作業2: 看我用方程式算命

請將您的process存檔為學號-1, 如: 106AB001\_1.rmp檔

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%)

一張含有 文字, 螢幕擷取畫面, 室內, 電腦 的圖片

自動產生的描述

* 資料說明（對於資料的了解）

這個資料集主要是利用回歸來預測波士頓地區的房價。

|  |  |  |
| --- | --- | --- |
| 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(Bk − 0.63)2 其中Bk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：

一張含有 圖表 的圖片

自動產生的描述

（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。

* 敘述統計

一張含有 資料表 的圖片

自動產生的描述

|  |  |  |
| --- | --- | --- |
|  | 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 |

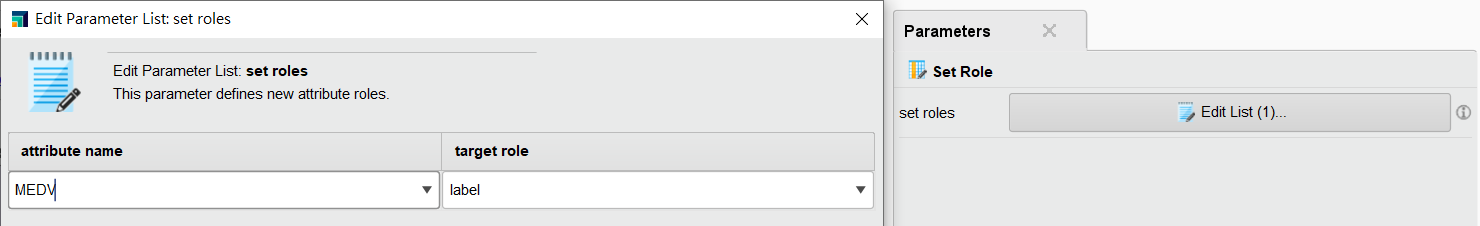
* 平均數>中位數
* 有25
* 資料前處理

一張含有 圖表 的圖片

自動產生的描述

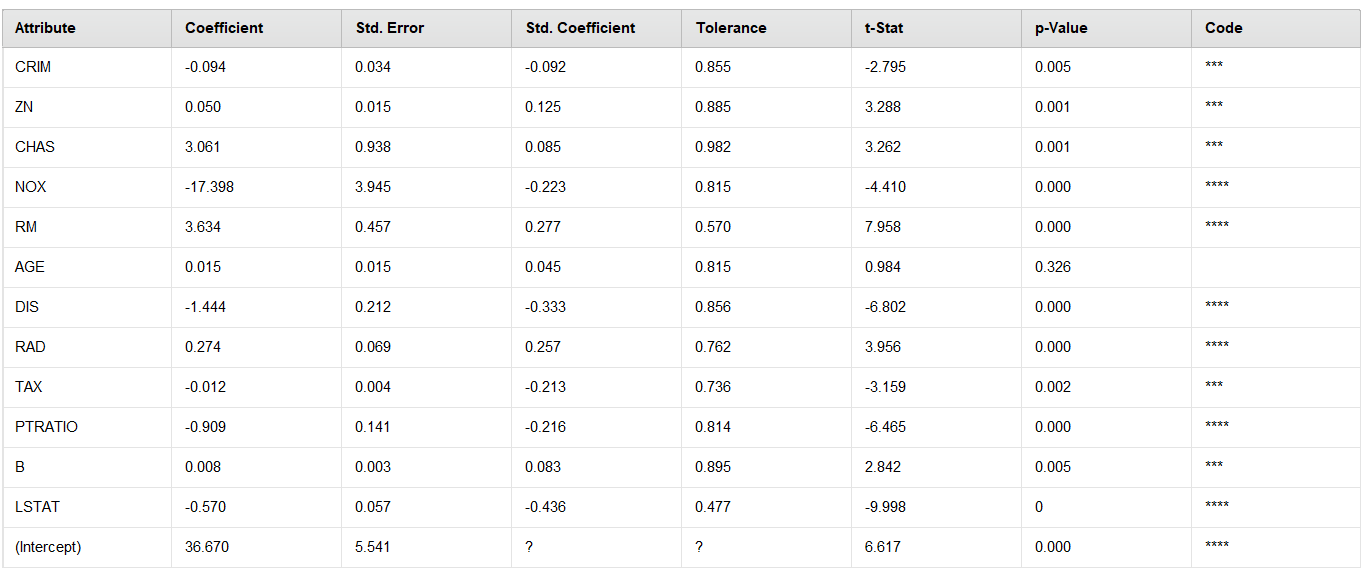
1. 讀入Data（Read CSV）

* 檔案路徑：C:\Users\acer0\OneDrive\桌面\boston.csv
* 總共506筆資料

1. 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。

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



2. 請找出cross validation, 並利用它進行模型的建置.何謂cross validation? (10%)

一張含有 文字 的圖片

自動產生的描述

* 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。

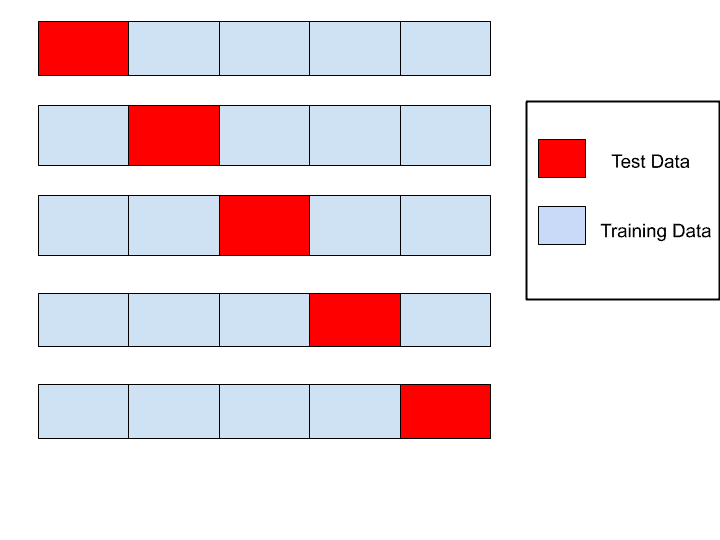
（Break the Data Set into 10 smaller subsets of data. Train and build the model on 9 of those. Keeping one of them for testing and measure the performance. Then iterate, Finally average the score.）

* 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Lable是Nominal則用Stratified，否則就用Shuffled。在這邊我選擇Automatic。

一張含有 文字, 信 的圖片

自動產生的描述

* 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中進行回歸的建構，並請貼出您的流程圖，並請說明相關之參數設定 (10%)

一張含有 圖表 的圖片

自動產生的描述

Linear Regression

一張含有 文字 的圖片

自動產生的描述

* 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）  
  一張含有 文字 的圖片

  自動產生的描述  
  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 |

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）

一張含有 文字 的圖片

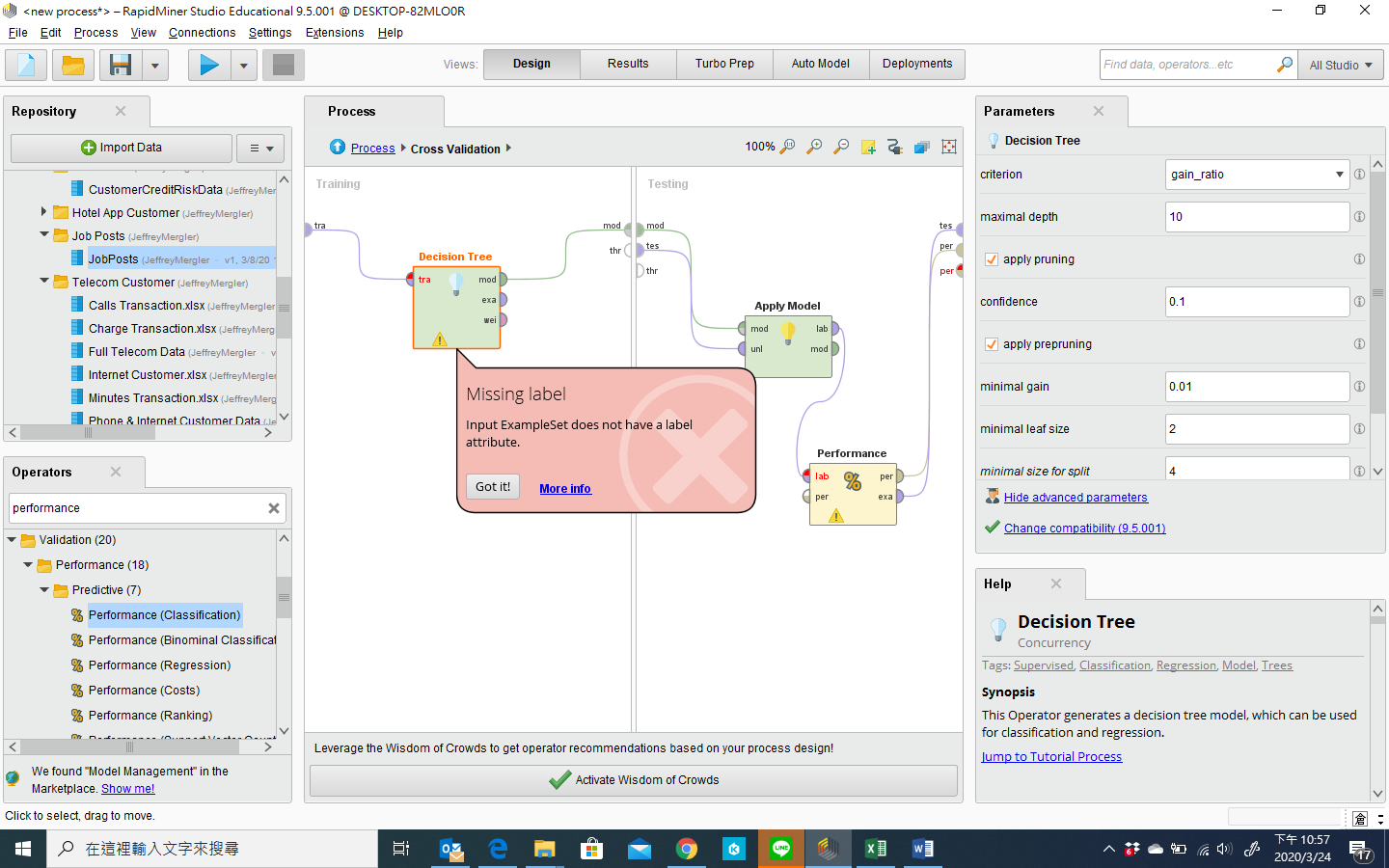
自動產生的描述

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

一張含有 文字 的圖片

自動產生的描述

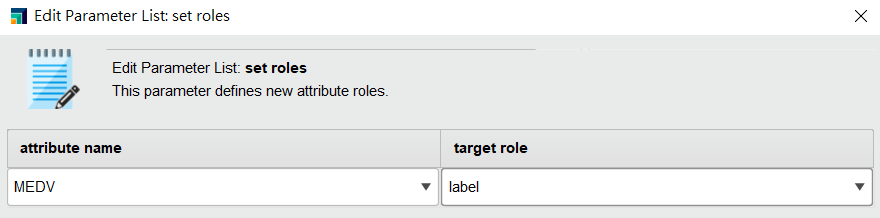
4. 請問以下錯誤訊息是指? 該如何解決(5%)

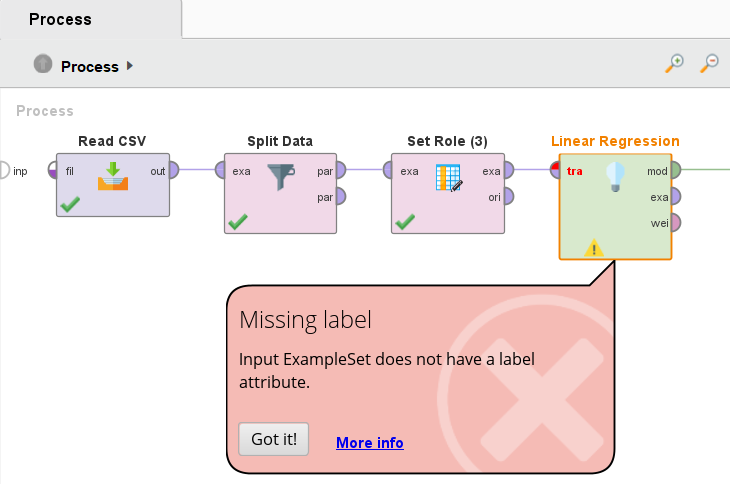
 一張含有 文字 的圖片

自動產生的描述

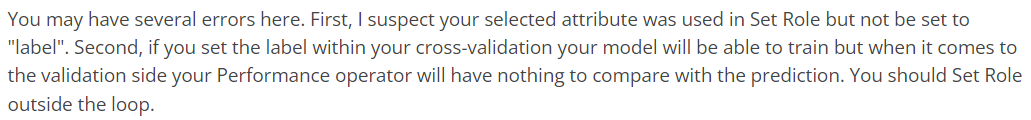
* Missing Label
* 訊息白話文：沒有標註Label 或 Label設定錯誤
* 推測原因（三種）：
  1. 沒有指定要預測哪一個欄位及沒有利用Set Role設定標籤。  
     →示意圖：

一張含有 圖表 的圖片

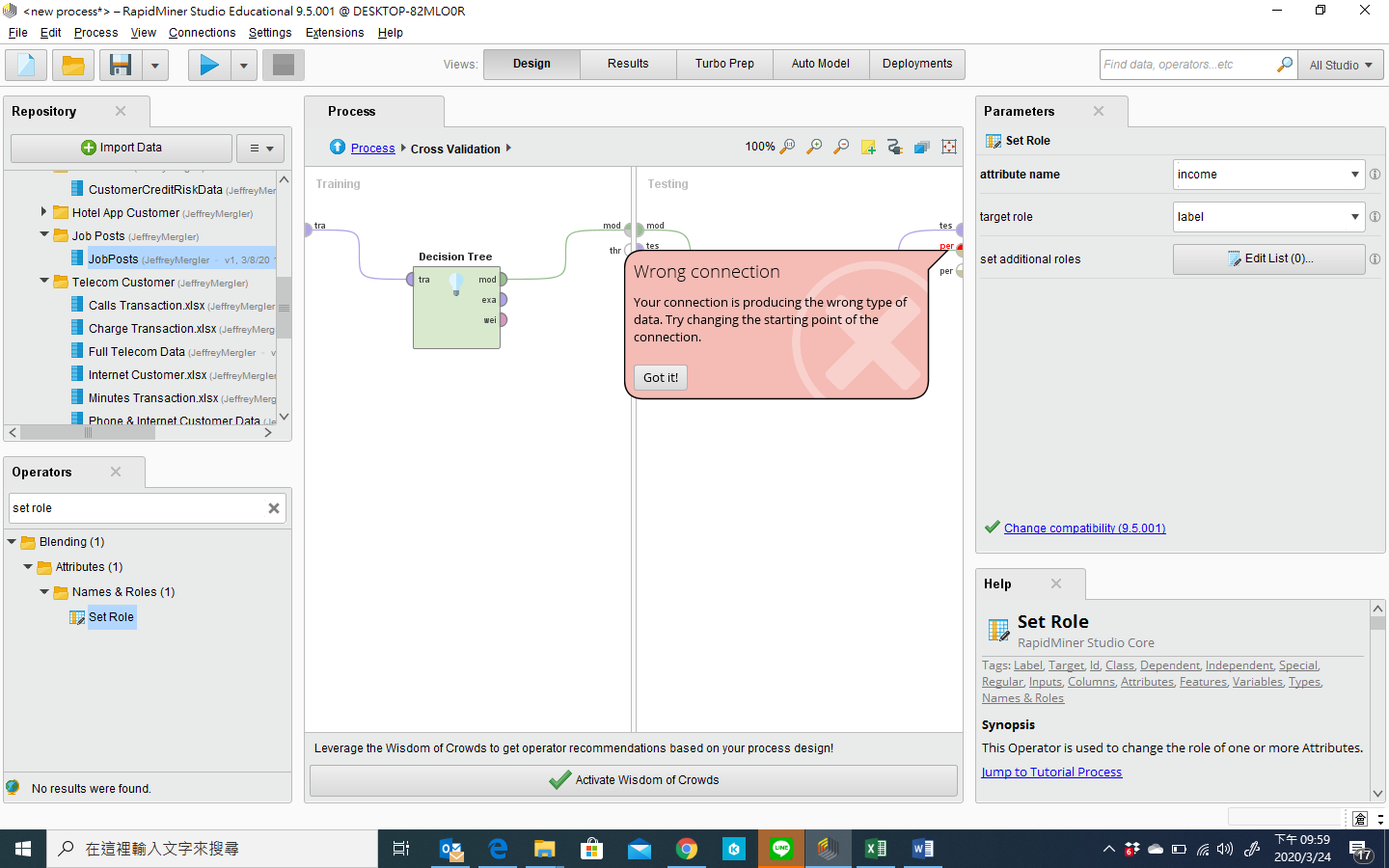
自動產生的描述  
→解決方法：利用Set Role設定標籤（Label）。在這份作業中要預測的是房價，所以要將MEDV設定為label。  


* 1. 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。
  2. 已經放置Set Role且將MEDV設定為label，仍然出現錯誤。  
     →示意圖：一張含有 圖表 的圖片

     自動產生的描述

→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。

5. 請問以下錯誤訊息是指? 該如何解決(5%)



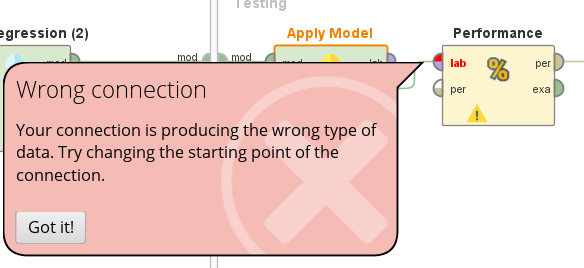
資料連接錯誤，下圖是我自行產生的錯誤：

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

一張含有 圖表 的圖片

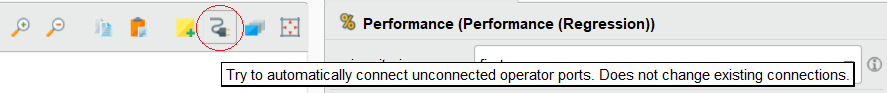
自動產生的描述

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



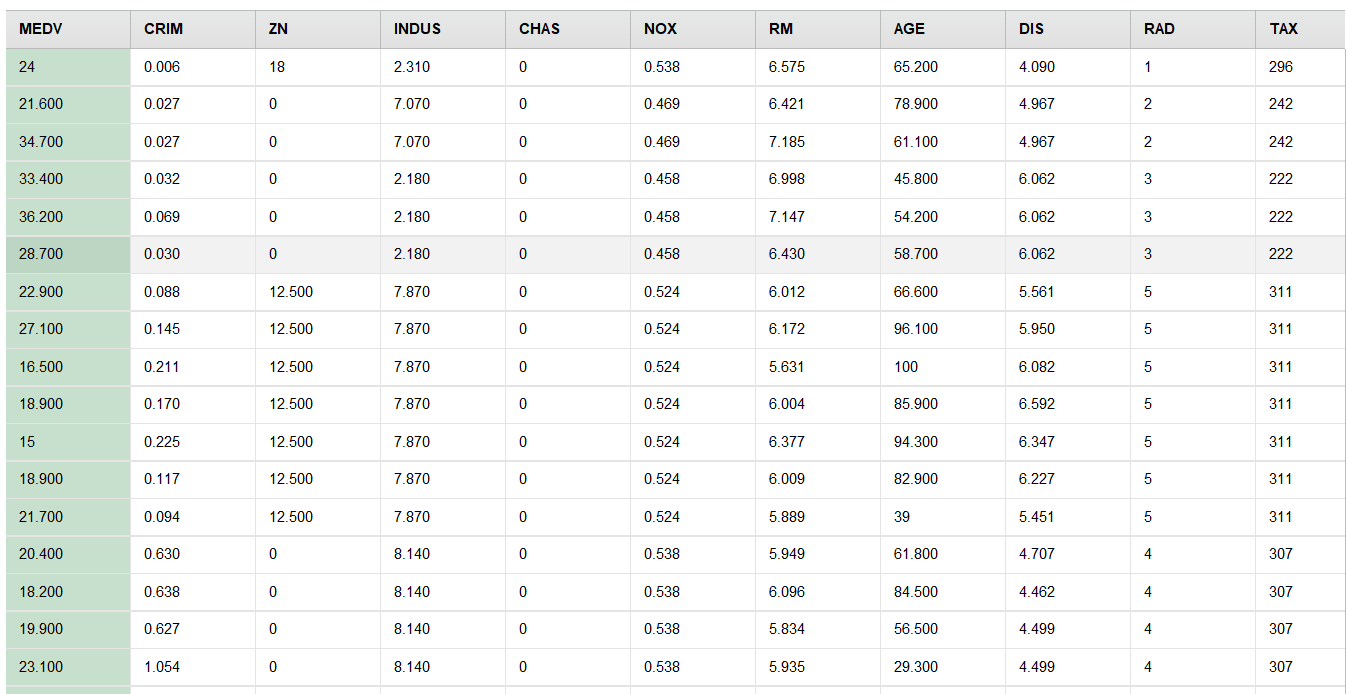
解決方式：

* 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。



* 方程式意涵

一張含有 資料表 的圖片

自動產生的描述

各個數值乘上權重係數後得出的預測值。

一張含有 文字 的圖片

自動產生的描述

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圖表畫出回歸方程式

一張含有 圖表 的圖片

自動產生的描述

橫軸代表MEDV的實際數值，縱軸代表MEDV的預測數值

一張含有 文字 的圖片

自動產生的描述

當模型越準確（Perfect Match），實際值=預測值，y=x，則X的係數會是1，斜率為1，呈45度角。

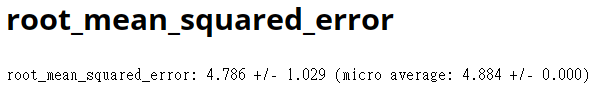
7.) 列出模型的指標，並請簡略說明他們的計算方式及含意(10%)



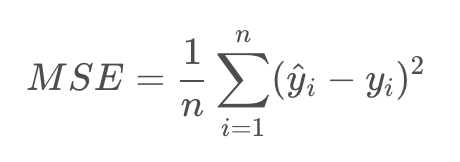
Accuracy、ClassificationError→分類問題才有

因此只要顯示root mean squared error及squared error

* root mean squared error （RMSE）



* Mean Squared Error（MSE） 均方誤差



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

一張含有 文字 的圖片

自動產生的描述

* MAE（Mean Absolute Error）平均絕對誤差

一張含有 圖表, 圖解 的圖片

自動產生的描述

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

一張含有 文字, 白板 的圖片

自動產生的描述

假設正確值是6（True=6）

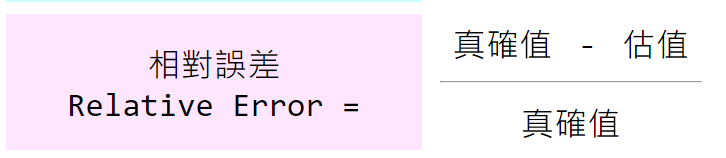
* 狀況一：預測值為8及4，則MAE=2、RMSE=2
* 狀況二：預測值為10及6，則MAE=2、RMSE= 2√2

→由此可知，RMSE對於越偏離正確值的極值越敏感。

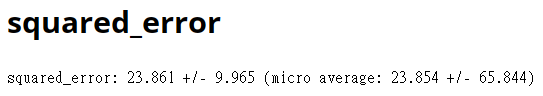
* Relative Error 相對誤差

一張含有 文字 的圖片

自動產生的描述



* squared\_error **→ 需要再請教老師**



The averaged squared error. 平均平方誤差

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.

〔Feature Selection〕

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

一張含有 文字 的圖片

自動產生的描述

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

一張含有 文字 的圖片

自動產生的描述

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 |

一張含有 文字, 室內, 白色, 充滿 的圖片

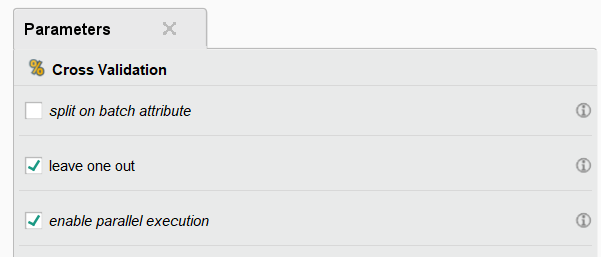
自動產生的描述

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%)

→ 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，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，仍以None最佳。

|  |  |
| --- | --- |
|  | RMSE |
| None | 3.665 |
| M5 | 3.667 |
| Greedy | 3.697 |
| T-test | 3.669 |

* RMSE=3.665

一張含有 文字 的圖片

自動產生的描述

* 迴歸方程式

一張含有 文字 的圖片

自動產生的描述

* 以Scatter圖表畫出回歸方程式

一張含有 圖表 的圖片

自動產生的描述

一張含有 文字 的圖片

自動產生的描述

Reference

〔Boston資料集簡述〕

* sklearn.datasets.load\_boston

<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-%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%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=_K4ZwIs1qyw>

* 應用統計學I22 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>

〔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>

* 回歸模型的衡量標準：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>