

計算思維與人工智慧

TA Class #05

RapidMiner2

主講者: 程至榮



政大
NATIONAL CHENGCHI UNIVERSITY



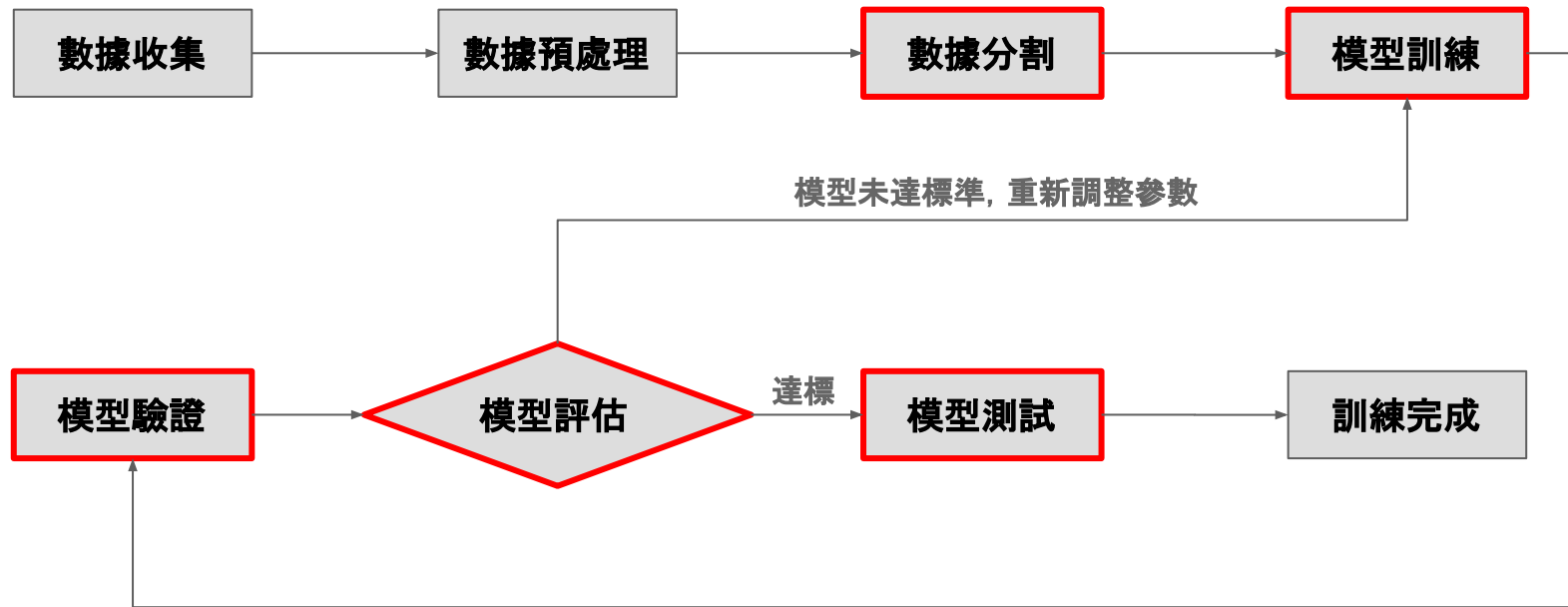
政大資訊科學系
Department of Computer Science, National Chengchi University

參考書目

大數據驅動商業決策:13 個 RapidMiner 商業預測操作實務

今日重點

模型訓練流程



常見的模型目標 & 今日重點

- The most common data science modeling tasks are these:
 - Classification—Deciding if something belongs to one category or another
 - Scoring—Predicting or estimating a numeric value, such as a price or probability
 - Ranking—Learning to order items by preferences
 - Clustering—Grouping items into most-similar groups
 - Finding relations—Finding correlations or potential causes of effects seen in the data
 - Characterization—Very general plotting and report generation from data

數據分割

Idea #1: Choose hyperparameters that work best on the data

BAD: $K = 1$ always works perfectly on training data



Idea #2: Split data into **train** and **test**, choose hyperparameters that work best on test data

BAD: No idea how algorithm will perform on new data

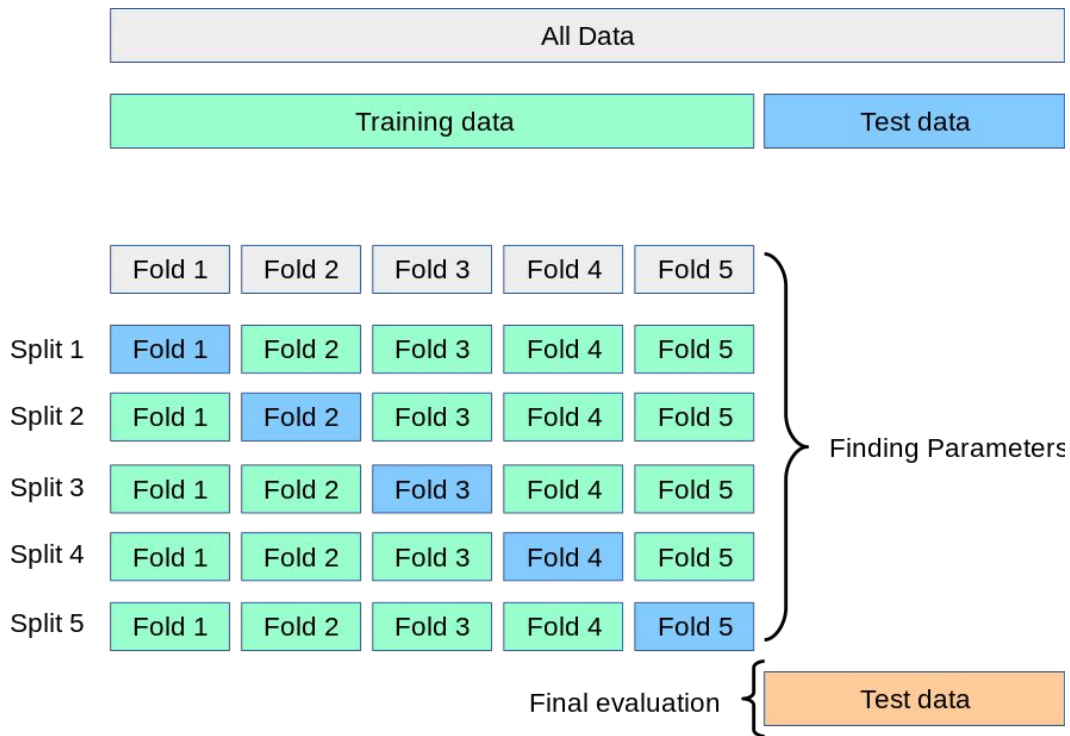


Idea #3: Split data into **train**, **val**, and **test**; choose hyperparameters on val and evaluate on test

Better!



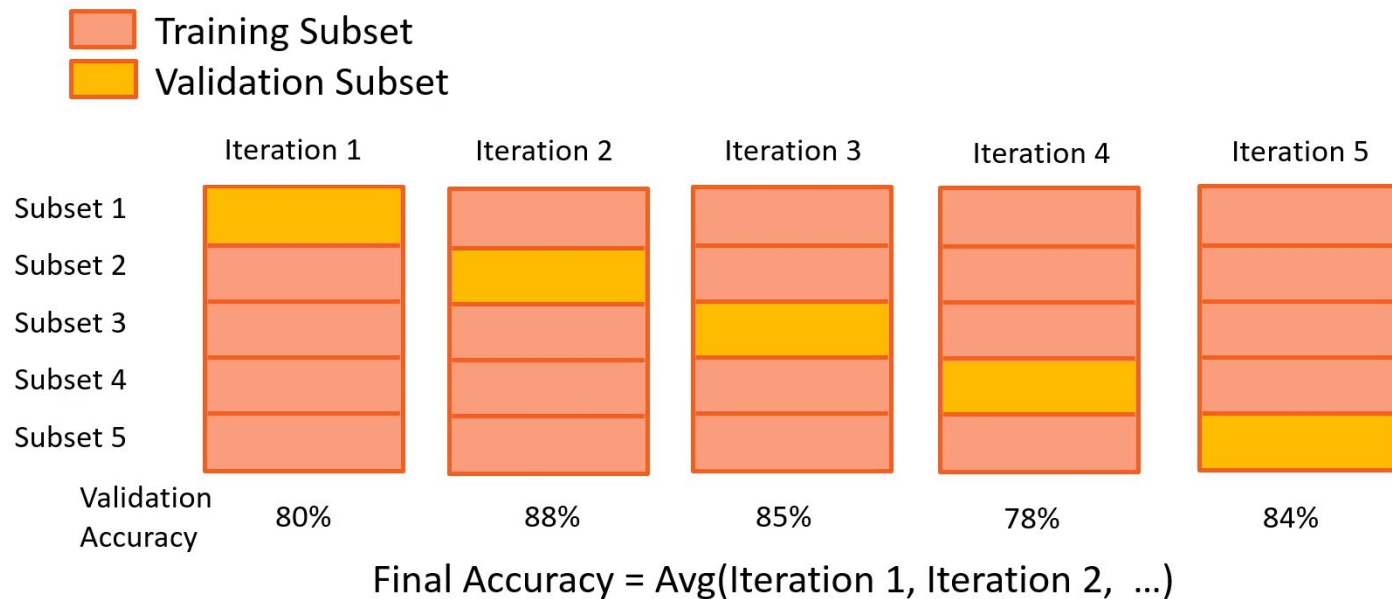
數據分割 Cross-validation



同一組模型參數訓練 k 次, 再對 k 個驗證集 Performance 取平均, 算出一個 Performance 平均值

Cross-validation: evaluating estimator performance

數據分割 Cross-validation



模型訓練 & 驗證

- **模型訓練**: 選定機器學習演算法以後, 經過設定恰當的**超參數(Hyperparameters)**、**擬合(Fitting)** 訓練資料以產生可用於後續相關預測的**模型(Model)**

Note: 通常機器學習相關的軟體、函式庫都會有**最佳化超參數**的功能

- **模型驗證**: 將驗證資料中的自變數 X 輸入訓練好的模型, 得到預測值 Y
- 這邊以**房價預測**為例, 大數據可以協助企業預測哪些因素 X 會影響客戶心中認定的房屋價值、影響的程度多寡, 進而決定合理房價 Y 。可能的因素 X 有交通、屋齡、生活機能、學校遠近、景觀.....

模型訓練 & 驗證 - Linear Regression

- Prediction

- $\hat{Y} = \hat{f}(X)$
- \hat{f} : our estimate for f
- \hat{Y} : resulting prediction for Y

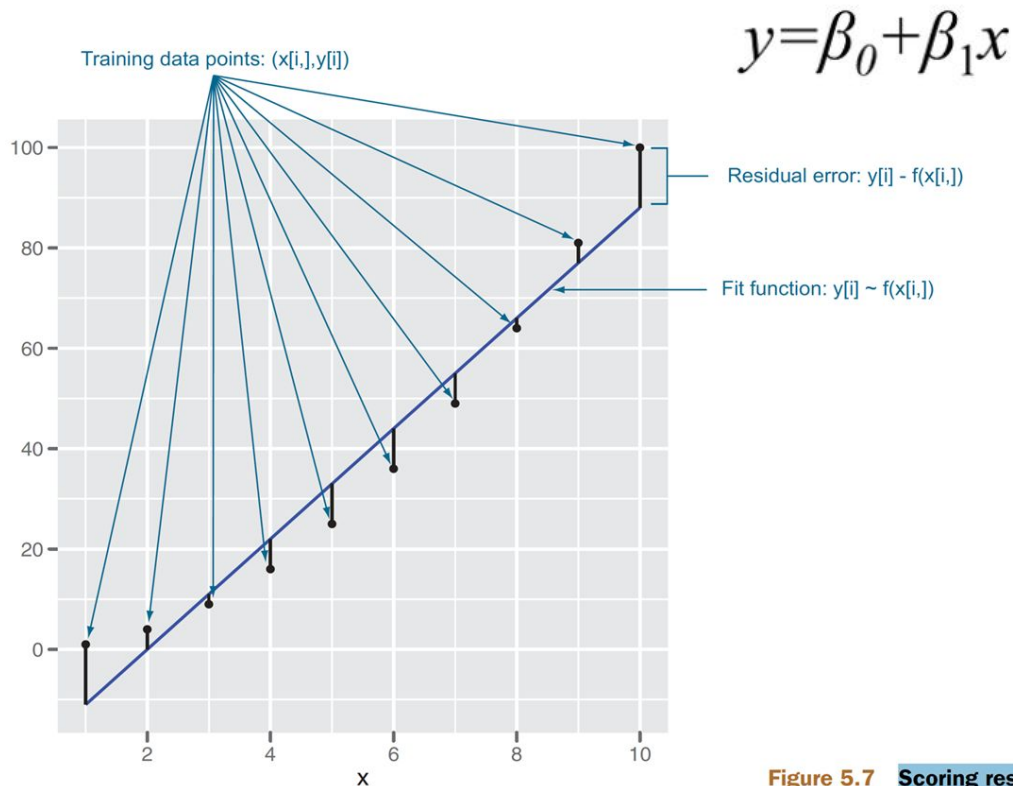


Figure 5.7 Scoring residuals

模型訓練 & 驗證 - Linear Regression

- Prediction

- $\hat{Y} = \hat{f}(X)$
- \hat{f} : our estimate for f
- \hat{Y} : resulting prediction for Y

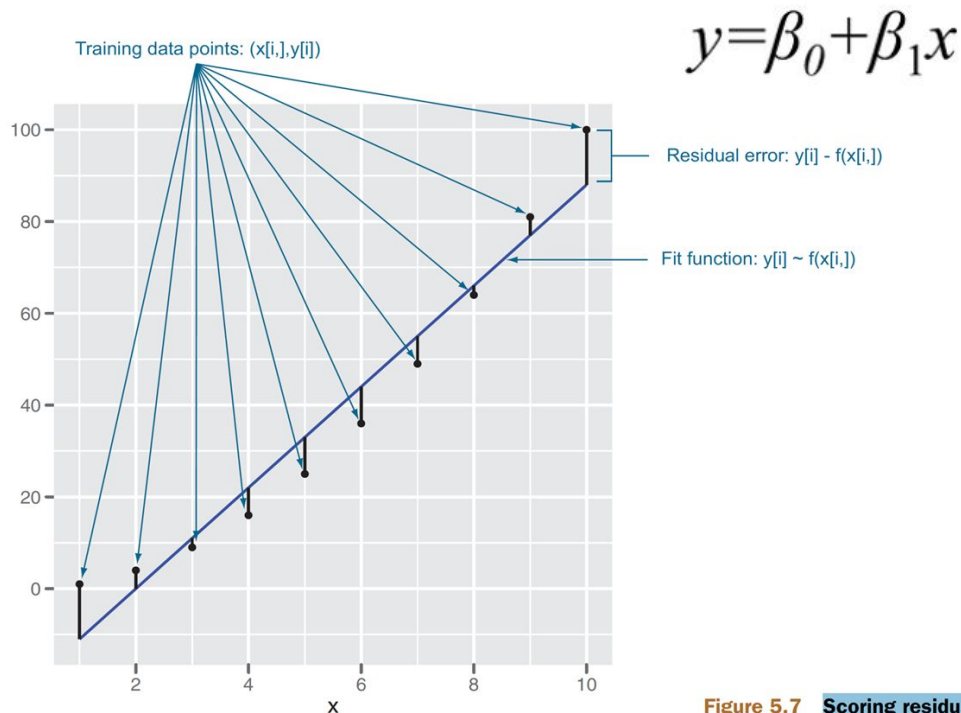


Figure 5.7 Scoring residuals

- Linear Regression 的目標是要找到一條直線, 可以讓資料點與線的殘差 (Residual) 最小

模型訓練 & 驗證 - Linear Regression

- Prediction

- $\hat{Y} = \hat{f}(X)$

- \hat{f} : our estimate for f

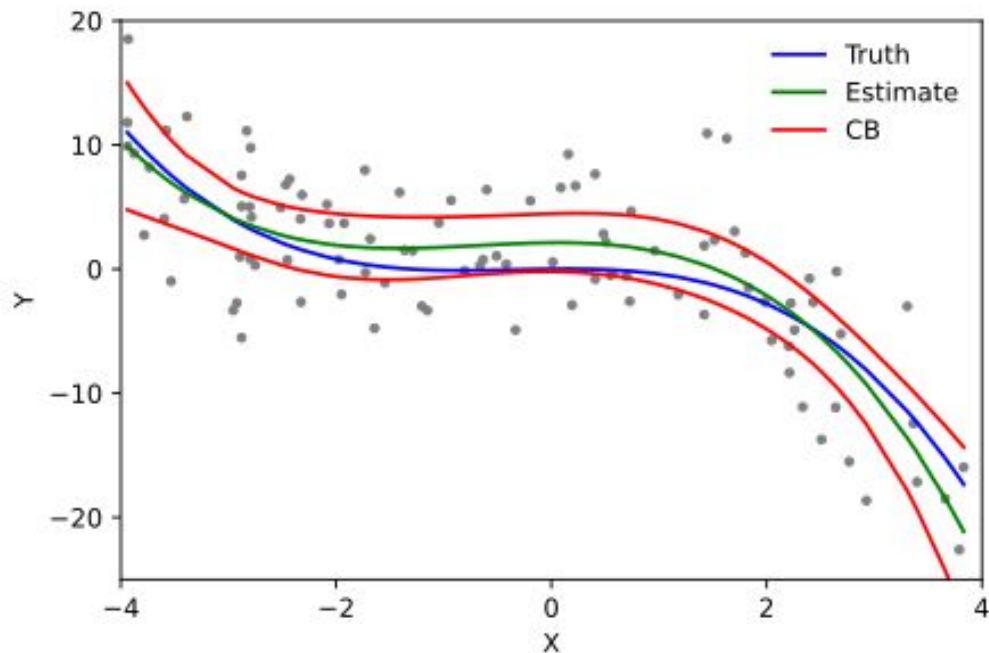
- \hat{Y} : resulting prediction for Y

以房價預測為例，可能的因素 X 有交通、屋齡、生活機能、學校遠近、景觀..... 等，而它們各自的影響程度又有所不同(權重)，故預測 Y 的公式可以改寫如下：

$$\hat{Y} = \beta_0 + \sum_{i=1}^n (\beta_i X_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

模型訓練 & 驗證 - Polynomial Regression (補充說明)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$



模型評估 - Regression task

- $MSE = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2$ 愈小愈好

- $RMSE = \sqrt{MSE}$ 愈小愈好

殘差是指實際值與估計值 (擬合值) 之間的差。迴歸模型的殘差愈小, 表示模型愈好, 而常用來衡量殘差的標準為 **Root Mean Square Error (RMSE)**

模型評估 - Regression task

$$\bullet R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

← Model error
← Variance in the dependent variable

1 – “how much unexplained variance your model leaves”

愈大愈好

R Squared 也是一種衡量迴歸模型表現的指標, R^2 愈高表示模型對原始數據的解釋能力愈好

模型評估 - Overfitting 問題

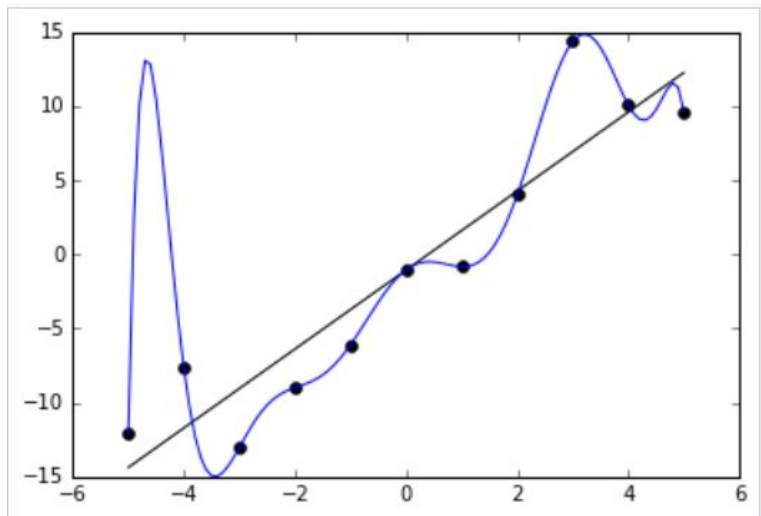
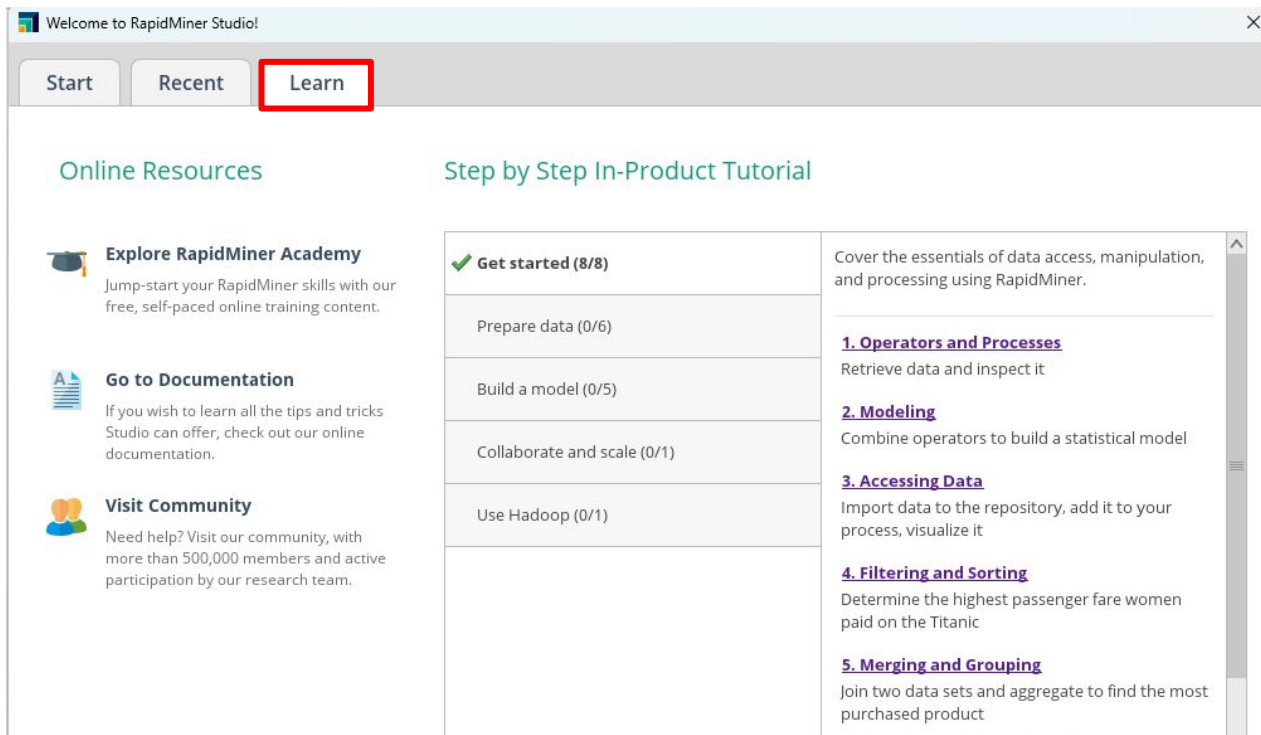


Figure 2. Noisy (roughly linear) data is fitted to a linear function and a polynomial function. Although the polynomial function is a perfect fit, the linear function can be expected to generalize better: if the two functions were used to extrapolate beyond the fitted data, the linear function should make better predictions.



https://twitter.com/jason_mayes/status/1296599149748551680

請大家從 Help -> Tutorials 打開內建的使用教學, 接下來將會帶大家拆解練習



Welcome to RapidMiner Studio!

Start Recent **Learn**

Online Resources

- Explore RapidMiner Academy**
Jump-start your RapidMiner skills with our free, self-paced online training content.
- Go to Documentation**
If you wish to learn all the tips and tricks Studio can offer, check out our online documentation.
- Visit Community**
Need help? Visit our community, with more than 500,000 members and active participation by our research team.

Step by Step In-Product Tutorial

✓ Get started (8/8)	Cover the essentials of data access, manipulation, and processing using RapidMiner.
Prepare data (0/6)	
Build a model (0/5)	1. Operators and Processes Retrieve data and inspect it
Collaborate and scale (0/1)	2. Modeling Combine operators to build a statistical model
Use Hadoop (0/1)	3. Accessing Data Import data to the repository, add it to your process, visualize it
	4. Filtering and Sorting Determine the highest passenger fare women paid on the Titanic
	5. Merging and Grouping Join two data sets and aggregate to find the most purchased product

Build a model (5/5)

Kaggle 競賽

白酒等級評估

這次的作業將利用迴歸分析進行白酒等級評估，預測目標為 quality (型別為 Integer)

競賽連結會放在 Moodle，請同學透過連結加入班級競賽

Community Prediction Competition · Private

1121 W2_234 Computational Thinking - RapidMiner2

Wine Quality Prediction - Classification Prediction

Host

Overview

Data

Discussion

Leaderboard

Rules

Team

Submissions

...

Overview

Wine Quality Prediction - Classification Prediction

The dataset used are from Kaggle & UCI machine learning repository.

<https://archive.ics.uci.edu/ml/datasets/wine+quality>

Start

Set start date via the [launch checklist](#).

Close

18 days to go

Competition Host

cheng-zhi-rong

Prizes & Awards

Kudos

Does not award Points or Medals

Participation

0 Competitors

0 Teams

0 Entries

Tags

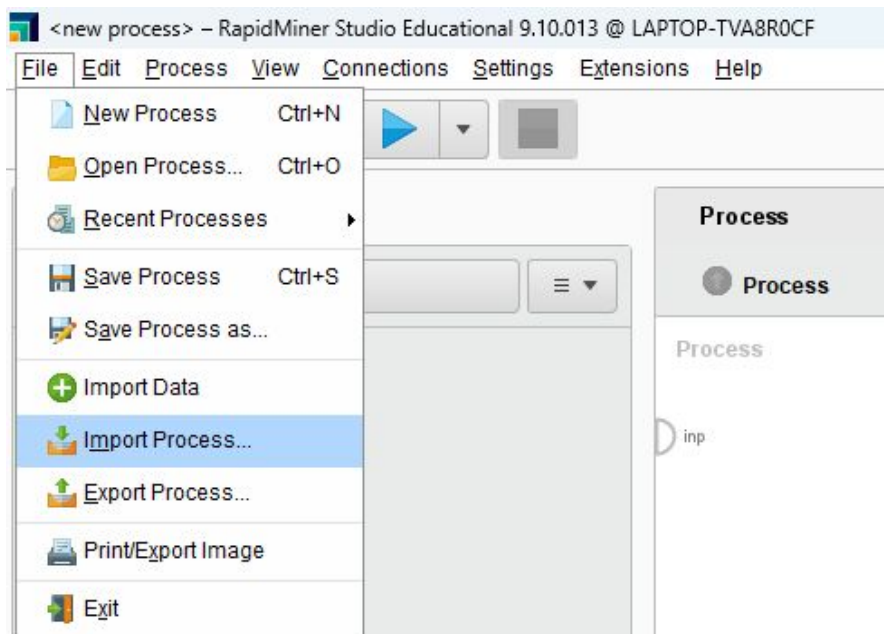
白酒等級評估 - 匯入資料

請從 Kaggle 下載 train.csv、test.csv >> 匯入 RapidMiner 並做以下設定 (記得勾選 Replace errors with missing values)

- **train.csv:** 使用 Change Role 將 Id 改為 id、使用 Change Role 將 quality 改為 label
- **test.csv:** 使用 Change Role 將 Id 改為 id

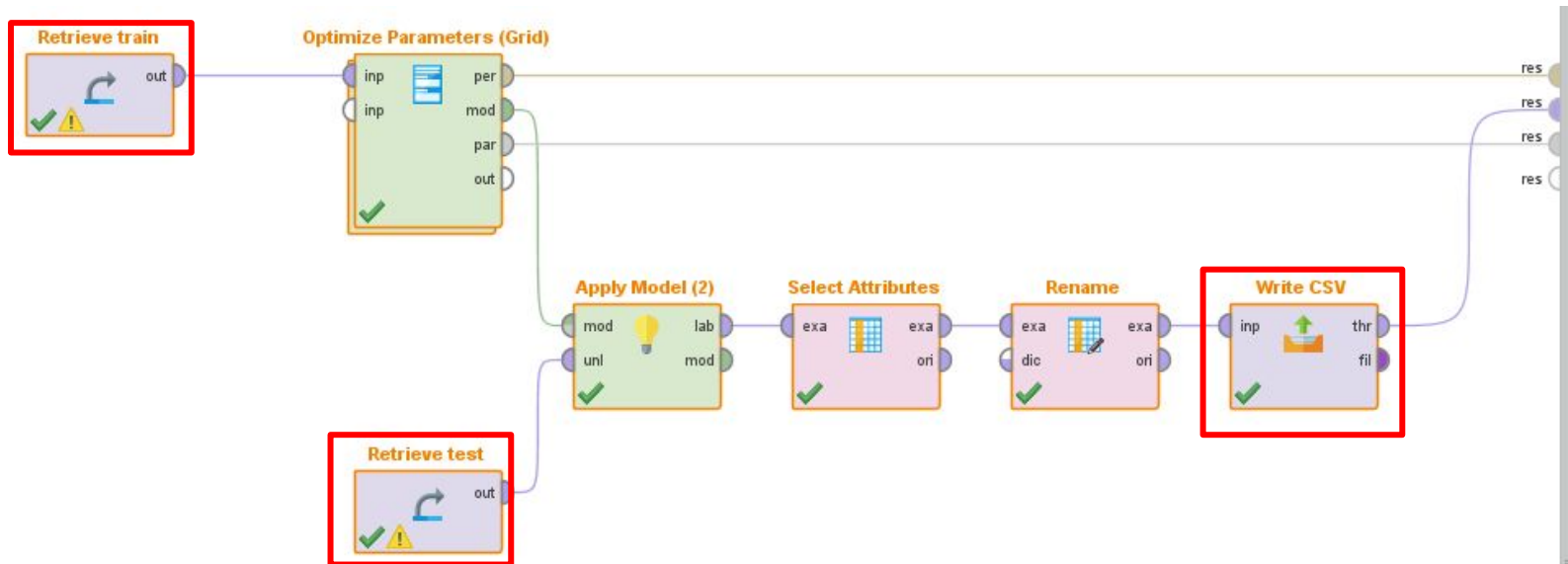
白酒等級評估 - 設計流程

請從 Moodle 下載 RapidMiner2_Wine_Quality.rmp 並匯入 RapidMiner



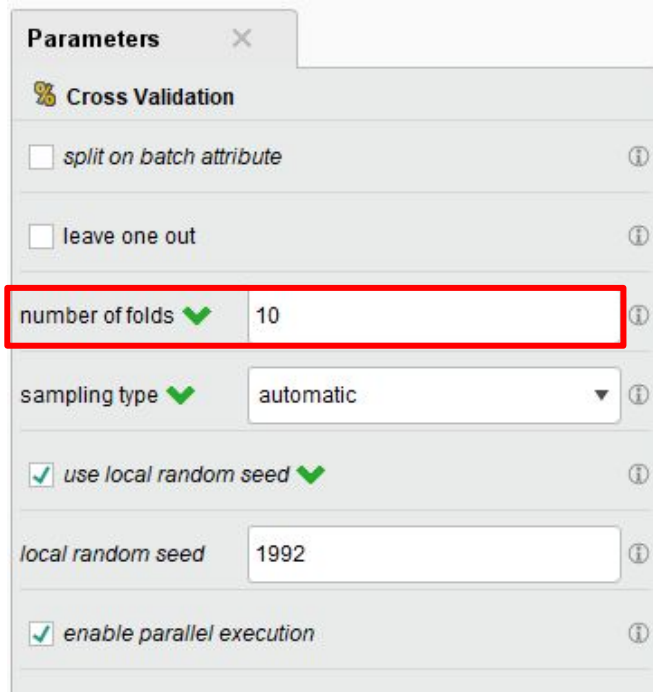
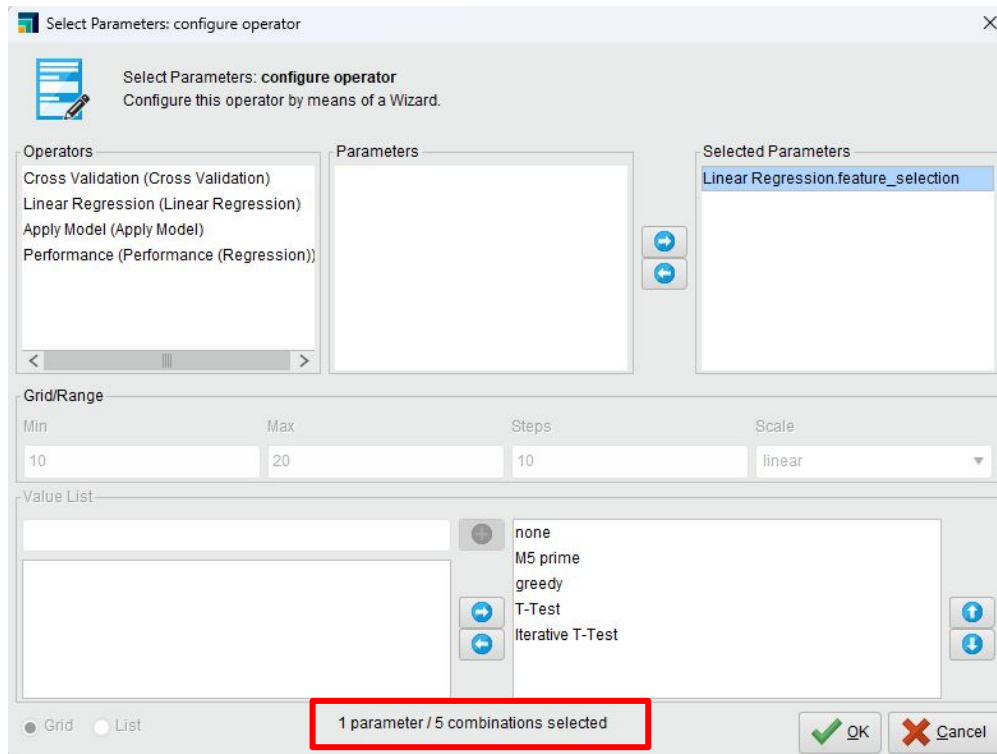
白酒等級評估 - 設計流程

記得按照之前教學的方式更改檔案路徑、CSV 輸出位置



Optimize Parameters (Grid) 觀念

如果 Optimize Parameters (Grid) 有 5 種參數組合, 而 Cross Validation 固定為 10-Fold



Optimize Parameters (Grid) 觀念

只會算出 5 種參數組合的 Performance, 並從其中挑一個最好的當做模型參數

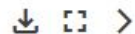
Optimize Parameters (Grid) × % PerformanceVector (Performance)		
Optimize Parameters (Grid) (5 rows, 3 columns)		
iteration	Linear Regression.feature_selection	root_mean_squared_error
2	M5 prime	0.714
5	Iterative T-Test	5.784
1	none	0.713
4	T-Test	0.719
3	greedy	0.713

- 5 種參數組合各自進行 10-Fold Cross Validation, 並各自計算出一個平均 Performance
- 總共的計算量為 $5 * 10 = 50$ 回合

白酒等級評估 - 上傳格式

預測目標為 quality (型別為 Integer)

sample_submission.csv (11.77 kB)



Detail Compact Column

2 of 2 columns ▾

About this file

This file does not have a description yet.



Id	# quality
3428	5
3429	5
3430	5

Data Explorer

335.97 kB

- sample_submission.csv
- test.csv
- train.csv

更改 Kaggle Team Name

請注意 Team name 務必改成以下格式 學號-系級-名字

Overview Data Code Discussion Leaderboard Rules Team

Submissions

Submit Predictions

...

Your Team

Everyone that competes in a Competition does so as a team - even if you're competing by yourself. [Learn more.](#)

General

TEAM NAME

111753151-資碩一-程至榮

This name will appear on your team's leaderboard position.

Evaluation - Mean Squared Error

上傳分數愈低愈好

Evaluation



The Evaluation metrics used are Mean Squared Error

Lower score is better

白酒等級評估- Baseline

以下是今天範例的 Public 和 Private score, 要拿到 "加分題" 的同學你的 **Private score** 必須優於 Baseline

✓ Sandbox Submissions

Upload a Submission CSV and make sure it produces the expected score. These submissions are private unless tagged as a Benchmark, which appears on the Leaderboard.

Create sandbox submission

Submission and Description

Private Score ⓘ

Public Score ⓘ

Benchmark ⓘ



Wine_Baseline.csv

Complete · 25m ago

0.60641

0.69841



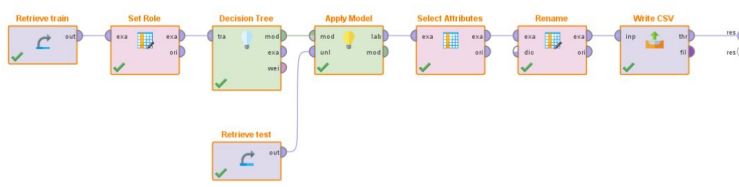
作業要求

基本題 4 分

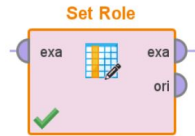
1. 不限定模型，成功上傳 kaggle(即 Leaderboard 有你的名字，且**格式正確**),
2. **Public Score** 優於 Baseline
3. 在 Moodle **上傳至少 1 頁 PDF**，說明你的設計流程、參數設定，並附上截圖(請參考簡報前面的教學是怎麼做的)
4. **上傳你的 Process file** (檔名: 學號_RapidMiner2.rmp, 範例: 111753151_RapidMiner2.rmp)

基本設計流程

以下為基本的設計流程，請透過 RapidMiner 的搜尋功能找出以下元件並排好。接著會說明參數設定，沒有特別講就是不用改設定。



參數設定



Parameters	
attribute name	Survived
target role	label
set additional roles	Edit List (0)...

加分題 1 分

1. 結算後在 **Leaderboard** 排名前 50% 且 **Private Score** 優於 **Baseline** 的同學可以得到額外的分數。

Public ☒ Private

The private leaderboard is calculated with approximately 70% of the test data.

#	△	Team	Members	Score	Entries	Last
1		Wine_Baseline.csv		0.60641		

透過上課教過的方法 + 上網查詢、調整資料前處理的方法 or 模型參數。

接下來的作業都沒有標準答案, 請大家盡可能的去嘗試!

作業注意事項

- 為了公平起見，且大家的**期末專題海報**預計會與 micro:bit 或 RapidMiner 有關，此作業**限定使用 RapidMiner 產生的 Submission 參賽**，請確認繳交的 Process file 可以產生正確的 Submission 檔案
- **請不要抄襲、或是直接拿別人的 Submission 檔案上傳**，助教會隨機抽查是否有排名分數與 Process file 不一致的問題。



Bonus

房價分析

資料集

1. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
2. INDUS: proportion of non-retail business acres per town
3. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
4. NOX: nitric oxides concentration (parts per 10 million)

1<https://archive.ics.uci.edu/ml/datasets/Housing>

123

20.2. Load the Dataset 124

5. RM: average number of rooms per dwelling
6. AGE: proportion of owner-occupied units built prior to 1940
7. DIS: weighted distances to five Boston employment centers
8. RAD: index of accessibility to radial highways
9. TAX: full-value property-tax rate per \$10,000
10. PTRATIO: pupil-teacher ratio by town 12. B: $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town 13. LSTAT: % lower status of the population
11. MEDV: Median value of owner-occupied homes in \$1000s

We can see that the input attributes have a mixture of units.

<https://www.kaggle.com/datasets/vikrishnan/boston-house-prices>

匯入資料

請從 Moodle 下載 C9_HousingData.csv 並匯入 RapidMiner

Import Data - Format your columns.

Format your columns.

Date format

☒ Replace errors with missing values ⓘ

		DIS real	RAD integer	TAX integer	PTRATIO real	B real	LSTAT real	MEDV real label
1	00	4.090	1	296	15.300	396.900	4.980	24.000
2	00	4.967	2	242	17.800	396.900	9.140	21.600
3	00	4.967	2	242	17.800	392.830	4.030	34.700
4	00	6.062	3	222	18.700	394.630	2.940	33.400
5	00	6.062	3	222	18.700	396.900	5.330	36.200
6	00	6.062	3	222	18.700	394.120	5.210	28.700
7	00	5.561	5	311	15.200	395.600	12.430	22.900
8	00	5.950	5	311	15.200	396.900	19.150	27.100
9	000	6.082	5	311	15.200	386.630	29.930	16.500
10	00	6.592	5	311	15.200	386.710	17.100	18.900
11	00	6.347	5	311	15.200	392.520	20.450	15.000
12	00	6.227	5	311	15.200	396.900	13.270	18.900
13	00	5.451	5	311	15.200	390.500	15.710	21.700
14	00	4.707	4	307	21.000	396.900	8.260	20.400
15	00	4.462	4	307	21.000	380.020	10.260	18.200
16	00	4.499	4	307	21.000	395.620	8.470	19.900
17	00	4.499	4	307	21.000	386.850	6.580	23.100

no problems.

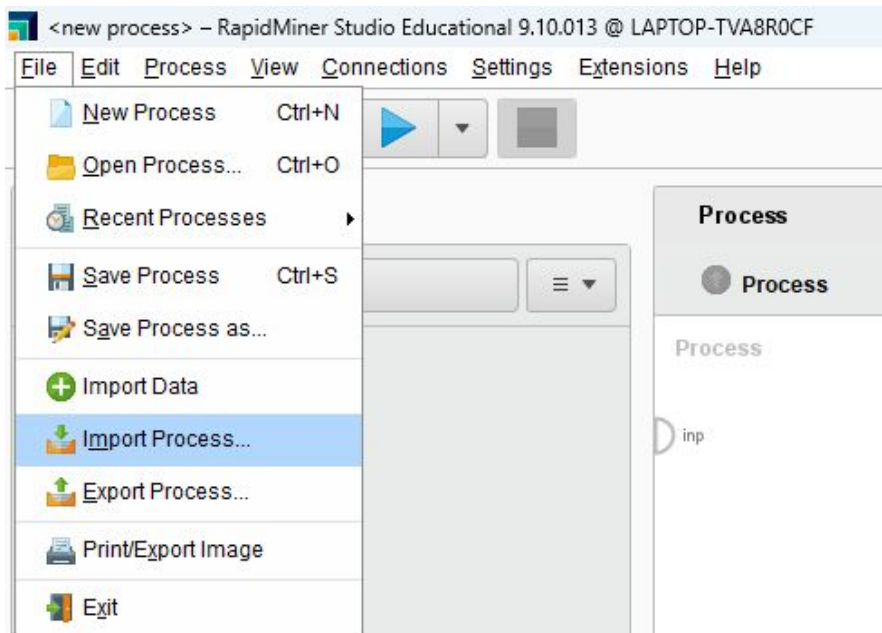
Previous Next Cancel

Change Type
Change Role
Rename column
Exclude column

MEDV >> Change Role (label)

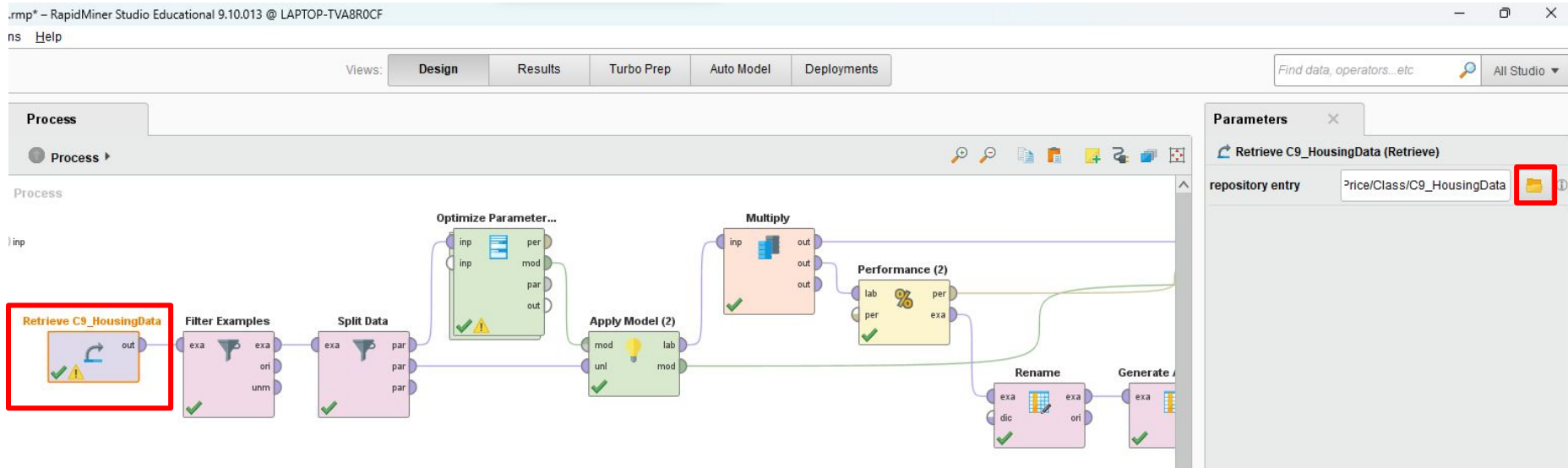
房價分析 - 設計流程

請從 Moodle 下載 F2309_ch09_03.rmp 並匯入 RapidMiner



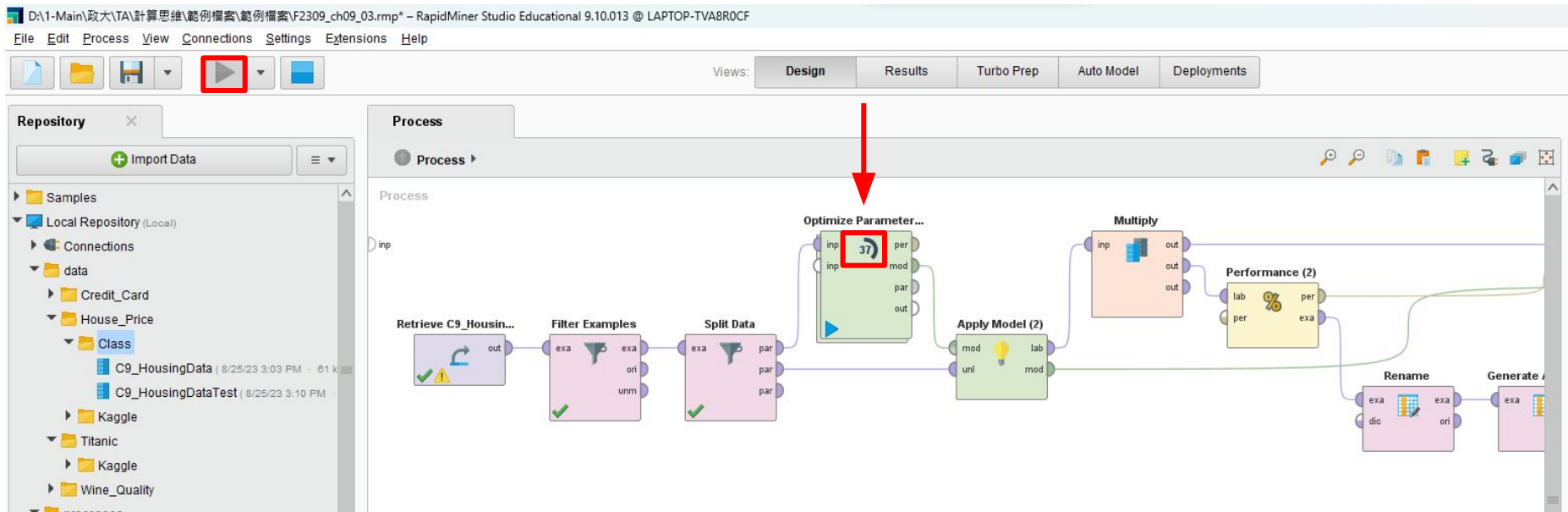
房價分析 - 設計流程

請將以下元件的路徑重新指定為你剛剛匯入的那個 C9_HousingData.csv



房價分析 - 設計流程

按下執行，如果步驟都正確應該可以看到 Process 有在運作



房價分析 - 研究假設的種類與寫法

- **虛無假設 H_0**

虛無假設用來主張**自變項**的效果**不存在**、不同的組別間不會因為加入了自變項而受到影響，通常用符號 H_0 表示；不接受虛無假設則表示接受了對立假設。

- **對立假設 H_1**

對立假設用來主張**自變項**的效果**存在**，換句話說對立假設闡述研究結果中組別之間會因為加入了自變項而受到影響；不接受對立假設則表示接受了虛無假設。

房價分析 - 研究假設的種類與寫法

舉例:

- **虛無假設 H0**

守望相助隊的成立(自變項)**不會**影響社區的犯罪率

- **對立假設 H1**

守望相助隊的成立(自變項)**會**影響社區的犯罪率

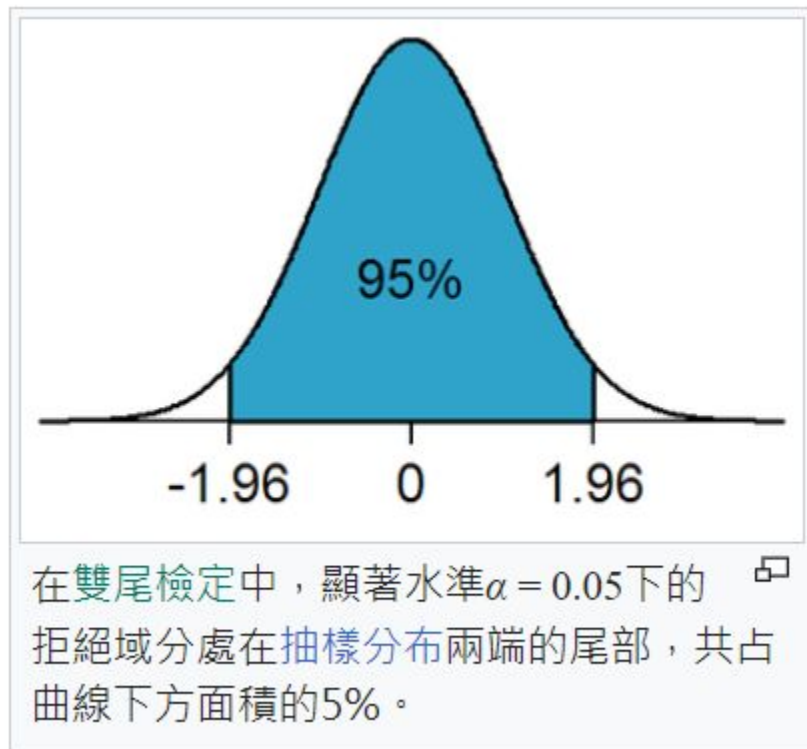
房價分析 - 顯著水準與決策規則

- **P 值定義:** 抽出一組樣本, 在**虛無假說 H_0 為真**的情況下, 得到這組樣本觀察結果的**機率值**為何?
- **顯著水準 α 值:** 事先設定好、為了與 P 值比較的一個門檻 (機率值), 用於決定接受或拒絕虛無假說 H_0

房價分析 - 顯著水準與決策規則

- 若分析結果機率 P 值 \leq 顯著水準 α 值:
拒絕虛無假設、接受對立假設
- 若分析結果機率 P 值 $>$ 顯著水準 α 值:
保留虛無假設

房價分析 - 顯著水準與決策規則



房價分析 - 顯著性檢定

按一下 p-Value 那格讓他由小到大排列。當顯著水準為0.05 時, **紅色框**表示該 Attribute 的影響非常顯著 ($P < 0.001$)、**藍色框**表示顯著 ($p < 0.01$)、**綠色框**表示不顯著 ($p \geq 0.05$)

Views: Design Results Turbo Prep Auto Model Deployments

Result History: ExampleSet (Normalize) × **LinearRegression (Linear Regression)** × PerformanceVector (Performance (2)) × ExampleSet (Multiply) × Optimize Parameters (Grid) ×

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value ↑	Code
LSTAT	-0.597	0.066	-0.414	0.606	-9.120	0	****
RM	4.328	0.545	0.327	0.587	7.935	0.000	****
DIS	-1.439	0.238	-0.343	0.947	-6.056	0.000	****
PTRATIO	-0.924	0.166	-0.226	0.853	-5.586	0.000	****
(Intercept)	29.572	6.834	?	?	4.327	0.000	****
NOX	-15.184	4.512	-0.201	0.923	-3.365	0.001	****
RAD	0.225	0.076	0.197	0.929	2.969	0.003	***
CHAS	2.956	1.021	0.092	0.992	2.895	0.004	***
B	0.014	0.005	0.096	0.963	2.815	0.005	***
TAX	-0.008	0.004	-0.142	0.904	-1.968	0.050	**
ZN	0.031	0.017	0.082	0.926	1.794	0.074	*

房價分析 - 迴歸係數 & 截距

D:\1-Main\政大\TA\計算思維\範例檔案\範例檔案\F2309_ch09_03.rmp* - RapidMiner Studio

File Edit Process View Connections Settings Extensions Help

Result History LinearRegression (Linear Regression) X

LinearRegression

Data

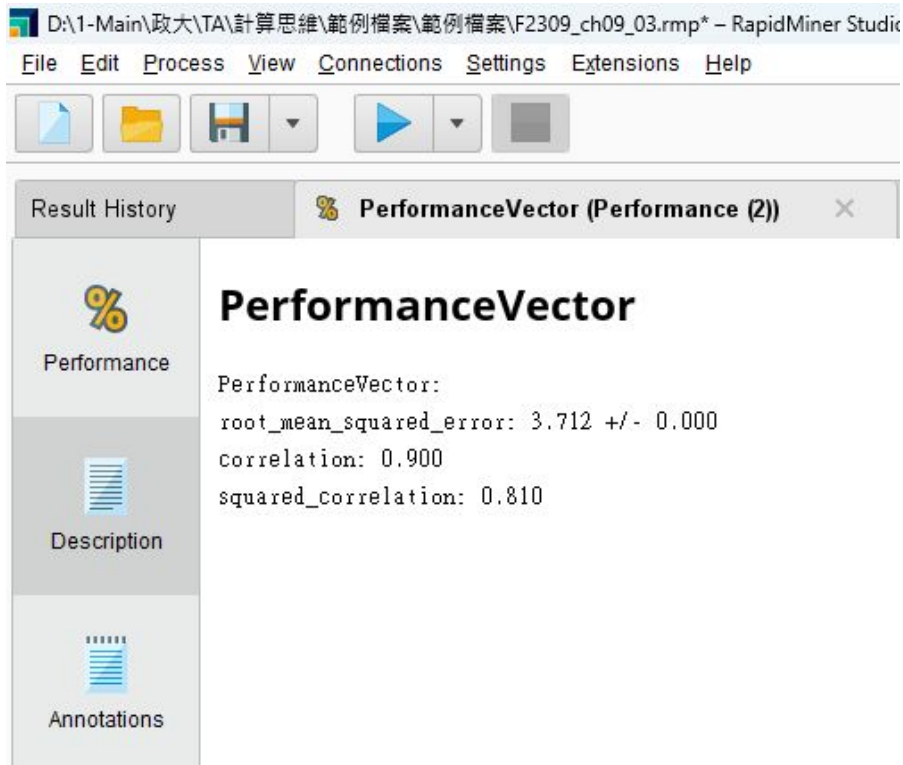
Description

Annotations

$$\begin{aligned} &0.031 * ZN \\ &+ 2.956 * CHAS \\ &- 15.184 * NOX \\ &+ 4.328 * RM \\ &- 1.439 * DIS \\ &+ 0.225 * RAD \\ &- 0.008 * TAX \\ &- 0.924 * PTRATIO \\ &+ 0.014 * B \\ &- 0.597 * LSTAT \\ &+ 29.572 \end{aligned}$$

房價分析 - PerformanceVector

RMSE 在 95% 的信心水準下, 最大誤差為3.712 千美元, squared_correlation (迴歸方程式對於資料的詮釋能力) 為 81%








The screenshot shows the RapidMiner Studio interface. The title bar indicates the file path: D:\1-Main\政大\TA\計算思維\範例檔案\範例檔案\F2309_ch09_03.rmp* - RapidMiner Studio. The menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. Below the menu is a toolbar with icons for opening files, saving, and running processes. The main workspace displays the 'PerformanceVector (Performance (2))' node. The left sidebar has three tabs: 'Performance' (selected), 'Description', and 'Annotations'. The 'Performance' tab shows the following results:

PerformanceVector

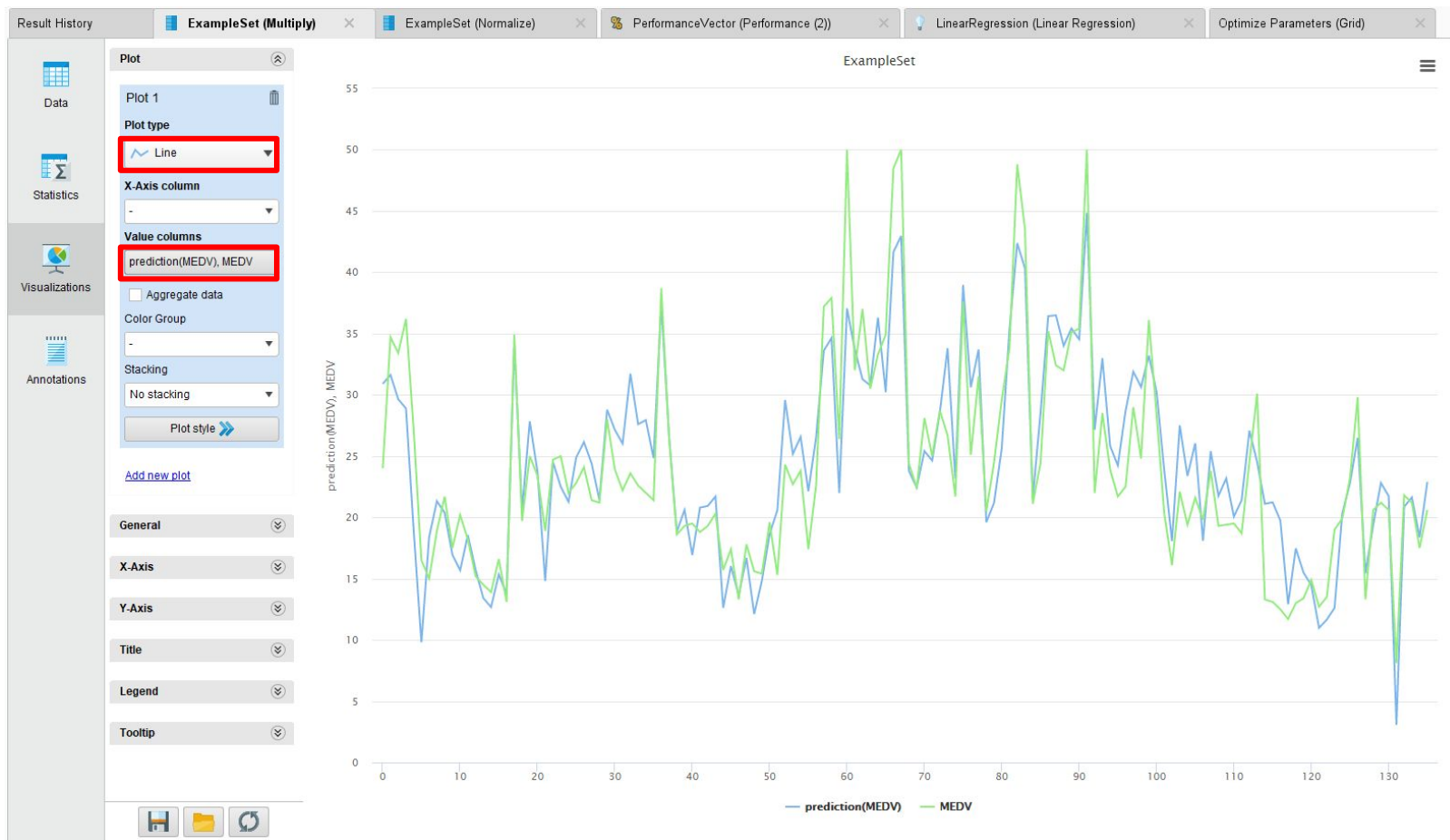
PerformanceVector:
root_mean_squared_error: 3.712 +/- 0.000
correlation: 0.900
squared_correlation: 0.810

房價分析 - 真實值 Vs 預測值

Result History		ExampleSet (Multiply) ×	
 Data	Open in		 Turbo Prep
			 Auto Model
 Statistics			
 Visualizations			

Row No.	MEDV	prediction(M...
1	24	30.895
2	34.700	31.624
3	33.400	29.639
4	36.200	28.889
5	27.100	18.920
6	16.500	9.802
7	15	18.398

房價分析 - 真實值 Vs 預測值



房價分析 - 殘差

Result History

ExampleSet (Normalize) × ExampleSet (M

Data

Statistics

Visualizations

Annotations

Open in

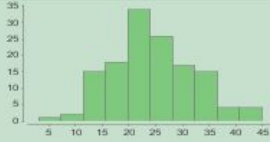
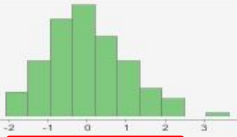
Turbo Prep

Auto Model

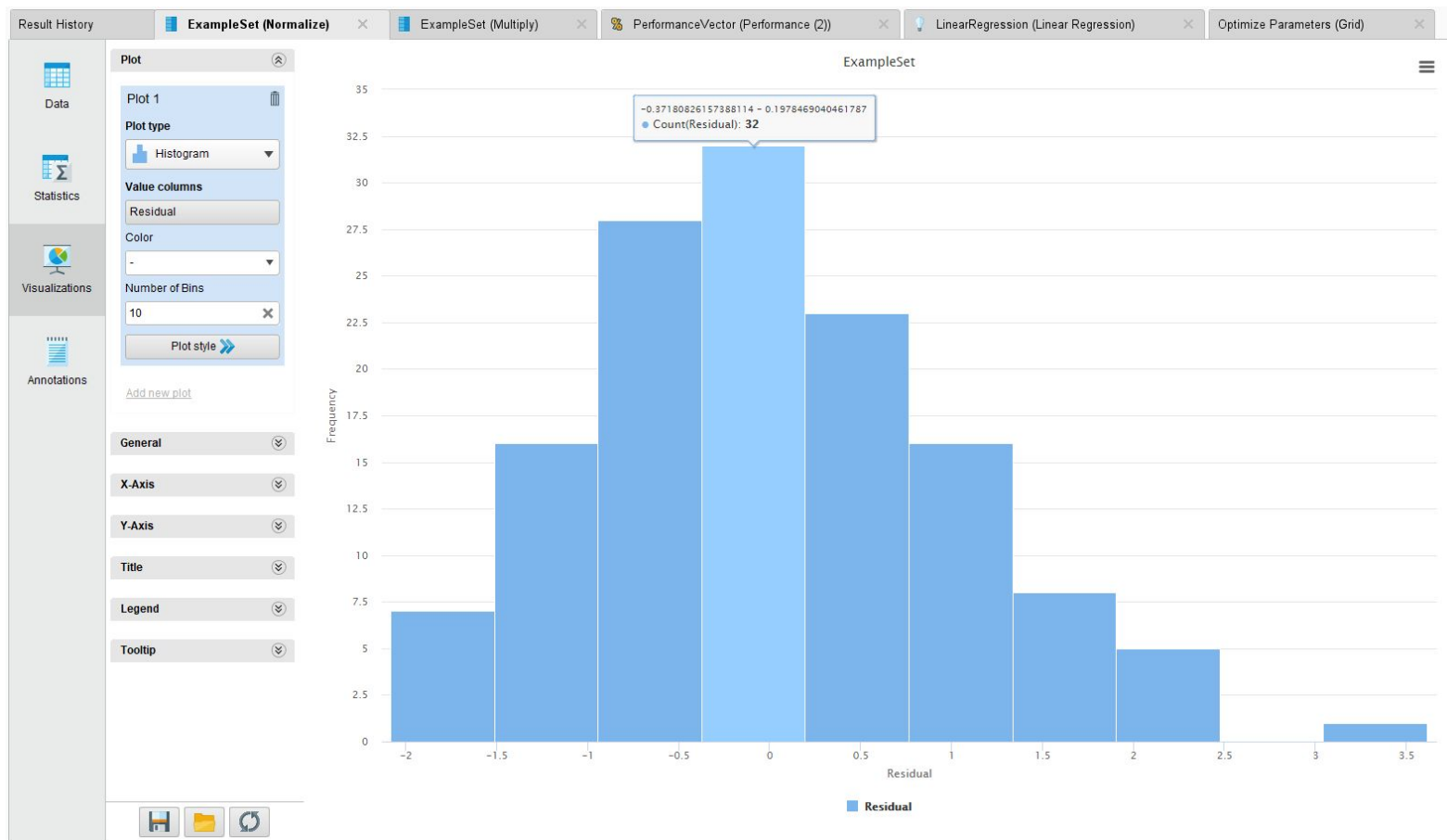
Row No.	MEDV	predictedMEDV	Residual
1	24	30.895	-1.747
2	34.700	31.624	0.947
3	33.400	29.639	1.132
4	36.200	28.889	2.091
5	27.100	18.920	2.326
6	16.500	9.802	1.925
7	15	18.398	-0.802
8	18.900	21.329	-0.541
9	21.700	20.377	0.473
10	17.500	16.942	0.267

房價分析 - 殘差

標準化殘差的平均值為 0、標準差為 1、接近常態分佈

Result History									
ExampleSet (Normalize) × ExampleSet (Multiply) × PerformanceVector (Performance (2)) × LinearRegression (Linear Regression) ×									
Data	Name	Type	Missing	Statistics					
	Label			Min	Max	Average			
	✓ MEDV	Real	0	8.100	50	24.022			
	^ Prediction								
Statistics	^ predictedMEDV	Real	0			Min	Max	Average	Deviation
						3.077	44.866	24.451	7.732
Visualizations				Open visualizations					
	^ Residual	Real	0			Min	Max	Average	Deviation
						-2.081	3.616	-0	1.000
				Open visualizations					

房價分析 - 殘差



房價分析 - 離群值



Reference

- [大數據驅動商業決策: 13 個 RapidMiner 商業預測操作實務](#)
- [RapidMiner 人工智慧機器學習軟體](#)
- [Data Science course by professor Jia-Ming Chang](#)
- [基礎統計名詞介紹網頁](#)
- [2021 iThome 鐵人賽 - 全民瘋 AI 系列 2.0](#)
- [Dr. Fish 漫遊社會統計](#)

Tools

- [ZoomIt - Sysinternals - Microsoft Learn](#)

