https://tmrmds.co/iti-ds-109b-ml/

Python 機器學習商務實戰

Day01

臺灣行銷研究有限公司

日期：2021/8/30 (一) 9:10 ~ 16:00

地點：ITI 新竹光復 109B

講者：鍾皓軒，
howard32180900@gmail.com

評量方式

1. 課堂Workshop（40%）
2. 學習心得報告（見下頁規範）
3. Bonus：個人舉手發言，納入學習心得報告加分（60%）

評量項目	課堂Workshop	學習心得報告
成績佔比	40%	60%

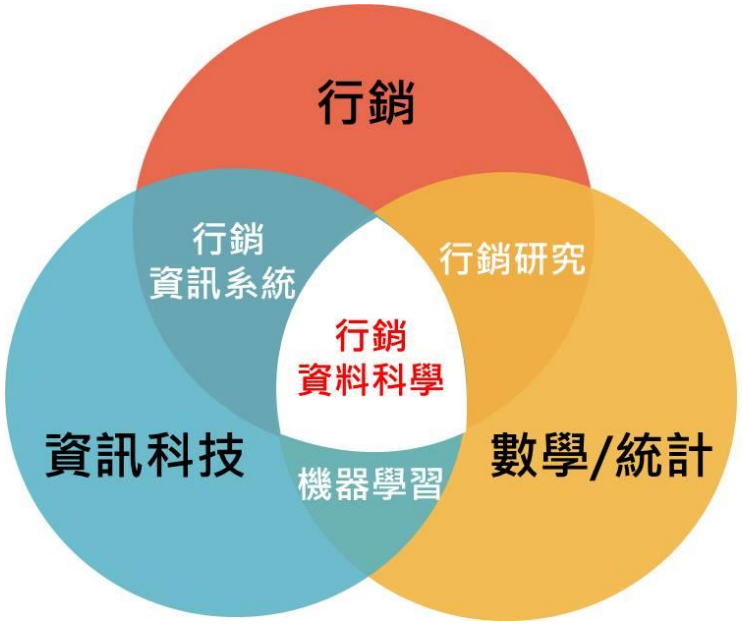
學習心得報告（60%）

1. 三日 Python **機器學習商務應用**課程的心得
2. 可交付成果：Word
3. 檔名：
   1) 班級_學號_姓名_貿協ML心得.docx
   2) 舉例：109B_B10108017_鍾皓軒_貿協ML心得.docx

日期	時間
2021/8/30(一)	9:10-16:00
2021/9/6(一)	9:10-16:00
2021/9/13(一)	9:10-16:00

這堂課程想要培養大家的能力是什麼？

TMR 預測性分析能力



商業價值

同時串聯Python與商業應用

背後的思維模式？

課前準備

環境安裝

TMR

課程介紹網站與前導安裝

https://tmrmds.co/iti-ds-109b-ml/

安裝視覺化套件graphviz

1. Windows：從這裡下載套件，一路點擊安裝到底
2. Mac：
   1) 60%的Mac同學不用做任何動作
   2) 40%的Mac同學請先測試「 00_環境安裝檔與IDE 」資料夾裡面下述檔案

是否產出

__pycache__	8/19/2021 10:57 AM	File folder	
00_請先讀我.txt	2/25/2021 11:05 AM	Text Document	1 KB
01_安裝.ipynb	8/17/2021 1:58 PM	IPYNB File	24 KB
02_安裝視覺化套件graphviz .pdf	11/24/2020 2:15 PM	Adobe Acrobat D...	164 KB
03_環境測試.ipynb	8/19/2021 11:46 AM	IPYNB File	5 KB
contract.csv	10/16/2019 11:02 AM	Microsoft Excel Co...	514 KB
marketing	8/19/2021 11:46 AM	File	2 KB
marketing.pdf	8/19/2021 11:46 AM	Adobe Acrobat D...	23 KB
util.py	6/30/2020 11:59 PM	PY File	49 KB

若MacOS環境測試發生錯誤

TMR
臺灣行銷研究有限公司

線上文件鏈接

1. # step1. 在spotlight 搜尋 「terminal」開啟終端機

2. # step2. 在終端機上輸入以下代碼 進行安裝
   1) /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"

3. # step3. 安裝完畢後，在終端機上輸入以下代碼 進行安裝
   1) brew install xgboost

4. # step4. 安裝完畢後，在終端機上輸入以下代碼 進行安裝
   1) brew install graphviz

5. # step5. 安裝完畢後，回到01_安裝.ipynb輸入以下代碼 並執行以進行安裝

6. # 回到03_環境測試，便可順利完成測試

三天預授大綱

Day	授課內容	目的	範疇
Day01	商務AI模型實戰	1. 機器學習概念與實作 2. 顧客推薦清單	機器學習
Day02	利潤評估模型簡介與 總體行銷調整分析	1. 老闆溝通容易 2. 最適利潤顧客推薦清單 3. 最好的利潤模型預測顧客購買行為 4. 以資料為依據做行銷調整	機器學習
Day03	基礎A/B行銷測試分析實作	廣告總體效益評估	統計學習

Day01 大綱

授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法

大綱

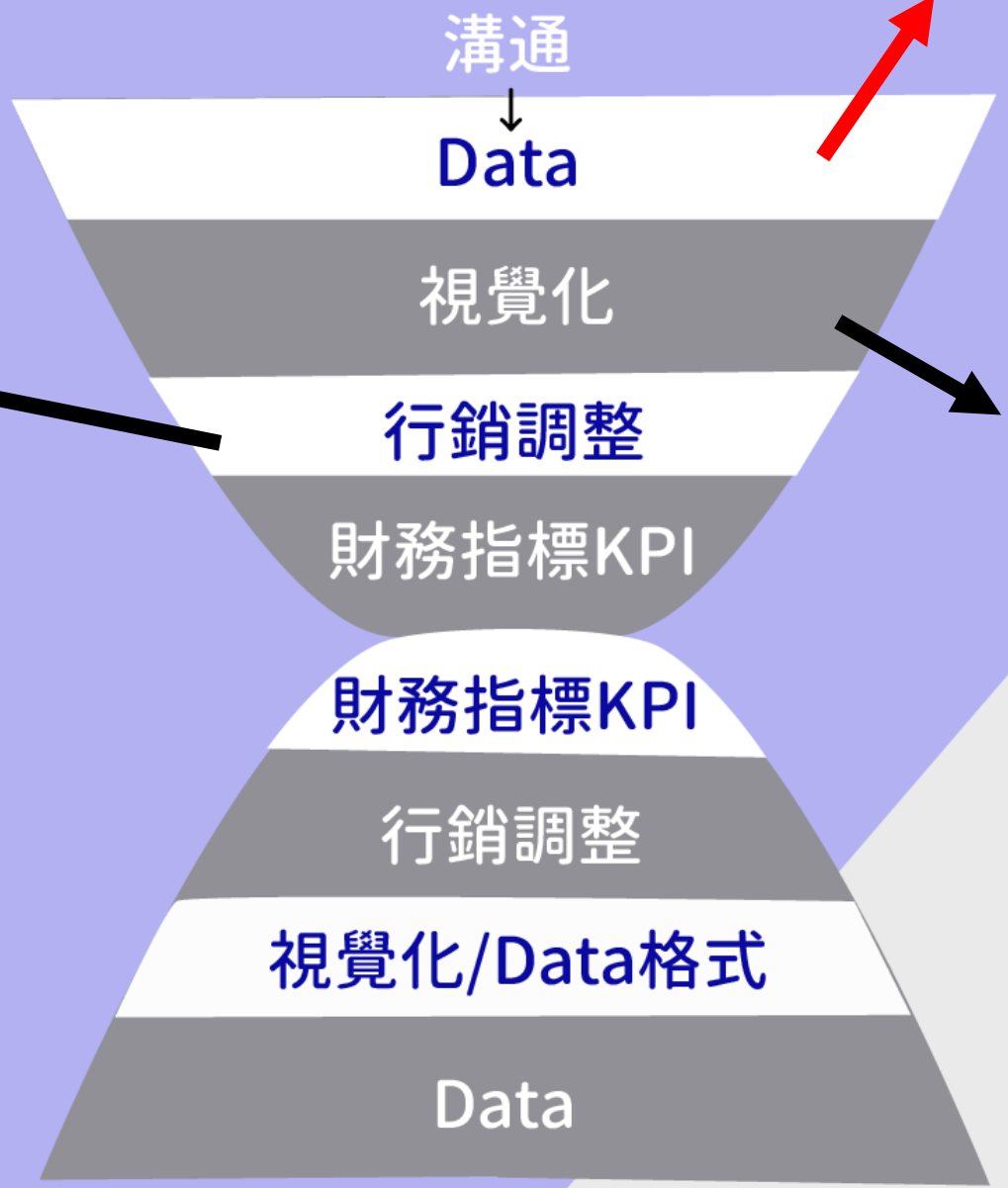
授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法

機器學習與統計學習的比較

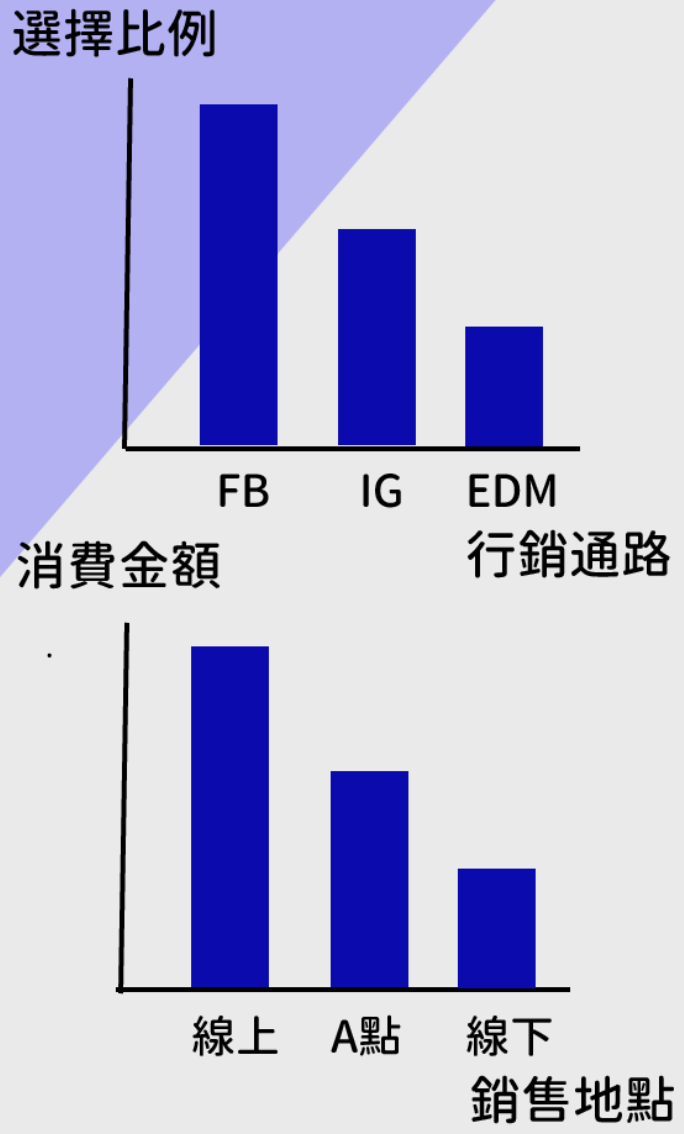
項目	統計學習	機器學習
目的	從樣本推論總體平均效應或效益 （average effect）	從樣本預測個體效果 （individual effect）
方法	從樣本推論母體	從樣本直接學習相關模式，套用到未來的預測資料
注重在	參數之間的因果關係	最終的預測結果
潛在服務產業	總體性質的服務產業 低價零售業 民生用品產業	個體性質的服務產業 高價零售業 專屬服務產業
產業範例	棒球體育賽事 IP周邊零售產品 行銷人手不足的產業	航空業 電信業 行銷人手較足的產業
常用的評估方法	P-value, confidence interval, R-squared等	RMSE, logloss等
常用的數據工具	logistic regression, Linear regression	XGBoost, Random Forest, LightGBM, Neural Network

資料洞見解析圖

VID	A產品消費者購買機率
V01	99.8%
V02	97.7%
V03	95.7%
⋮	
V100	0.001%

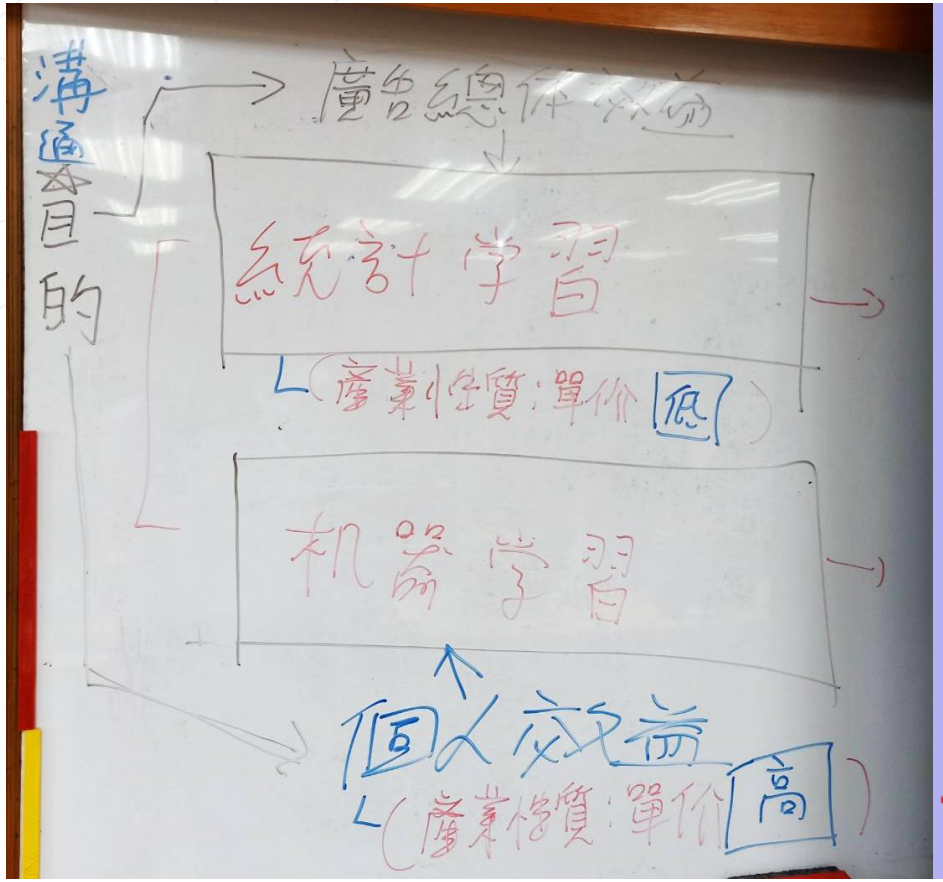


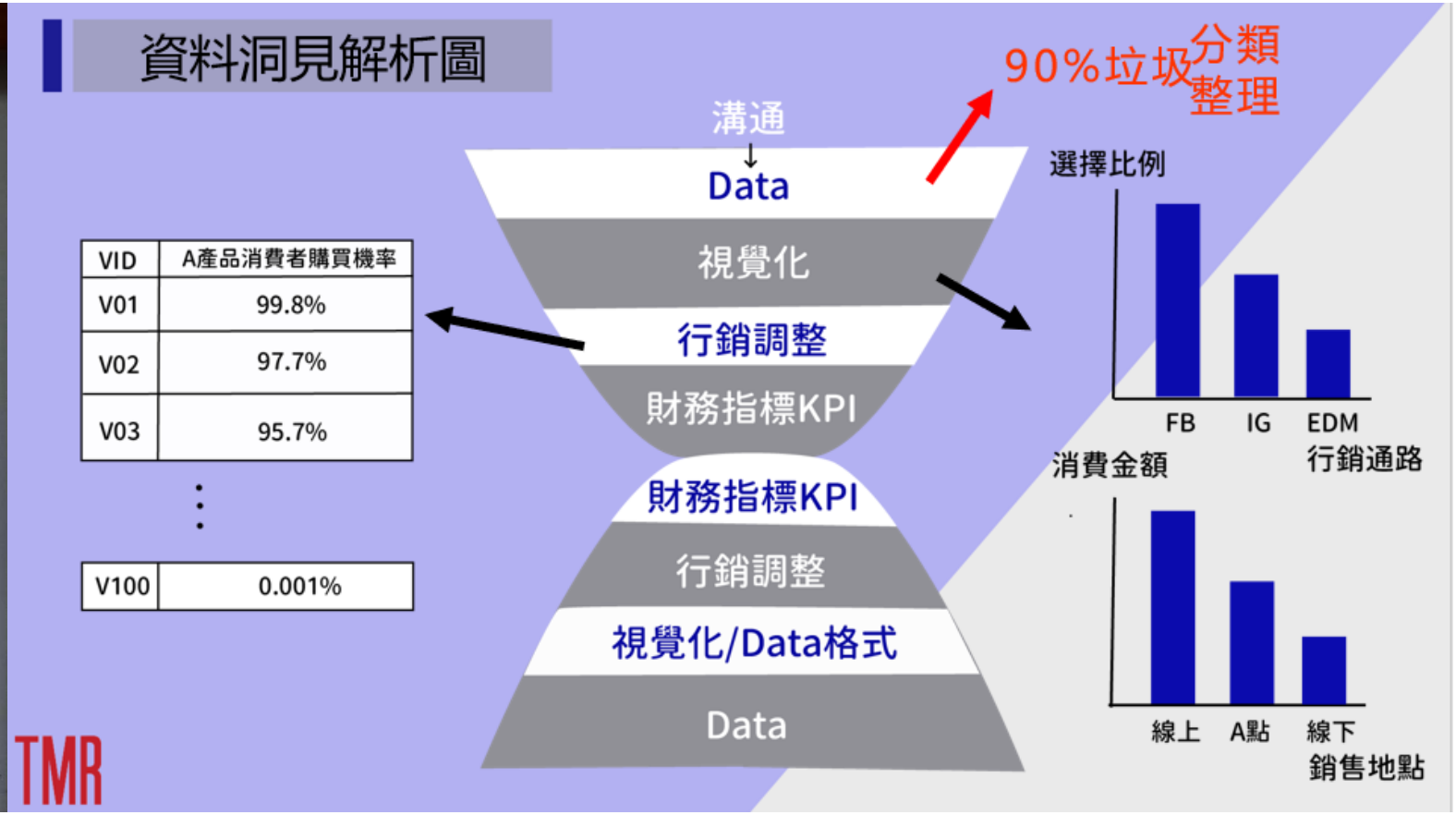
90%垃圾分類整理



TMR

Data Insight 延伸版





溫故知新：行銷資料科學數據思維

行銷資料科學的應用

老闆的思維

對於機器學習，老闆希望的是...

1. 投資可以好好利用
2. 機器學習結合獲利指標，讓主管快速了解分析成果
3. 挑選最佳的機器學習模型，創造營收最大化
4. 從顧客體驗分析中訂定產品開發策略，成為下一個獲利關鍵

綜整：對於機器學習的商務應用來說

1. 資料科學家的思維是...
   1) 如何將模型精準度調到最好
   2) 參數怎麼調到最好
   3) Deep learning 的解決方法

2. 老闆的思維是...

行銷資料科學的應用案例

什麼是精準行銷下的商業價值呢?

係數：影響目標變數大小

$$Y = a\ X = f(x)$$

開源獲利

依變數：「開源獲利」的變數
ex： 業績、按讚數

自變數：各種可能的數字。
ex：人口、價格、成本、使用者行為
(eg. 運動內衣使用行為、連續掃貨)

什麼是精準行銷下的商業價值呢?

係數：影響目標變數大小

$$Y = a\ X = f(x)$$

節流省錢

依變數：「節流省錢」的變數
ex： 詐欺偵測、變相節省經費

自變數：各種可能的數字。
ex：人口、價格、成本、使用者行為
(eg. 運動內衣使用行為、連續掃貨)

行銷資料科學的應用

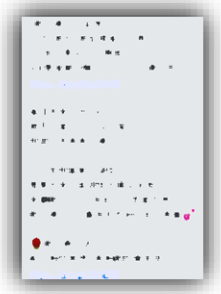
1. 個人精準行銷

$f($ 影像資料集

ID	產品	標準化停留時間
1	1	5
1	2	3
1	3	4
1	4	3
1	5	3
1	7	4
1	8	1
1	9	5
1	11	2
1	13	5
1	15	5
1	16	5
1	18	4
1	19	5
1	21	1
1	22	4
1	25	4

$)=$ UID001 要推薦
A牌紫色洋裝 – 購買機率 0.83
B紅色休閒帽 – 購買機率 0.77

2. 寫手機器人

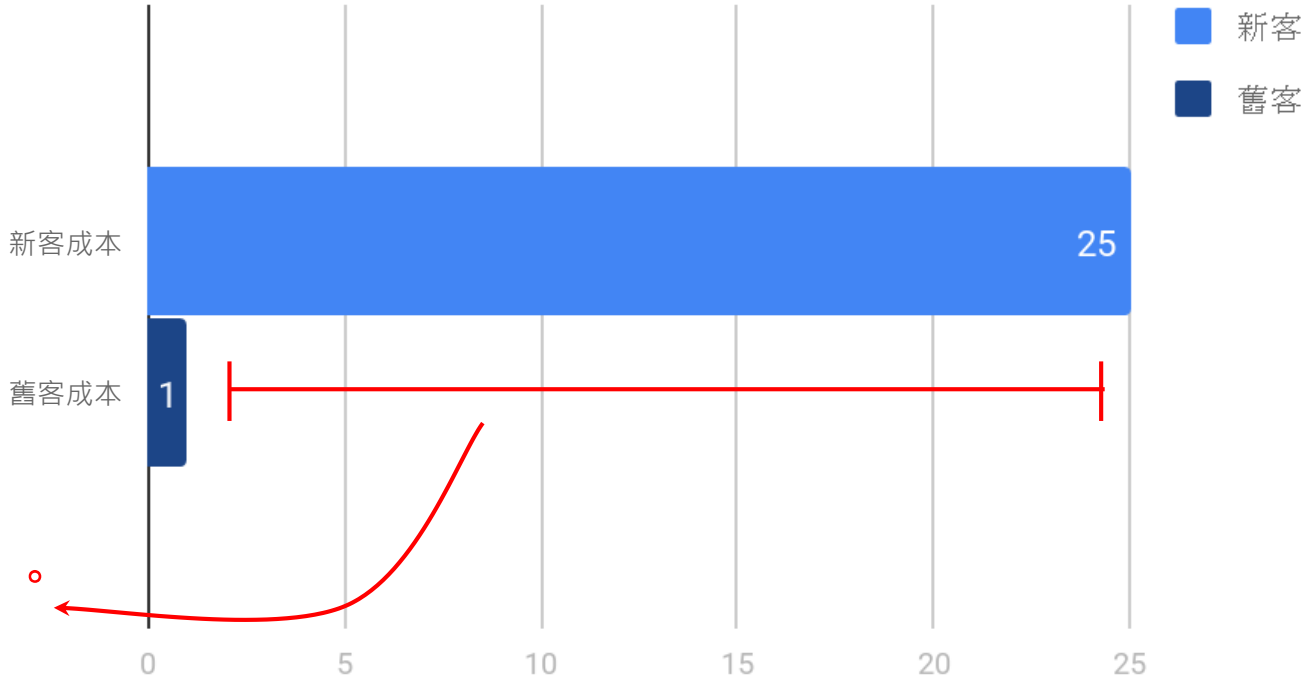
$f($  $)=$

自動給分	預測客單價： $ 54-79
文章建議	1. 建議瀏覽這3篇文章 2. 加入優惠、塑身2種字詞，預計客單價可達 $ 65-120

開發新客與留住舊客費用比較

哈佛商業評論曾說：

「留住一個客人，遠比帶進一個新客更划算！」



實際新客費用是舊客5～25 倍。
對舊客推薦商品更划算！

數據來源：https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

解決方案-分析舊客推薦

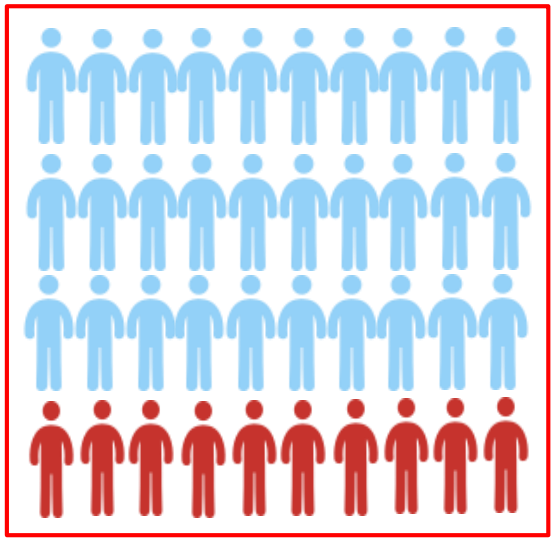
對現有的**顧客關係管理**（CRM）資料並結合顧客**訂單**資料分析，找出相似顧客的購買習慣。

達到精準推薦！

為什麼需要精準名單？

不買A商品的顧客

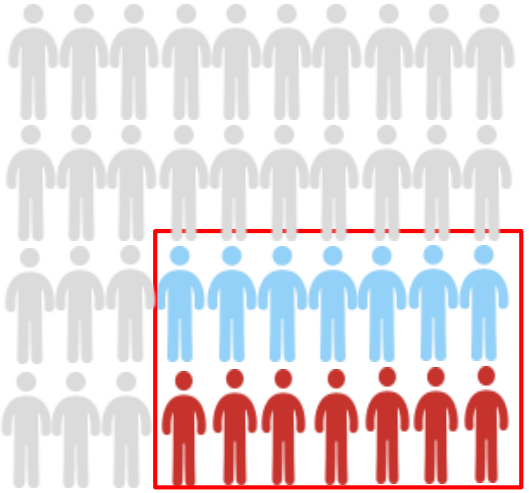
會買A商品的顧客



對**所有會員**發送
A商品促銷

不買A商品的顧客

會買A商品的顧客



只對**精準名單**發送
A商品促銷

區塊代表會員人數

精準名單人數

用更低的宣傳成本達到轉換效果。

老闆問：「10個人業績是10萬元，但用了精準名單後，業績只有7萬元耶」怎麼回答這件事情？

目標客群精準推薦
結果演示

客戶ID	商品預測購買機率	商品【實際】購買狀況
客戶ID5093	0.9642585	1
客戶ID8850	0.9635527	1
客戶ID9232	0.96198136	1
客戶ID3959	0.9610572	1
客戶ID4520	0.9566384	1
客戶ID376	0.9552811	1
客戶ID776	0.9541129	1
客戶ID6375	0.95382756	1
客戶ID5607	0.9523173	0
客戶ID7242	0.9515154	1

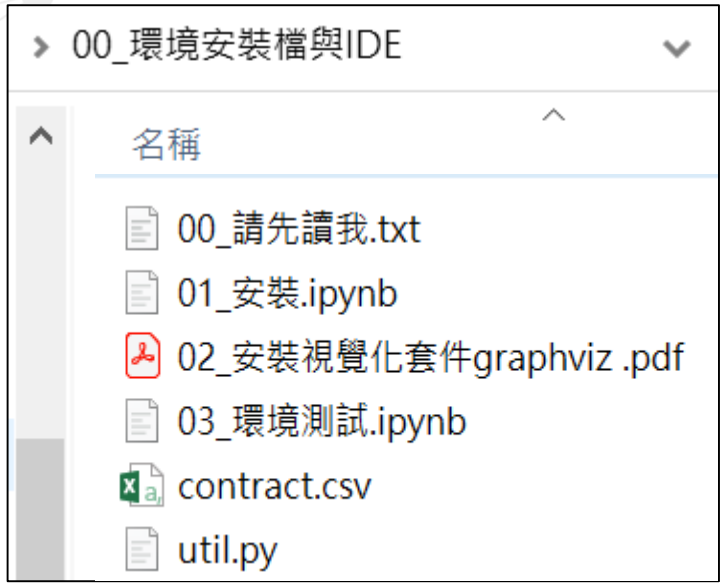
精準名單

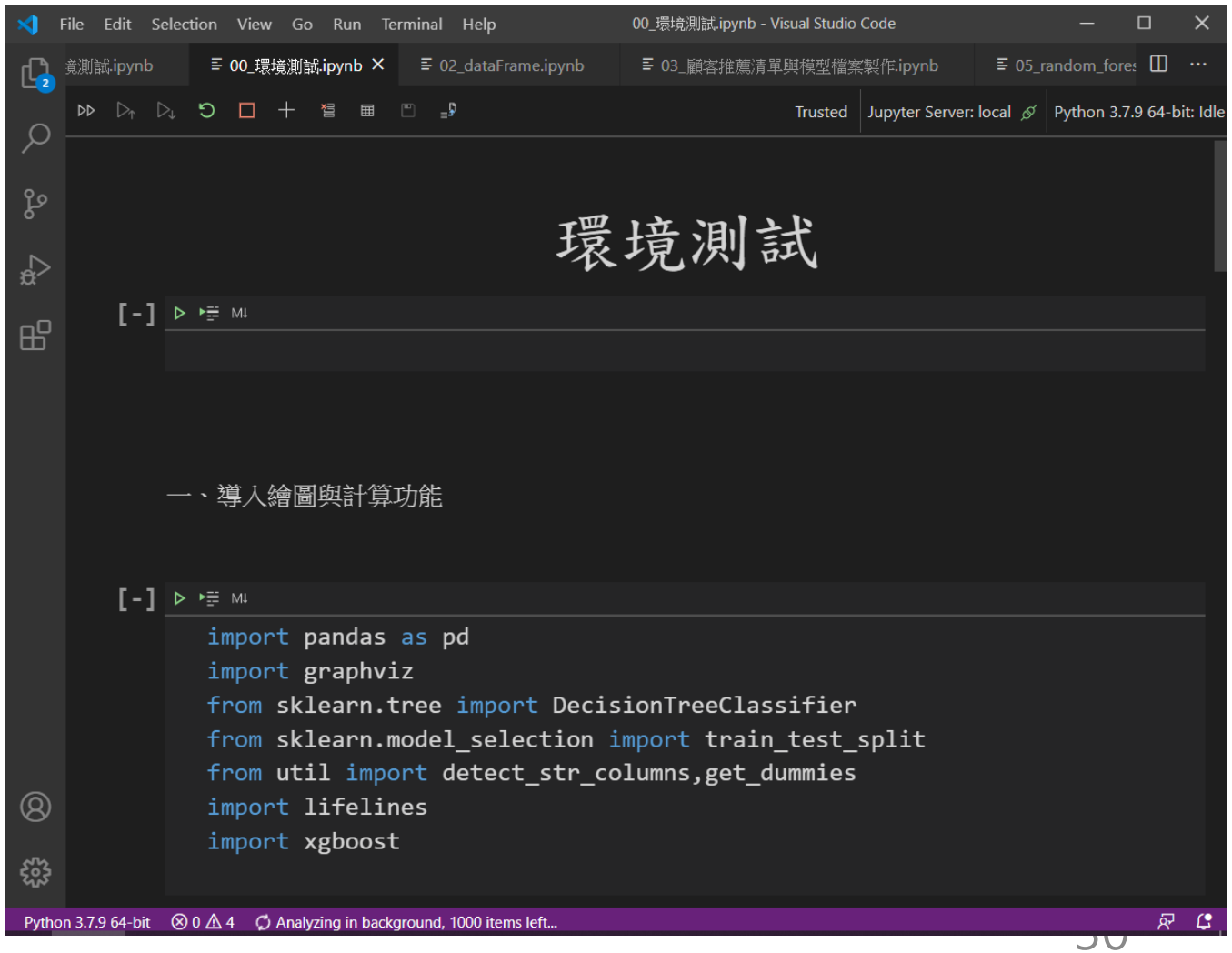
簡訊推送

1.6%購買轉換率

從0.4%提升4倍

測試環境


00_環境安裝檔與IDE
名稱
00_請先讀我.txt
01_安裝.ipynb
02_安裝視覺化套件graphviz .pdf
03_環境測試.ipynb
contract.csv
util.py


File Edit Selection View Go Run Terminal Help
00_環境測試.ipynb - Visual Studio Code
00_環境測試.ipynb
02_dataFrame.ipynb
03_顧客推薦清單與模型檔案製作.ipynb
Trusted
Jupyter Server: local
Python 3.7.9 64-bit: Idle
環境測試
[-]
一、導入繪圖與計算功能
[-]
import pandas as pd
import graphviz
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from util import detect_str_columns,get_dummies
import lifelines
import xgboost
Python 3.7.9 64-bit
Analyzing in background, 1000 items left...

大綱

授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法

Python回顧小複習

You are in Restricted Mode

Code is in a restricted mode intended for safe code browsing.
Configure your settings or learn more.

In a Trusted Window

You trust the authors of the files in the current window. All features are enabled:

- ✓ Tasks are allowed to run
- ✓ Debugging is enabled
- ✓ All extensions are enabled

Trust

In Restricted Mode

You do not trust the authors of the files in the current window. The following features are disabled:

- ✗ Tasks are not allowed to run
- ✗ Debugging is disabled
- ✗ 13 extensions are disabled or have limited functionality

Trusted Folders & Workspaces

You trust the following folders, their subfolders, and workspace files.

Host	Path
Local	c:\Users\howar\Desktop\STP系統可能性評估\2.第二頁：市場區隔分析

Add Folder

Security › Workspace › Trust: **Banner**

Controls when the restricted mode banner is shown.

untilDismissed

Security › Workspace › Trust: **Empty Window**

Controls whether or not the empty window is trusted by default within VS Code. When used with Security › Workspace › Trust: Untrusted Files, you can enable the full functionality of VS Code without prompting in an empty window.

Security › Workspace › Trust: **Enabled**

Controls whether or not workspace trust is enabled within VS Code.

Security › Workspace › Trust: **Startup Prompt**

Controls when the startup prompt to trust a workspace is shown.

never

Security › Workspace › Trust: **Untrusted Files**

Controls how to handle opening untrusted files in a trusted workspace. This setting also applies to opening files in an empty window which is trusted via Security › Workspace › Trust: Empty Window.

open

Jupyter: **Allow Unauthorized Remote Connection**

Allow for connecting the Interactive window to a https Jupyter server that does not have valid certificates. This can be a security risk, so only use for known and trusted servers.

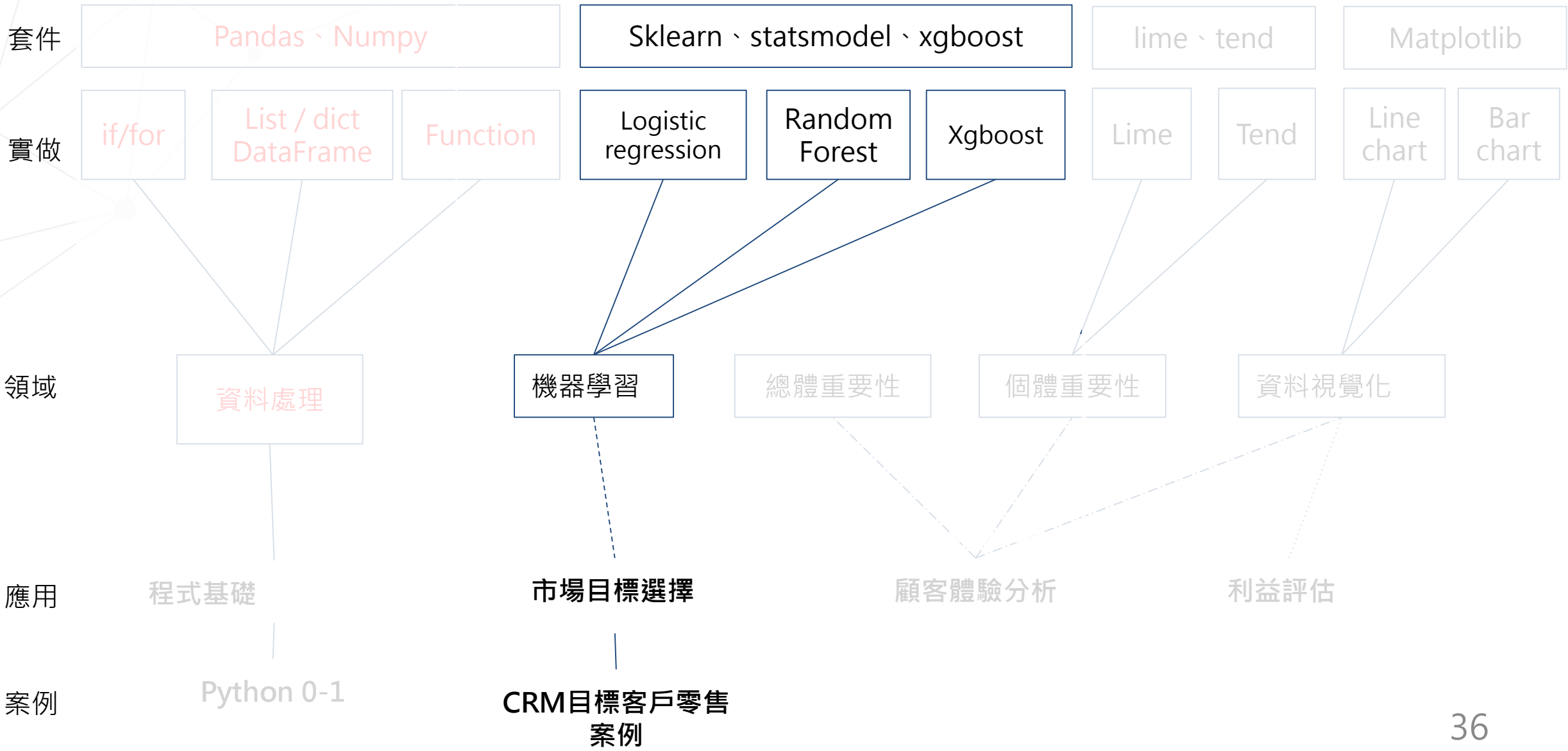
Show matching extensions

TMR
臺灣行銷研究有限公司

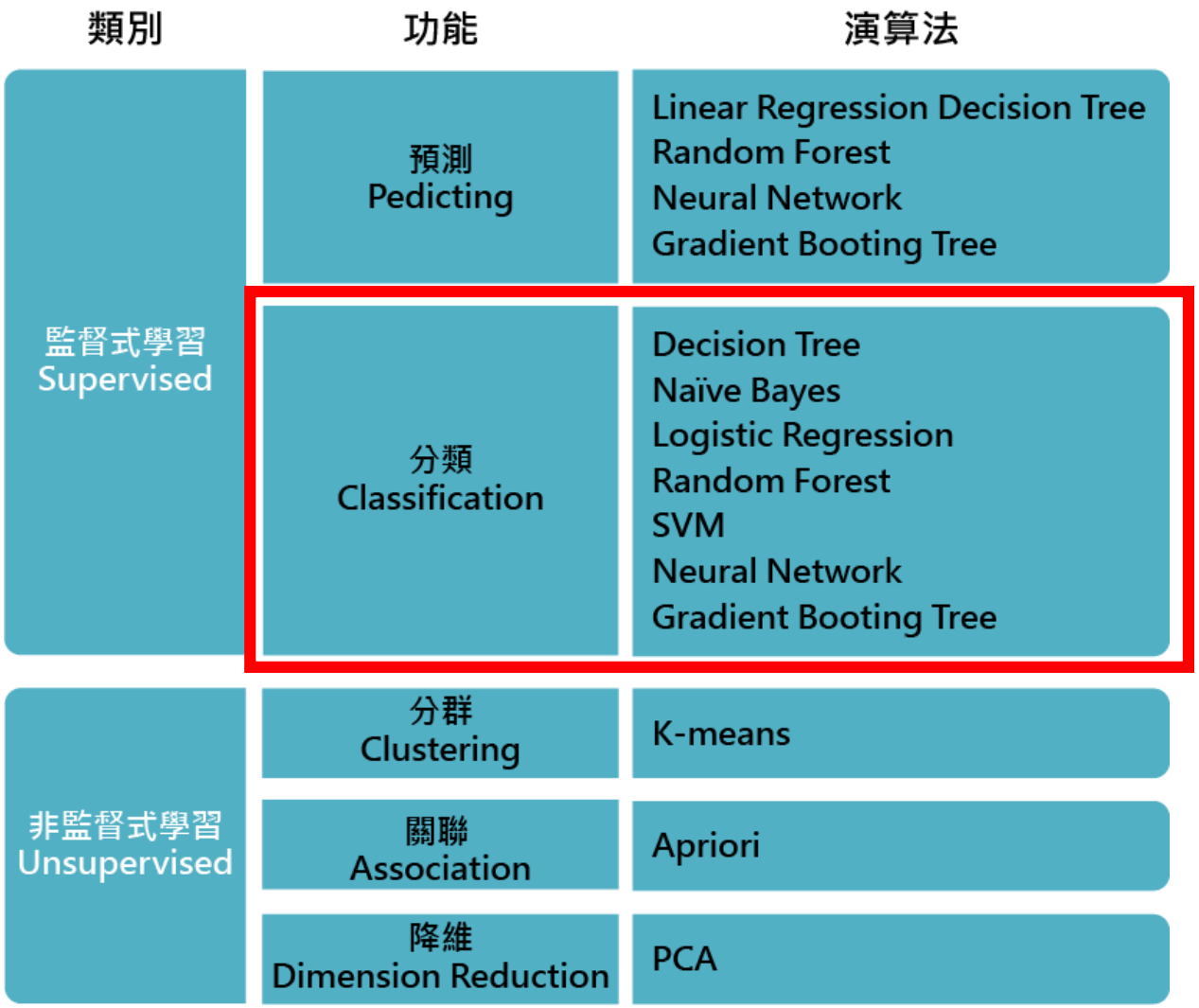
大綱

授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法

Python機器學商務實戰 – 學習地圖



機器學習概論－「分類」



類別	功能	演算法
監督式學習 Supervised	預測 Pedicting	Linear Regression Decision Tree Random Forest Neural Network Gradient Booting Tree
	分類 Classification	Decision Tree Naïve Bayes Logistic Regression Random Forest SVM Neural Network Gradient Booting Tree
非監督式學習 Unsupervised	分群 Clustering	K-means
	關聯 Association	Apriori
	降維 Dimension Reduction	PCA

為什麼我們要探討【分類】模型?

1. 商業經營上，常探討...【非1及0的課題】
2. 金融、電商零售、人資...分析主題 → 商業上賺錢的課題
   1) 金融 Input: 收入、年齡、經濟狀況... **output: 核發信用卡與否**
   2) 電商 Input: 產品使用狀況、網站停留時間... **output: 會不會買**
   3) 人資 Input: 年資、星座、遲到時數... **output: 會不會離職**

3. 最重要的是....
4. **管理者/老闆能不能第一時間就知道我們在說甚麼**

分類 vs 迴歸理解程度

1. 以老闆角度來思考... 哪個直觀?

2. 分類
   1) 分類模型準確率 = 90%
   2) A消費者會購買P產品機率為 80%

分類

管理者第一時間就知道購買P產品的機會很高，非常直觀

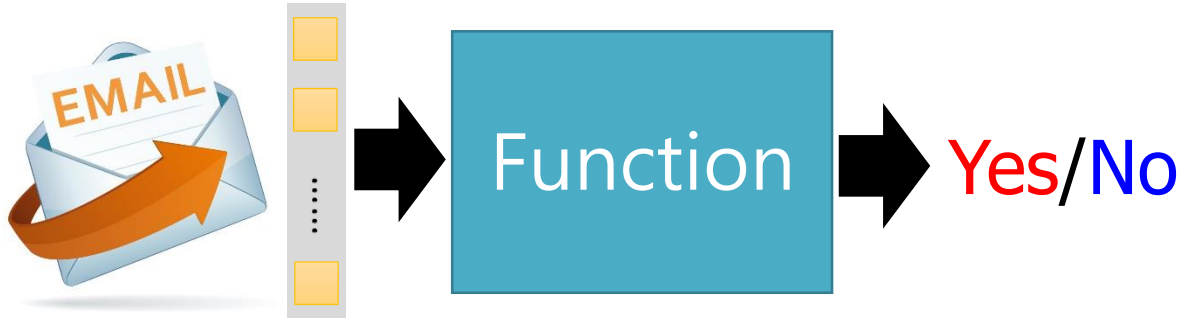
3. 迴歸
   1) 迴歸模型損失函數 RMSE = 12.56
   2) A消費者估計會花100元購買P產品，且平均誤差為11.15.....

迴歸

管理者首先要有RMSE的相關知識，再來要知道P產品的價格，也就是說，管理者要了解每一產品價格，如果有...千種商品..

簡單說分類

Spam filtering



(http://spam–filter–review.toptenreviews.com/)

分類的最大利器 – Confusion Matrix

1. 準確率 = (TP+TN)/所有測試資料集 =(TP+TN) / (TP+TN+FN+FP)

混淆矩陣		預測結果		
	全體測試樣本	預期為會買($\hat{p}$)	預期為不買($\hat{n}$)	
實際結果	實際為會買(P)	**真的會買** **True Positive (TP)**	假的不買 False Negative (FN)	召回率 (recall) = $\frac{TP}{TP+FN}$
	實際為不買(N)	假的買 False Positive (FP)	**真的不買** **True Negative (TN)**	
		精確率 (Precision) = $\frac{TP}{TP+FP}$		準確率 (accuracy)= $\frac{TP + \mathrm{TN}}{all\ Test\ data}$

分類的優化指標

1. 準確率（Accuracy）:
   1) 挑選模型基本的好壞法則

2. 精確率 (Precision) :
   1) 時常會使用在保安系統等高行銷成本之商業活動

3. 召回率 (recall) :
   1) 時常會使用在低行銷成本之商業活動

4. 不懂? 沒關係!
   1) 您不懂，老闆也不懂，所以後面我們會以【獲利】來挑選模型!

你愛數學嗎？想認真知道原理嗎？

$$P(C_1|x) = \sigma(z)$$

$$z = \ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}x^T(\Sigma^1)^{-1}x + (\mu^1)^T(\Sigma^1)^{-1}x - \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1$$
$$+\frac{1}{2}x^T(\Sigma^2)^{-1}x - (\mu^2)^T(\Sigma^2)^{-1}x + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + \ln\frac{N_1}{N_2}$$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$z = \underbrace{(\mu^1 - \mu^2)^T\Sigma^{-1}}_{\boldsymbol{w^T}}x \underbrace{- \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + \ln\frac{N_1}{N_2}}_{\mathrm{b}}$$

$$P(C_1|x) = \sigma(w \cdot x + b)$$

In generative model, we estimate $N_1, N_2, \mu^1, \mu^2, \Sigma$

如果想要知道「分類」等機器學習模型更深的應用...

1. 李宏毅 – Machine learning – 數學完整介紹
   1) 分類模型課程
   2) Machine learning全課程

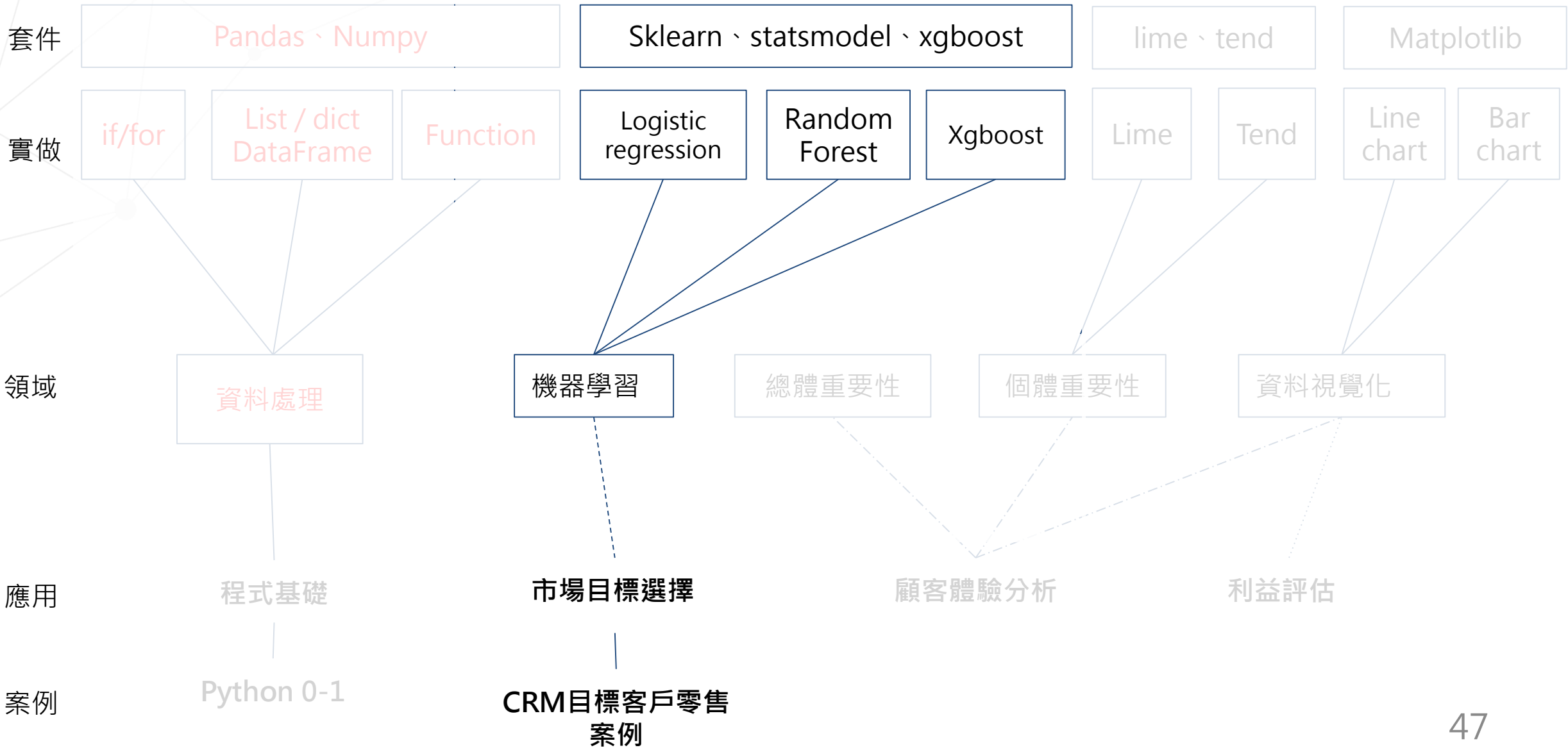
大綱

授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo

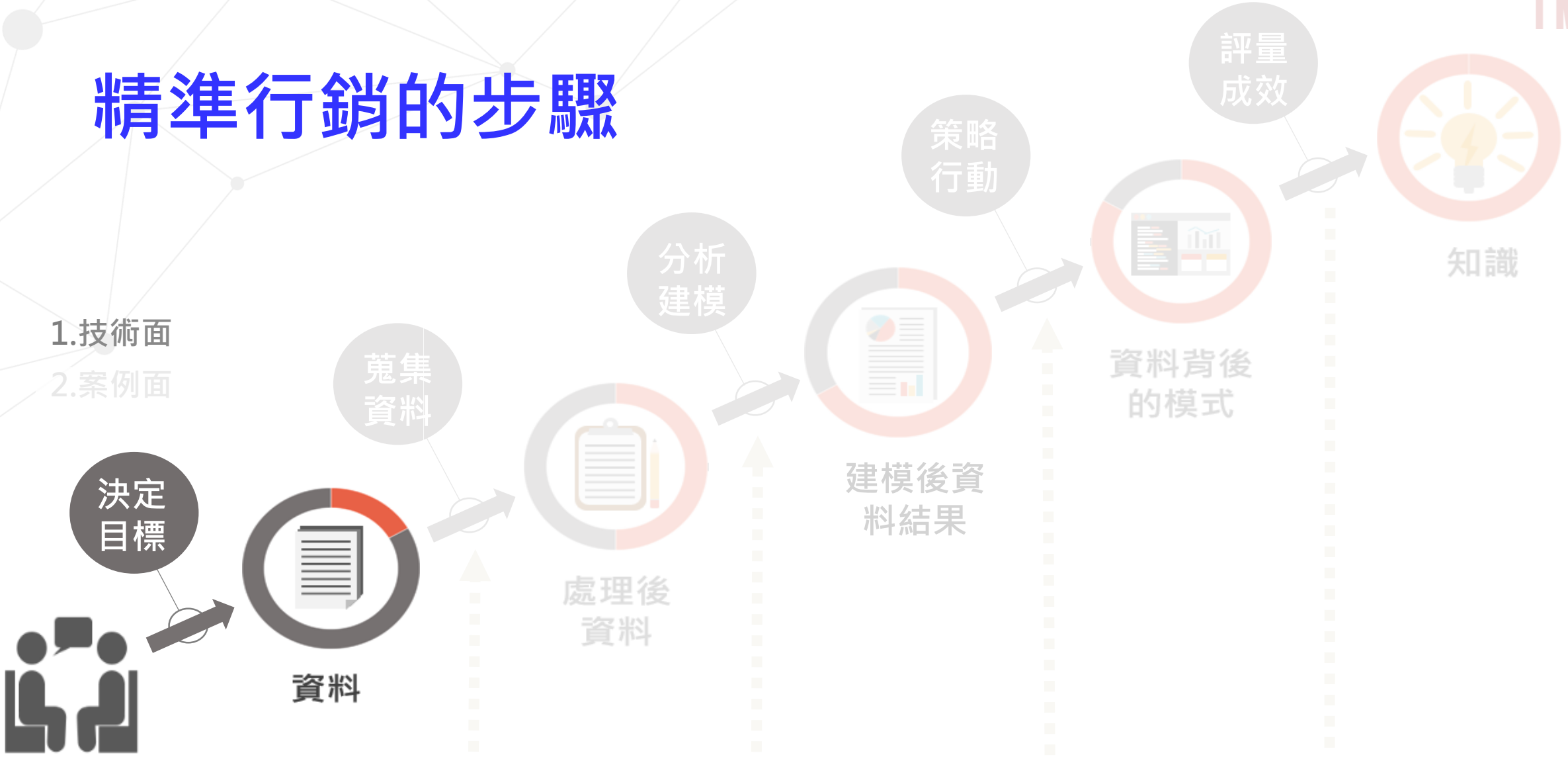
Python實戰 – 商業分類模型

1. 如何挑選機器學習演算法
2. 模型實戰演練

Python機器學商務實戰 – 學習地圖



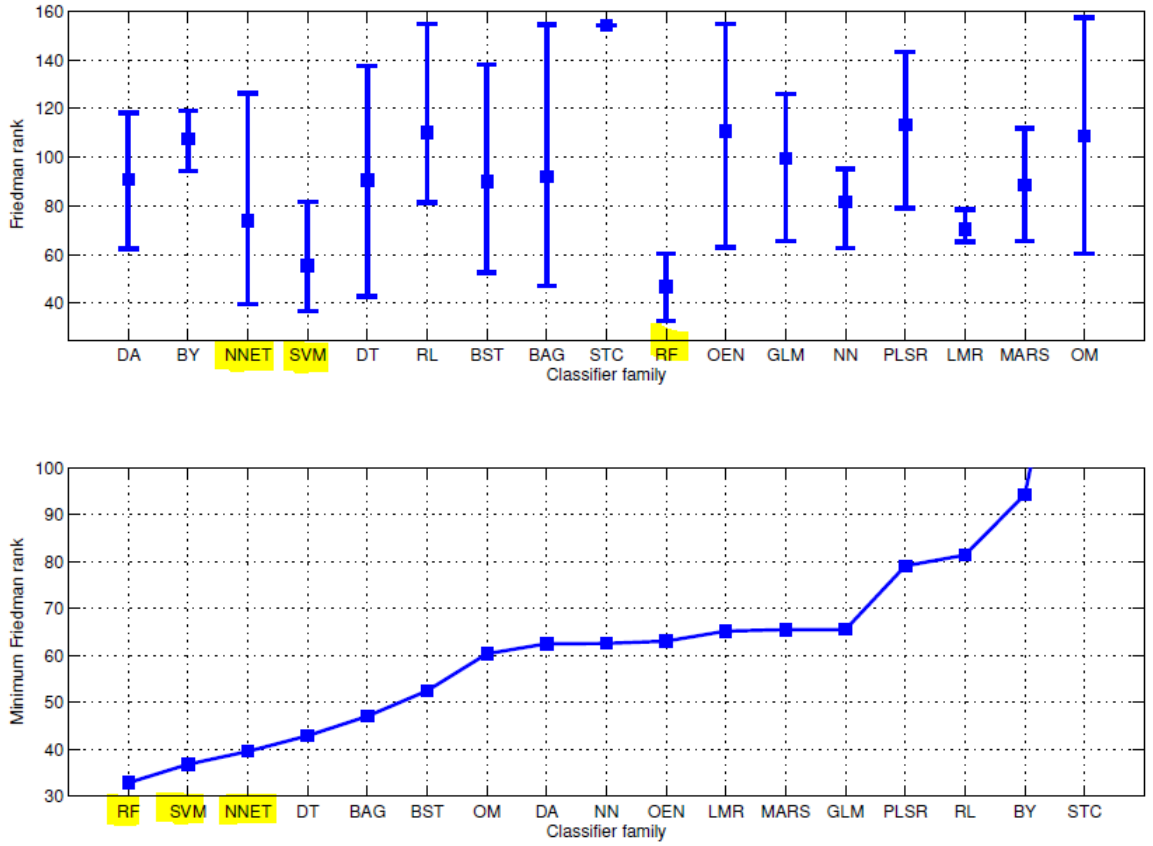
精準行銷的步驟



繪圖者：廖庭儀、鍾皓軒

問題：
該找何種機器學習演算法?

2014年的3大最好模型

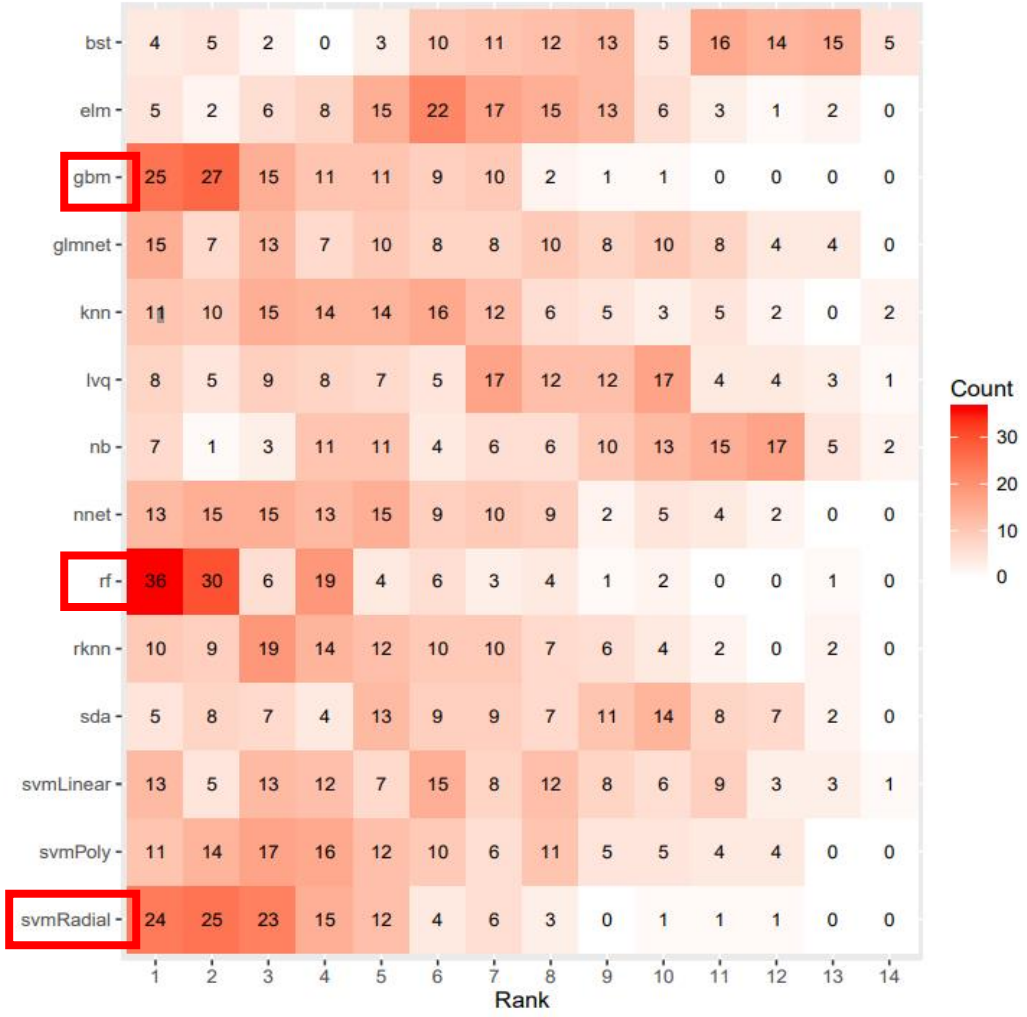


2014研究指出
三大最好的預測模型：RF、SVM、ANN
參數最易控制排序：RF、SVM、ANN

Rank	Acc.	κ	Classifier
32.9	82.0	63.5	parRF_t (RF)
33.1	**82.3**	**63.6**	rf_t (RF)
36.8	81.8	62.2	svm_C (SVM)
38.0	81.2	60.1	svmPoly_t (SVM)
39.4	81.9	62.5	rforest_R (RF)
39.6	82.0	62.0	elm_kernel_m (NNET)
40.3	81.4	61.1	svmRadialCost_t (SVM)
42.5	81.0	60.0	svmRadial_t (SVM)
42.9	80.6	61.0	C5.0_t (BST)
44.1	79.4	60.5	avNNet_t (NNET)

Source： Fernandez–Delgado, M.; Cernades, E.; Barro, S.; Amorim, D. A. （2014）, Do we need hundreds of classifiers to solve real world problems? J. Machine. Learning. Res., 15, 3133−3181

2016年的3大最好模型



alg	mean rank	count
rf	3.04	36
svmRadial	3.36	24
gbm	3.41	25
nnet	4.88	13
rknn	5.04	10
svmPoly	5.14	11
knn	5.32	11
svmLinear	6.15	13
glmnet	6.16	15
elm	6.55	5
lvq	6.96	8
sda	7.05	5
nb	8.23	7
bst	9.08	4

Source： Jacques Wainer （2016）, Comparison of 14 different families of classification algorithms on 115 binary datasets, arXiv：1606.00930v1

2014與2016年最好模型比較

2014年的研究

Rank	Acc.	κ	Classifier
32.9	82.0	63.5	parRF_t (RF)
33.1	**82.3**	**63.6**	rf_t (RF)
36.8	81.8	62.2	svm_C (SVM)
38.0	81.2	60.1	svmPoly_t (SVM)
39.4	81.9	62.5	rforest_R (RF)
39.6	82.0	62.0	elm_kernel_m (NNET)
40.3	81.4	61.1	svmRadialCost_t (SVM)
42.5	81.0	60.0	svmRadial_t (SVM)
42.9	80.6	61.0	C5.0_t (BST)
44.1	79.4	60.5	avNNet_t (NNET)

2016年的研究

alg	mean rank	count
rf	3.04	36
svmRadial	3.36	24
gbm	3.41	25
nnet	4.88	13
rknn	5.04	10
svmPoly	5.14	11
knn	5.32	11
svmLinear	6.15	13
glmnet	6.16	15
elm	6.55	5
lvq	6.96	8

2020個人paper

1. gbm
2. Rf

Source： Nguyen, L. T. K., Chung, H.-H., Tuliao, K. V., & Lin, T. M. Y. (2020). Using XGBoost and Skip-Gram Model to Predict Online Review Popularity. SAGE Open. https://doi.org/10.1177/2158244020983316

常用類別模型

請注意!! 所有演算法都是獨立產出結果
箭頭只代表他們是相關的演算法

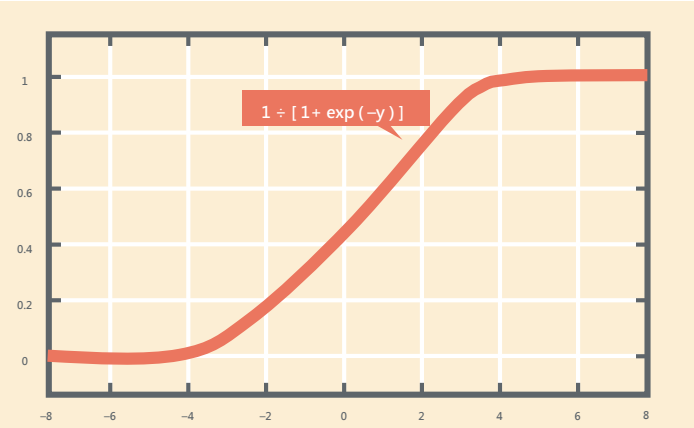
線性迴歸

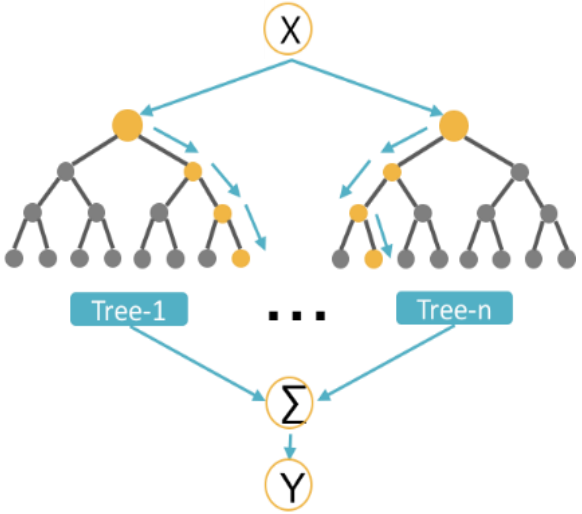
羅吉斯迴歸

決策樹

隨機森林

XGBoost





dmlc XGBoost

單一模型判斷

多棵決策樹一起「獨立」
投票決定

多棵決策樹一起「共同影響」
投票決定

常用類別模型

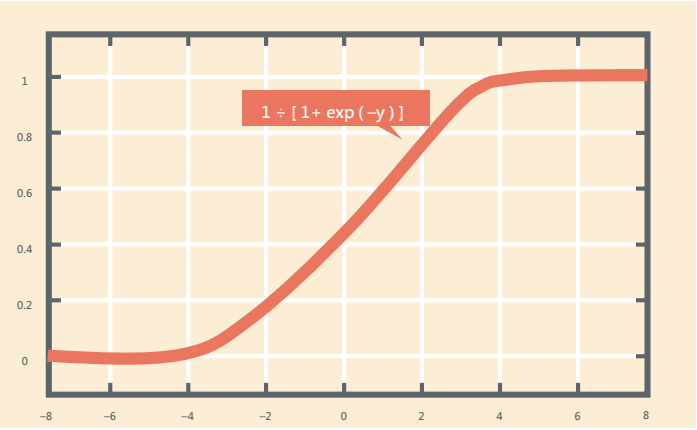
線性迴歸

羅吉斯迴歸

決策樹

隨機森林

XGBoost



dmlc XGBoost

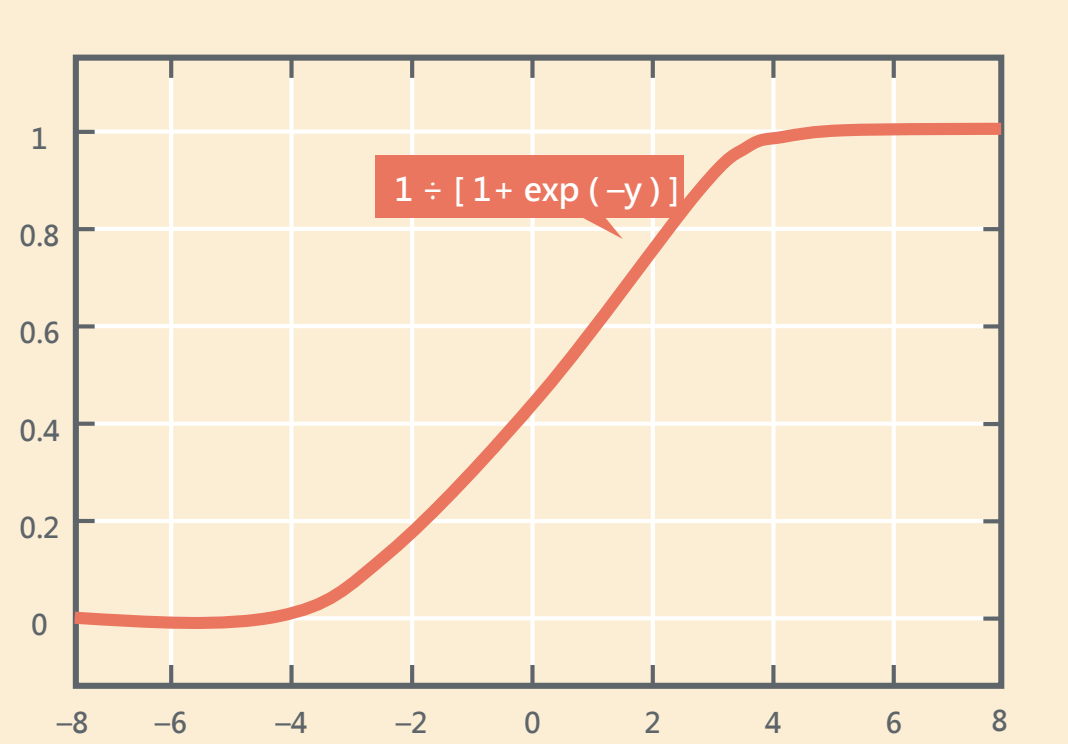
單一模型判斷

多棵決策樹一起「獨立」投票決定

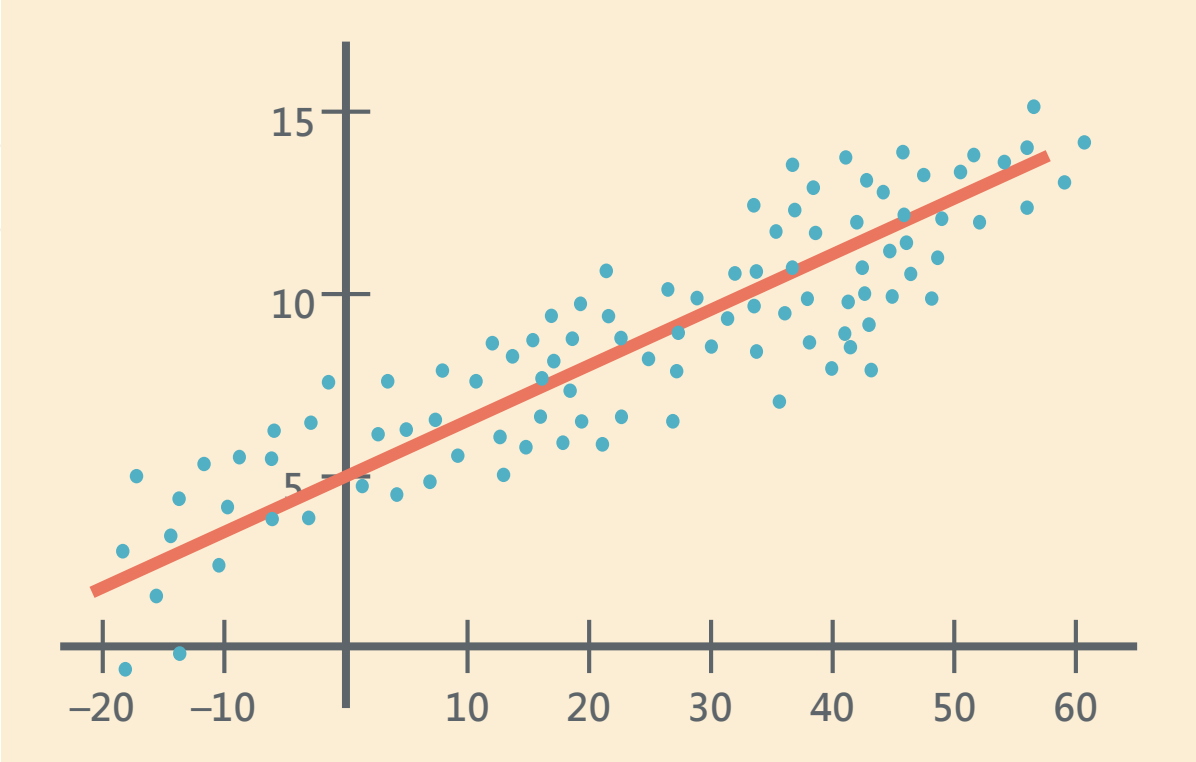
多棵決策樹一起「共同影響」投票決定

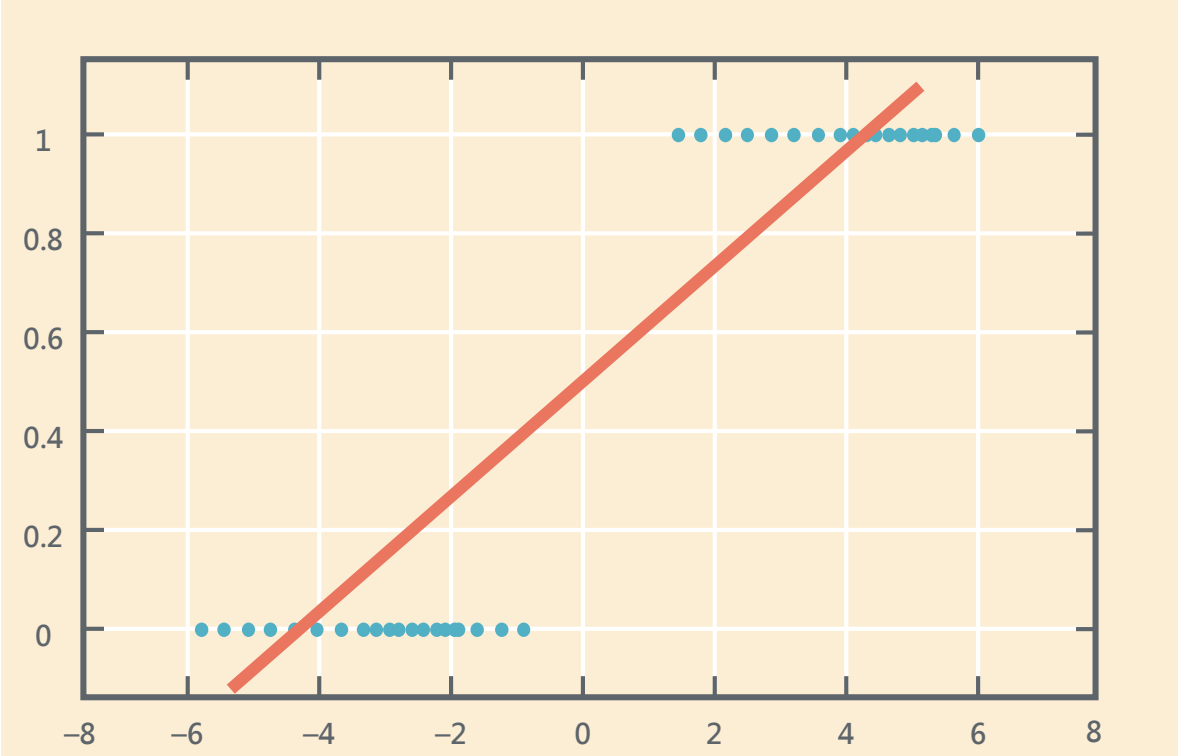
羅吉斯迴歸

1. 也稱「邏輯迴歸」(Logistic regression)
2. 可以思考為線性迴歸進階版
3. 可以判斷類別變數 (0 or 1)
4. Sigmoid函數



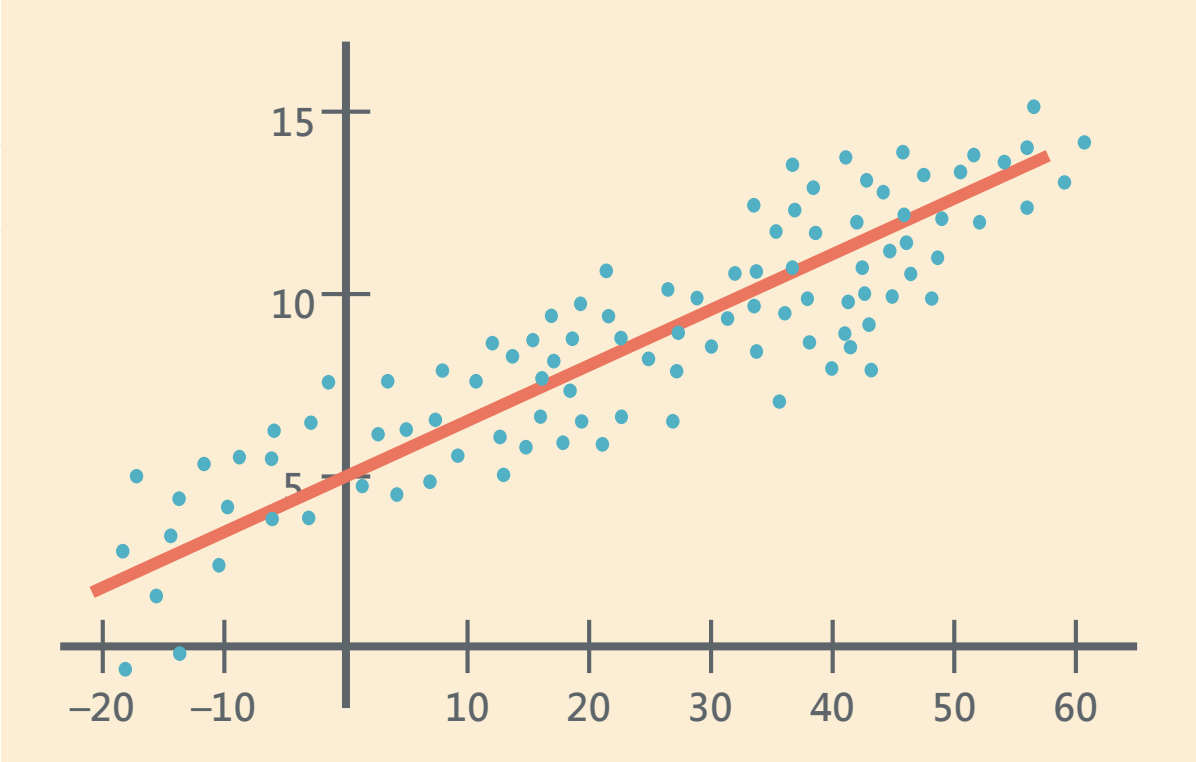
羅吉斯迴歸
為什麼不用linear regression做分類預測？



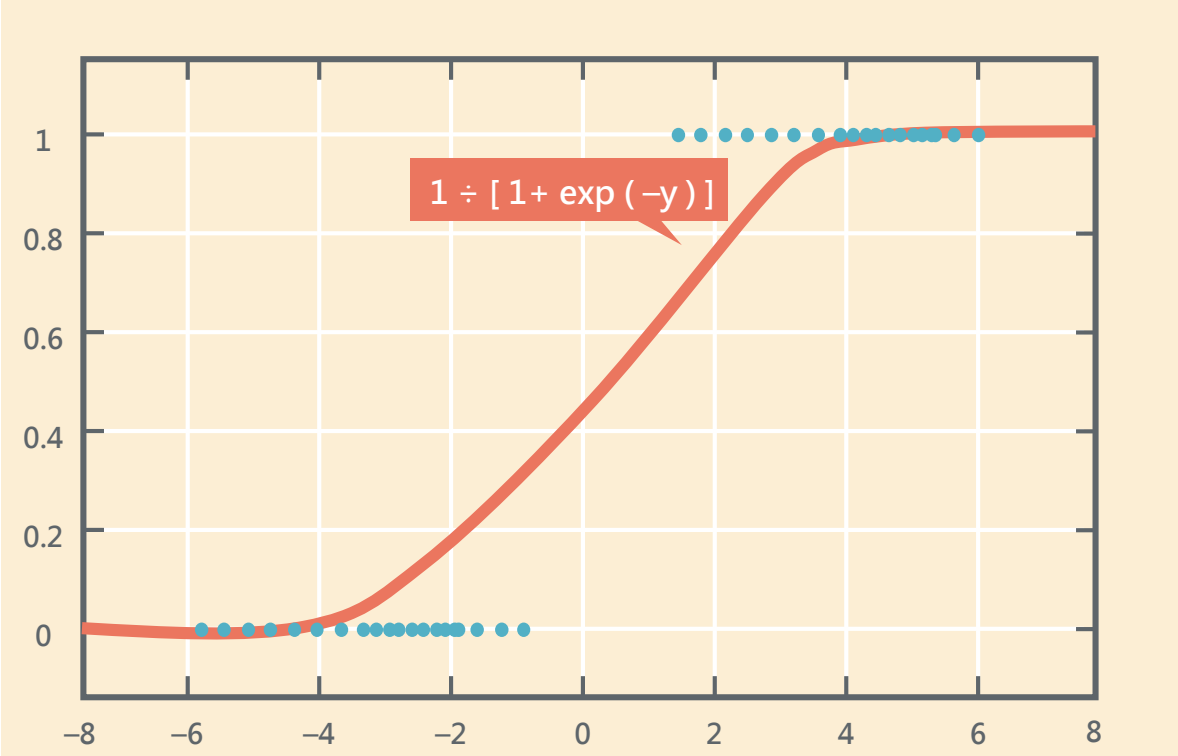


羅吉斯迴歸

$$y = b_0 + b_1 * x$$



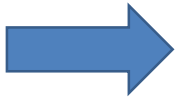
$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



範例資料 -目標欄位(Label 、target) 與 特徵欄位

Age	Buy
10	0
13	0
15	0
18	0
40	1
52	1
53	1
56	1
57	1

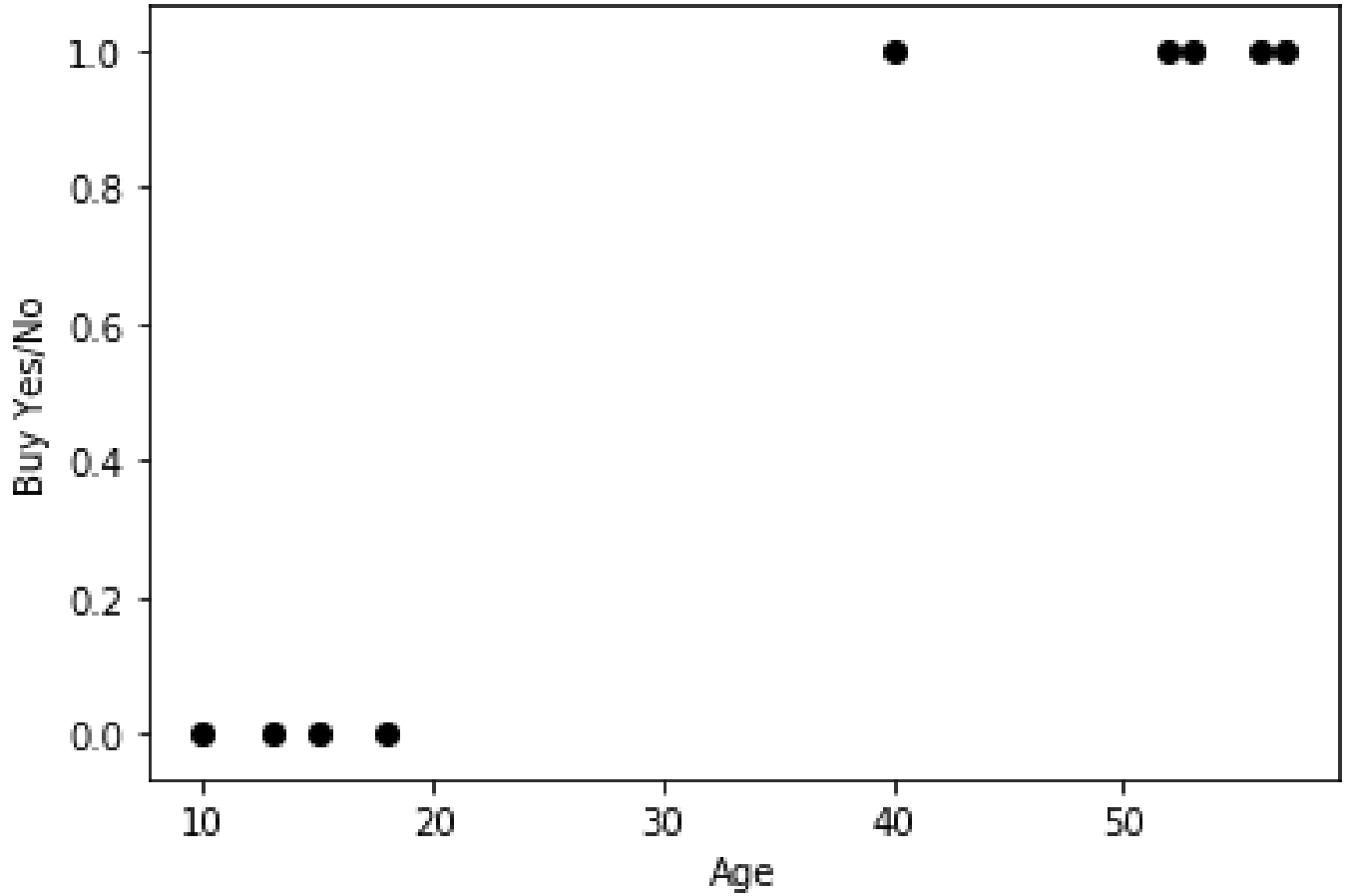
目標欄位



特徵欄位

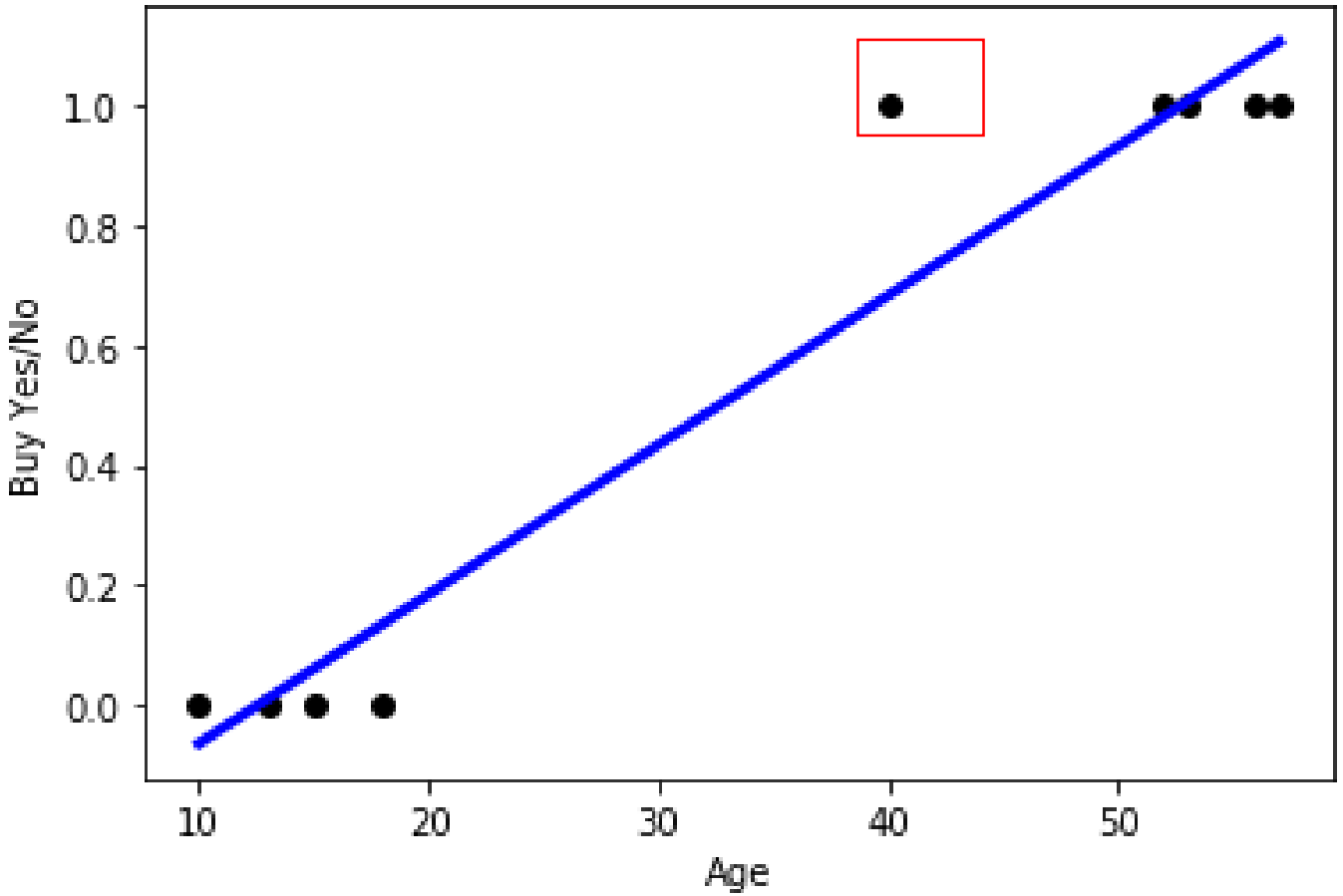
調查公司的商品在不同年紀的購買狀況?

Age	Buy
10	0
13	0
15	0
18	0
40	1
52	1
53	1
56	1
57	1



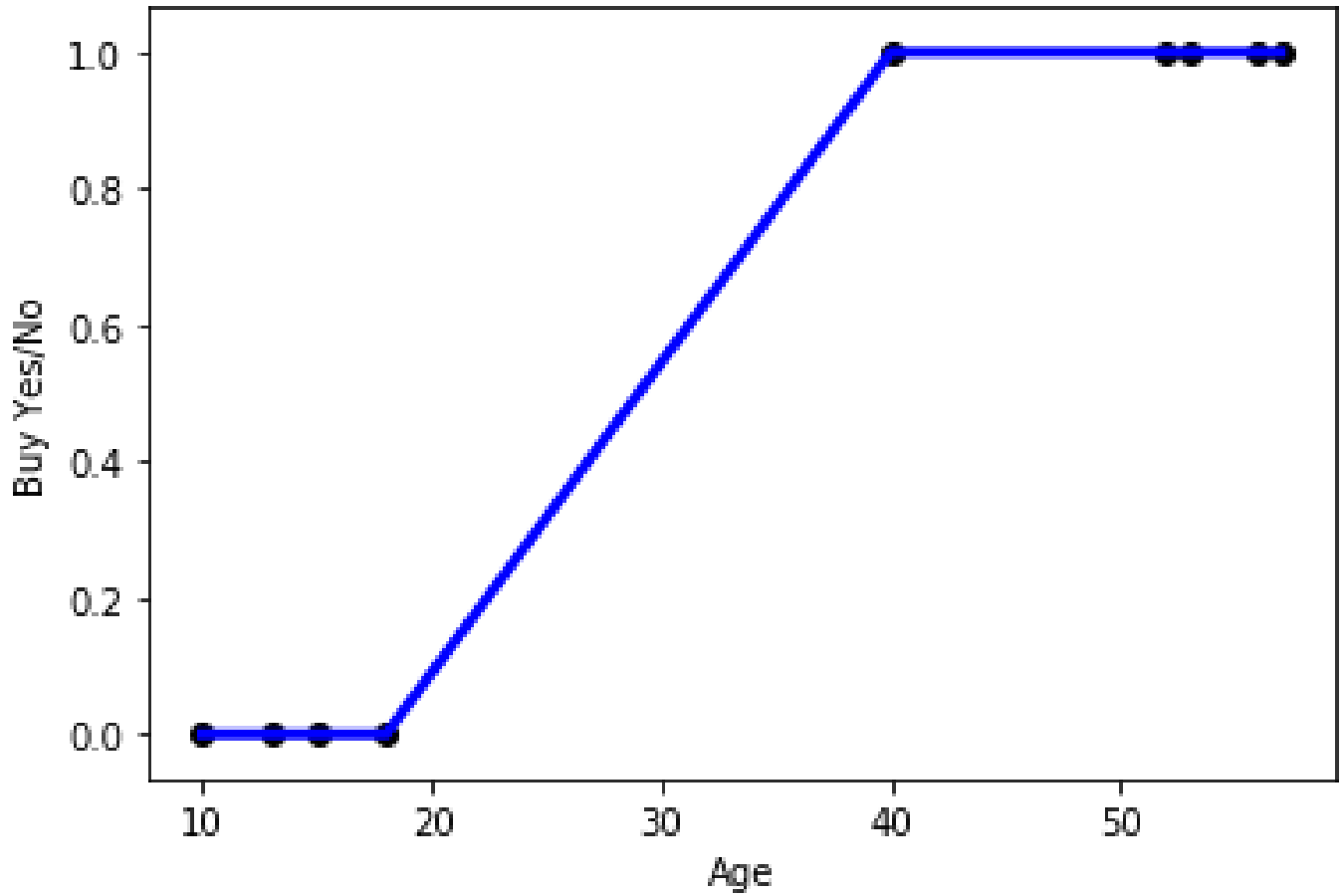
使用線性迴歸的話

Age	Buy
10	0
13	0
15	0
18	0
40	1
52	1
53	1
56	1
57	1



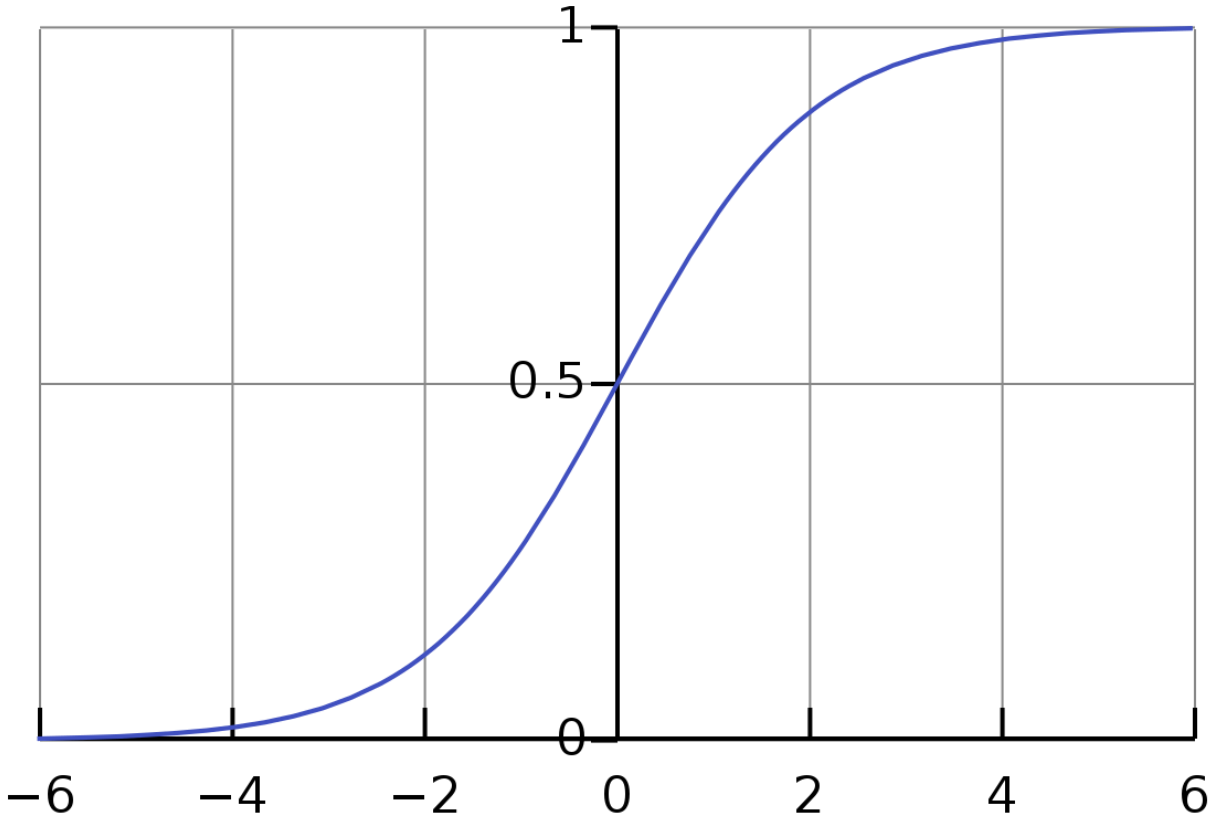
使用羅吉斯迴歸(Logistic regression)

Age	Buy
10	0
13	0
15	0
18	0
40	1
52	1
53	1
56	1
57	1



01_logistic_regression.py

Sigmoid函數



重點在於讓輸出的結果 可以保持在 **0** 跟 **1**

問題：
如何評估目前模組的好壞?

機器學習概論
分類式的監督式學習

rod_output_nun	locations	gender	age	click_on_prod	balance	registry_to_use_time	credit_card_paid	active_member	estimated_sa ry	buy
619	Taipei	Female	42	2	0	1	1	0	101349	1
608	Tainan	Female	41	1	83807.9	1	0	0	112543	0
502	Taipei	Female	42	8	159661	3	1	1	113932	1
699	Taipei	Female	39	1	0	2	0	1	93826.6	0
850	Tainan	Female	43	2	125511	1	1	0	79084.1	0
645	Tainan	Male	44	8	113756	2	1	1	149757	1
822	Taipei	Male	50	7	0	2	1	0	10062.8	0
376	Taichung	Female	29	4	115047	4	1	1	119347	1
501	Taipei	Male	44	4	142051	2	0	0	74940.5	0
684	Taipei	Male	27	2	134604	1	1	0	71725.7	0
528	Taipei	Male	31	6	102017	2	0	1	80181.1	0

預測結果
1
1
0
0

把20%當作 未知資料去預測後跟原本的答案比較
這樣就可以評估模型

問題：
如何讓準度變高?

標準化變數 (Standard Scaler or Z-score)

- Z值的量代表著原始分數和母體平均值之間的距離，是以標準差為單位計算。在原始分數低於平均值時Z則為負數，反之則為正數。換句話說，Z值是從感興趣的點到均值之間有多少個標準差。
- 公式 = (每個數字 - 平均) / 標準差

重點在於 會把數字轉換到 0的附近
可以幫助 model 計算加快、**更準**！
但又不失去原本資料的特性

Z-score轉換

重點在於 會把數字轉換到 0 的附近
可以幫助 model 計算加快、**更準**！
但又不失去原本資料的特性

10	20	30	40	50

Z-score

-1.4	-0.7	0	0.7	1.4

Z-score轉換

平均 ：3

1	2	3	4	5

Z-score

-1.4	-0.7	0	0.7	1.4

代表3 是平均

代表1是距離平均值 -1.4 個標準差

代表4 是距離平均值0.7 個標準差

來實作一下！


02_1_logistic_regression_準確度.py

```
# 標準化變數
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```


02_1_logistic_regression_標準化.py

```
In [45]: from sklearn.metrics import accuracy_score
    ...:
    ...: accuracy = accuracy_score(y_test, y_pred)
    ...: print(accuracy)
0.725
```

```
In [47]: from sklearn.metrics import accuracy_score
    ...:
    ...: accuracy = accuracy_score(y_test, y_pred)
    ...: print(accuracy)
    ...:
0.9125
```

準確度：**0.725**

準確度： **0.913**

來實作一下【精準名單】

03_logistic_regression_
推薦清單.py

TMR
臺灣行銷研究有限公司

UID	客戶對A商品“實際”購買狀態	客戶對A商品“預測”購買狀態
2872	1	0.99897707
555	1	0.998776
2092	1	0.99848217
3152	1	0.9981318
2154	1	0.9980094
7257	1	0.9979923
4166	1	0.99740416
9323	1	0.99729663
7457	1	0.99740416
2499	1	0.99729663
4563	1	0.99694985
1427	1	0.9957709

常用類別模型

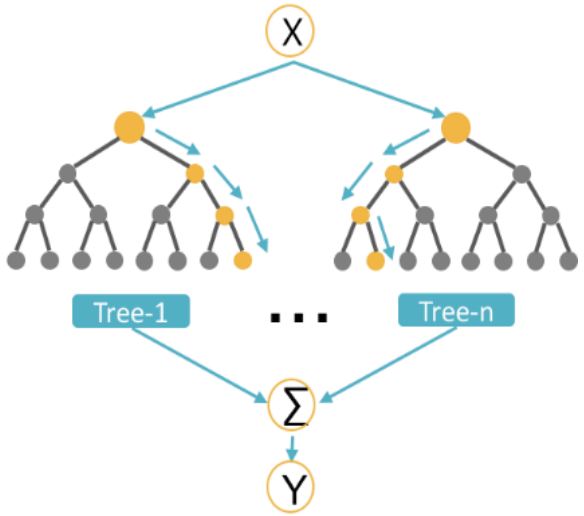
線性迴歸

羅吉斯迴歸

決策樹

隨機森林

XGBoost



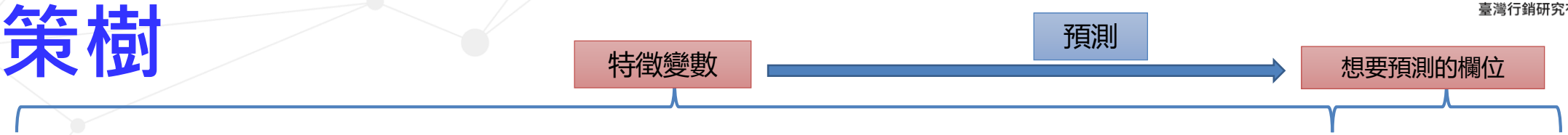
dmlc XGBoost

單一模型判斷

多棵決策樹一起「獨立」投票決定

多棵決策樹一起「共同影響」投票決定

決策樹



編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
1	低	中	是	否	否
2	中	低	否	是	否
3	高	高	是	否	是
4	高	高	是	是	是
5	中	高	是	否	否
6	低	低	否	否	否
7	高	低	否	否	否
8	低	低	否	是	是
9	高	中	是	否	是
10	低	中	否	否	否
11	高	中	否	是	否
12	中	中	是	是	否

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
年所得級距	高–>是	5	2	12	3
	中–>否	3	0		
	低–>否	4	1		

編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
3	高	高	是	否	☑是
4	高	高	是	是	☑是
7	高	低	否	否	☒否
9	高	中	是	否	☑是
11	高	中	否	是	☒否

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
年所得級距	高–>是	5	2	12	3
	中–>否	3	0		
	低–>否	4	1		

編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
2	中	低	否	是	☑否
5	中	高	是	否	☑否
12	中	中	是	是	☑否

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
年所得級距	高–>是	5	2	12	3
	中–>否	3	0		
	低–>否	4	1		

編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
1	低	中	是	否	☑否
6	低	低	否	否	☑否
8	低	低	否	是	☒是
10	低	中	否	否	☑否

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
資產級距	高–>是	?	?	12	3
	中–>否	5	1		
	低–>否	4	1		

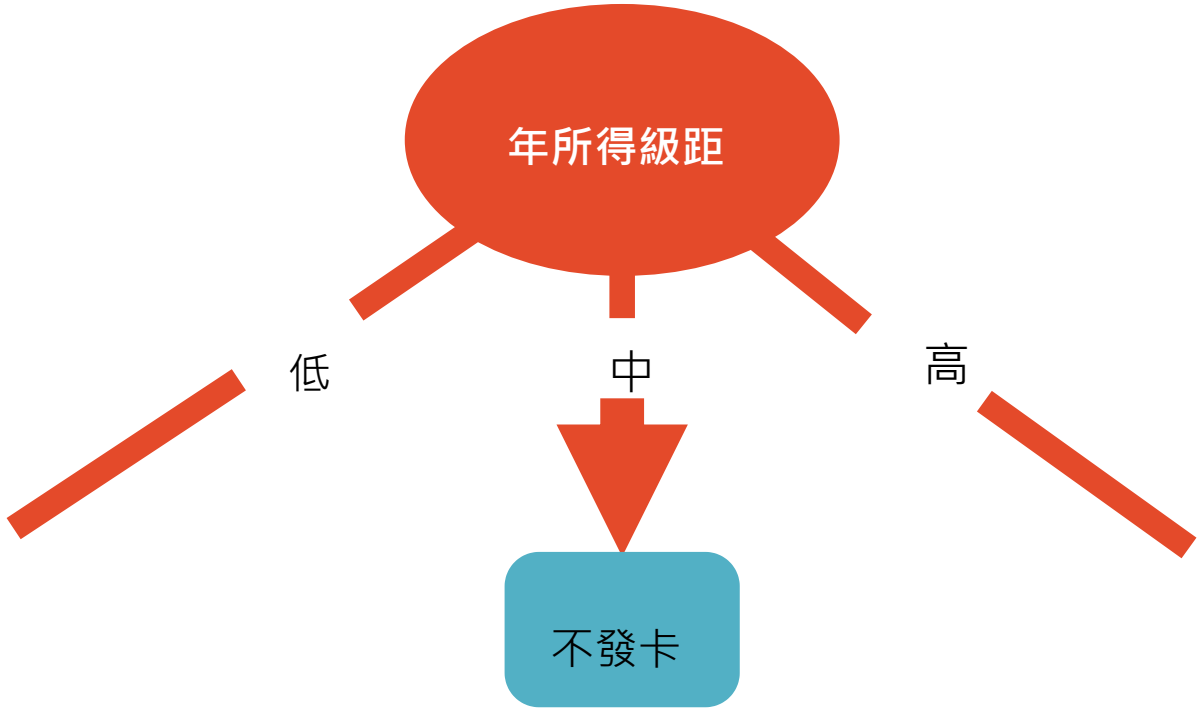
編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
3	高	高	是	否	是
4	高	高	是	是	是
5	中	高	是	否	否

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
年所得級距	高–>是	5	2	12	3
	中–>否	3	0		
	低–>否	4	1		
資產級距	高–>是	3	1	12	3
	中–>否	5	1		
	低–>否	4	1		
是否擁有信用卡	是–>是	6	3	12	4
	否–>否	6	1		
存款是否超過20萬	是–>否	5	2	12	4
	否–>否	7	2		

決策樹

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
年所得級距	高–>是	5	2	12	3
	中–>否	3	0		
	低–>否	4	1		



繪圖者：鄭雅馨

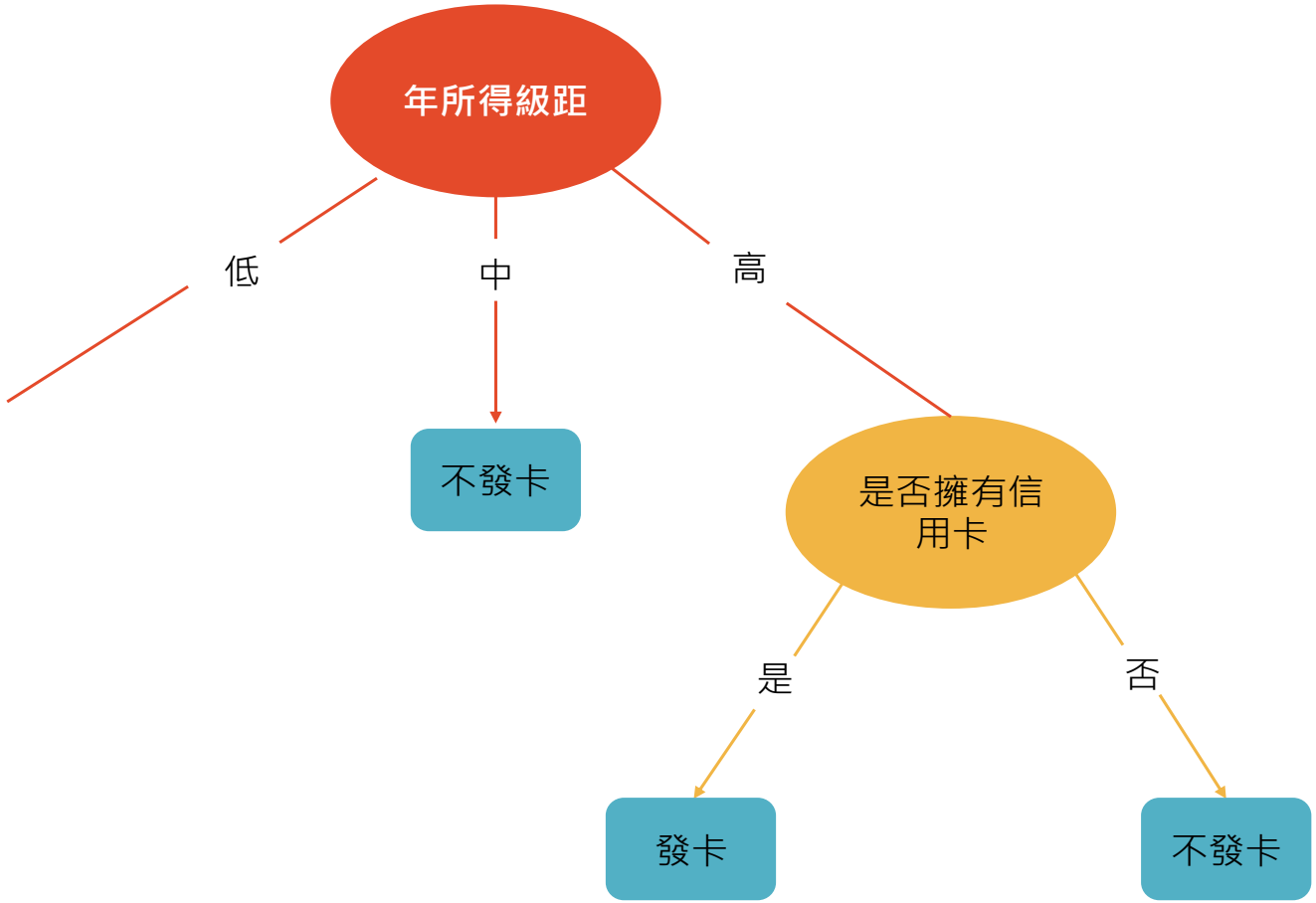
決策樹

下表如何解釋？

編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
3	高	高	☑是	否	☑是
4	高	高	☑是	是	☑是
7	高	低	☑否	否	☑否
9	高	中	☑是	否	☑是
11	高	中	☑否	是	☑否

變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
資產級距	高–>是	2	0	5	1
	中–>否	2	1		
	低–>否	1	0		
是否擁有信用卡	是–>是	3	0	5	0
	否–>否	2	0		
存款是否超過20萬	是–>是	2	1	5	2
	否–>否	3	1		

決策樹



繪圖者：鄭雅馨

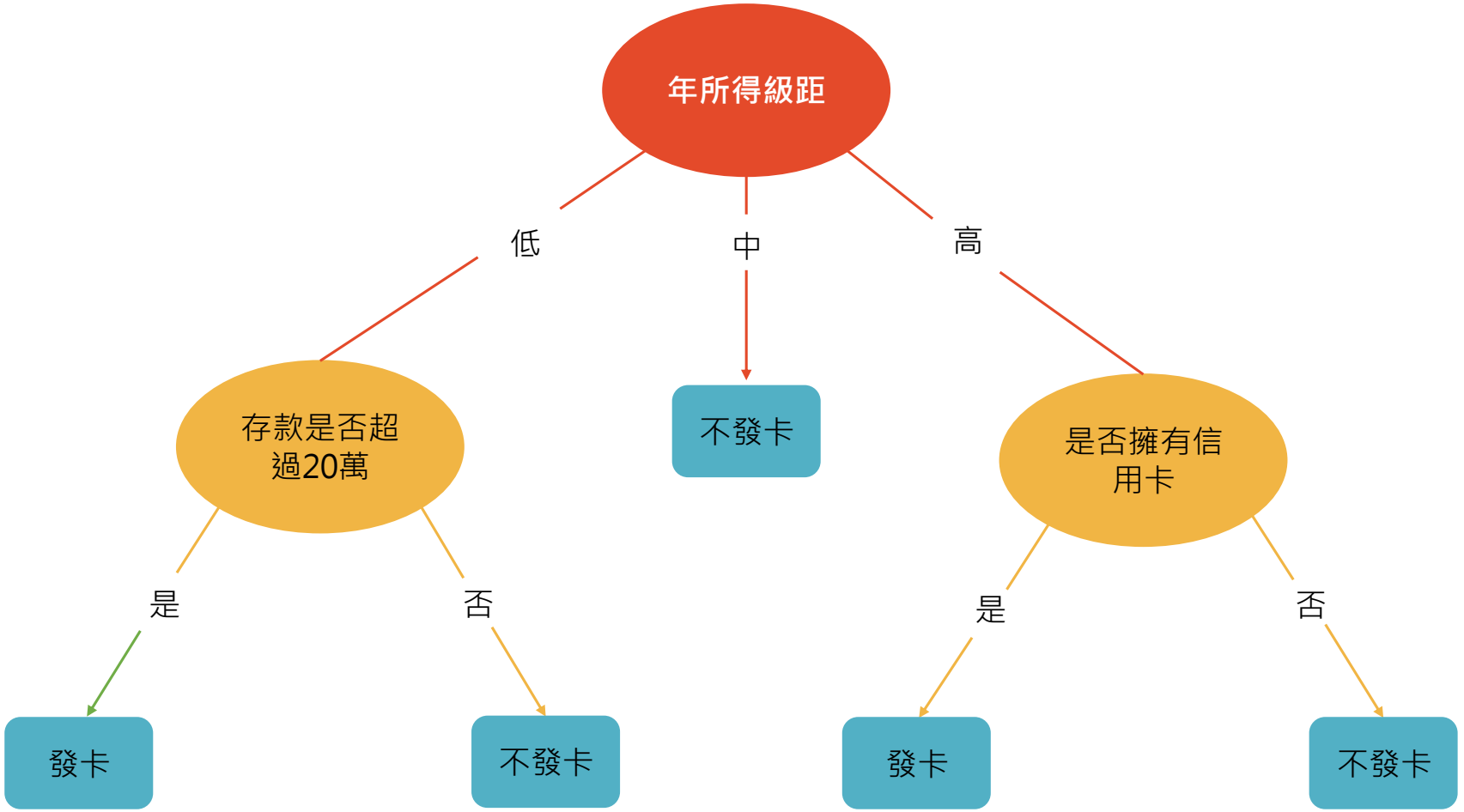
決策樹

下表如何解釋？

編號	年所得級距	資產級距	是否擁有信用卡	存款是否超過20萬	是否核卡
1	低	中	是	☑否	☑否
6	低	低	否	☑否	☑否
8	低	低	否	☑是	☑是
10	低	中	否	☑否	☑否

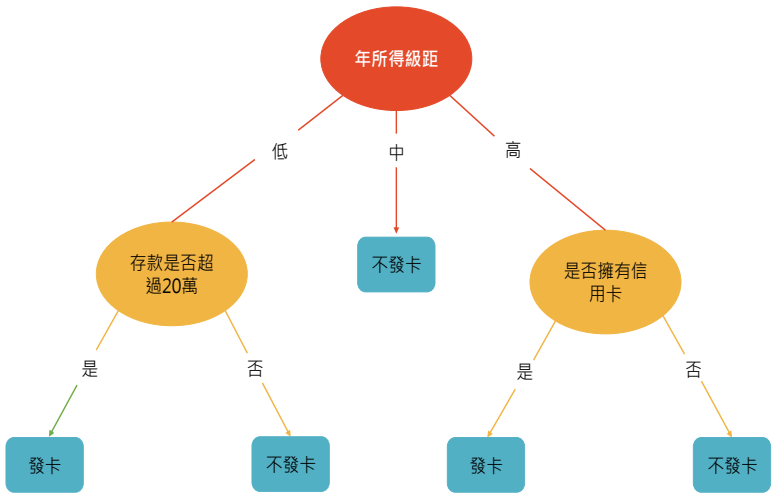
變數	規則	錯誤		總錯誤	
		次數	錯誤	次數	錯誤
資產級距	高–>是	-	-	4	1
	中–>否	2	0		
	低–>否	2	1		
是否擁有信用卡	是–>否	1	0	4	1
	否–>否	3	1		
存款是否超過20萬	是–>是	1	0	4	0
	否–>否	3	0		

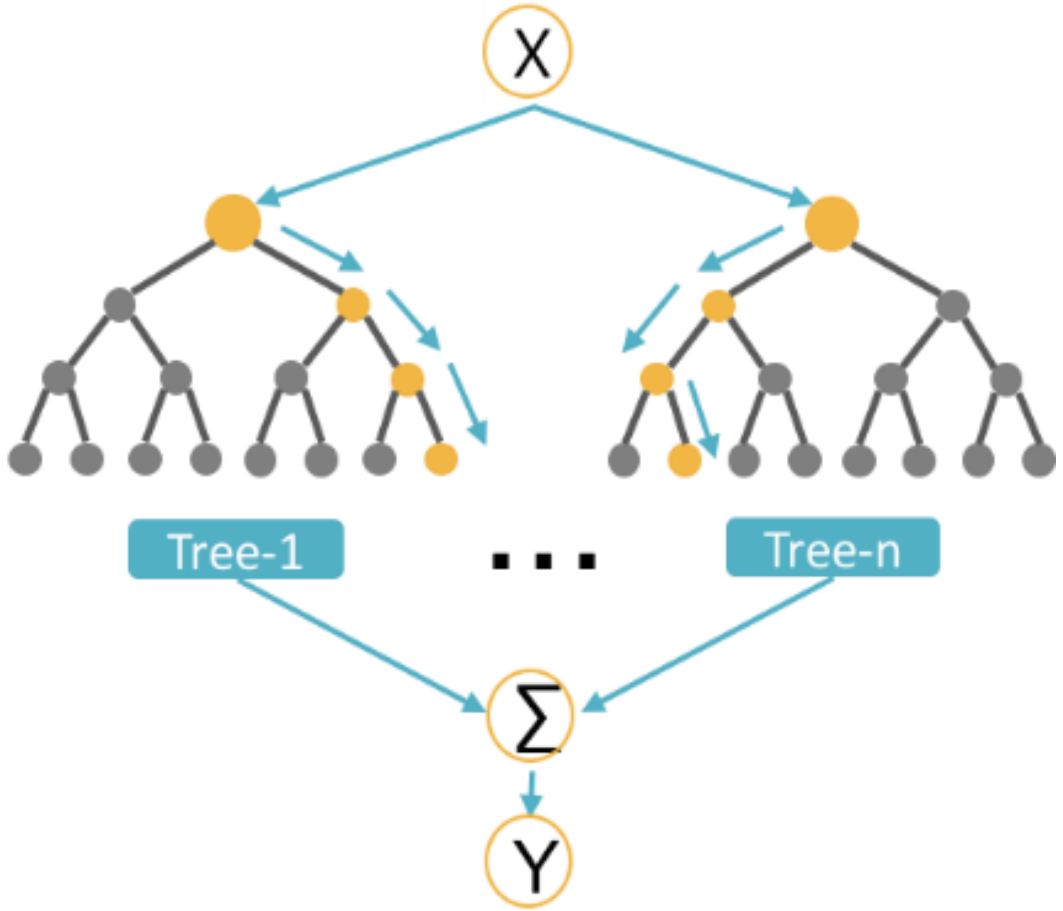
決策樹



繪圖者：鄭雅馨

隨機森林


年所得級距
低
中
高
存款是否超過20萬
不發卡
是否擁有信用卡
是
否
是
否
發卡
不發卡
發卡
不發卡


X
Tree-1
Tree-n
Σ
Y

繪圖者：鄭雅馨

常用類別模型

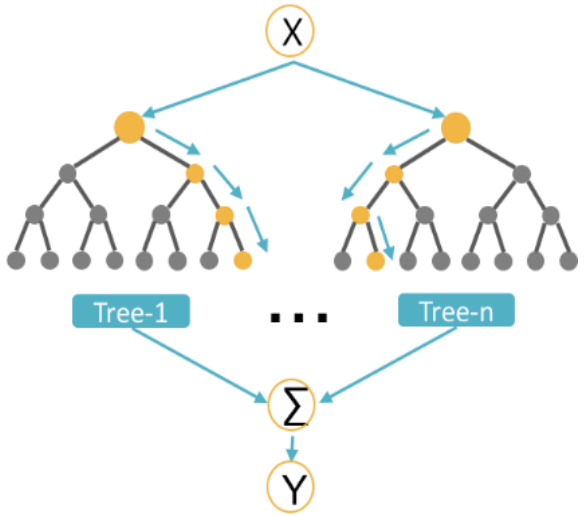
線性迴歸

羅吉斯迴歸

決策樹

隨機森林

XGBoost



dmlc XGBoost

單一模型判斷

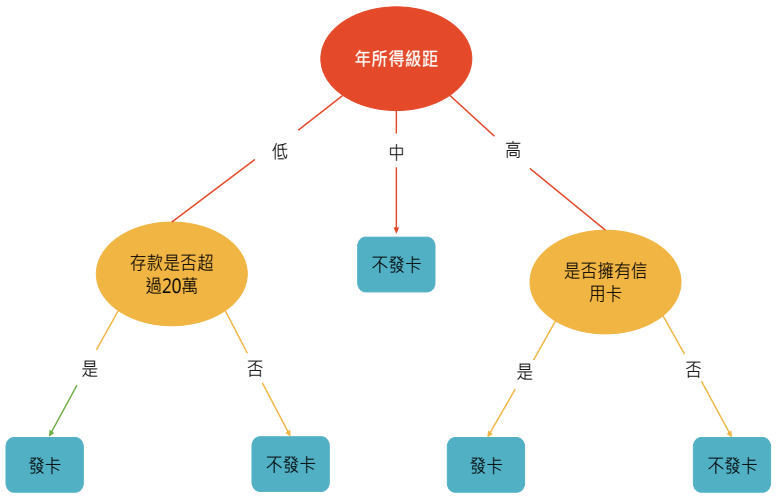
多棵決策樹一起「獨立」投票決定

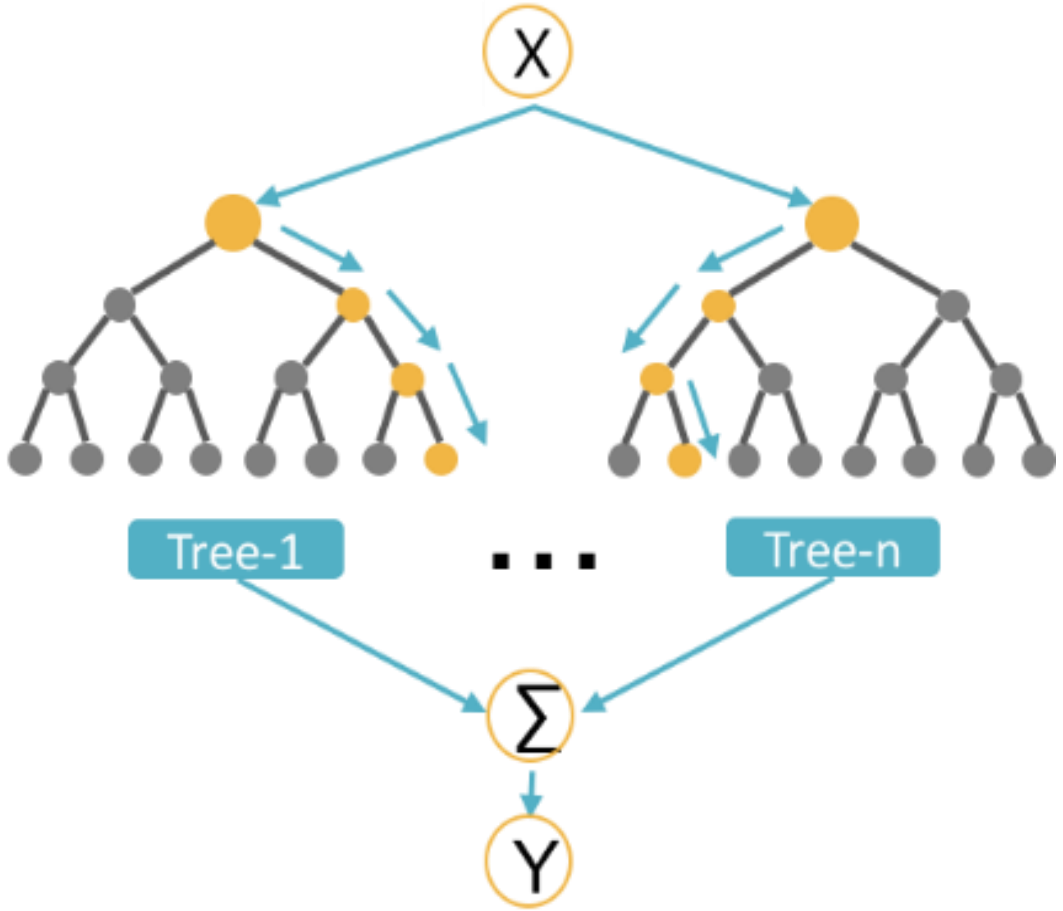
多棵決策樹一起「共同影響」投票決定

隨機森林

1. 電影「魔戒」裡，有一個讓人印象深刻的場景，會說話的樹人（Ent）向來因為固定在地面上，表面上看似軟弱無力，但是在法貢森林裡的樹人首領樹鬍，卻能率領樹人群一舉攻下了薩魯曼的半獸人要塞「艾辛格」，原來樹人是魔戒中強悍且唯一的植物兵種
2. 有趣的是，在機器學習中，由多棵「決策樹」構成的「隨機森林」分類器，也和魔戒的樹人也有異曲同工之妙，集樹成林扮演眾志成城的強大演算角色。
3. https://v.qq.com/x/page/w0348o3wsr7.html

隨機森林


年所得級距
低
中
高
存款是否超過20萬
不發卡
是否擁有信用卡
是
否
是
否
發卡
不發卡
發卡
不發卡


X
Tree-1
Tree-n
Σ
Y

繪圖者：鄭雅馨

隨機森林

1. 電影「魔戒」裡，有一個讓人印象深刻的場景，會說話的樹人（Ent）向來因為固定在地面上，表面上看似軟弱無力，但是在法貢森林裡的樹人首領樹鬍，卻能率領樹人群一舉攻下了薩魯曼的半獸人要塞「艾辛格」，原來樹人是魔戒中強悍且唯一的植物兵種
2. 有趣的是，在機器學習中，由多棵「決策樹」構成的「隨機森林」分類器，也和魔戒的樹人也有異曲同工之妙，集樹成林扮演眾志成城的強大演算角色。
3. https://v.qq.com/x/page/w0348o3wsr7.html

常用類別模型

線性迴歸

羅吉斯迴歸

決策樹

隨機森林

XGBoost

dmlc XGBoost

單一模型判斷

多棵決策樹一起「獨立」投票決定

多棵決策樹一起「共同影響」投票決定

XGBoost

1. 目前最熱門演算法之一。
2. 可以認為是目前樹狀模型最先進演算法。
3. 訓練快速、準確率又高。
4. 與RF 每棵樹獨立不同，每顆樹間會傳承經驗。

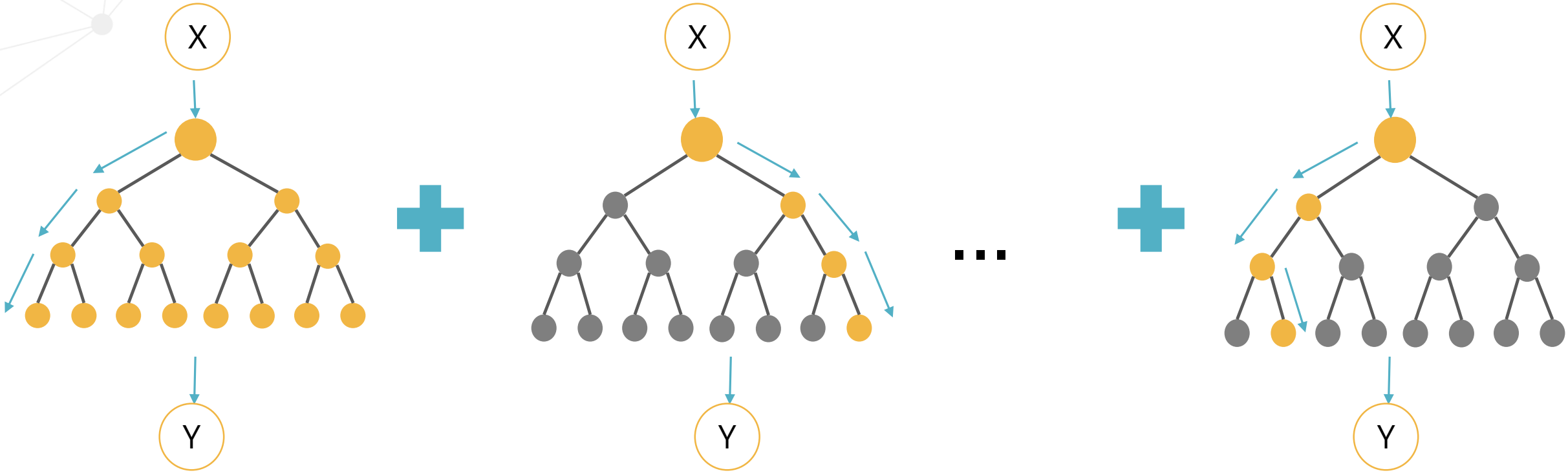
此演算法原理過度數學。

為了讓學生容易理解，些許將會與實際演算法不同。

XGBoost – 重要參數 - 圈數

1. n_estimators：
   代表疊代幾圈，通常數字大一點會比較精確。
   但過多可能訓練到某幾圈就停止精確了。

XGBoost



大綱

授課內容	備註
溫故知新：行銷資料科學數據思維	行銷資料科學的應用與課程案例demo
Python回顧小複習	For、if 與Pandas操作
商業分類模型概念探討	了解商業分類模型的最大利器為何及有哪些商業獲利的優化指標可以使用
商業分類模型介紹與實戰	介紹Logistic Regression、Decision Tree、Random Forest與XGBoost之概念與實戰，最後實作推薦清單
Day2：顧客推薦清單與模型檔案製作	知道具體應該推薦哪一位客戶；揭秘業界模型製作方法

精準行銷的步驟

1.技術面

2.案例面

繪圖者：廖庭儀、鍾皓軒

情境主題

1. 產品：A公司主打的線上服務P產品
2. 通路：100%的網路媒介
3. 價格：$ 3,500元新台幣
4. 成本：$ 1,650元新台幣
5. 行銷成本：$ 300元新台幣 / 每位顧客
6. 銷售：A公司提供30天的免費線上服務—P產品，期望期限內顧客能正式簽約購買P產品之服務
7. 資料蒐集：購買 = 1；無購買 = 0，**提供的為10,000位顧客使用第7天時購買狀況之資料集**
8. 面對難題：
   1) **不曉得到底應該使用廣告全投放還是機器學習模型來做投放?**
   2) **每隔1–2天便對數以萬計的顧客發送電子行銷文宣，不但購買率低下，甚至造成諸多客訴**

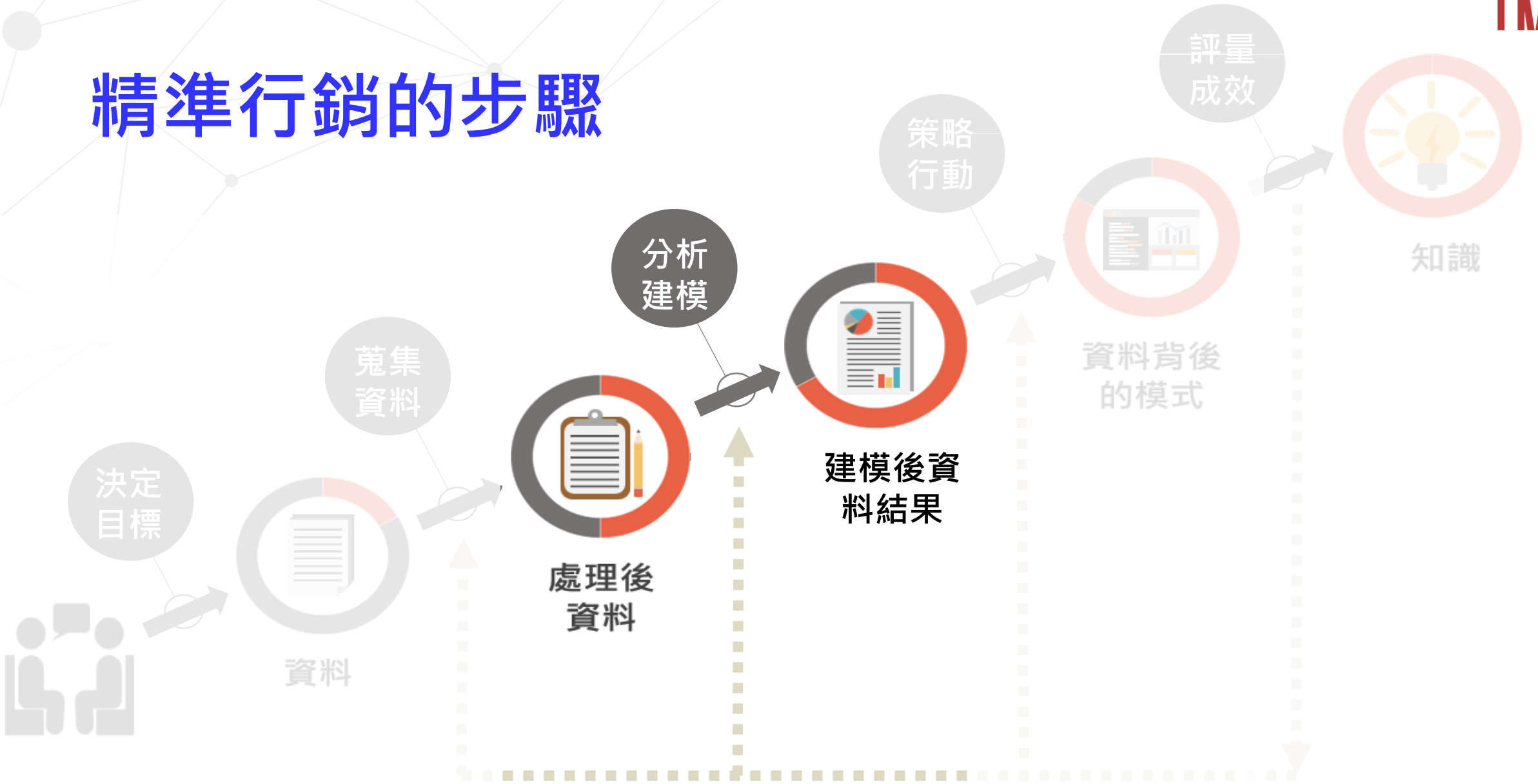
變數解釋

1. 特徵變數
   1) prod_output_num : 產品結果輸出次數
   2) locations : 使用者地區
   3) gender : 性別
   4) age : 年齡
   5) click_on_prod : 產品點擊次數
   6) balance : 點數餘額
   7) registry_to_use_time : 產品註冊到使用的時間
   8) credit_card_paid : 是否使用信用卡付月費
   9) active_member : 是否為活躍用戶
   10) estimated_salary : 估計薪資

04_A線上服務公司案例.py

2. 目標變數
   1) buy:購買與否

精準行銷的步驟



繪圖者：廖庭儀、鍾皓軒

資料處理與轉換

1. 標籤編碼（Label Encoding）：該方法將各類別轉換成數字，在商業界及社會科學中，最常見的就是李克特5點量表的轉換，這也是「順序尺度」之概念，也就是具有大小之分的類別，建議適用該法。
2. 獨熱編碼（One–Hot Encoding）：這代表將類別轉換成欄位格式，並且以1（有該類別）與0（沒有該類別）的形式表示之。

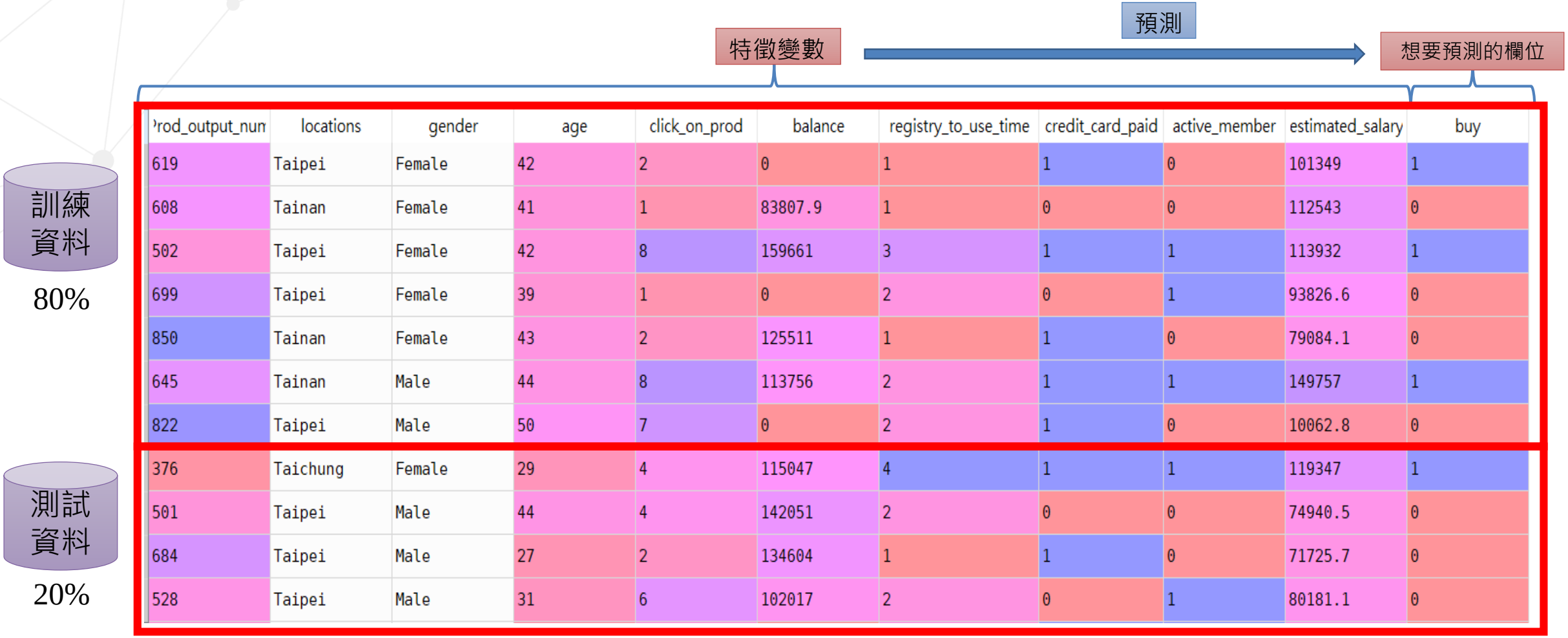
rod_output_nun	locations	gender	age	click_on_prod	balance	registry_to_use_time	credit_card_paid	active_member	estimated_salary	buy
619	Taipei	Female	42	2	0	1	1	0	101349	1
608	Tainan	Female	41	1	83807.9	1	0	0	112543	0
502	Taipei	Female	42	8	159661	3	1	1	113932	1
699	Taipei	Female	39	1	0	2	0	1	93826.6	0
850	Tainan	Female	43	2	125511	1	1	0	79084.1	0
645	Tainan	Male	44	8	113756	2	1	1	149757	1

資料處理與轉換

›rod_output_nun	locations	gender	age	click_on_prod	balance	registry_to_use_time	credit_card_paid	active_member	estimated_salary	buy
619	Taipei	Female	42	2	0	1	1	0	101349	1
608	Tainan	Female	41	1	83807.9	1	0	0	112543	0
502	Taipei	Female	42	8	159661	3	1	1	113932	1
699	Taipei	Female	39	1	0	2	0	1	93826.6	0
850	Tainan	Female	43	2	125511	1	1	0	79084.1	0
645	Tainan	Male	44	8	113756	2	1	1	149757	1

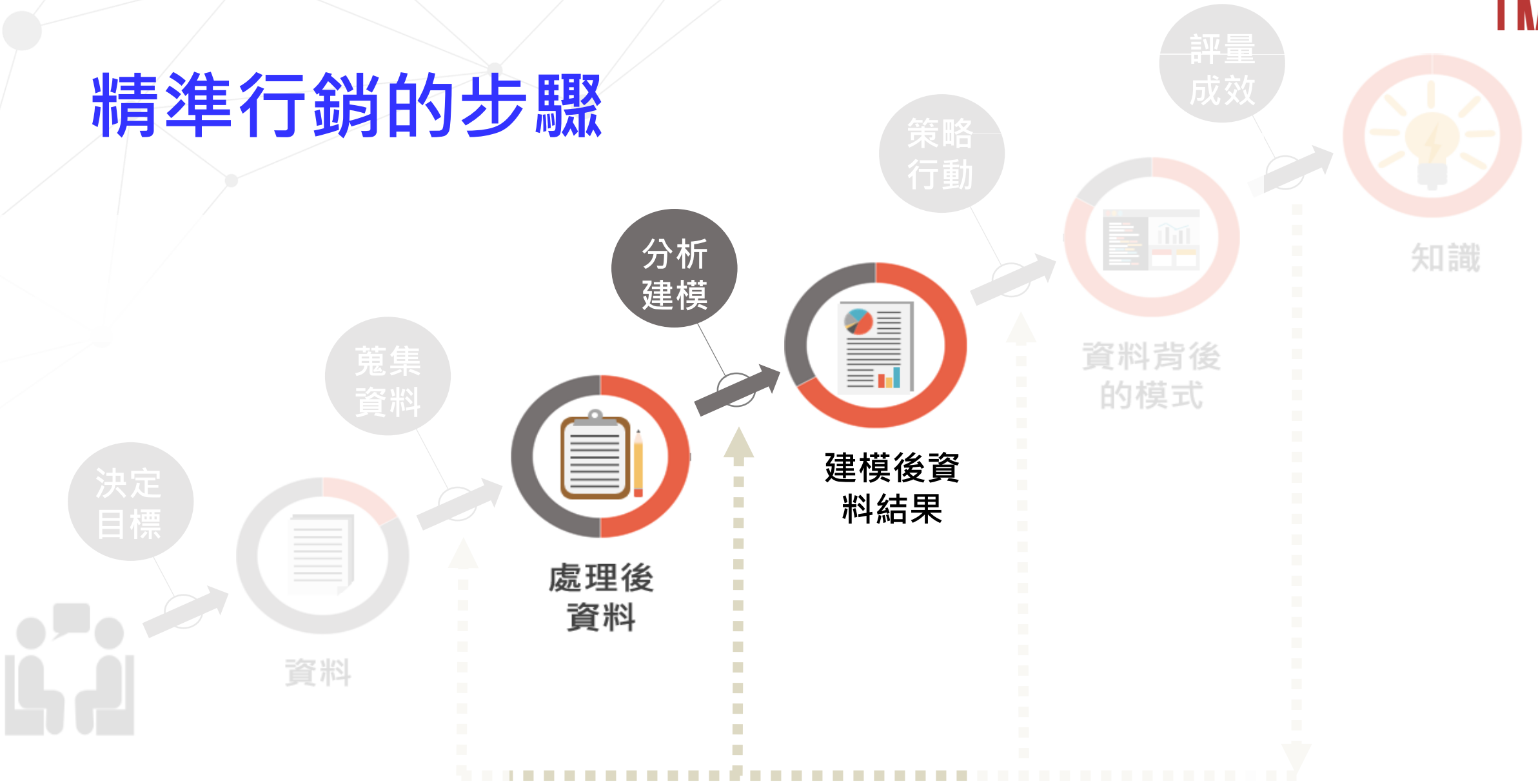
estimated_salary	buy	locations_Taichung	locations_Tainan	locations_Taipei	gender_Female	gender_Male
101349	1	0	0	1	1	0
112543	0	0	1	0	1	0
113932	1	0	0	1	1	0
93826.6	0	0	0	1	1	0
79084.1	0	0	1	0	1	0

機器學習概論
分類式的監督式學習



rod_output_nun	locations	gender	age	click_on_prod	balance	registry_to_use_time	credit_card_paid	active_member	estimated_salary	buy
619	Taipei	Female	42	2	0	1	1	0	101349	1
608	Tainan	Female	41	1	83807.9	1	0	0	112543	0
502	Taipei	Female	42	8	159661	3	1	1	113932	1
699	Taipei	Female	39	1	0	2	0	1	93826.6	0
850	Tainan	Female	43	2	125511	1	1	0	79084.1	0
645	Tainan	Male	44	8	113756	2	1	1	149757	1
822	Taipei	Male	50	7	0	2	1	0	10062.8	0
376	Taichung	Female	29	4	115047	4	1	1	119347	1
501	Taipei	Male	44	4	142051	2	0	0	74940.5	0
684	Taipei	Male	27	2	134604	1	1	0	71725.7	0
528	Taipei	Male	31	6	102017	2	0	1	80181.1	0

精準行銷的步驟



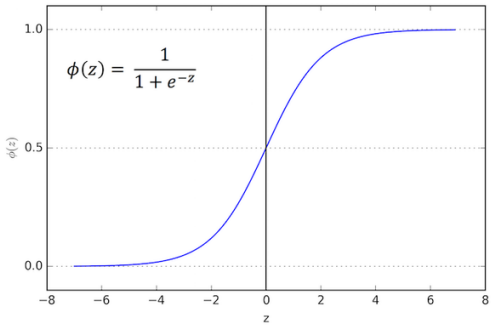
繪圖者：廖庭儀、鍾皓軒

羅吉斯迴歸
A公司的P產品目標市場預測結果

80%

訓練資料

	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	loan
19	1	1	42	2	0.00	1	1	1	101348.88	1
02	1	1	42	8	159660.80	3	1	0	113931.57	1
50	2	1	43	2	125510.82	1	1	1	79084.10	0
645	2	2	44	8	113755.78	2	1	0	149756.71	1
822	1	2	50	7	0.00	2	1	1	10062.80	0
501	1	2	44	4	142051.07	2	0	1	74940.50	0
684	1	2	27	2	134603.88	1	1	1	71725.73	0
528	1	2	31	6	102016.72	2	0	0	80181.12	0
635	2	1	35	7	0.00	2	1	1	65951.65	0
616	3	2	45	3	143129.41	2	0	1	64327.26	0
549	2	1	24	9	0.00	2	1	1	14406.41	0
587	2	2	45	6	0.00	1	0	0	158684.81	0
726	1	1	24	6	0.00	2	1	1	54724.03	0
636	2	1	32	8	0.00	2	1	0	138555.46	0
510	2	1	38	4	0.00	1	1	0	118913.53	1
846	1	1	38	5	0.00	1	1	1	187616.16	0
577	1	2	25	3	0.00	2	0	1	124508.29	0
756	3	2	36	2	136815.64	1	1	1	170041.95	0
571	1	2	44	9	0.00	2	0	0	38433.35	0
574	3	1	43	3	141349.43	1	1	1	100187.43	0
411	1	2	29	0	59697.17	2	1	1	53483.21	0
591	2	1	39	3	0.00	3	1	0	140469.38	1
553	3	2	41	9	110112.54	2	0	0	81898.81	0
722	2	1	29	9	0.00	2	1	1	142033.07	0
490	2	2	31	3	145260.23	1	0	1	114066.77	0
804	2	2	33	7	76548.60	1	0	1	98453.45	0
850	1	2	36	7	0.00	1	1	1	40812.90	0
465	1	1	51	8	122522.32	1	0	0	181297.65	1
556	1	1	61	2	117419.35	1	1	1	94153.83	0
834	1	1	49	2	131394.56	1	0	0	194365.76	1



羅吉斯迴歸

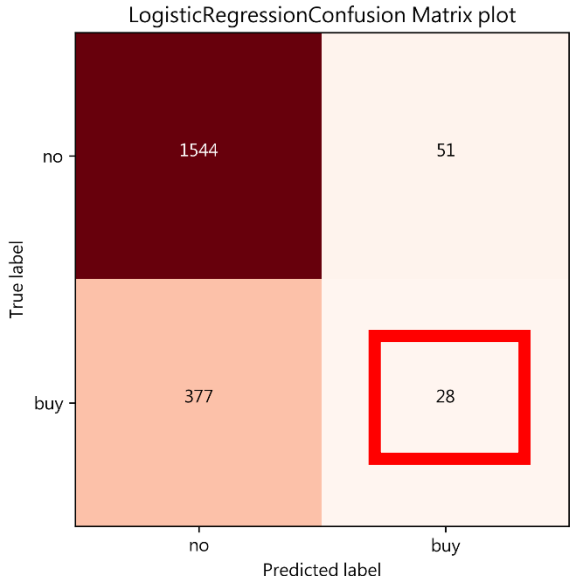
預測結果

loan	y_pred
0	0
0	0
1	1
0	0
0	0
0	0
1	1
0	0
0	0
0	0
1	0
0	0
0	0
0	0
1	0
0	0
0	0
0	0
1	1

20%

測試資料

Test Accuracy = 0.811

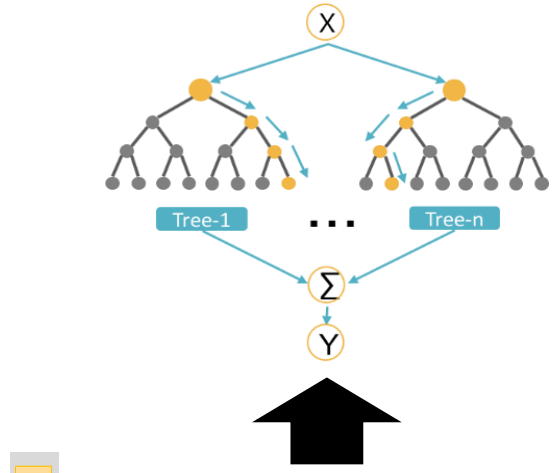


隨機森林
A公司的P產品目標市場預測結果

80%

訓練資料

	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	loan
19	1	1	42	2	0.00	1	1	1	101348.88	1
02	1	1	42	8	159660.80	3	1	0	113931.57	1
50	2	1	43	2	125510.82	1	1	1	79084.10	0
6	645	2	2	44	8	113755.78	2	1	0	149756.71	1
7	822	1	2	50	7	0.00	2	1	1	10062.80	0
9	501	1	2	44	4	142051.07	2	0	1	74940.50	0
10	684	1	2	27	2	134603.88	1	1	1	71725.73	0
11	528	1	2	31	6	102016.72	2	0	0	80181.12	0
15	635	2	1	35	7	0.00	2	1	1	65951.65	0
16	616	3	2	45	3	143129.41	2	0	1	64327.26	0
18	549	2	1	24	9	0.00	2	1	1	14406.41	0
19	587	2	2	45	6	0.00	1	0	0	158684.81	0
20	726	1	1	24	6	0.00	2	1	1	54724.03	0
22	636	2	1	32	8	0.00	2	1	0	138555.46	0
23	510	2	1	38	4	0.00	1	1	0	118913.53	1
25	846	1	1	38	5	0.00	1	1	1	187616.16	0
26	577	1	2	25	3	0.00	2	0	1	124508.29	0
27	756	3	2	36	2	136815.64	1	1	1	170041.95	0
28	571	1	2	44	9	0.00	2	0	0	38433.35	0
29	574	3	1	43	3	141349.43	1	1	1	100187.43	0
30	411	1	2	29	0	59697.17	2	1	1	53483.21	0
31	591	2	1	39	3	0.00	3	1	0	140469.38	1
33	553	3	2	41	9	110112.54	2	0	0	81898.81	0
35	722	2	1	29	9	0.00	2	1	1	142033.07	0
37	490	2	2	31	3	145260.23	1	0	1	114066.77	0
38	804	2	2	33	7	76548.60	1	0	1	98453.45	0
39	850	1	2	36	7	0.00	1	1	1	40812.90	0
42	465	1	1	51	8	122522.32	1	0	0	181297.65	1
43	556	1	1	61	2	117419.35	1	1	1	94153.83	0
44	834	1	1	49	2	131394.56	1	0	0	194365.76	1



隨機森林

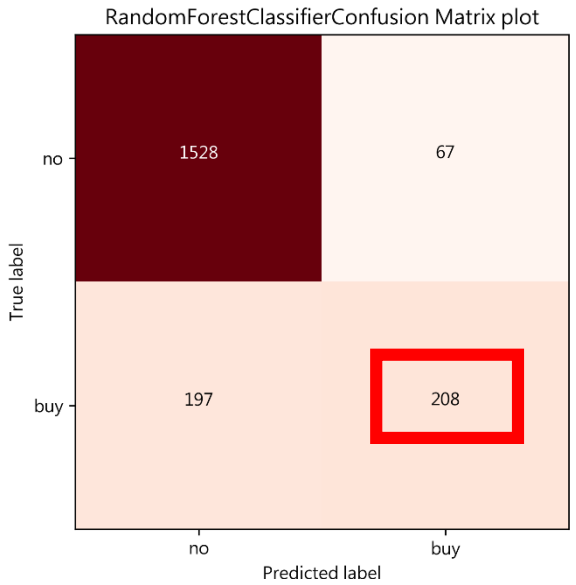
預測結果

loan	y_pred
0	0
0	0
1	1
0	0
0	0
0	0
1	1
0	0
0	0
0	0
1	0
0	0
0	0
0	0
1	0
0	0
0	0
1	1

20%

測試資料

Test Accuracy = 0.868



隨機森林
A公司的P產品目標市場預測結果

80%

訓練資料

	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	loan
19	1	1	42	2	0.00	1	1	1	101348.88	1
02	1	1	42	8	159660.80	3	1	0	113931.57	1
50	2	1	43	2	125510.82	1	1	1	79084.10	0
645	2	2	44	8	113755.78	2	1	0	149756.71	1
822	1	2	50	7	0.00	2	1	1	10062.80	0
501	1	2	44	4	142051.07	2	0	1	74940.50	0
684	1	2	27	2	134603.88	1	1	1	71725.73	0
528	1	2	31	6	102016.72	2	0	0	80181.12	0
635	2	1	35	7	0.00	2	1	1	65951.65	0
616	3	2	45	3	143129.41	2	0	1	64327.26	0
549	2	1	24	9	0.00	2	1	1	14406.41	0
587	2	2	45	6	0.00	1	0	0	158684.81	0
726	1	1	24	6	0.00	2	1	1	54724.03	0
636	2	1	32	8	0.00	2	1	0	138555.46	0
510	2	1	38	4	0.00	1	1	0	118913.53	1
846	1	1	38	5	0.00	1	1	1	187616.16	0
577	1	2	25	3	0.00	2	0	1	124508.29	0
756	3	2	36	2	136815.64	1	1	1	170041.95	0
571	1	2	44	9	0.00	2	0	0	38433.35	0
574	3	1	43	3	141349.43	1	1	1	100187.43	0
411	1	2	29	0	59697.17	2	1	1	53483.21	0
591	2	1	39	3	0.00	3	1	0	140469.38	1
553	3	2	41	9	110112.54	2	0	0	81898.81	0
722	2	1	29	9	0.00	2	1	1	142033.07	0
490	2	2	31	3	145260.23	1	0	1	114066.77	0
804	2	2	33	7	76548.60	1	0	1	98453.45	0
850	1	2	36	7	0.00	1	1	1	40812.90	0
465	1	1	51	8	122522.32	1	0	0	181297.65	1
556	1	1	61	2	117419.35	1	1	1	94153.83	0
834	1	1	49	2	131394.56	1	0	0	194365.76	1

dmlc XGBoost

XGBoost

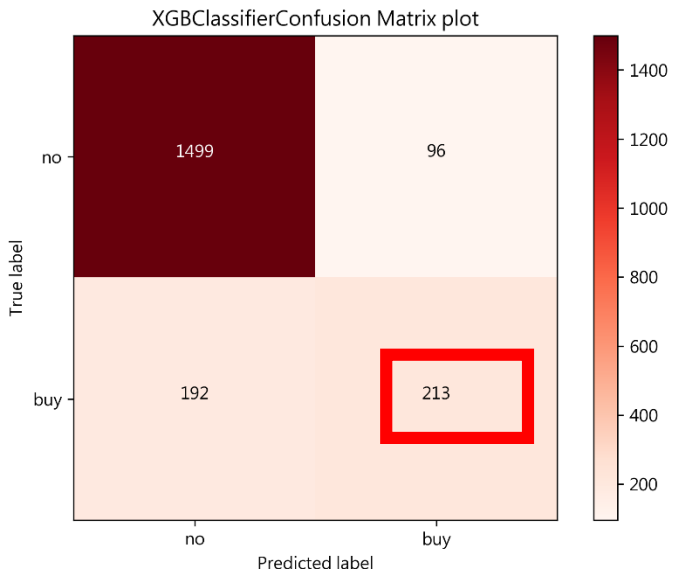
預測結果

loan	y_pred
0	0
0	0
1	1
0	0
0	0
0	0
1	1
0	0
0	0
0	0
0	0
1	0
0	0
0	0
0	0
1	0
0	0
0	0
0	0
1	1

20%

測試資料

Test Accuracy = 0.856

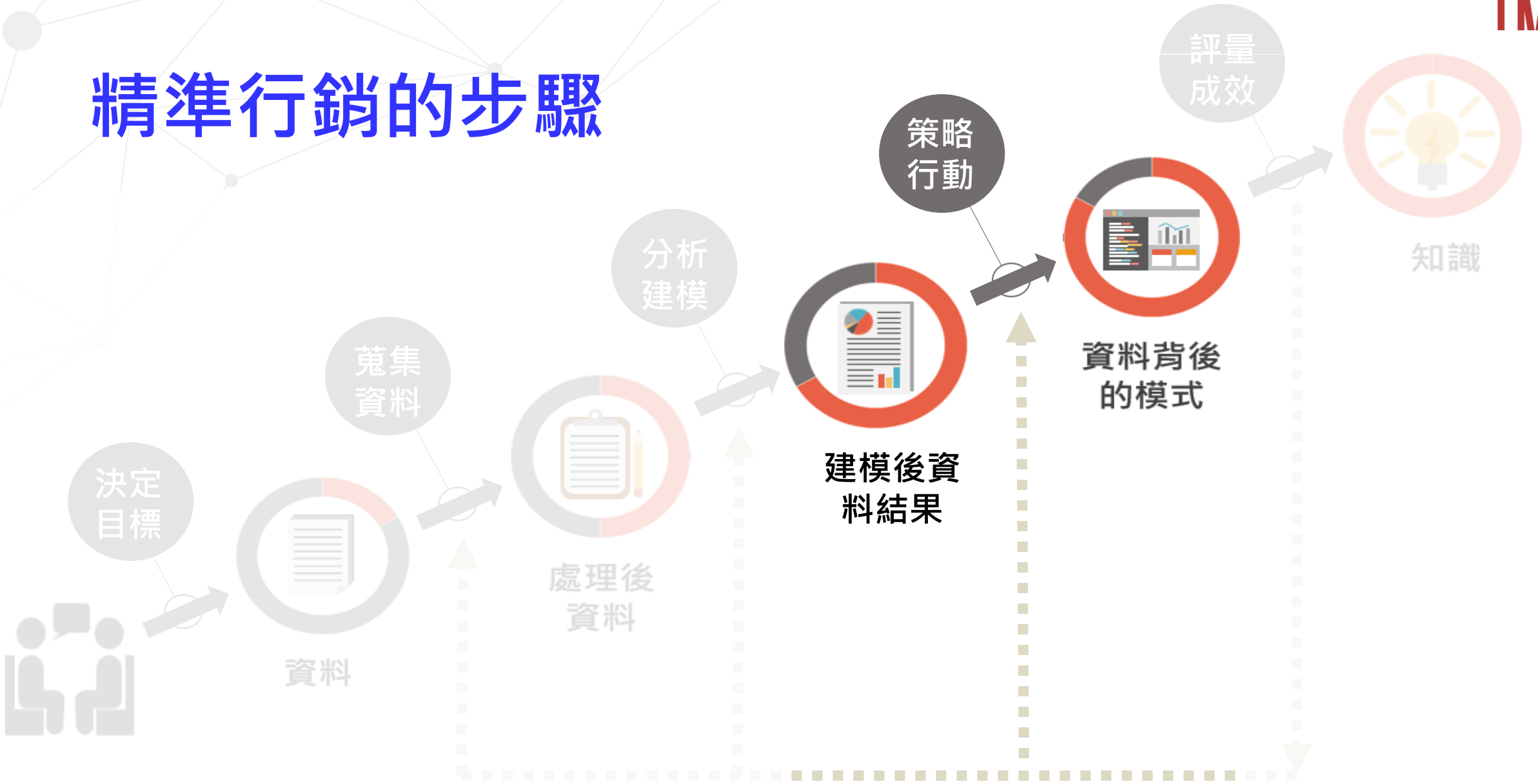


問題：

預測出來了，感覺很厲害!

但… So what? 真的有【效果】?!

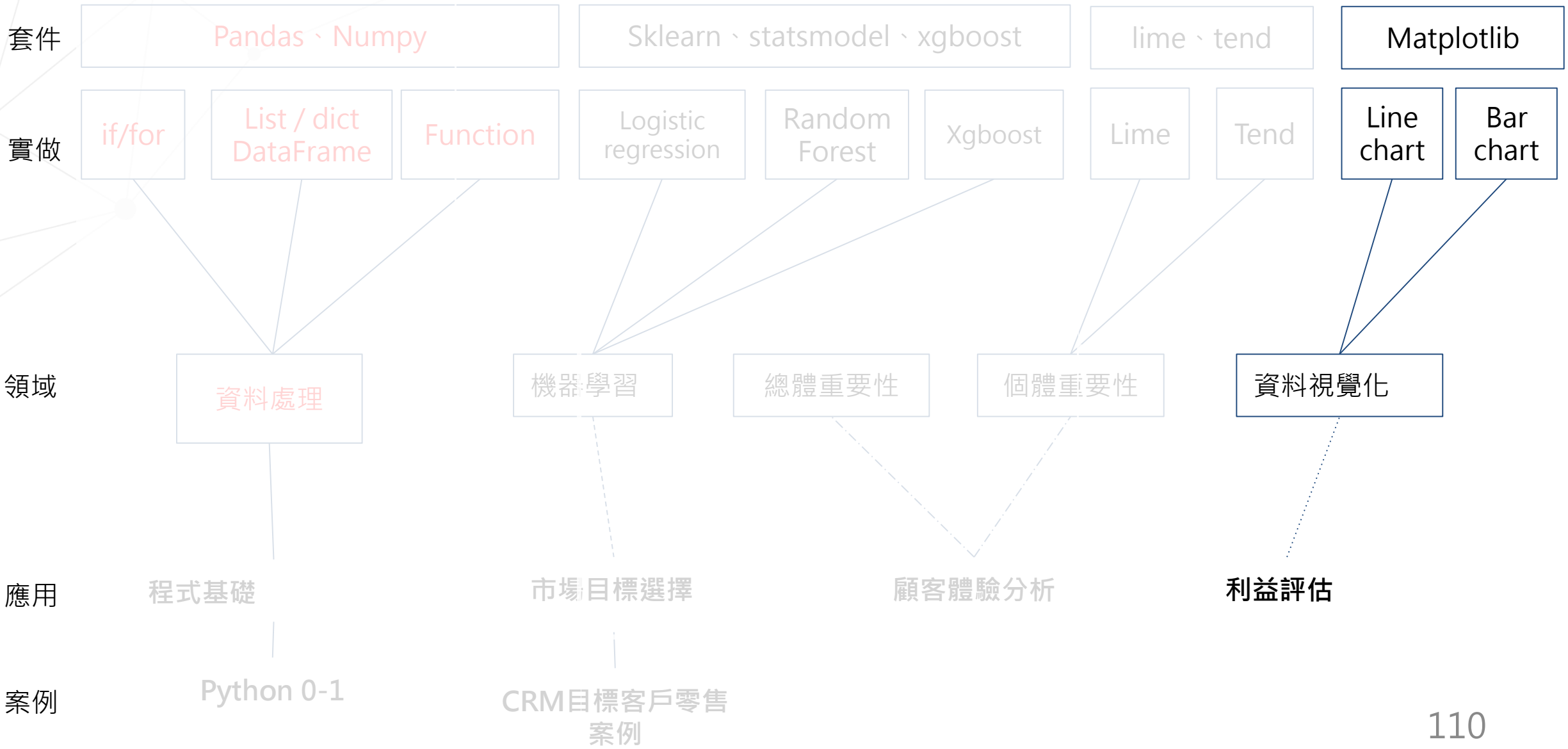
精準行銷的步驟



繪圖者：廖庭儀、鍾皓軒

利潤評估模型

Python機器學商務實戰 – 學習地圖



機器學習實戰

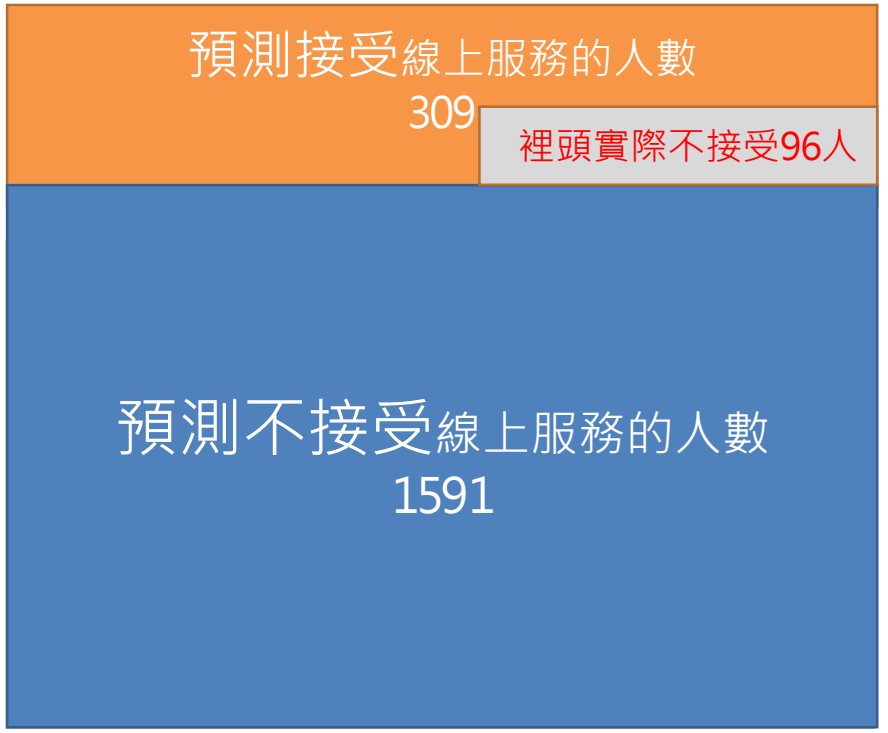
A公司目標市場的實際狀況與預測結果比較

思考時間：請問全市場的利潤與預測市場的利潤？

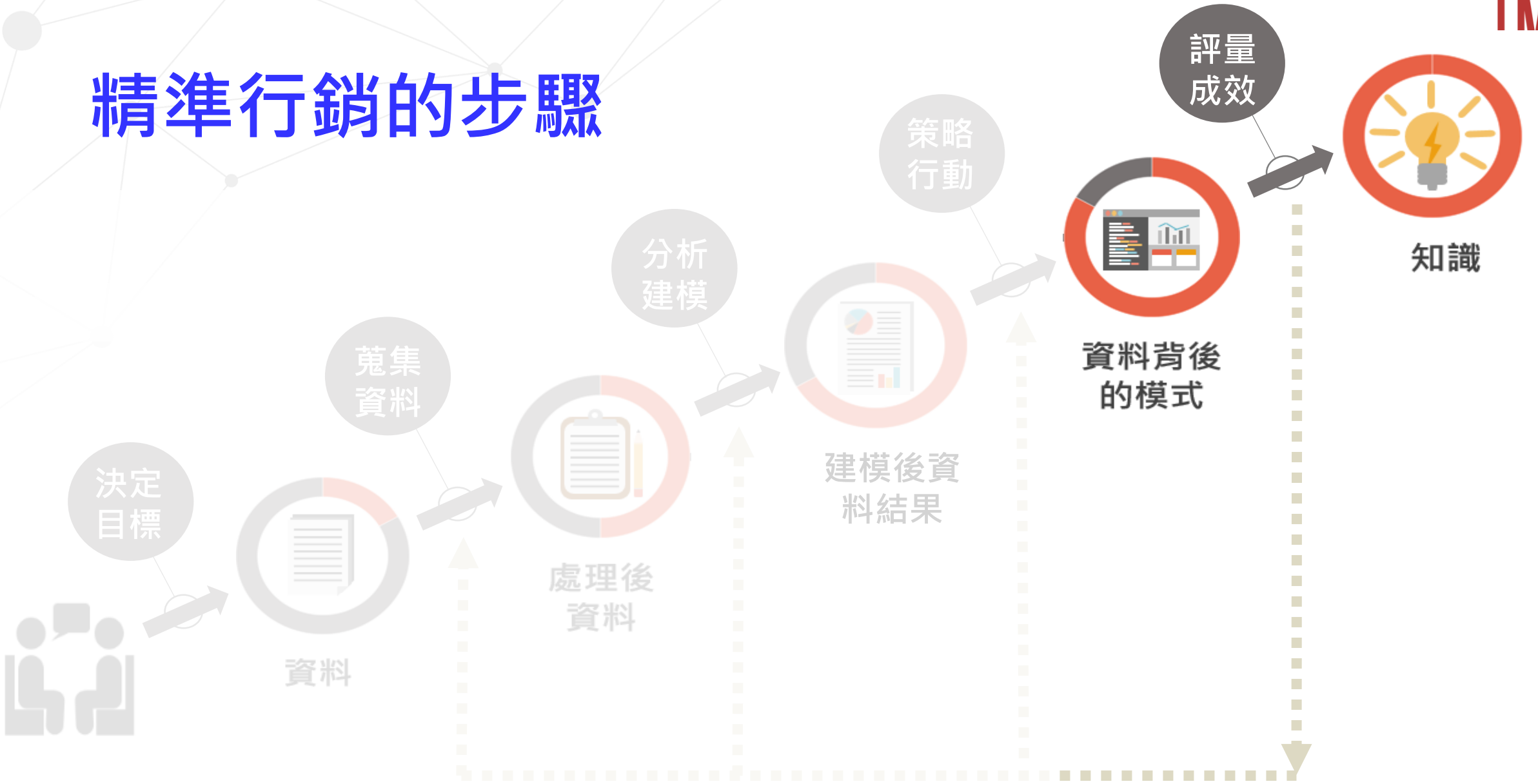
測試資料

20%
2000人

項目	金額
單品價格	$3,500
單品營業成本	$1,650
單品行銷費用	$300



精準行銷的步驟



繪圖者：廖庭儀、鍾皓軒

謝謝！敬請指教！