

作業 2: 看我用方程式算命

是否很期待呢,我們終於要開始第一個分析專案了。

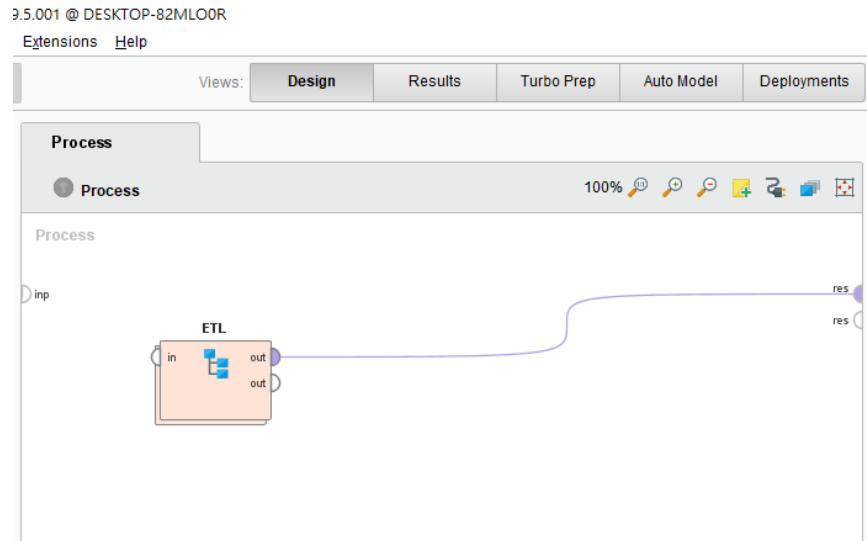
講到分析，不得不來了解一下大名鼎鼎的狠角色：迴歸方程式.

以下作業請先讀入 **baston** 檔案,並開始分析:

1. 讀入檔案後,請大概說明一下您對於資料的了解(10%),並進行「必要」之資料

前處理,並請將相關前處理存於 ETL 子流程如下圖(請說明您做了什麼,為什麼)

(10%)



對於資料的了解→須要以敘述統計表達

- 資料說明 (對於資料的了解)

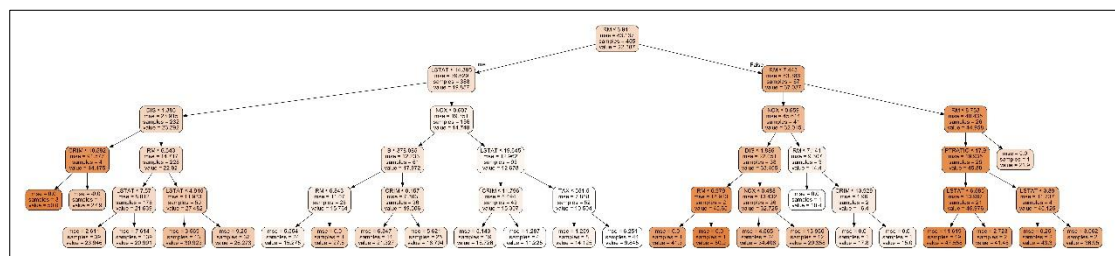
這個資料集主要是利用迴歸 (Regression) 來預測波士頓地區的房價，以下說

明各欄位：

欄位	內容	資料型態
CRIM	人均犯罪率	Real
ZN	25,000 平方英尺以上民用土地的比例	Real
INDUS	城鎮非零售業商用土地比例	Real
CHAS	是否鄰近查爾斯河，1 是鄰近，0 是不鄰近。	Integer
NOX	一氧化氮濃度 (千萬分之一)	Real

RM	住宅的平均房間數	Real
AGE	自住且建於 1940 年前的房屋比例	Real
DIS	到 5 個波士頓就業中心的加權距離	Real
RAD	到高速公路的便捷度指數	Integer
TAX	每萬元的房產稅率	Real
PTRATIO	城鎮學生教師比例	Real
B	$1000(B_k - 0.63)^2$ 其中 B_k 是城鎮中黑人比例	Real
LSTAT	低收入人群比例	Real
MEDV	自住房中位數價格，單位是千元	Real

分析結果大致呈現如下：



(上圖是我用 Python 及 Sklearn 做出來的分類樹)

然而這份 Boston 資料集因為存在道德爭議 (ethical problem)，例如 B 欄位城鎮中黑人比例，已經在 Sklearn 1.2 版本被移除了。

(The Boston housing prices dataset has an ethical problem. The scikit-learn maintainers therefore strongly discourage the use of this dataset unless the purpose of the code is to study and educate about ethical

issues in data science and machine learning.)



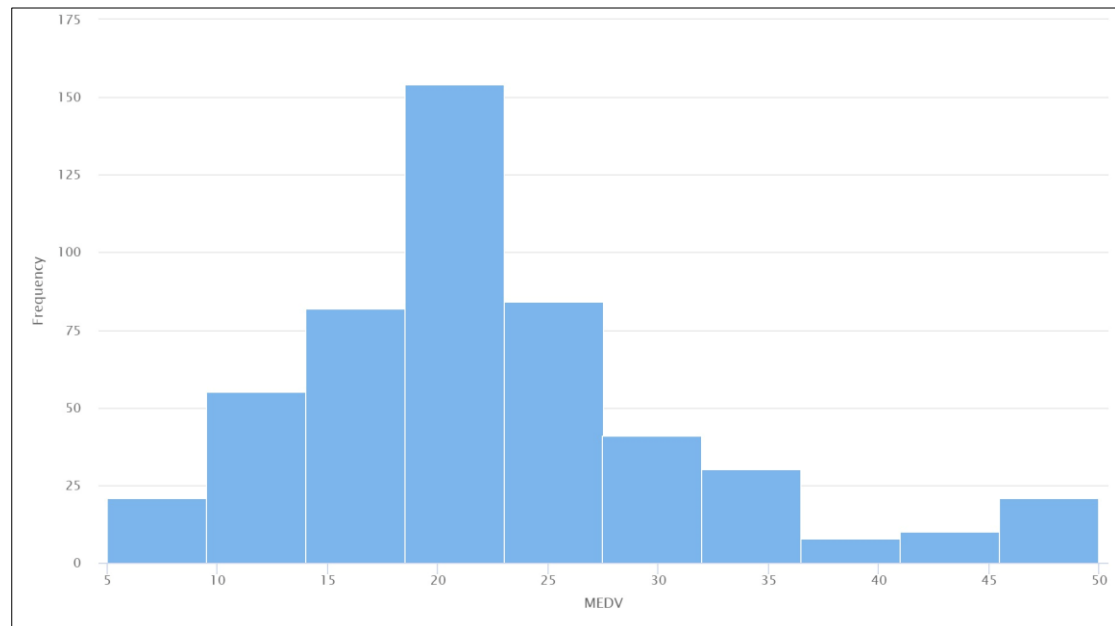
HomeWork_敘述
統計.xlsx

● 敘述統計

	平均數	中間值	標準差	偏度	偏態	最小值	最大值	Q1四分位數	Q3四分位數
CRIM	3.613523557	0.25651	8.601545105	5.223148798	右偏	0.00632	88.9762	0.08196	3.6819425
ZN	11.36363636	0	23.32245299	2.225666323	右偏	0	100	0	12.5
INDUS	11.13677866	9.69	6.860352941	0.295021568	右偏	0.46	27.74	5.175	18.1
CHAS	0.06916996	0	0.253994041	3.405904172	右偏	0	1	0	0
NOX	0.554695059	0.538	0.115877676	0.729307923	右偏	0.385	0.871	0.449	0.624
RM	6.284634387	6.2085	0.702617143	0.403612133	右偏	3.561	8.78	5.88475	6.626
AGE	68.57490119	77.5	28.14886141	-0.59896264	左偏	2.9	100	44.85	94.1
DIS	3.795042688	3.20745	2.105710127	1.011780579	右偏	1.1296	12.1265	2.09705	5.212575
RAD	9.549407115	5	8.707259384	1.004814648	右偏	1	24	4	24
TAX	408.2371542	330	168.5371161	0.669955942	右偏	187	711	279	666
PTRATIO	18.4555336	19.05	2.164945524	-0.802324927	左偏	12.6	22	17.375	20.2
B	356.6740316	391.44	91.29486438	-2.890373712	左偏	0.32	396.9	375.3	396.2325
LSTAT	12.65306324	11.36	7.141061511	0.906460094	右偏	1.73	37.97	6.9275	16.9925
MEDV	22.53280632	21.2	9.197104087	1.108098408	右偏	5	50	16.95	25

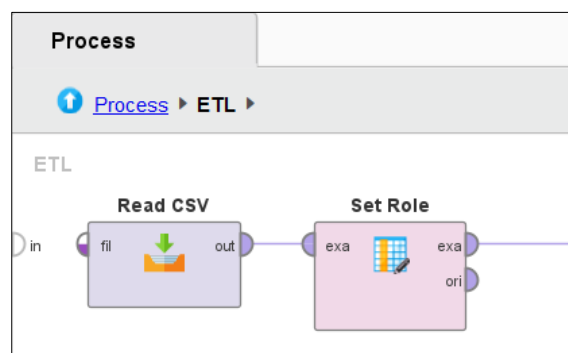
	MEDV	欄位說明
平均數	22.53280632	平均數 > 中位數
中位數	21.2	
標準差	9.197104087	越大代表大部分的數值和其平均值之間差異較大，資料較分散，越小代表大部分的數值和其平均值之間差異較小，資料較集中。
偏度	1.108098408	大部份的數值落在平均數的哪一邊，若資料分配較多集中在低數那方，稱為正偏態分配。
偏態	右偏	偏度 > 0，右偏
最小值	5	
最大值	50	

範圍	45	最大值-最小值
Q1 四分位數	16.95	將資料分為四等分後，第 25%的數值
Q3 四分位數	25	將資料分為四等分後，第 75%的數值



(↑ 初步畫圖觀察，MEDV 房價或許不是常態分布，有點左傾，資料分配較多集中在低數那方。)

● 資料前處理



1. 讀入 Data (Read CSV)

➤ 檔案路徑：C:\Users\acer0\OneDrive\桌面\boston.csv

➤ 總共 506 筆資料

2. 將 MEDV 設定為 label



因為 MEDV 是要預測的值，所以要將 MEDV 設定為 label。否則會發生

Missing Label 的錯誤，詳細請見第四題。

3. 資料欄位彼此皆相關，無須再做像是 Select Attribute 等處理。

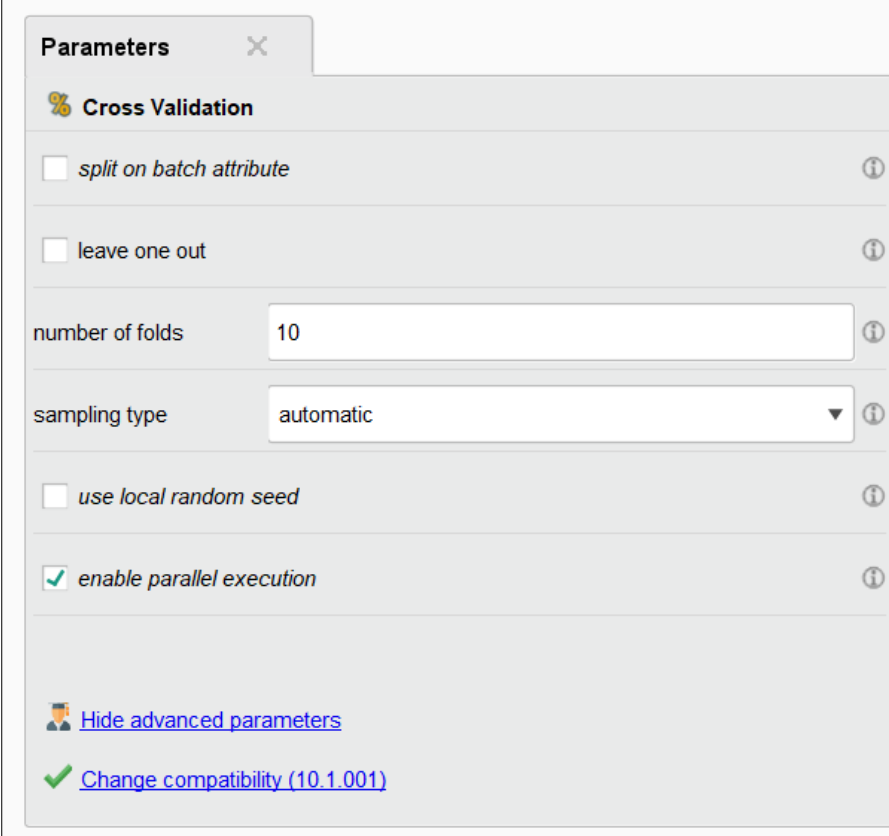
→但流程內有 *Select Attribute*，主要是為了第八題而做的步驟。

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
CRIM	-0.094	0.034	-0.092	0.855	-2.795	0.005	***
ZN	0.050	0.015	0.125	0.885	3.288	0.001	***
CHAS	3.061	0.938	0.085	0.982	3.262	0.001	***
NOX	-17.398	3.945	-0.223	0.815	-4.410	0.000	****
RM	3.634	0.457	0.277	0.570	7.958	0.000	****
AGE	0.015	0.015	0.045	0.815	0.984	0.326	
DIS	-1.444	0.212	-0.333	0.856	-6.802	0.000	****
RAD	0.274	0.069	0.257	0.762	3.956	0.000	****
TAX	-0.012	0.004	-0.213	0.736	-3.159	0.002	***
PTRATIO	-0.909	0.141	-0.216	0.814	-6.465	0.000	****
B	0.008	0.003	0.083	0.895	2.842	0.005	***
LSTAT	-0.570	0.057	-0.436	0.477	-9.998	0	****
(Intercept)	36.670	5.541	?	?	6.617	0.000	****

(↑ 讀取資料後的結果示意圖)

2. 請找出 cross validation, 並利用它進行模型的建置.何謂 cross validation?

(10%)



Parameters

Cross Validation

☐ split on batch attribute

☐ leave one out

number of folds: 10

sampling type: automatic

☐ use local random seed

☒ enable parallel execution

[Hide advanced parameters](#)

☒ [Change compatibility \(10.1.001\)](#)

- Number of folds：要將資料切成的份數，在這邊我寫 10 份，代表我要將資料切成 10 份。

(Break the Data Set into 10 smaller subsets of data. Train and build the model on nine of those. Keeping one of them for testing and measure the performance. Then iterate, Finally average the score.)

- Sampling Type：訓練集與測試集如何採樣。
 - ✓ Linear→Ex.前面 800 筆訓練，後面 200 筆測試。適合時間訓練分析，例如預測台積電股價。

✓ Shuffled→Ex.抽 8 筆跳過 2 筆 (80%, 20%) · 8 筆 (80%) 訓練、2 筆 (20%) 測試。

✓ Stratified→分層抽樣，抽取的比例盡量與母體相近。

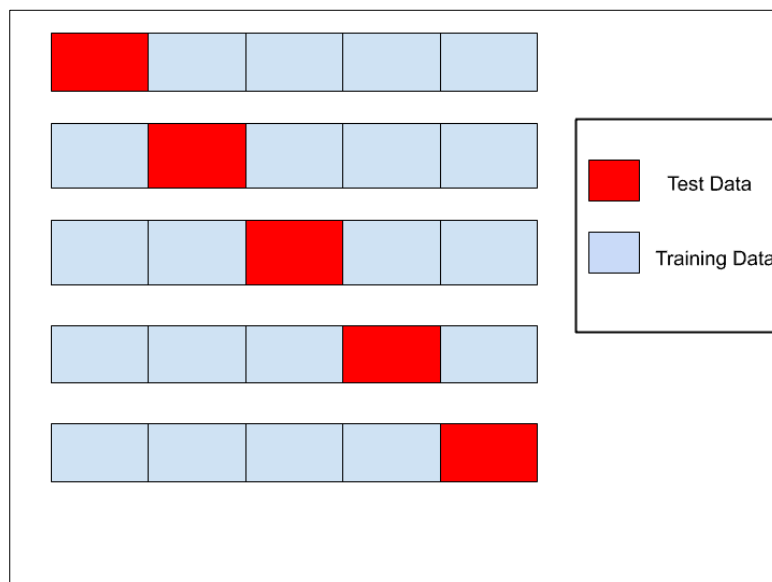
✓ Automatic→若 Label 是 Nominal 則用 Stratified，否則就用

Shuffled。在這邊我選擇 Automatic。

The Split Data operator can use several types of sampling for building the subsets. Following options are available:

- Linear sampling: Linear sampling simply divides the ExampleSet into partitions without changing the order of the examples i.e. subsets with consecutive examples are created.
- Shuffled sampling: Shuffled sampling builds random subsets of the ExampleSet. Examples are chosen randomly for making subsets.
- Stratified sampling: Stratified sampling builds random subsets and ensures that the class distribution in the subsets is the same as in the whole ExampleSet. For example in the case of a binominal classification, Stratified sampling builds random subsets such that each subset contains roughly the same proportions of the two values of the class labels.
- Automatic: Uses stratified sampling if the label is nominal, shuffled sampling otherwise.

● Cross Validation 交叉驗證：



K-fold Cross-Validation

交叉驗證指的是每次把一部分的 samples 拿出來當訓練，另外一部分拿來當

測試，經過多次訓練，每次的訓練資料皆不同，最後算出每次驗證的平均。

EX. 如上圖，將資料切成五份，並且跑五次，第一次用第一批次資料做測試集、第二次用第二批次資料做測試集...

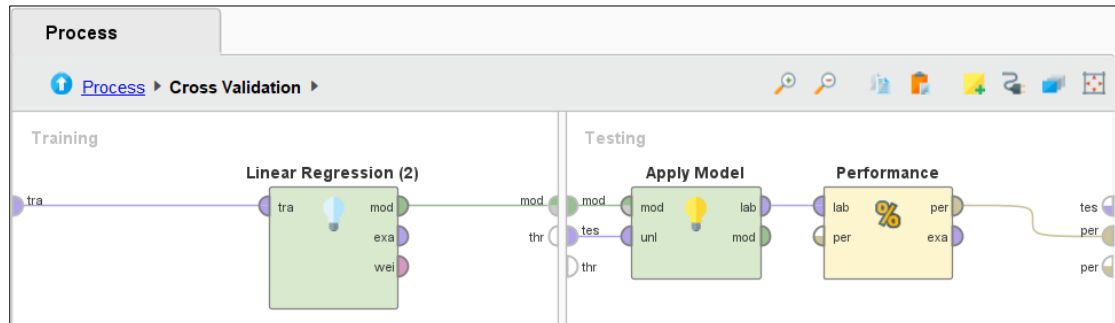
- 那為什麼不用 Train_Test_Split：

因為一次性的 Train_Test_Split 太過偏頗，太靠運氣。在少量樣本的狀況下，可能會抽到某些資料驗證出來覺得模型訓練得還不錯，但換抽另一批資料來驗證就又覺得模型訓練的很糟糕。而為了避免這個狀況，可以比較有效的來評估模型的好壞，這時候我們就會採用「交叉驗證 Cross-Validation」的方法來做驗證。

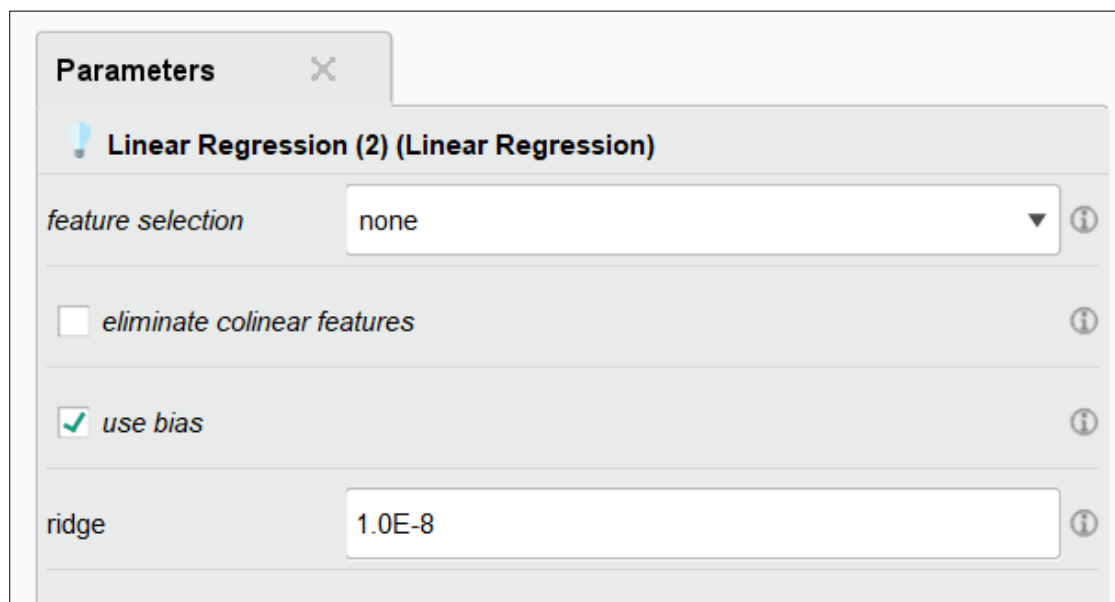
- [補充] 留一法交叉驗證 Leave-One-Out Cross Validation：

資料切成的份數等於數據集中數據的個數，並且每次只使用一筆作為測試集，剩下的全部作為訓練集。假設只有十筆資料，切分成十份(number of folds=10)，一個批次只有一筆測試資料，預測準確率僅 100%或 0% (只有對或錯，沒有答對率 90%這種中間值)。雖然這樣的方式可以讓我們瞭解哪筆資料明顯造成我們模型的偏差，但是這樣一筆一筆驗證的方式非常的消耗計算資源和時間，成本過於龐大。

3. 請在 cross validation 中進行迴歸的建構，並請貼出您的流程圖，並請說明相關之參數設定 (10%)



Linear Regression



➤ Feature Selection :

當有眾多欄位時，應當如何選取重要欄位特徵。(strategies for reducing the number of features)

當有眾多欄位時，進行降維，篩除不需要的欄位，降低雜訊。(In general it's a good idea to have as few influence factors as possible for your model, so it's less susceptible for noise and errors.)

然而我們也不希望因為降維而篩除了潛在資訊。(On the other hand, you don't want to lose potential information. So it's always a trade off between selecting the right amount of features.)

因此並沒有絕對的準則，應該要用哪一個條件。(There's no golden rule which selection strategy gives you the best results)

M5 is also called M5 Prime, selects a subset of attributes, which improves the Akaike information criterion the most.

T-test performs the statistical test of the same name to consider if a feature has a significant influence on the target class.

Greedy is a forward selection strategy, where each round the attribute with the lowest contribution (again based on the Akaike information criterion) is deselected.

(↑ 上圖為 Feature Selection 的各參數說明。)

	RMSE
None	4.786
M5	4.801
Greedy	4.773
T-test	4.773

(↑ 各參數執行後的 RMSE 數值結果。)

我認為資料欄位彼此皆相關，所以我會選擇 None，先暫且不要任何 Feature Selection。

➤ eliminate colinear features :

This is an important protection when we do have those attributes that

are highly correlated with each other or identical.

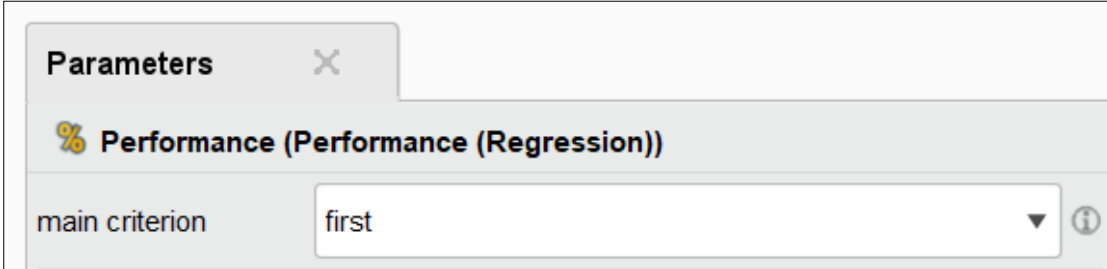
➤ Use Bias :

Allow us to determine whether an intercept value will be calculated or not.

➤ Ridge Regression :

As the value here approaches 0, it approaches a true least square regression model, and the larger the value will be the bigger the regularization constant, so that it will be more robust to linearity.

Performance (Regression)



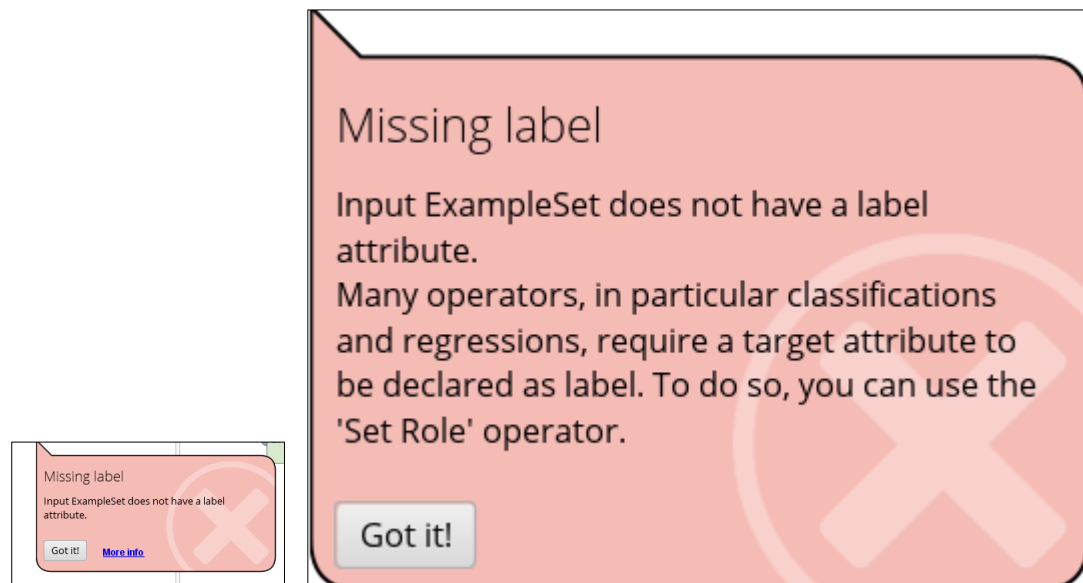
The screenshot shows a 'Parameters' window with a close button (X). Below the title bar, there is a section titled 'Performance (Performance (Regression))' with a yellow icon. Under this section, there is a 'main criterion' label and a dropdown menu currently set to 'first'. An information icon (i) is visible to the right of the dropdown.

➤ main criterion : 衡量比較向量，在這邊直接選用預設值，first。

- **main_criterion**

The main criterion is used for comparisons and needs to be specified only for processes where performance vectors are compared, e.g. attribute selection or other meta optimization process setups. If no *main criterion* is selected, the first criterion in the resulting performance vector will be assumed to be the *main criterion*.

4. 請問以下錯誤訊息是指？該如何解決(5%)



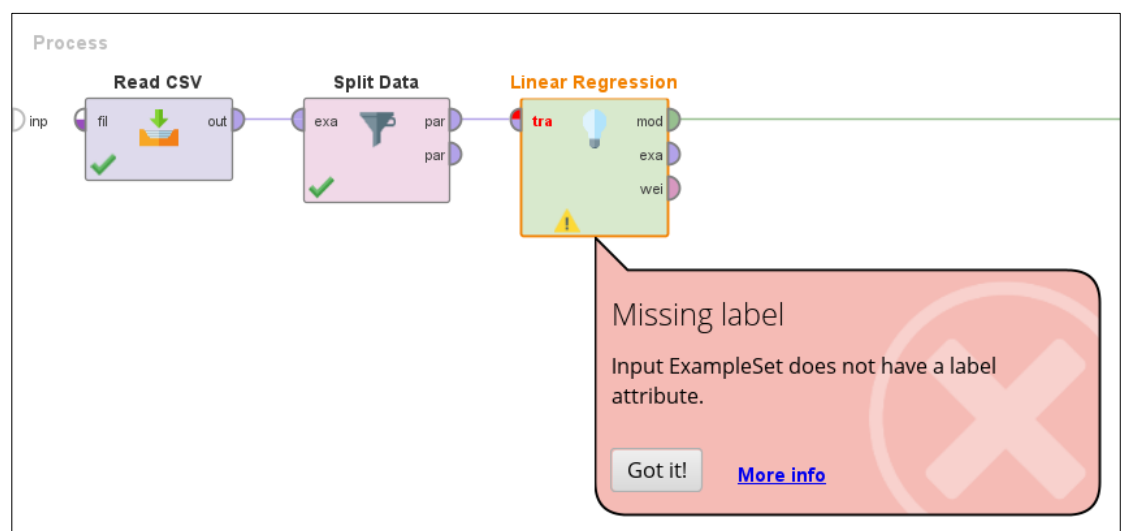
- Missing Label

- 訊息白話文：沒有標註 Label 或 Label 設定錯誤

- 推測原因 (三種)：

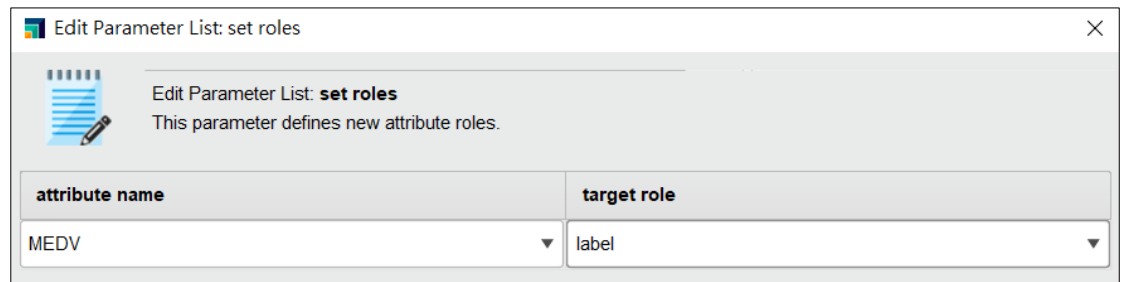
1. 沒有指定要預測哪一個欄位及沒有利用 Set Role 設定標籤。

→示意圖：



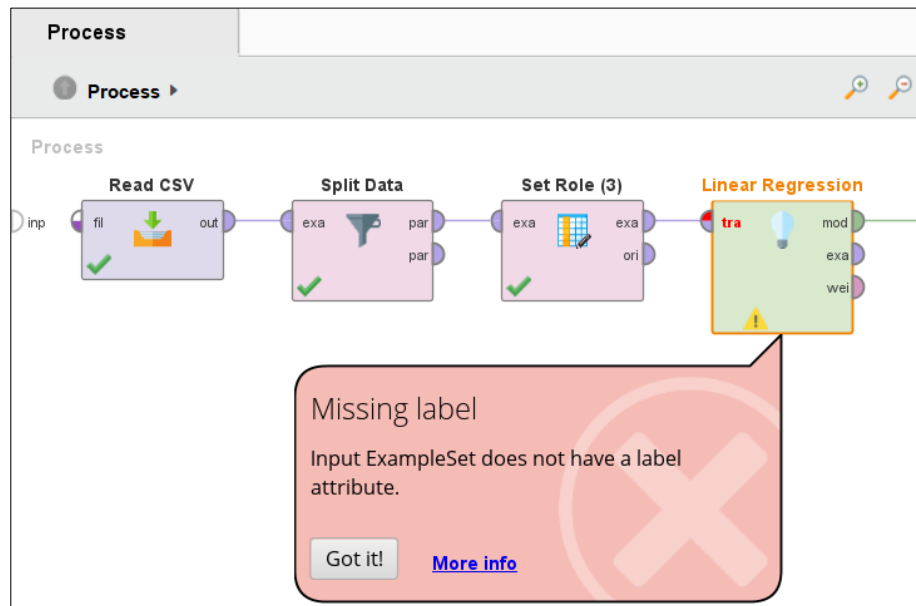
→解決方法：利用 Set Role 設定標籤 (Label)。在這份作業中要預測

的是房價，所以要將 MEDV 設定為 label。



2. 已經放置 Set Role，卻仍出現錯誤。

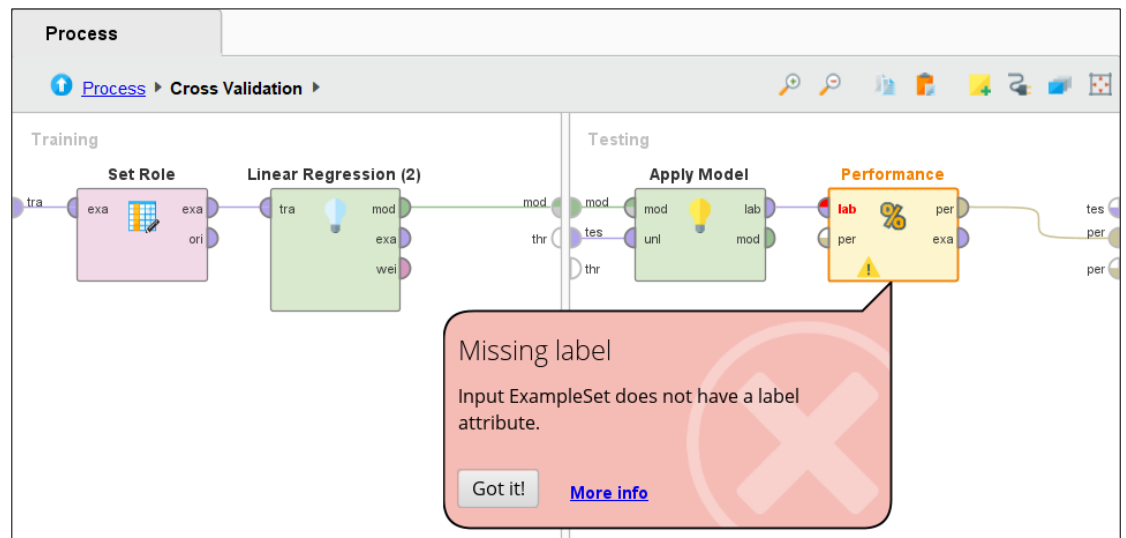
→示意圖：



→解決方法：Set Role 設定錯誤，請參照上面圖片的設定，將 MEDV 設定為 label。

3. 已經放置 Set Role 且將 MEDV 設定為 label，仍然出現錯誤。

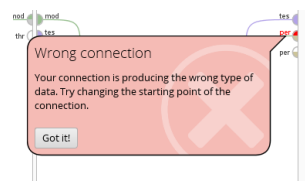
→示意圖：



→解決方法：Set Role 放置位置錯誤，因為將 Set Role 放在 Cross Validation 裡了，這將導致有資料可以訓練，但到了要驗證模型的時候，卻沒有資料可以與預測值比較。(if you set the label within your cross-validation your model will be able to train but when it comes to the validation side your Performance operator will have nothing to compare with the prediction.) 應該將 Set Role 放在外層。

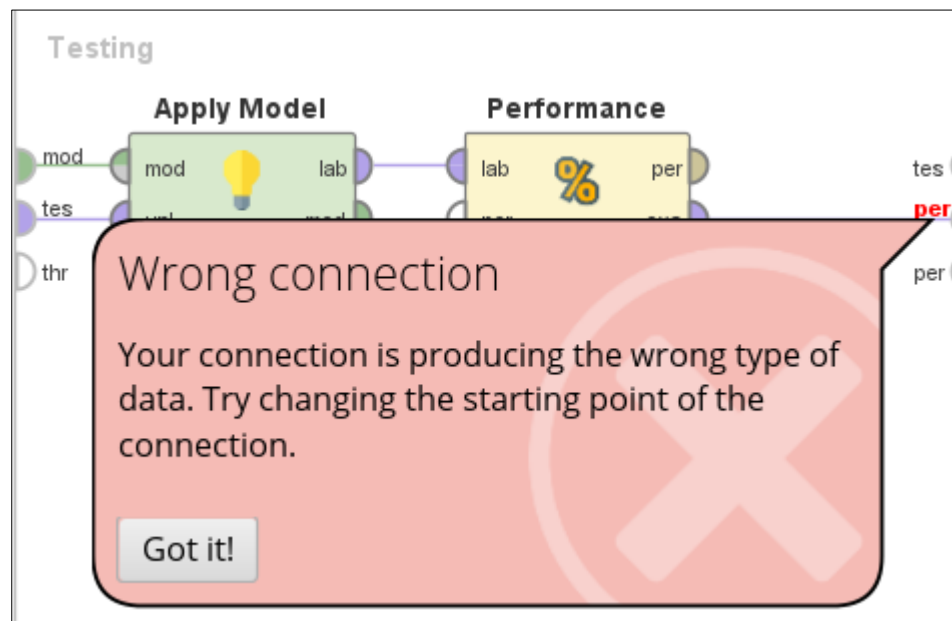
You may have several errors here. First, I suspect your selected attribute was used in Set Role but not be set to "label". Second, if you set the label within your cross-validation your model will be able to train but when it comes to the validation side your Performance operator will have nothing to compare with the prediction. You should Set Role outside the loop.

5. 請問以下錯誤訊息是指？該如何解決(5%)

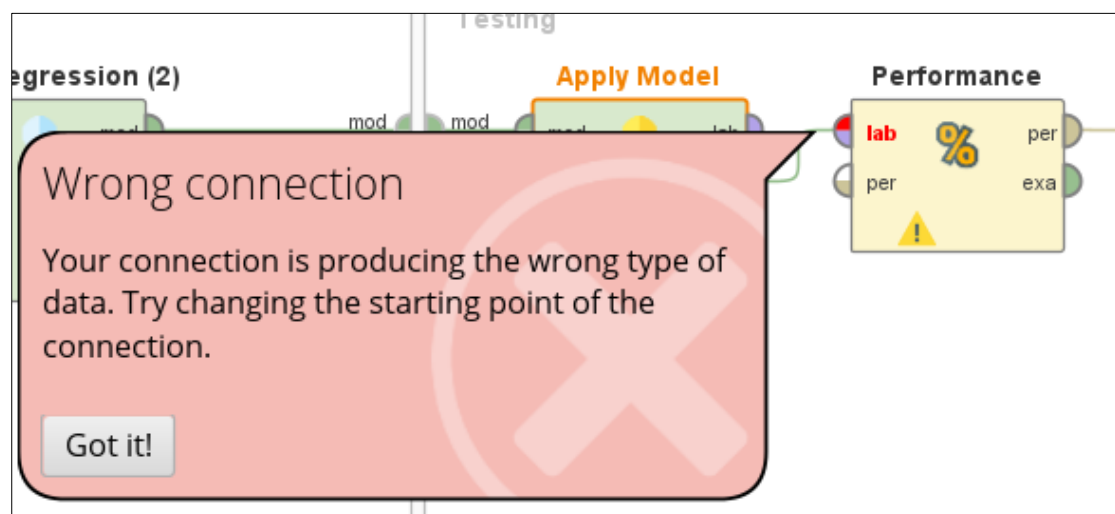


資料連接錯誤，下圖是我自行產生的錯誤：

- Performance 應當與 Performance 連接，而不是與 Example 連接。

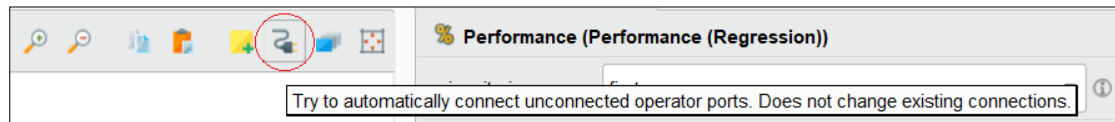


- Labelled Data 應當與 Labelled Data 連接，而不是與 Model 連接。



解決方式：

- 仔細查看資料類型後，重新連接。
- 或利用系統自動連接 (Try to automatically connect unconnected operator ports.)



6. 若您的模型在建構過程中有上述問題，請試著解決並以 MEDV 為預測標

的，並加以說明您的方程式意涵(10%)

- 我在建構模型的過程中，並無發生任何問題。
- 以 MEDV 為預測標的，建構模型後如下表格。

MEDV	prediction(MEDV)	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
24	30.004	0.006	18	2.310	0	0.538	6.575	65.200	4.090
21.600	25.026	0.027	0	7.070	0	0.469	6.421	78.900	4.967
34.700	30.568	0.027	0	7.070	0	0.469	7.185	61.100	4.967
33.400	28.607	0.032	0	2.180	0	0.458	6.998	45.800	6.062
36.200	27.944	0.069	0	2.180	0	0.458	7.147	54.200	6.062
28.700	25.256	0.030	0	2.180	0	0.458	6.430	58.700	6.062
22.900	23.002	0.088	12.500	7.870	0	0.524	6.012	66.600	5.561
27.100	19.536	0.145	12.500	7.870	0	0.524	6.172	96.100	5.950
16.500	11.524	0.211	12.500	7.870	0	0.524	5.631	100	6.082
18.900	18.920	0.170	12.500	7.870	0	0.524	6.004	85.900	6.592
15	18.999	0.225	12.500	7.870	0	0.524	6.377	94.300	6.347
18.900	21.587	0.117	12.500	7.870	0	0.524	6.009	82.900	6.227
21.700	20.907	0.094	12.500	7.870	0	0.524	5.889	39	5.451
20.400	19.553	0.630	0	8.140	0	0.538	5.949	61.800	4.707
18.200	19.283	0.638	0	8.140	0	0.538	6.096	84.500	4.462
19.900	19.297	0.627	0	8.140	0	0.538	5.834	56.500	4.499
23.100	20.528	1.054	0	8.140	0	0.538	5.935	29.300	4.499

- 方程式意涵

LinearRegression

```

- 0.108 * CRIM
+ 0.046 * ZN
+ 0.021 * INDUS
+ 2.687 * CHAS
- 17.767 * NOX
+ 3.810 * RM
+ 0.001 * AGE
- 1.476 * DIS
+ 0.306 * RAD
- 0.012 * TAX
- 0.953 * PTRATIO
+ 0.009 * B
- 0.525 * LSTAT
+ 36.459

```

各個數值乘上權重係數後得出的預測值。

LinearRegression

```

- 0.108 * CRIM
+ 0.046 * ZN
+ 0.021 * INDUS
+ 2.687 * CHAS
- 17.767 * NOX
+ 3.810 * RM
+ 0.001 * AGE
- 1.476 * DIS
+ 0.306 * RAD
- 0.012 * TAX
- 0.953 * PTRATIO
+ 0.009 * B
- 0.525 * LSTAT
+ 36.459
          
```

- 1.) 權重係數*資料數值
- 2.) 相加
- 3.) 再加上Bias=36.459
- 4.) 得出預測值

權重係數*資料數值

例如以下這筆資料

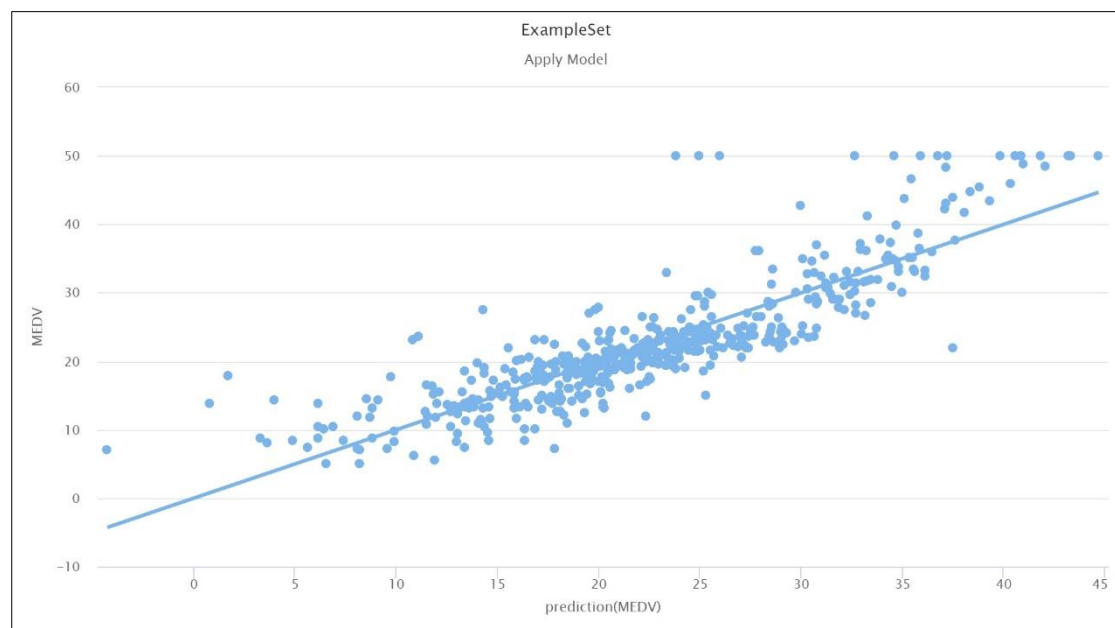
資料標題	原始資料數值
CRIM	14.0507
ZN	0
INDUS	18.1
CHAS	0
NOX	0.597
RM	6.657
AGE	100
DIS	1.5275
RAD	24
TAX	666

PTRATIO	20.2
B	35.05
LSTAT	21.22
MEDV	17.2

預測值 $=(-0.108*14.0507)+(0.046*0)+(0.021*18.1)+(2.687*0)+(-$
 $17.767*0.597)+(3.81*6.657)+(0.001*100)+(-1.476*1.5275)+(0.306*24)+(-$
 $0.012*666)+(-0.953*20.2)+(0.009*35.05)+(-$
 $0.525*21.22)+36.459=17.199655$

與實際值差異(Absolute Error)= ABS(預測值-實際值)= $17.199655-$
 $17.2=0.0003446$

- 以 Scatter 圖表畫出迴歸方程式



橫軸代表 MEDV 的預測數值 (Prediction) ，縱軸代表 MEDV 的實際數值 。

- Regression (linear) MEDV

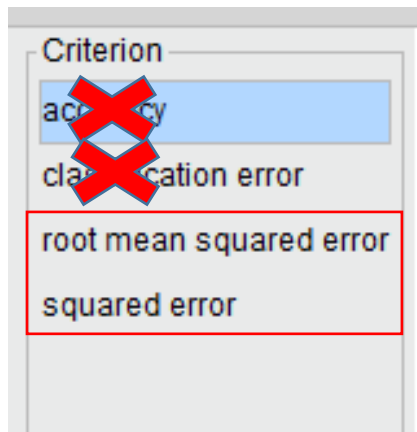
X: 44.67308338286236, Y: 44.67308338524772

Linear function: $1.0000000001077396x - 2.427700187563675e-9$

當模型越準確 (Perfect Match)，實際值=預測值，也就是 $y=x$ ，則 X 的係數

會是 1，斜率為 1，呈 45 度角。

7.) 列出模型的指標，並請簡略說明他們的計算方式及含意(10%)



Accuracy、ClassificationError 是分類問題才有的評估指標，因此只要顯示 root mean squared error 及 squared error。

- root mean squared error (RMSE)

root_mean_squared_error

root_mean_squared_error: 4.786 +/- 1.029 (micro average: 4.884 +/- 0.000)

- Mean Squared Error (MSE) 均方誤差

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

MSE 是計算模型預測值和實際值相差的平方總合除以資料數量，因為平方的特性，若預測值距離實際值誤差越大，MSE 也就越大，換句話說，當單一 bias 大的時候會有懲罰作用，對於極值 (outliers)會相對敏感。而 **RMSE 就是它的平方根，愈小表示模型愈準確。**

- Absolute Error

absolute_error

absolute_error: 3.390 +/- 0.586 (micro average: 3.390 +/- 3.516)

➤ MAE (Mean Absolute Error) 平均絕對誤差

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

MAE 是計算計算模型預測值和實際值相差的絕對值總合除以資料數量。

(抵銷正負誤差的方式，除了平方之外，還有取絕對值，MAE 就是取絕對值來計算平均誤差)，相對之下對極值比較不敏感，如果資料集裡面極值很多，那可以考慮用 MAE 來當作指標。

- 比較 RMSE 與 MAE

平均

預測值	MAE	RMSE
預測1: 8 & 4	$MAE = \frac{1}{2}(8-6 + 4-6) = 2$	$RMSE = \sqrt{\frac{1}{2}((8-6)^2 + (4-6)^2)} = 2$
預測2: 10 & 6	$MAE = \frac{1}{2}(10-6 + 6-6) = 2$	$RMSE = \sqrt{\frac{1}{2}((10-6)^2 + (6-6)^2)} = \sqrt{8} = 2.8$

$MAE_1 = MAE_2$
 $RMSE_1 < RMSE_2$
 (找幾個值來看)

假設正確值是 6 (True=6)

- 狀況一：預測值為 8 及 4，則 MAE=2、RMSE=2
- 狀況二：預測值為 10 及 6，則 MAE=2、RMSE= $2\sqrt{2}$

→由此可知，RMSE 對於越偏離正確值的極值越敏感。

- Relative Error 相對誤差

relative_error

relative_error: 16.99% +/- 3.15% (micro average: 16.98% +/- 19.17%)

相對誤差
Relative Error =

$$\frac{\text{真確值} - \text{估值}}{\text{真確值}}$$

- squared_error

squared_error

squared_error: 23.861 +/- 9.965 (micro average: 23.854 +/- 65.844)

The averaged squared error. 平均平方誤差

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2$$

公式： (← 已於 04/07 請教老師)

但是,您用了幾乎全部的變數.幾乎全部的變數. 幾乎全部的變數.

8.以下我們將限制您的變數使用量，

8-1 請問若只讓您使用 5 個變數進行預測,您會選擇(請列出變數名稱)? (5%)

8-2 請問若只讓您使用 3 個變數進行預測,您會選擇(請列出變數名稱)? (5%)

8-3 您的選擇依據是? (5%) 變數變少後,正確率的變化是? (5%)

→ 呈現答案就好

- 選擇依據：我會根據各資料欄位與預測值 (MEDV 房價) 之間的

Correlation 相關性來選擇變數。

Attribute	Correlation	取絕對值後的 Correlation
LSTAT	-0.73766273	0.737662726
RM	0.695359947	0.695359947
PTRATIO	-0.50778669	0.507786686
INDUS	-0.48372516	0.48372516
TAX	-0.46853593	0.468535934
NOX	-0.42732077	0.427320772
CRIM	-0.38830461	0.388304609
RAD	-0.38162623	0.381626231
AGE	-0.37695457	0.376954565
ZN	0.360445342	0.360445342

B	0.33346082	0.33346082
DIS	0.249928734	0.249928734
CHAS	0.175260177	0.175260177



- 若僅使用五個變數，則我會選擇 LSTAT、RM、PTRATIO、INDUS、TAX

root_mean_squared_error

root_mean_squared_error: 5.204 +/- 1.195 (micro average: 5.326 +/- 0.000)

RMSE 由原本的 4.786 變為 5.204 (Worsen)

- 若僅使用三個變數，則我會選擇 LSTAT、RM、PTRATIO

root_mean_squared_error

root_mean_squared_error: 5.189 +/- 1.174 (micro average: 5.307 +/- 0.000)

RMSE 由原本的 4.786 變為 5.189 (Worsen)

→ 這告訴我們隨意刪除資料欄位 (篩選特徵重要程度) 會影響模型的準確度 ;

即使再怎麼看起來不相關的資料欄位 , 但只要刪除後 , 就會影響模型的準確度

(通常是變差) 。因此不要自己隨意決定特徵的重要性 , 讓模型去判斷就好。

- [補充敘述] 另外一種可以選取重要特徵的方式 , 是根據 Code 的星星數量。

(One of the easiest ways to determine which attributes play the strongest role model is to look at the number of asterisks under code)

Attribute	Coefficient	Code	有多少*
ZN	0.046420458	****	4
NOX	-17.7666111	****	4
RM	3.809865208	****	4
DIS	-1.47556684	****	4
RAD	0.306049479	****	4

PTRATIO	-0.95274723	****	4
B	0.009311683	****	4
LSTAT	-0.52475838	****	4
CRIM	-0.10801136	***	3
CHAS	2.686733818	***	3
TAX	-0.01233459	***	3
INDUS	0.020558626		0
AGE	6.92E-04		0

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code ↓
ZN	0.046	0.014	0.118	0.870	3.382	0.001	****
NOX	-17.767	3.820	-0.224	0.806	-4.651	0.000	****
RM	3.810	0.418	0.291	0.574	9.116	0	****
DIS	-1.476	0.199	-0.338	0.842	-7.398	0.000	****
RAD	0.306	0.066	0.290	0.758	4.613	0.000	****
PTRATIO	-0.953	0.131	-0.224	0.794	-7.283	0.000	****
B	0.009	0.003	0.092	0.904	3.467	0.001	****
LSTAT	-0.525	0.051	-0.407	0.475	-10.347	0	****
(Intercept)	36.459	5.103	?	?	7.144	0.000	****
CRIM	-0.108	0.033	-0.101	0.850	-3.287	0.001	***
CHAS	2.687	0.862	0.074	0.985	3.118	0.002	***
TAX	-0.012	0.004	-0.226	0.732	-3.280	0.001	***
INDUS	0.021	0.061	0.015	0.679	0.334	0.738	
AGE	0.001	0.013	0.002	0.807	0.052	0.958	

但因為有 8 個資料欄位 (ZN、NOX、RM、DIS、RAD、PTRATIO、B、LSTAT) 的 Code 星星數量皆為 4，並且 RapidMiner 軟體並沒有將各欄位的 Code 以確切的數值呈現，無法知道這 8 個欄位彼此之間的大小關係。若以這方式再從這 8 個中選取 5 個或 3 個，我認為是有爭議的，所以我才改用 Correlation 相關係數來觀察特徵影響的幅度，並作為我的選擇依據。

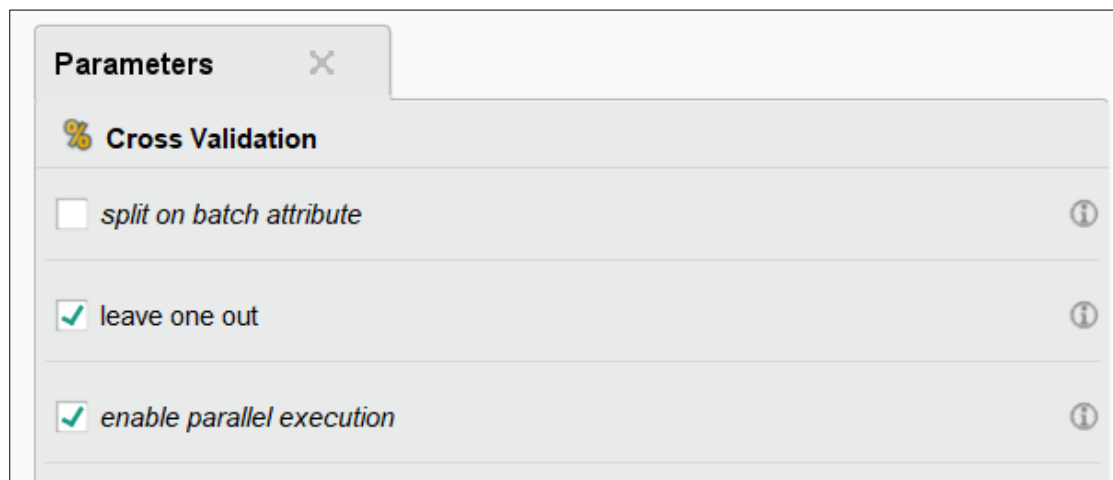
9.最後，在 5 個變數的限制下，請調整參數，找出一條正確率最高的迴歸方程

式，並請說明您試過那些參數的調整？ (10%)

請將您的 process 存檔為學號-1, 如: 106AB001_1.rmp 檔

→ explain 解釋

- 主要調整參數：



為了最大化準確率，將 CV 切割資料的方法改為留一法交叉驗證 Leave-One-Out Cross Validation。準確率的確明顯提升 (以 RMSE 作為判斷基準)，且因為僅有 506 筆資料，運算時間並沒有跟著大幅增加。

- 其它調整參數：

我還試過調整 Linear Regression 中的 feature selection，下面的表示是各參數調正結果，以 None 最佳。

	RMSE
None	3.665
M5	3.667

Greedy	3.697
T-test	3.669

- RMSE=3.665 (從 4.786 下降為 3.665)

root_mean_squared_error

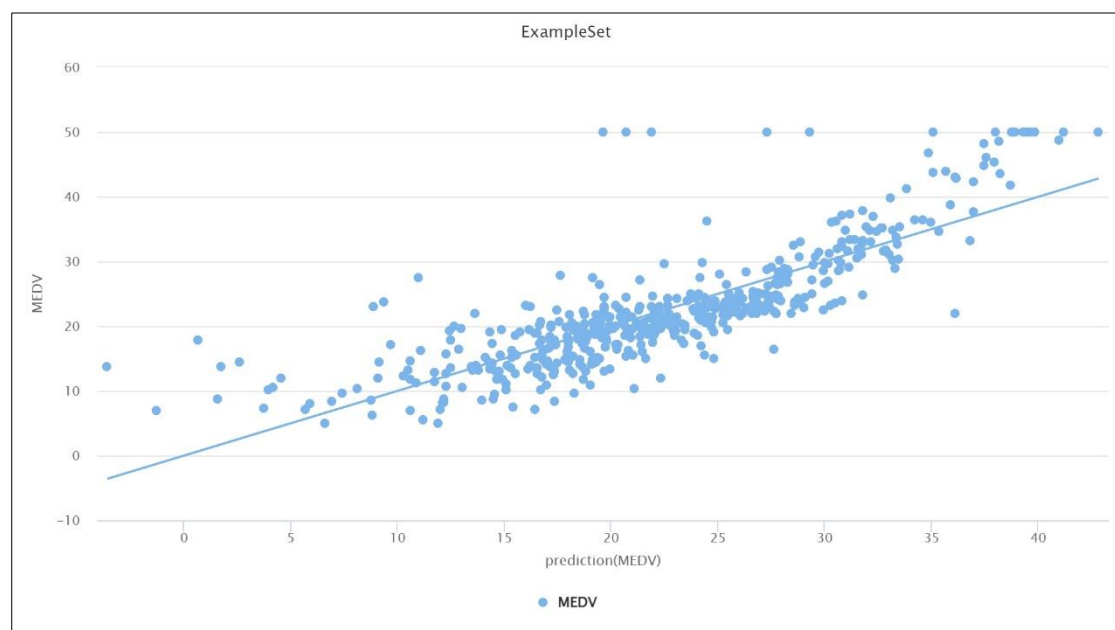
root_mean_squared_error: 3.665 +/- 3.836 (micro average: 5.303 +/- 0.000)

- 迴歸方程式

LinearRegression

```
0.057 * INDUS
+ 4.625 * RM
- 0.004 * TAX
- 0.876 * PTRATIO
- 0.559 * LSTAT
+ 17.518
```

- 以 Scatter 圖表畫出迴歸方程式



● Regression (linear) MEDV

X: 42.81551738181244, Y: 42.815517381969435

Linear function: $1.0000000000007741x - 1.7443964586376026e-10$

Reference

[Boston 資料集簡述]

- sklearn.datasets.load_boston

[https://scikit-](https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html)

[learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html](https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html)

- What impacts Boston Housing Prices

[https://medium.com/li-ting-liao-tiffany/python-](https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7)

[%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-](https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7)

[%90-boston-](https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7)

[housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B](https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7)

[9-9c535fb7ceb7](https://medium.com/li-ting-liao-tiffany/python-%E5%BF%AB%E9%80%9F%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-boston-housing%E6%B3%A2%E5%A3%AB%E9%A0%93%E6%88%BF%E5%83%B9-9c535fb7ceb7)

[敘述統計]

- 資料分析 03 統計學- 敘述統計

[https://medium.com/@momuschao/%E8%B3%87%E6%96%99%E5%88%](https://medium.com/@momuschao/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-03-%E7%B5%B1%E8%A8%88%E5%AD%B8-%E6%95%98%E8%BF%B0%E7%B5%B1%E8%A8%88-701acd6b6277)

[86%E6%9E%90-03-%E7%B5%B1%E8%A8%88%E5%AD%B8-](https://medium.com/@momuschao/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-03-%E7%B5%B1%E8%A8%88%E5%AD%B8-%E6%95%98%E8%BF%B0%E7%B5%B1%E8%A8%88-701acd6b6277)

[%E6%95%98%E8%BF%B0%E7%B5%B1%E8%A8%88-701acd6b6277](https://medium.com/@momuschao/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-03-%E7%B5%B1%E8%A8%88%E5%AD%B8-%E6%95%98%E8%BF%B0%E7%B5%B1%E8%A8%88-701acd6b6277)

- 呼叫 Excel 資料分析增益集工具

<https://www.youtube.com/watch?v=KoBqDuitcLQ>

- 利用 Excel 資料分析工具進行敘述統計分析

https://www.youtube.com/watch?v=_K4Zwls1qyw

- 應用統計學 敘述統計 EXCEL 操作

<https://www.youtube.com/watch?v=OZDXPF8S7g8>

[Cross Validation 相關資料]

- 【機器學習】交叉驗證 Cross-Validation

<https://jason-chen-1992.weebly.com/home/-cross-validation>

- 機器學習：交叉驗證！

<https://ithelp.ithome.com.tw/articles/10197461>

- 留一法交叉驗證 Leave-One-Out Cross Validation

<https://blog.csdn.net/baishuiniyaonulia/article/details/122052893>

- Validating a Model

<https://academy.rapidminer.com/learn/video/validating-a-model>

- Linear Regression demo

<https://academy.rapidminer.com/learn/video/linear-regression-demo>

- RapidMiner Tutorial - How to run a linear regression using cross-validation in RapidMiner

<https://www.youtube.com/watch?v=HouVO6mkcUA&t=439s>

[missing label 錯誤]

- How I get out of missing label error

<https://community.rapidminer.com/discussion/57235/how-i-get-out-of-missing-label-error>

- 機器學習_學習筆記系列(13)：交叉驗證(Cross-Validation)和 MSE、MAE、R2

<https://tomohiroliu22.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98%E7%B3%BB%E5%88%97-13-%E4%BA%A4%E5%8F%89%E9%A9%97%E8%AD%89-cross-validation-%E5%92%8Cmse-mae-r2-bc8fef393f7c>

- 迴歸模型的衡量標準：MSE. RMSE. MAE. MPE

<https://ithelp.ithome.com.tw/articles/10274551?sc=rss.iron>

- 什麼是平均絕對誤差 Mean Absolute Error, MAE ?

<https://staruphackers.com/%E4%BB%80%E9%BA%BC%E6%98%AF%E5%B9%B3%E5%9D%87%E7%B5%95%E5%B0%8D%E8%AA%A4%E5%B7%AE-mean-absolute-error-mae%EF%BC%9F/>

- What M5, greedy and T-test is meaning

<https://community.rapidminer.com/discussion/59224/what-m5-greedy-and-t-test-is-meaning>