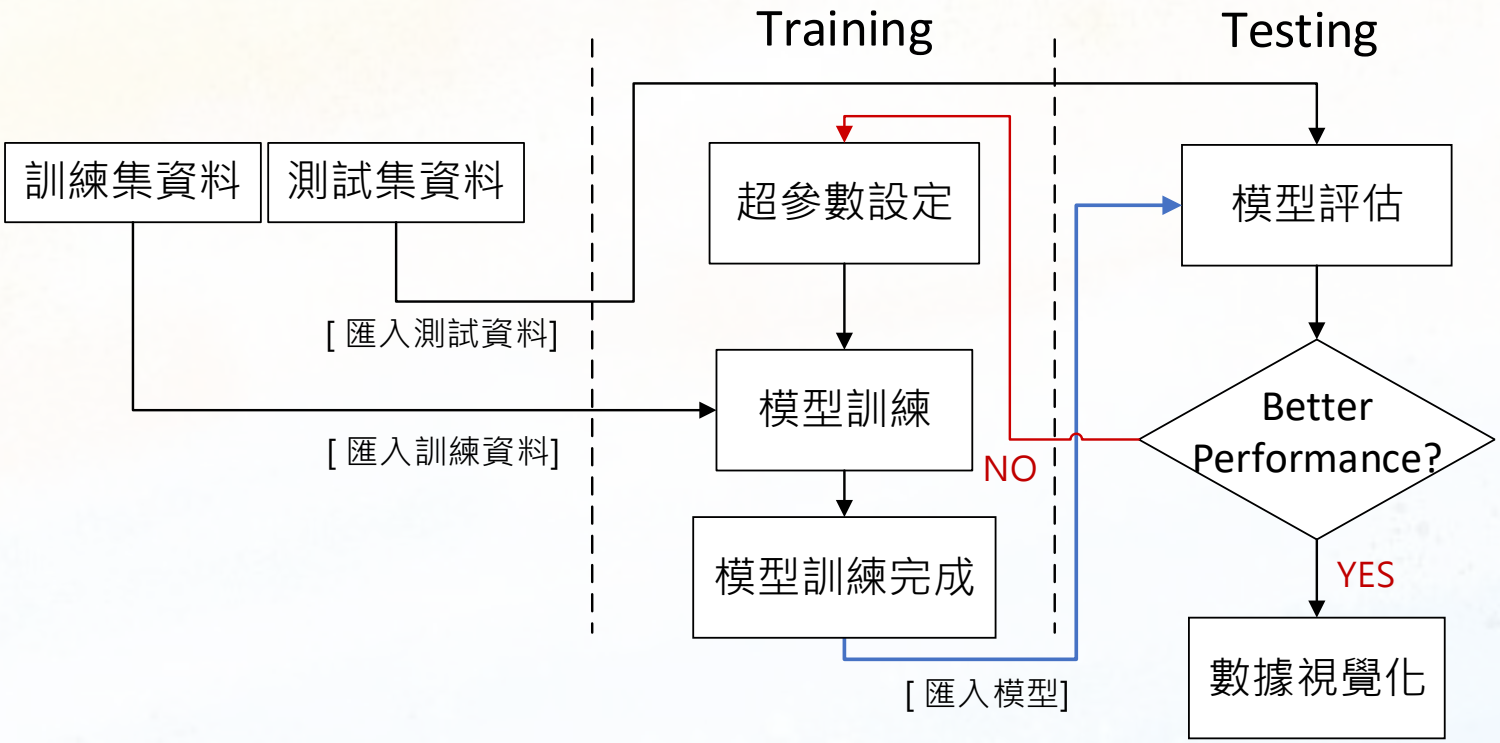
1. 門檻值：> 0.15
2. 門檻值： > 0.25
3. 門檻值： > 0.3
4. 特徵全保留



模型訓練與評估

模型訓練與評估

機器學習模型流程圖

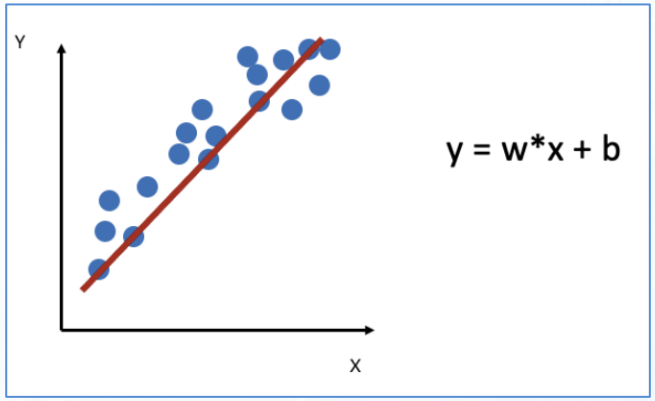


訓練集比例：測試集比例 = 8 : 2

評分指標：R-squared (r2_score)

Linear Regression

- Linear Regression (線性回歸)
  - 是統計上在找多個自變數和依變數之間的關係所建出來的模型
  - 透過計算出迴歸模型線的方程式，就能獲得預測值
  - 若遇到 underfitting，可提高迭代次數解決或提高模型複雜度
  - 若遇到 overfitting，確認是否訓練資料太少或迭代次數過多
  - 能適時使用正規化 (L1/L2) 處理



Linear Regression

- 模型測試結果：
  - 門檻值>0.15：

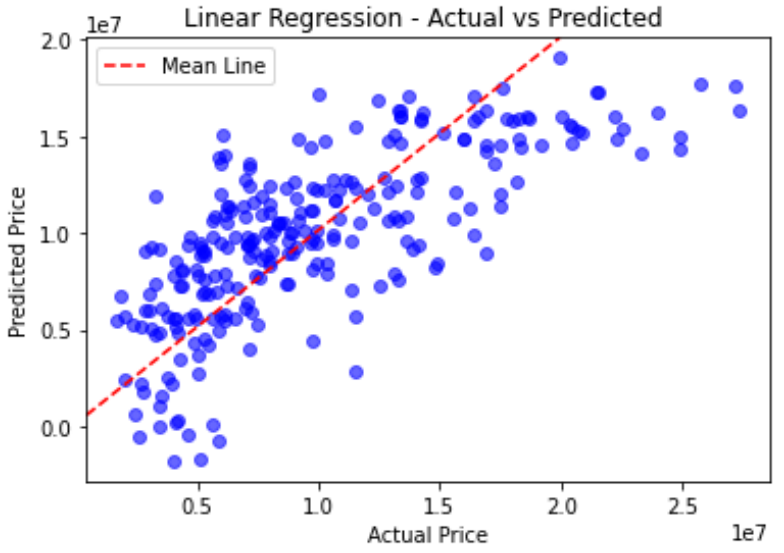
    R-squared Score: 53.2967
  - 門檻值>0.25：

    R-squared Score: 50.9247
  - 門檻值 > 0.33：

    R-squared Score: 40.1616
  - 全保留特徵：

    R-squared Score: 66.5357

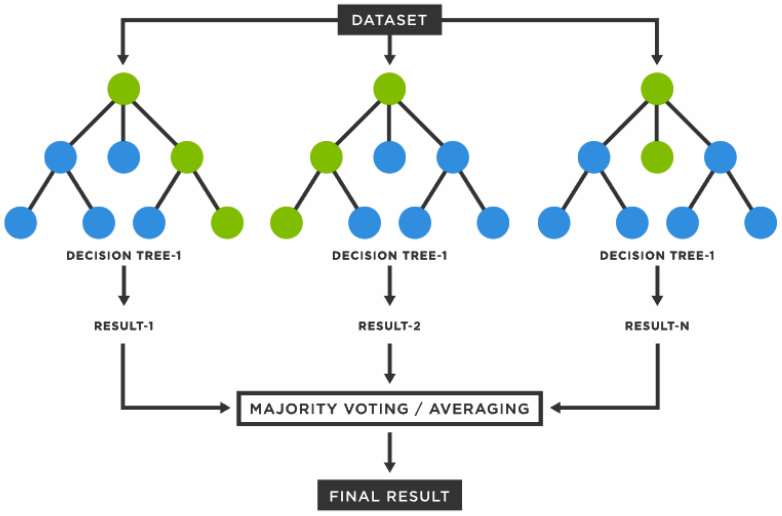
※ 使用預設模型參數



效果最好，因此決定使用未特徵選取的資料集

RandomForest

- Random Forest (隨機森林)
  - 由多棵決策樹所組成，進階版的決策樹
  - 使用 Bagging 加上隨機特徵採樣 (Bootstrap) 的方法所產生
  - 比較不易受極端值影響
  - 不足在於無法控制模型內部的執行



圖片來源： https://www.tibco.com/zh-hant/reference-center/what-is-a-random-forest

RandomForest

- **隨機取樣：**
  - 每棵樹的生成都會先從訓練集中隨機抽取 n 筆資料出來，而這 n 筆資料是可以被重複抽取的
- **隨機選取特徵：**
  - 每一棵樹都是從 n 筆資料中隨機挑選 k 個特徵做樣本
  - 隨機森林每一棵樹的特徵和資料可能都不同，所以最後決策出來的結果都會不一樣
- **回歸問題：**
  - M個決策樹模型的輸出 (預測結果) 取平均

RandomForest

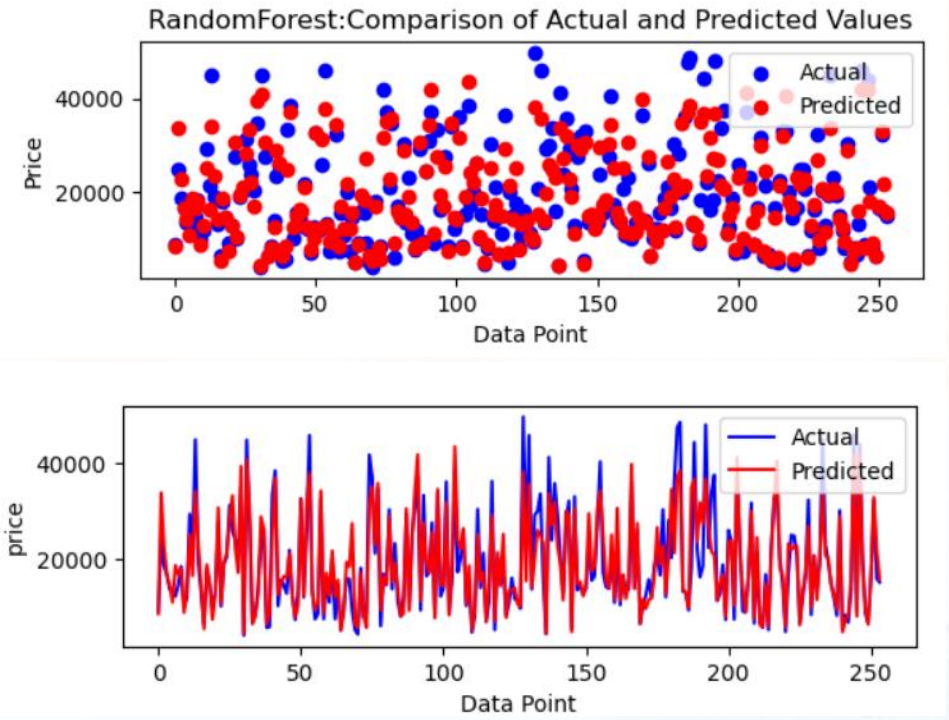
- **隨機森林模型參數**

1. n_estimators (決策樹數量)：隨機森林中決策樹的數量。
2. Criterion (分割標準)：用於衡量每個節點上分割的指標，常見的選項有 squared_error (均方誤差) 和 absolute_error (平均絕對誤差)。
3. max_depth (最大樹深度)：控制每棵決策樹的最大深度。
4. min_samples_split (節點分割所需的最小樣本數)：控制節點繼續分割所需的最小樣本數。
5. min_samples_leaf (葉節點所需的最小樣本數)：控制每個葉節點所需的最小樣本數。
6. max_features (拆分時考慮的特徵數量)：決定每個節點考慮的特徵數量。
7. Bootstrap (重複抽樣)：控制是否使用有放回的抽樣方式。當設置為 True 時，每個決策樹都會進行有放回的抽樣，從原始數據集中隨機抽取樣本。

RandomForest

- **參數設定與模型測試**

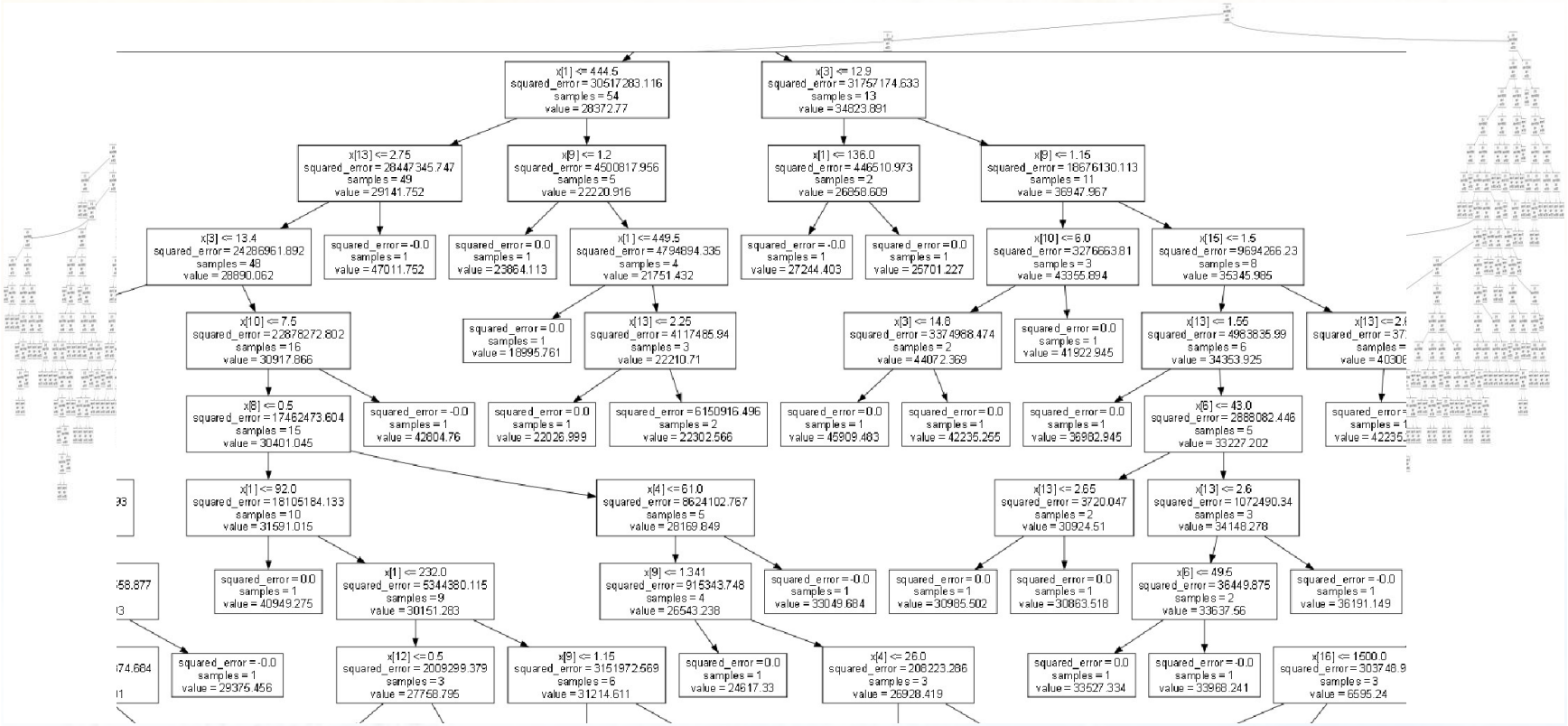
'bootstrap': True,
'ccp_alpha': 0.0,
'criterion': 'squared_error' ,
'max_depth': None,
'max_features': 1.0,
'max_leaf_nodes': None,
'max_samples': None,
'min_impurity_decrease': 0.0,
'min_samples_leaf': 1,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 100,
'n_jobs': None,
'oob_score': False,
'random_state': None,
'verbose': 0,
'warm_start': False



隨機森林R-squared Score: 83.04014

RandomForest

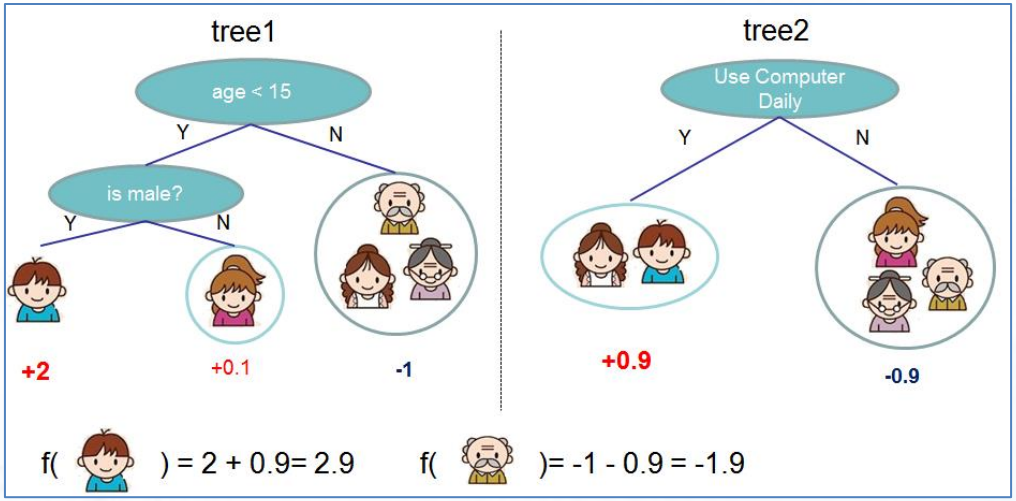
- 隨機森林的其中一棵樹：



XGBOOST

- eXtreme Gradient Boosting (XGBoost)
  - 基於 Gradient Boosting 做延伸
  - 結合 Bagging 和 Boosting 的優點
  - XGBoost 可想成是多個 CART 進行融合

dmlc XGBoost



圖片來源：https://ithelp.ithome.com.tw/articles/10301273

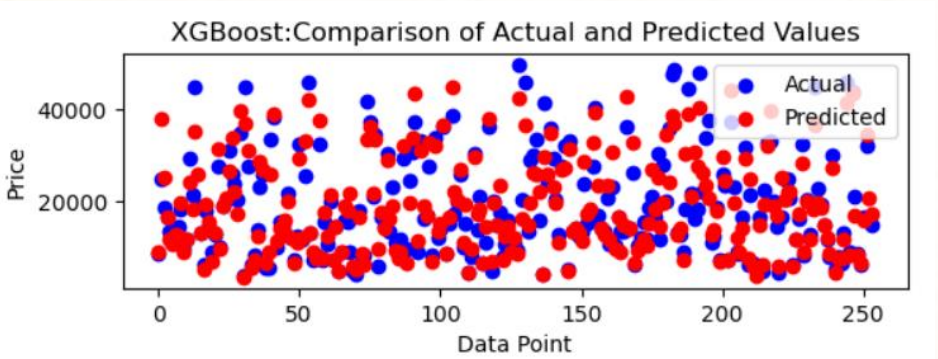
XGBOOST

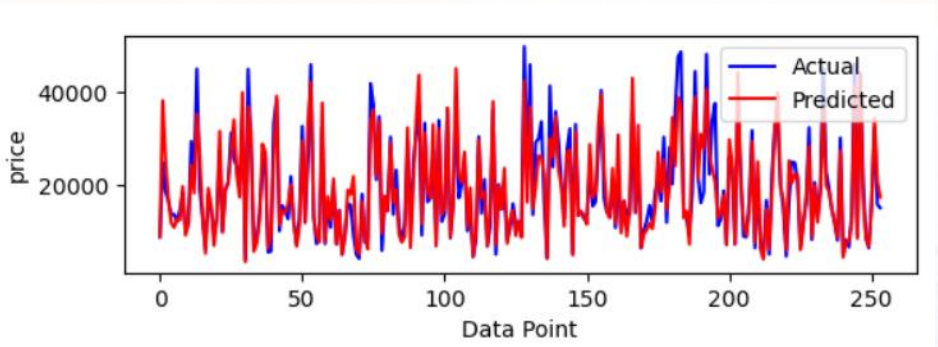
- RF 與 XGBOOST 差異：
- 隨機森林
  - 是從樣本中隨機抽出 n 筆資料，重複執行迴圈 m 次
  - 建立 m 顆完全不相關的決策樹，每一棵樹預測目標皆相同
  - 取所有預測結果的平均當作最後輸出結果
- XGBOOST
  - (實際結果 - 前 m - 1 顆樹的預測結果加總) 當作第 m 顆樹的預測目標

XGBOOST

- **參數設定與模型測試**

- learning_rate = 0.19
- n_estimators = 500
- max_depth = 8
- min_child_weight = 1





XGBOOST r2_score : 85.0658

結論與未來展望

結論

- 本研究使用 Kaggle 筆電價格資料集，且使用機器學習模型進行預測
  - Linear Regression
  - Random Forests
  - eXtreme Gradient Boosting
- 樹狀結構模型對於筆電價格預測有較佳的表現
- 測試集中最高評分結果是 XGBOOST 模型
  - r2_score : 85.0658

未來展望

- 針對資料集的前處理可以做的更好
  - 對各家廠牌的價格進行分析，將文書筆電與遊戲筆電分類
- 使用更佳的特徵選取演算法，提高準確率
- 使用其他機器學習模型進行預測
  - SVM、LightGBM

實驗設備與工作分配

- Device and tools
  - **硬體**
    - Windows 11 專業版 64位元
    - CPU：Intel(R) Core(TM) i5-13500 2.50GHZ
    - RAM：32G
  - **軟體**
    - 編譯工具：Spyder / Jupyter Notebook； 語言：python 3.7+
    - 套件：scikit-learn library、numpy、pandas、matplotlib、Xgboost、 regex
- 工作分配
  - 林垣志：資料前處理、XGBOOST 模型、Python 程式實作、期中&期末報告
  - 許至佑：特徵選取、線性回歸模型、Python程式實作、期末報告
  - 許楷俊：資料蒐集、隨機森林模型、Python 程式實作、期末報告

參考文獻

- Linear Regression

  [Python 實作] 迴歸模型 Regression – PyInvest (pyecontech.com)

- XGBoost 論文

  https://arxiv.org/pdf/1603.02754.pdf

  https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBClassifier

- XGBoost 實作

  https://ithelp.ithome.com.tw/articles/10301273

- Random Forest 【分類、回歸】

  https://ithelp.ithome.com.tw/articles/10303882?sc=iThelpR

- 相關 Python 功能操作

  https://blog.longwin.com.tw/2023/05/python-pandas-dataframe-append-no-attribute-2023/

感謝聆聽