Python 機器學習

廖家緯 Jiawei

Ph.D. Candidate in Computer Science
National Taiwan University

jwliao1209

Who am I?

Website

BS in Math

MS in Applied Math

5th NTU DAC

PhD Candidate in CSIE

AAAI 2025

2020 — Math Tutor

2022 — Research Intern

2023 — DA Intern

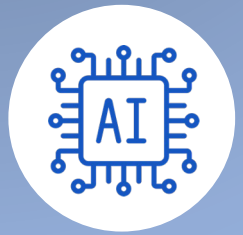
2024 — Research Assistant

2025 — AI Research Intern; SWE Intern; ML Research Intern

AI 逐漸普及化: 從判斷式到生成式

判斷式 AI

生成式 AI



人工智慧

1950s

機器學習

1980s

深度學習

2010s

生成式 AI

2020s

通用人工智慧

?

1. 對話機器人
2. 虛擬試穿 / 上妝
3. 圖片編輯
4. 音樂生成

不同行業如何用 AI 解決問題

[App] 哪些使用者會持續訂閱?

[電動車] 該在哪裡設點?

[資安] 如何判斷釣魚郵件?

[電商] 消費者對哪些產品有興趣?

[代言] 該找哪些網紅合作?

有了 DL / LLM 為什麼還要學 ML?

大多數電商、金融業依然以表格型資料為主

Why do tree-based models still outperform deep learning on typical tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
MLIA, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

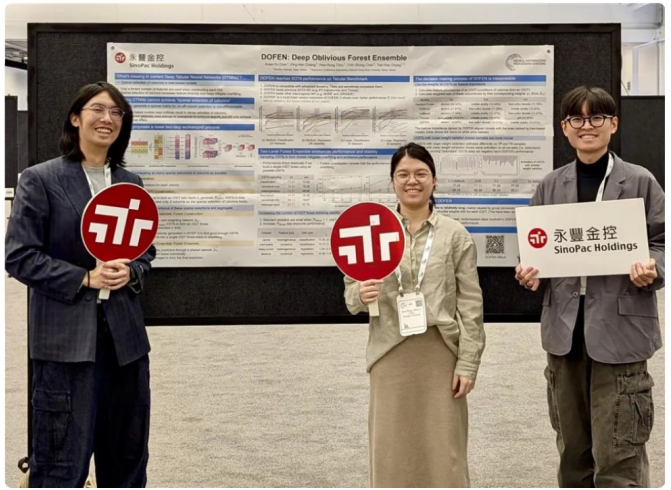
Abstract

While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data (~10K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: **1.** be robust to uninformative features, **2.** preserve the orientation of the data, and **3.** be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner.

永豐金AI科學家團隊 研發實力再度躍上國際舞台

工商時報 黃于庭
2024年12月24日

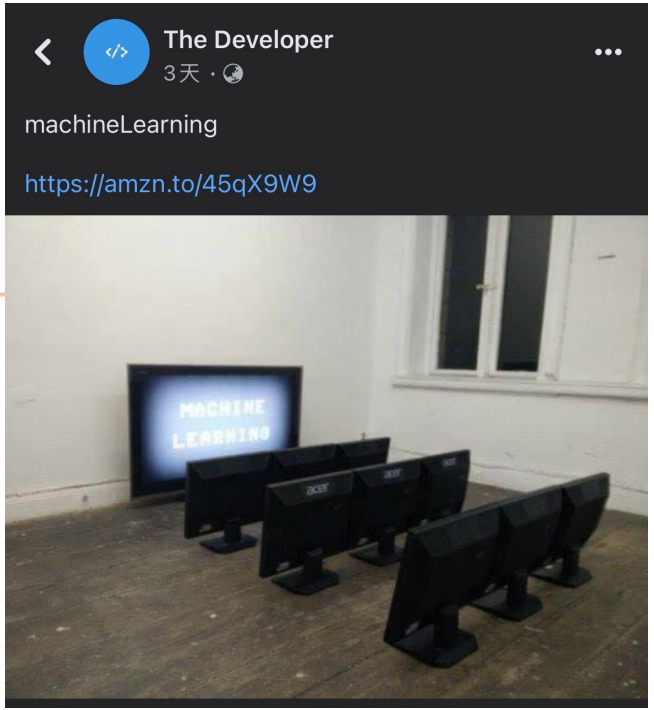
NEURAL INFORMATION PROCESSING SYSTEMS



永豐金控科技長張天豪領導的人工智慧資料科學家團隊，最新研究論文被知名AI國際論壇「神經資訊處理系統大會」所接受，於12 / 8～12 / 14至加拿大與國際頂尖研究團隊交流分享。圖 / 永豐金控提供

認識機器學習

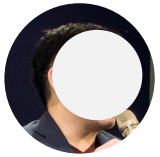
為什麼現在人人都要會 AI?



課程目標

1. 了解機器學習方法能解決什麼樣的問題
   - 定義問題 (適當的特徵 X 與標籤 y)
2. 掌握機器學習的完整流程
3. 熟悉經典機器學習模型
   - 包含使用時機、輸入、輸出、模型特性、優缺點比較
   - 知道如何透過參數調整來優化模型
4. 了解如何評估與分析模型
5. 培養運用機器學習解決企業問題的能力
   - 使用 scikit-learn 套件完整實作機器學習流程

機器學習是什麼?



網紅 CEO

大數據 + 大模型 !!

電資學院教授

尋找一個 **function (函式)**，當我們輸入一筆資料，這個 function 可以產出一個我們期望的結果。

來看看 AI 自己怎麼說

來看看 AI 自己怎麼說

什麼是機器學習?

機器學習（Machine Learning, ML）是一種人工智慧（AI）技術，它讓電腦能夠 **從數據中學習模式**，並在沒有明確指令的情況下做出預測或決策。機器學習的核心概念是讓電腦「學習」過去的經驗，以便在未來遇到類似情況時能夠 **自動適應**。…

機器學習 (Machine Learning, ML) 是人工智慧的一個分支，旨在 **透過資料訓練模型**，使電腦能夠從經驗中學習並做出 **預測** 或**決策** ，而無需明確程式設計。…

Question: 統計 vs 機器學習?　　統計用來找相關性、趨勢、因果關係，機器學習用來預測

學習機器學習三階段

第一階段

1. 了解模型的作用, input, output 是什麼
2. 了解重要超參數的意義
3. 了解模型的優缺點及限制
4. 了解將問題轉化為機器學習可解的形式

▼

第二階段

1. 能根據經驗、領域知識調整超參數，提昇模型效能
2. 能根據問題，修改機器學習模型

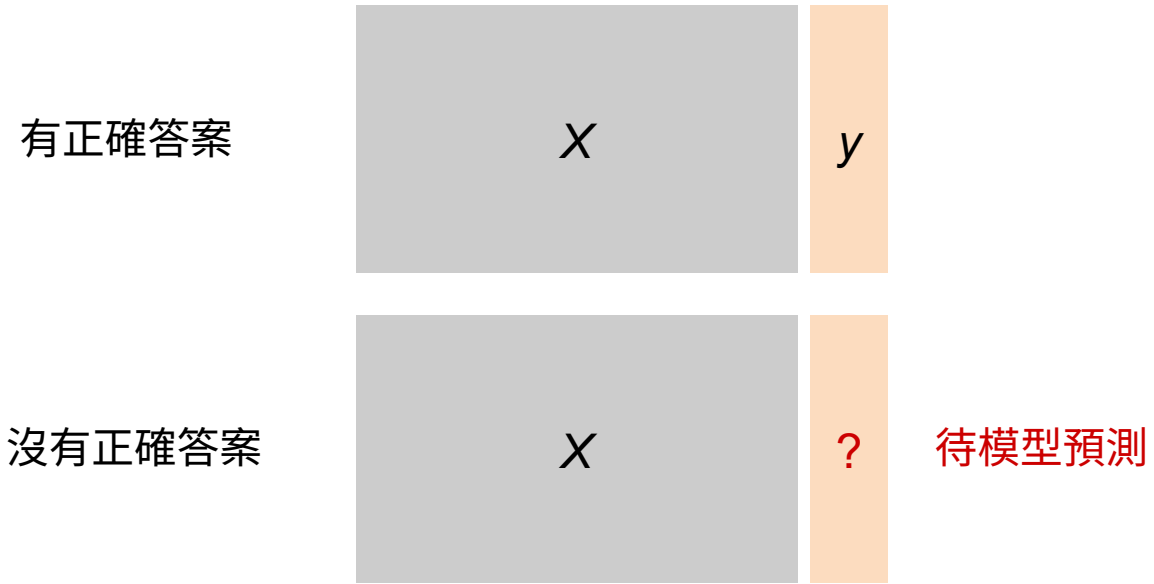
▼

第三階段

1. 設計新的機器學習模型或修改底層程式碼

機器學習模型在做什麼

定義問題的 **特徵 (feature) *X*** 和 **標籤 (label) *y***



機器學習可以依照任務屬性可以分為

監督式學習	有標籤的資料	• 分類: 客戶流失預測 • 回歸: 房價預測
非監督式學習	沒有標籤的資料	• 分群: 將顧客分成高價值/低價值/潛力群體 • 降維: 將高維顧客行為用二維視覺化
半監督式學習	部分資料有標籤	• 分類: 醫學影像 (少量資料) 疾病分類
強化學習	從環境中學習策略	• 控制: 自動駕駛 • 模仿: 拉麵機器人

機器學習專案流程

定義問題	模型的 input, output, 問題指標
處理資料	重複值移除、缺失值填補、特徵轉換
建立模型	SVM, 決策樹, XGBoost, DNN, ...
評估模型好壞	透過訂定的指標進行線下、線上評測; A/B testing
解釋模型現象	Feature Importance, SHAP
產出分析報告	簡報、故事線

SVM 分類模型

一條直線劃分楚河漢界

機器學習分類模型

準備資料

`(X, y)`

模型訓練

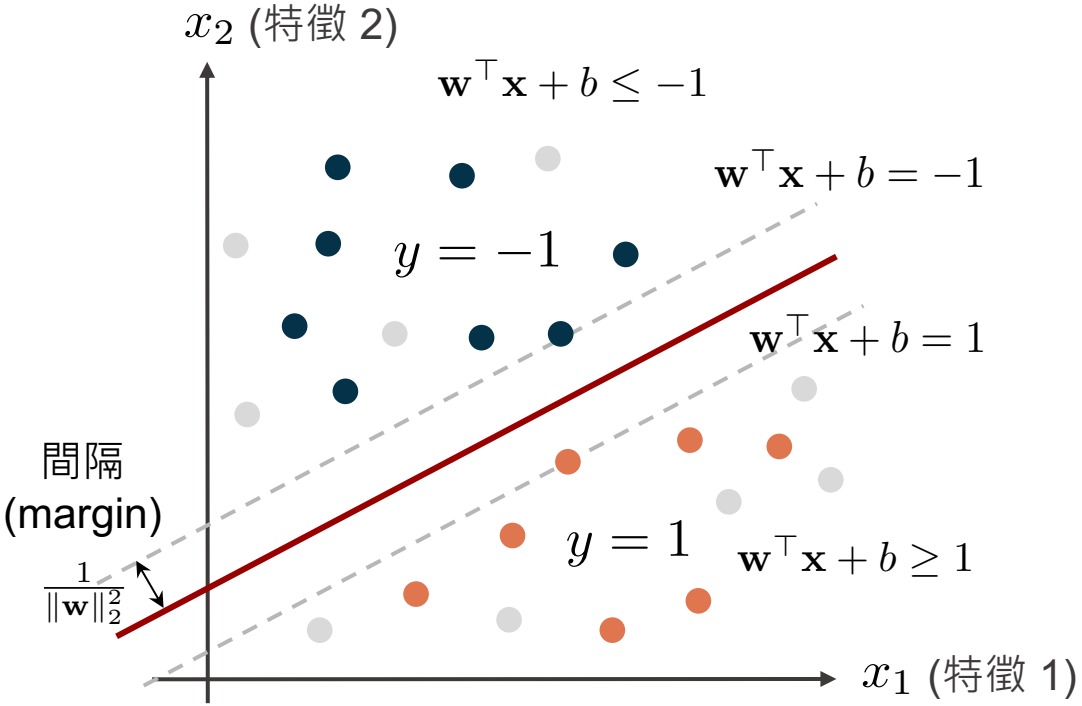
`model.fit(X, y)`

輸出

`model.predict(X)`

線性 (Linear) SVM

目標: 找到一條 **直線** 將兩類資料點分開



資料: $(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(N)}, y^{(N)})$

$\mathbf{x} = [x_1\ x_2]^\top$

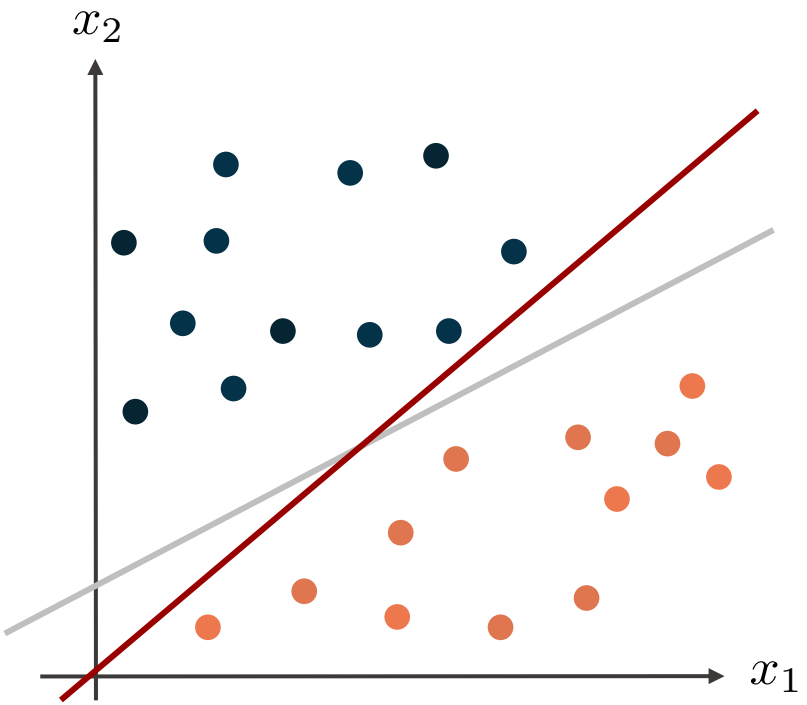
模型參數

$$y = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$$

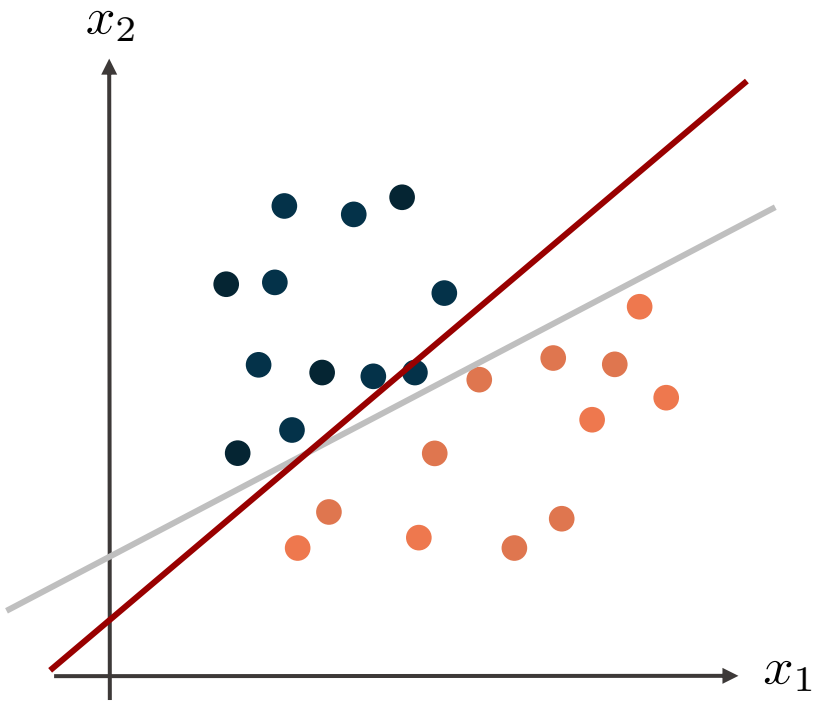
$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|_2^2}$$
$$\text{s.t. } y^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1, \forall i$$

SVM 的陷阱: 不做數據標準化模型很容易學爛掉

有做標準化把數據拉開



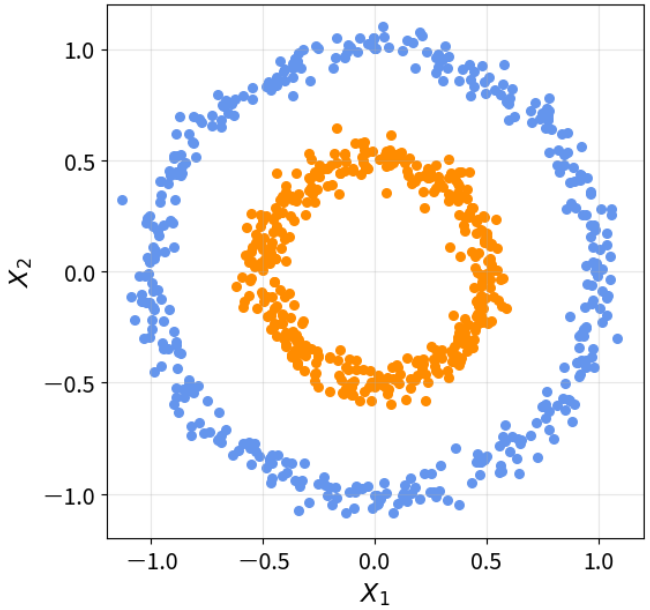
沒做標準化線稍微擾動很容易分錯



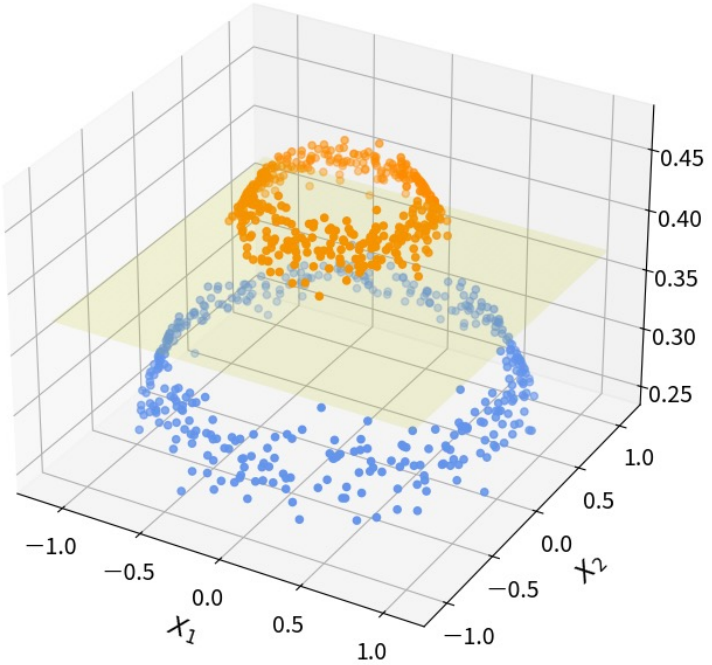
如果資料的決策邊界不是一條直線怎麼辦?

Kernel SVM: 把資料轉換至高維度讓資料變成線性可分

低維度無法解決的事情就到高維度解決



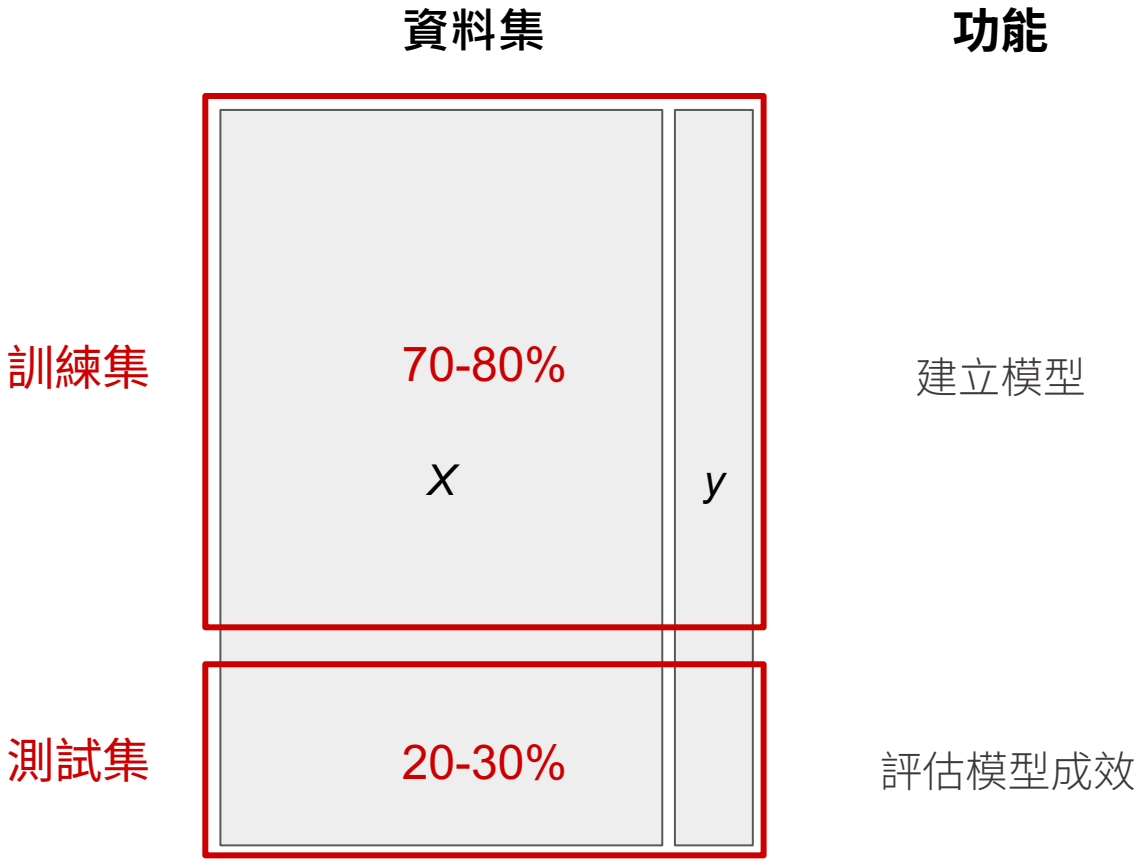
升到高維度



分割資料集: 訓練、測試

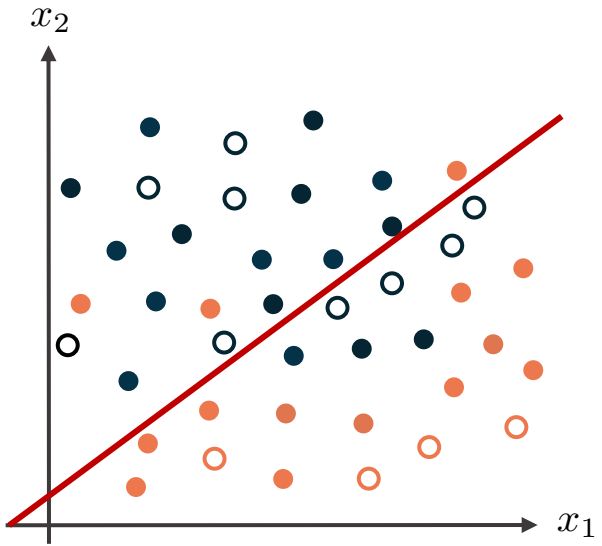
大家來分工，有人負責建模型，有人負責測試

分割資料集



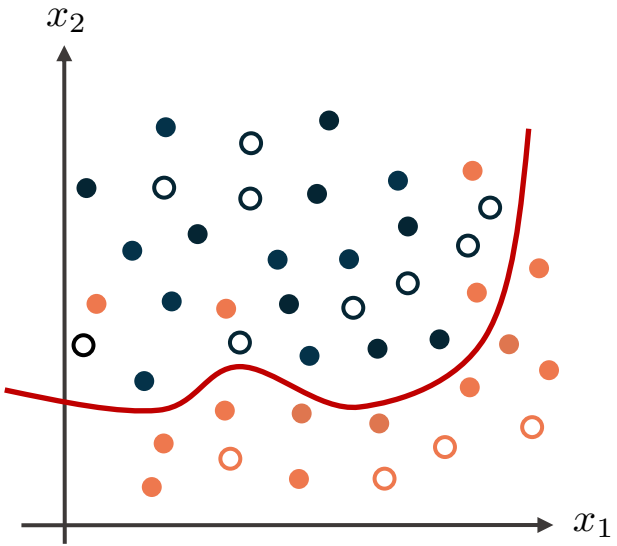
為什麼模型學不好?

低度擬合 (Underfitting)
訓練集準確率低

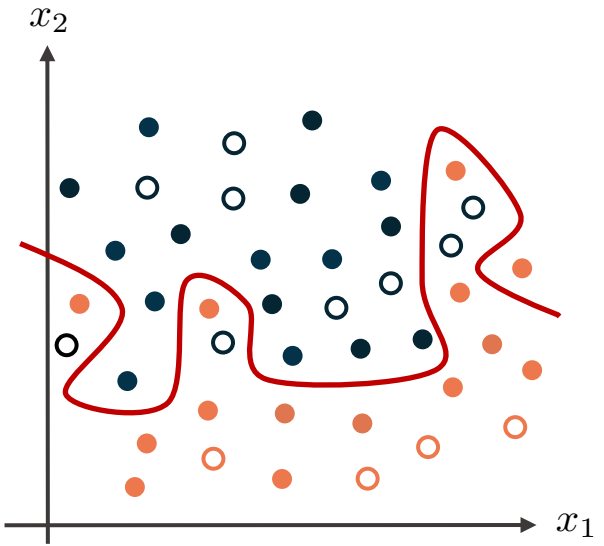


模型複雜度太低

好的模型



過度擬合 (Overfitting)
測試集準確率顯著比訓練集低



模型複雜度太高

過度討好訓練集的下場
會讓測試集的準確度很難看

訓練集 測試集

特徵工程

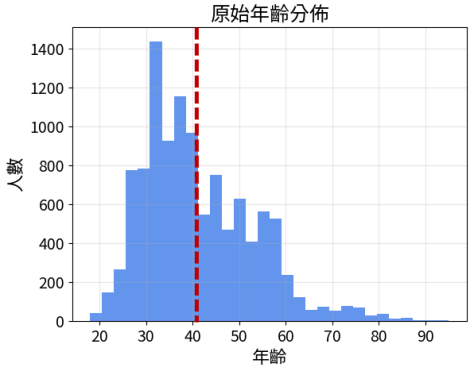
把資料特徵變成模型喜歡的樣子，提升模型準確度

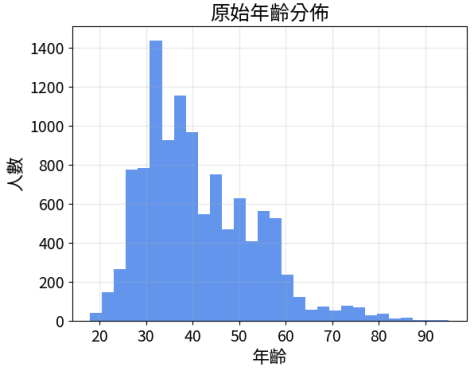
資料清理

- 移除重複值
- 填補缺失值
  - 填補統計值 e.g. 平均數, 中位數, 眾數
  - 填補出現次數最多的類別
  - 填補特殊符號: “None”
  - 使用統計或機器學習方法填補 e.g. KNN

特徵縮放: 線性變換

轉換前





標準化

`StandardScaler`

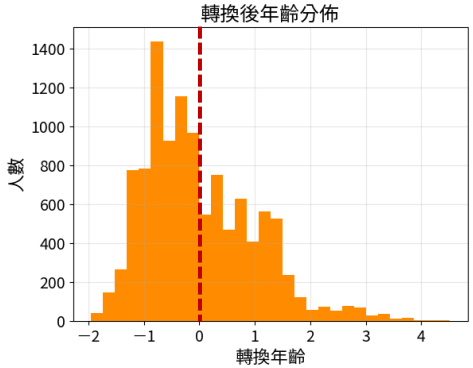
$$\frac{X - \mu}{\sigma}$$

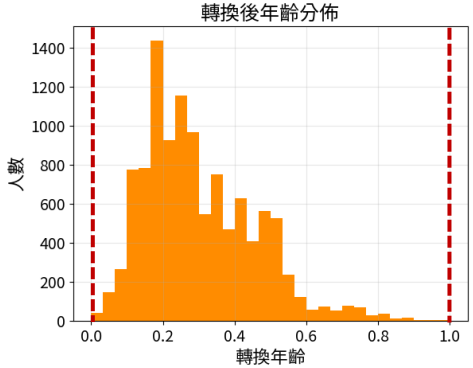
最大最小正規化

`MinMaxScaler`

$$\frac{X - \min}{\max - \min}$$

轉換後





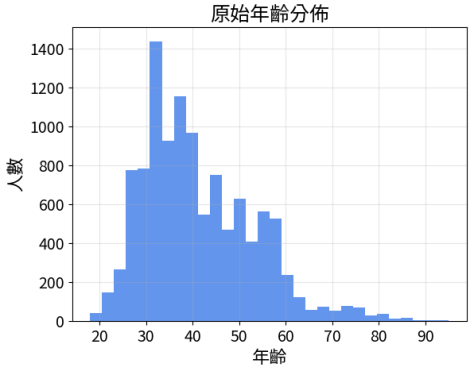
效果

1. 適用於特徵分布接近**常態分布**
2. 轉換後分布**平均為 0**，**標準差為 1**
3. 轉換後特徵**無特定範圍**

1. 適用於任何特徵分布
2. 轉換後保持原分布性質
3. 轉換後特徵範圍為 **[0, 1]**

特徵縮放: 非線性變換

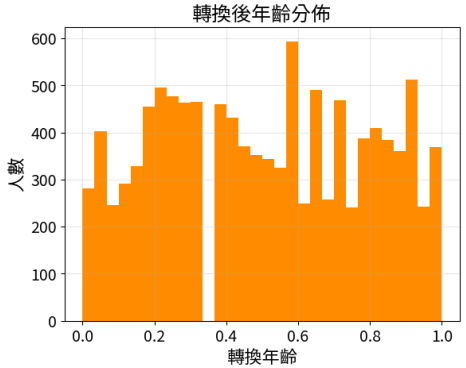
轉換前



分位數轉換

`QuantileTransformer`

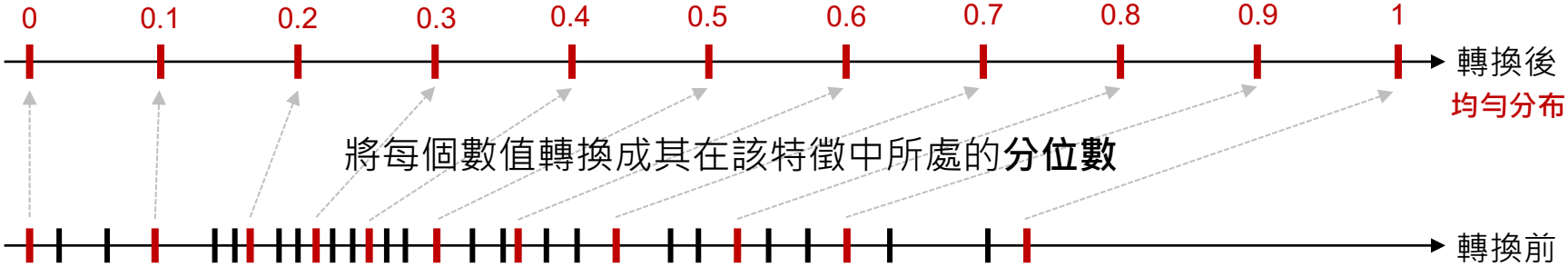
轉換後



效果

1. 適用於任何特徵分布
2. 可將分布轉換為**均勻分布**或**常態分布**
3. 轉換後會**破壞特徵間的相關性**

原理

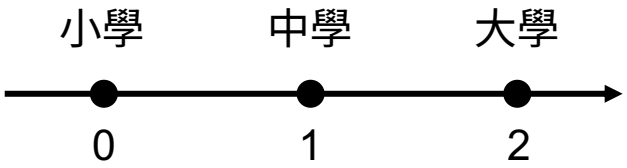


類別變數轉換

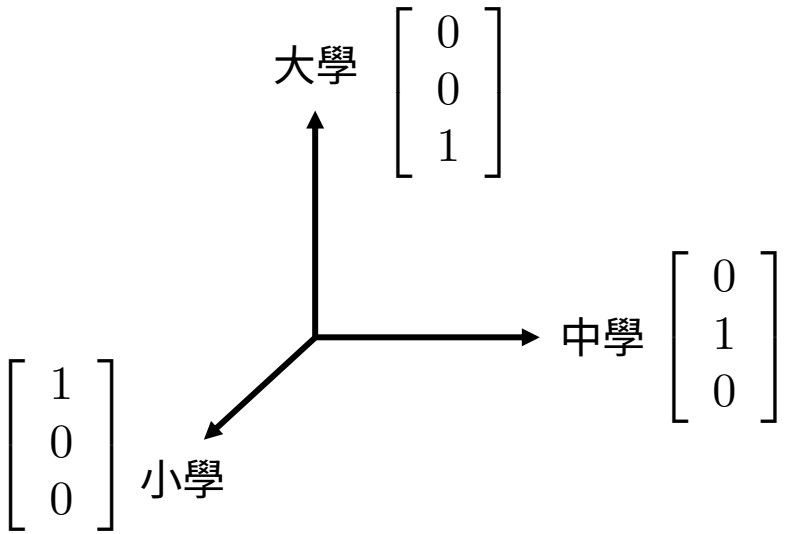
教育程度類別: {小學、中學、大學}

Label Encoding

`LabelEncoder`



One-hot Encoding



類別變數轉換

教育程度類別: {小學、中學、大學}

Q: 什麼時候不適合用 Target Encoding

Target Encoding

教育程度	標籤 y	
小學	0	⟶ mean = 0
中學	1	⟶ mean = 0.5
中學	0	
大學	1	
大學	1	⟶ mean = 0.75
大學	1	
大學	0	

Frequency Encoding

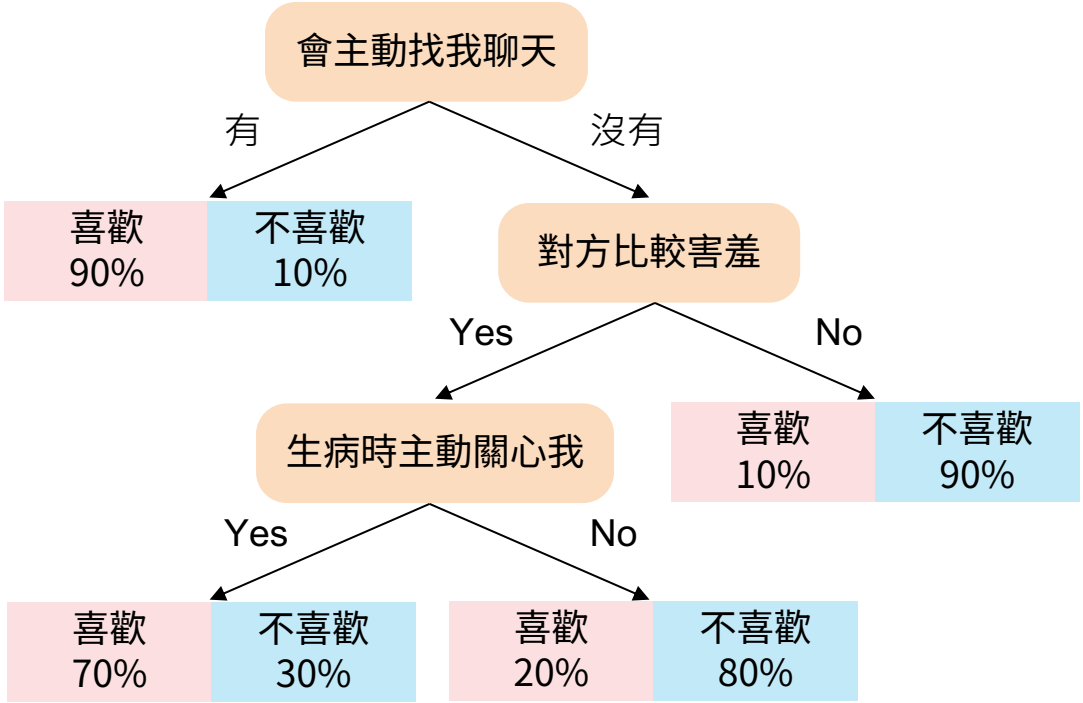
教育程度	
小學	⟶ frequency = 1 / 7
中學	⟶ frequency = 2 / 7
中學	
大學	
大學	⟶ frequency = 4 / 7
大學	
大學	

Tree-based 分類模型

讓機器像人一樣用樹的結構來做決策

人類如何決策: 這是一個暈船仔的故事

她到底喜不喜歡我?

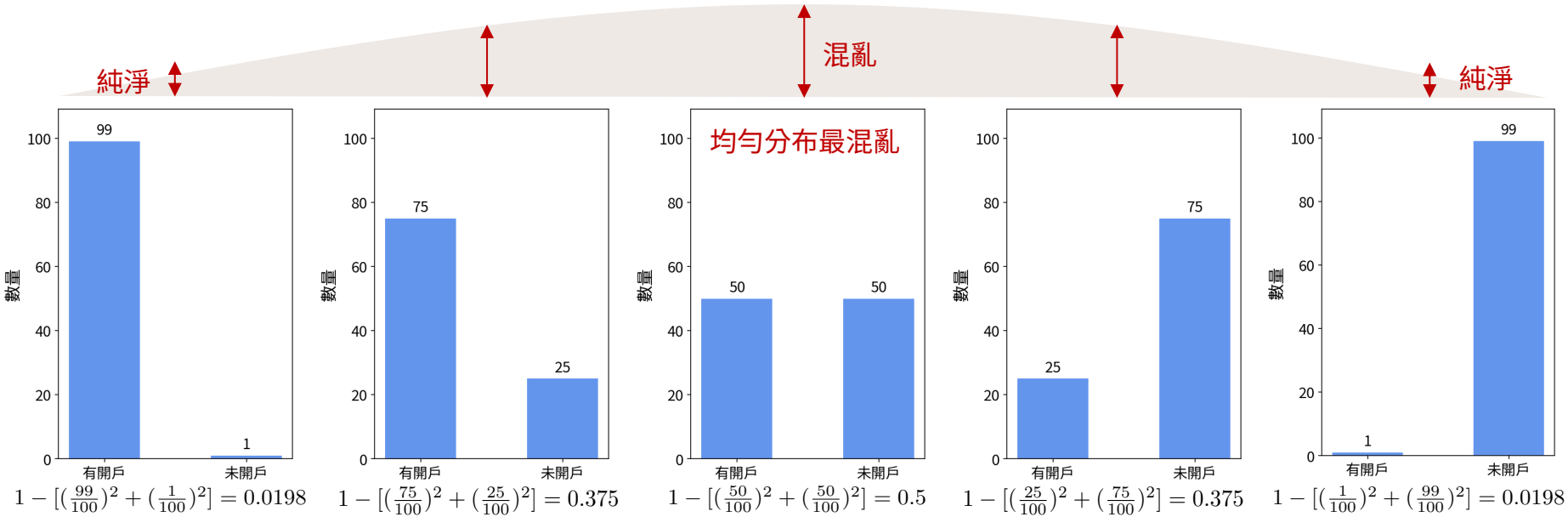


願每個還在暈船的人都能成功等到對方的訊息

定義 Gini Index 讓機器自動分類

我們需要定義一個標準來衡量資料的 **混亂** 程度

$$\text{Gini}(p) = 1 - (p_1^2 + p_2^2 + \cdots + p_C^2)$$



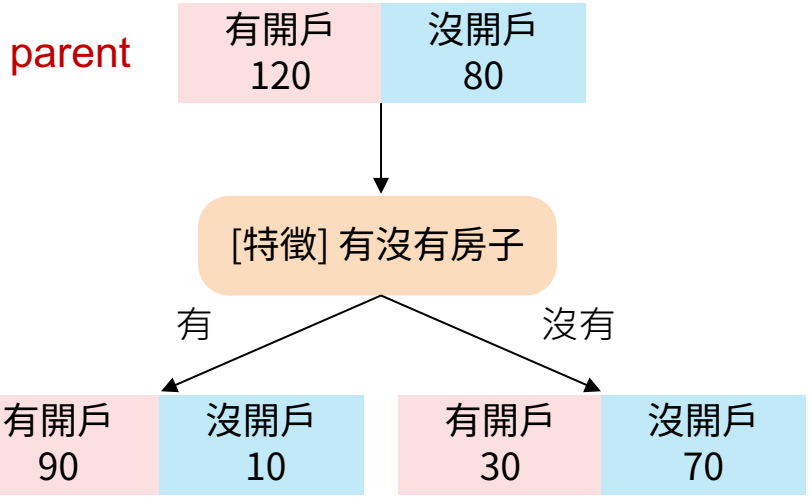
決策樹 (Decision Tree)

有了衡量混亂的指標，如何制定挑選特徵的順序？ 選 $\text{Gain} = \Delta\,\text{Gini}$ 最大的特徵

$$\text{Gain} = \text{Gini}_{\text{parent}} - \text{Gini}_{\text{split}}$$

→ 差異越大代表分完後 **越純淨**

`DecisionTreeClassifier`



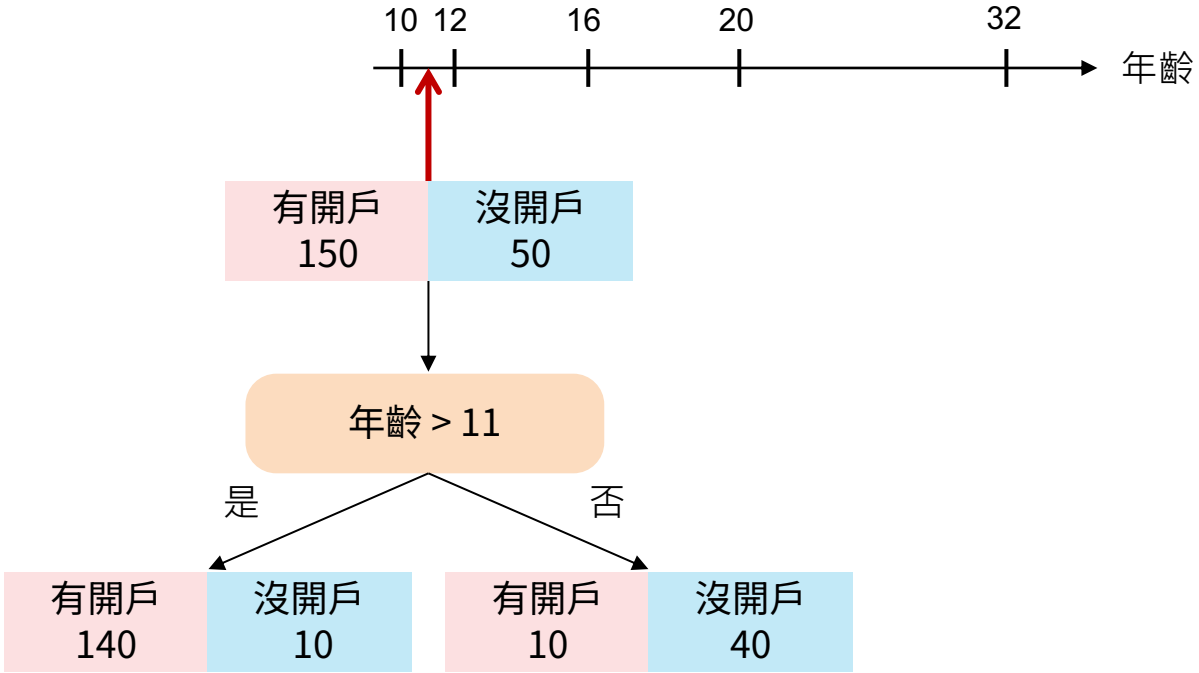
$$\text{Gini}_{\text{parent}} = 1 - [(\tfrac{120}{200})^2 + ((\tfrac{80}{200})^2]$$

$$\text{Gini}_{\text{split}} = \tfrac{1}{2}\text{Gini}_{\text{left}} + \tfrac{1}{2}\text{Gini}_{\text{right}}$$

$$1 - [(\tfrac{90}{100})^2 + (\tfrac{10}{100})^2] \qquad 1 - [(\tfrac{30}{100})^2 + (\tfrac{70}{100})^2]$$

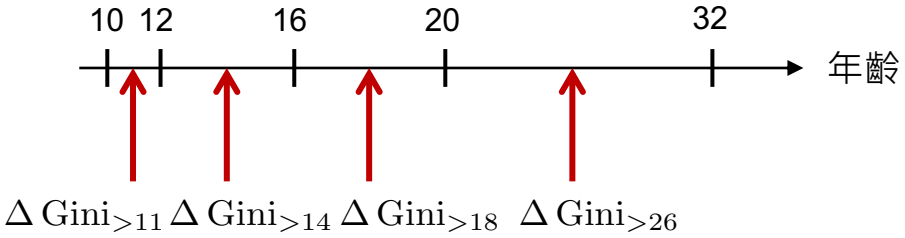
決策樹 (Decision Tree)

連續型特徵怎麼處理?



決策樹 (Decision Tree)

連續型特徵怎麼處理?



選 $\Delta\,\mathrm{Gini}_{>\mathrm{age}}$ 最大的當作切點

決策樹優點 vs 缺點

優點

1. 執行速度快
2. 可同時處理數值型、類別型特徵
3. 原理直觀，容易理解，能做 可解釋性分析

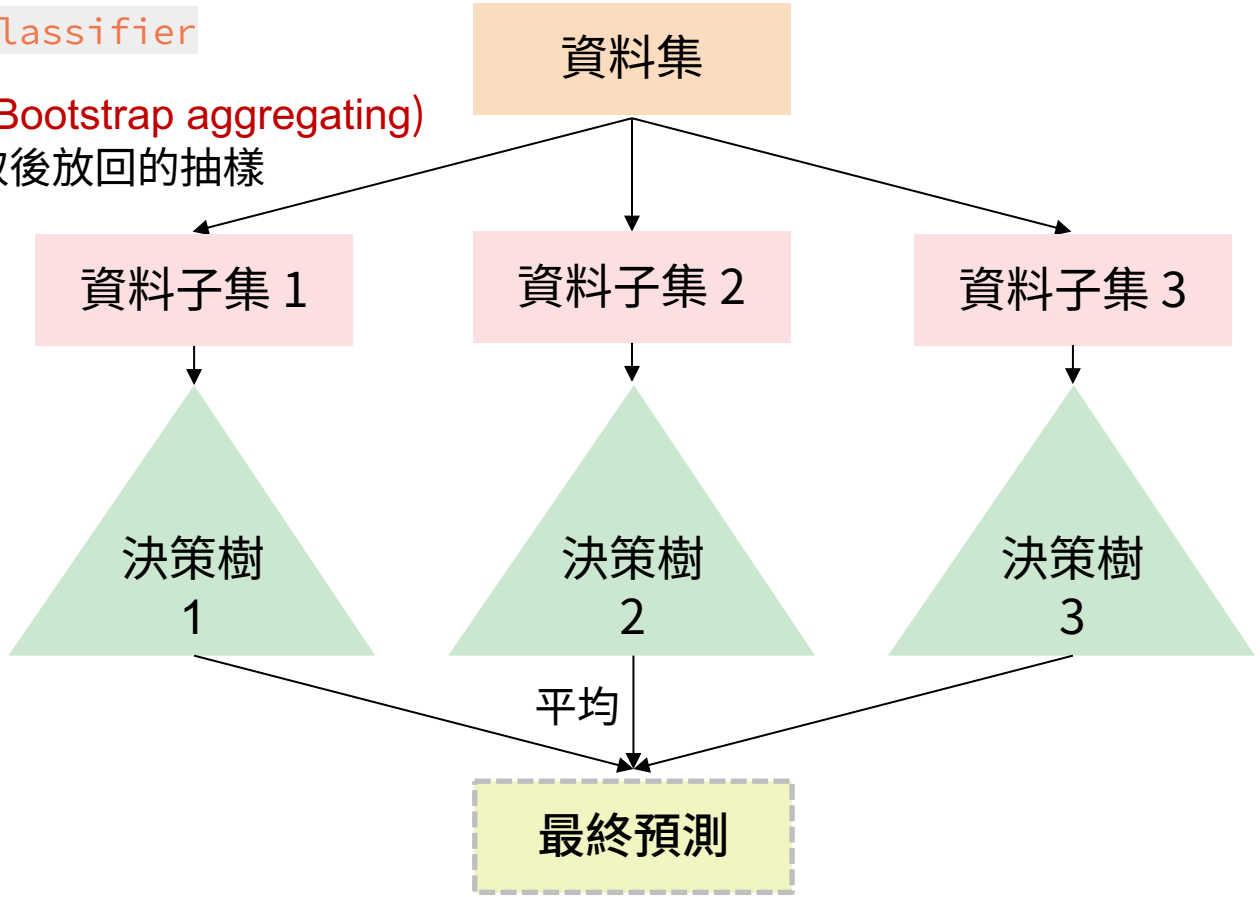
缺點

1. 容易發生 過擬合 (overfitting)
2. 忽略資料集中特徵的關聯性

隨機森林 (Random Forest)

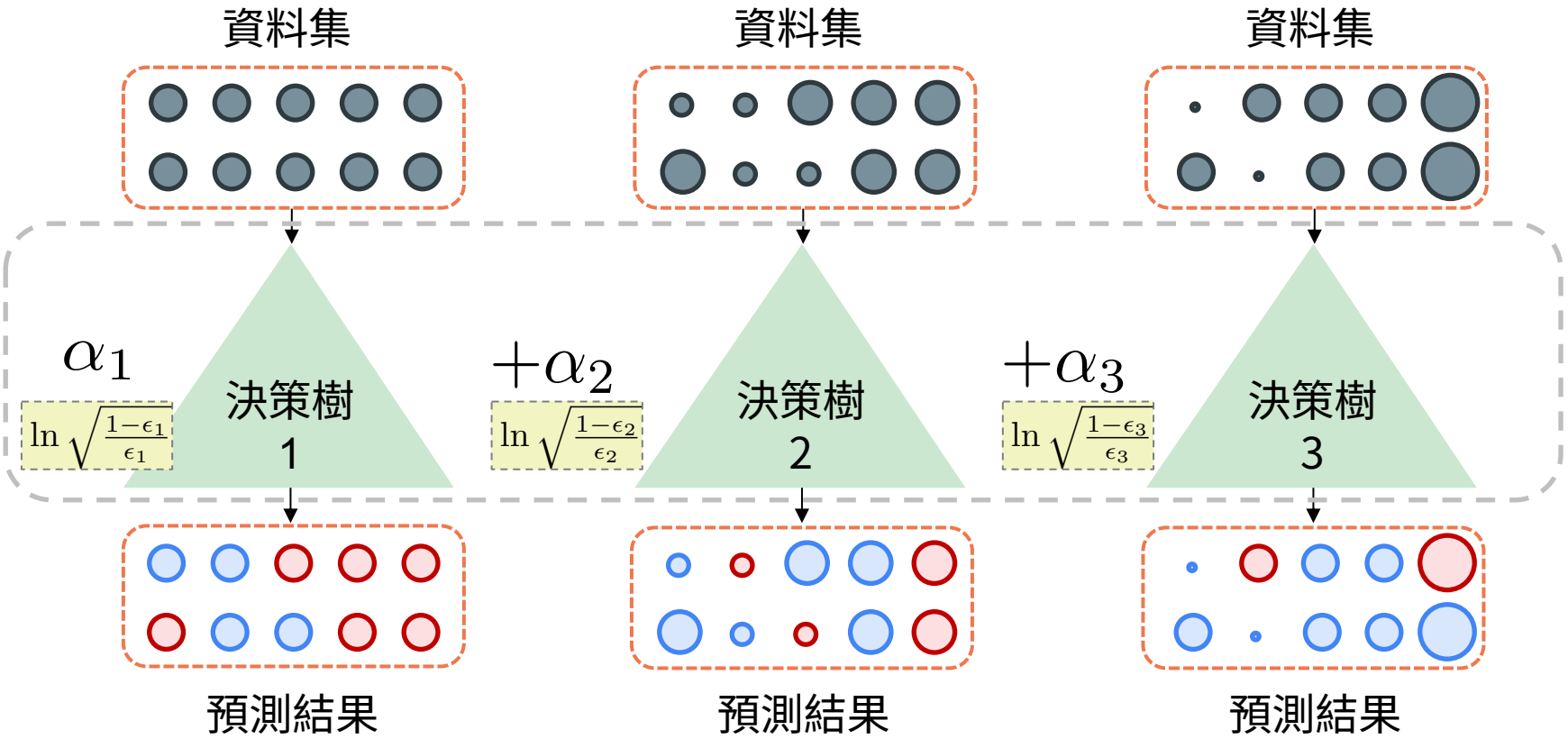
`RandomForestClassifier`

Bagging (Bootstrap aggregating)
取後放回的抽樣



自適應增強 (Adaptive Boosting; AdaBoost)

`AdaBoostClassifier`

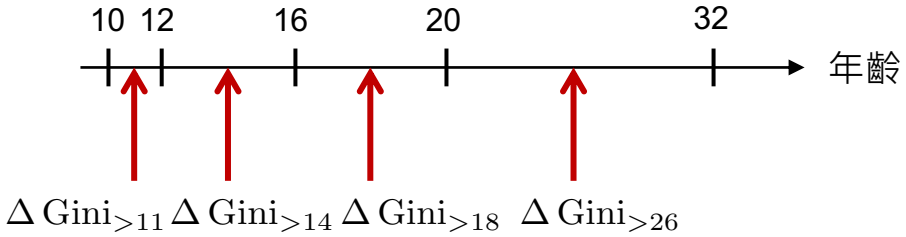


特徵工程對 Tree-based 模型的影響

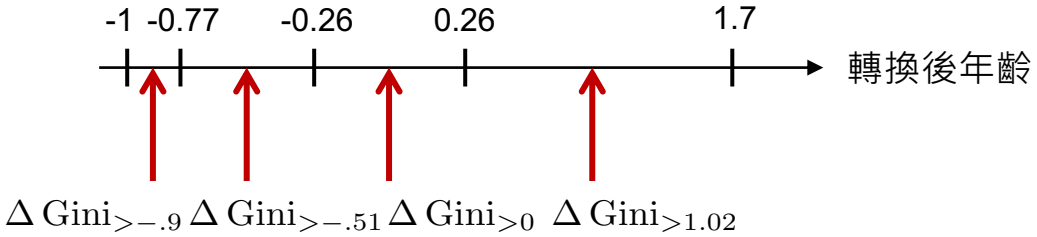
以數值標準化為例

$$\text{Gini}(p) = 1 - (p^2_{\{y=0\}} + p^2_{\{y=1\}})$$

原始年齡



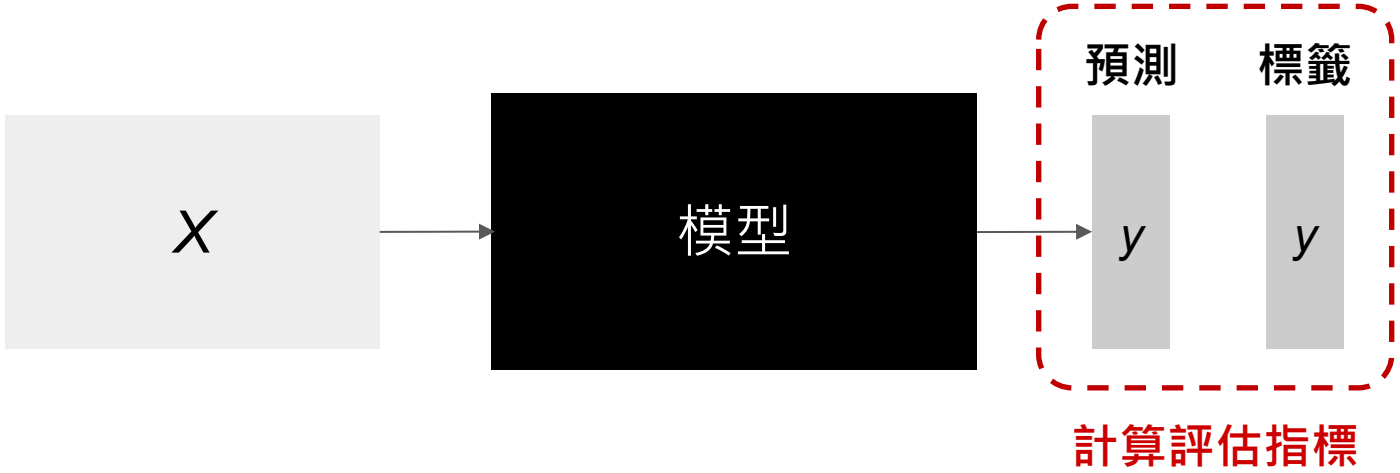
轉換後年齡



模型的評估指標

知道模型的好或壞

模型評估: 準確率 (Accuracy)

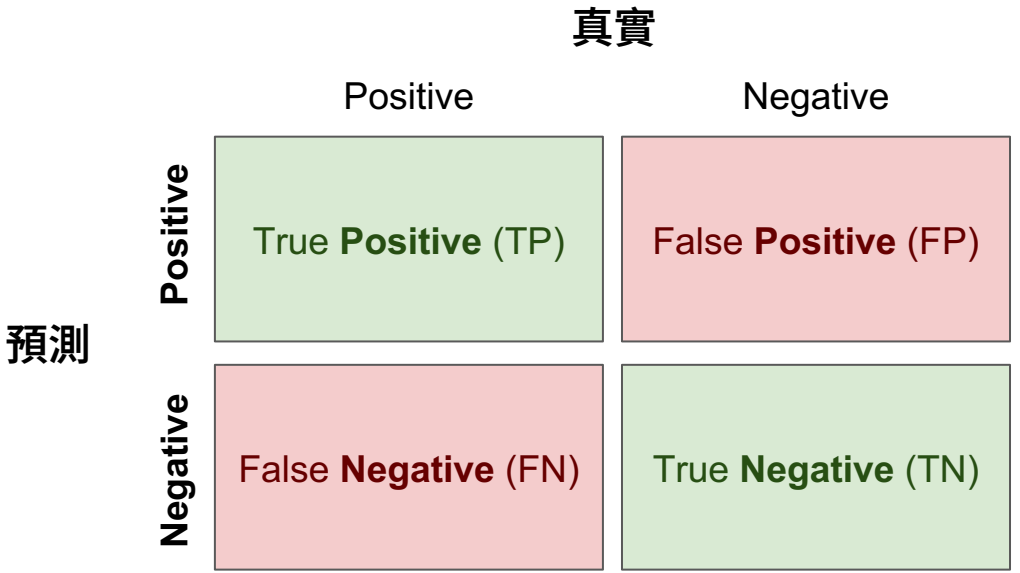


準確率 = 答對的數量 / 總數

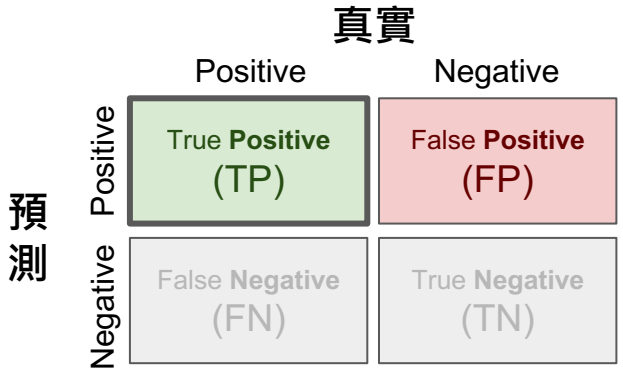
Question: 只看 Accuracy 會有什麼問題?

評估指標: 混淆矩陣 (Confusion Matrix)

- **真陽性** (True Positive，簡稱 **TP**):
  預測有，實際也有
- **偽陽性** (False Positive，簡稱 **FP**):
  預測有，實際卻沒有
- **真陰性** (True Negative，簡稱 **TN**):
  預測沒有，實際也沒有
- **偽陰性** (False Negative，簡稱 **FN**):
  預測沒有，實際卻有



評估指標: Precision vs Recall

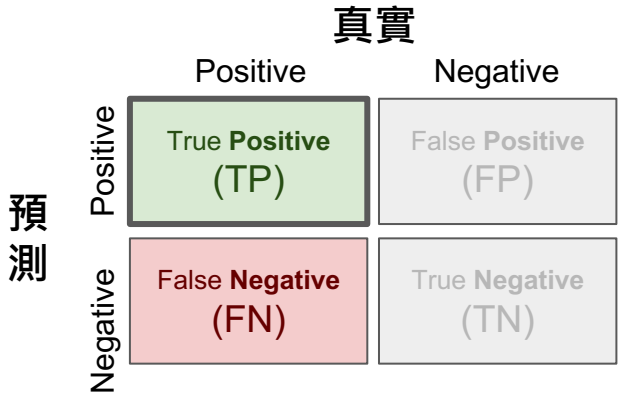


Precision = $\frac{TP}{TP + FP}$

`precision_score`

預測 為陽性中
預測 **正確** 的比例

1. 金融卡盜刷異常偵測
2. 癌症篩檢
3. 推薦系統推播



Recall = $\frac{TP}{TP + FN}$

`recall_score`

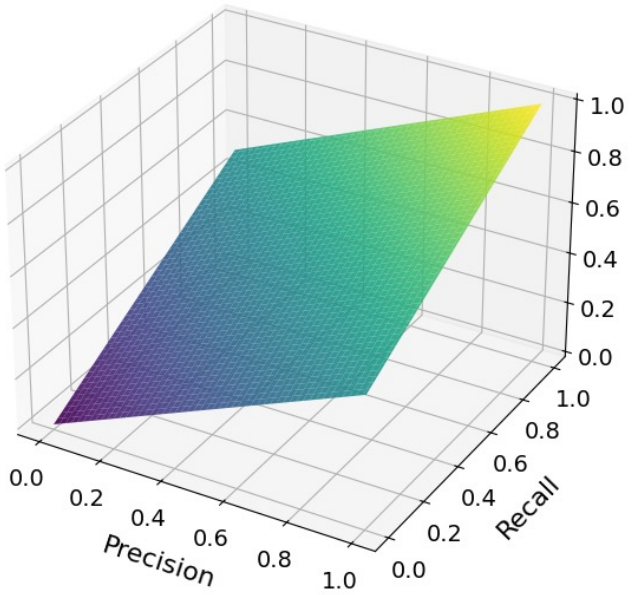
真實 為陽性中
預測 **正確** 的比例

1. 系統入侵偵測
2. 工廠異常預警
3. 失智症篩檢

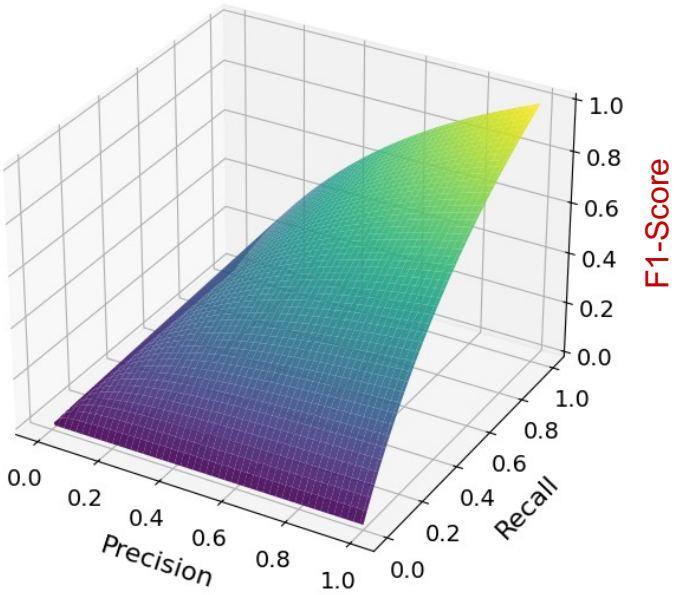
評估指標: F1-score (Precision, Recall 兩個都想要)

F1-Score 為 Precision 和 Recall 的**調和平均** `f1_score`

算數平均 $\frac{P+R}{2}$



調和平均 $\frac{2PR}{P+R}$



其他商業指標 (Business Metric)

- 根據目標訂定指標
  - 轉換率 (Conversion Rate) = 購買人數 / 瀏覽人數
  - 點擊率 (Click Through Rate, CTR) = 點擊次數 / 瀏覽人數
  - 收入 (revenue)

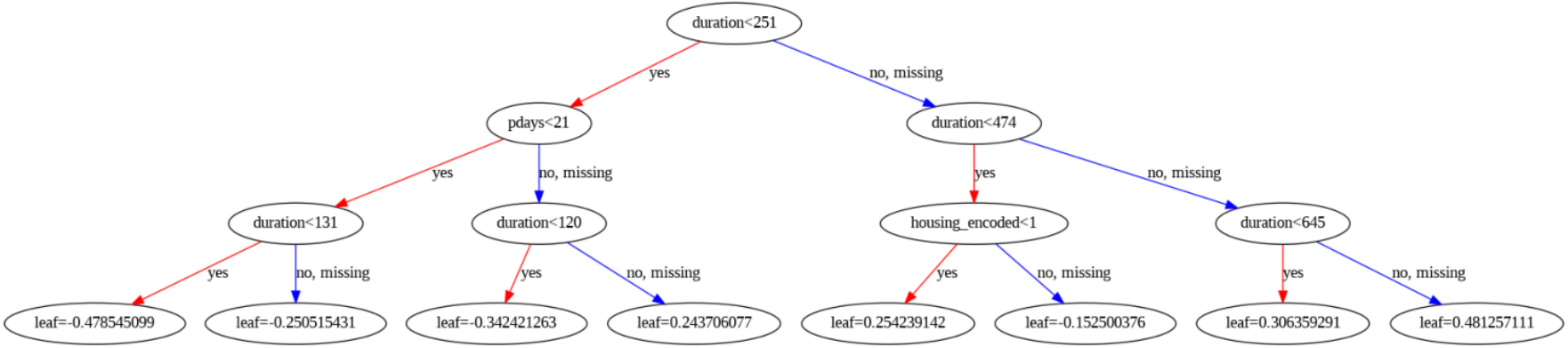
- 結合 A / B Testing 訂定指標
  - Control Group (A 組: 基準), Experiment Group (B 組: 實驗)
  - 提升率 = Experiment Group 指標 / Control Group 指標

Medium 推薦好文: 解讀 Appier 的 PizzaHut 案例，看人工智慧如何搭配策略執行 [戴士翔 Dennis]

模型解釋

2B 客戶最喜歡問的

視覺化 XGBoost



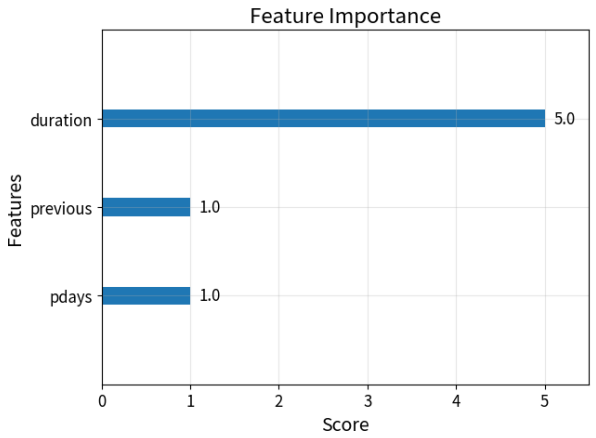
機率	0.38259574	0.43769664	0.41522144	0.56062675	0.56321962	0.46194862	0.57599636	0.61804468

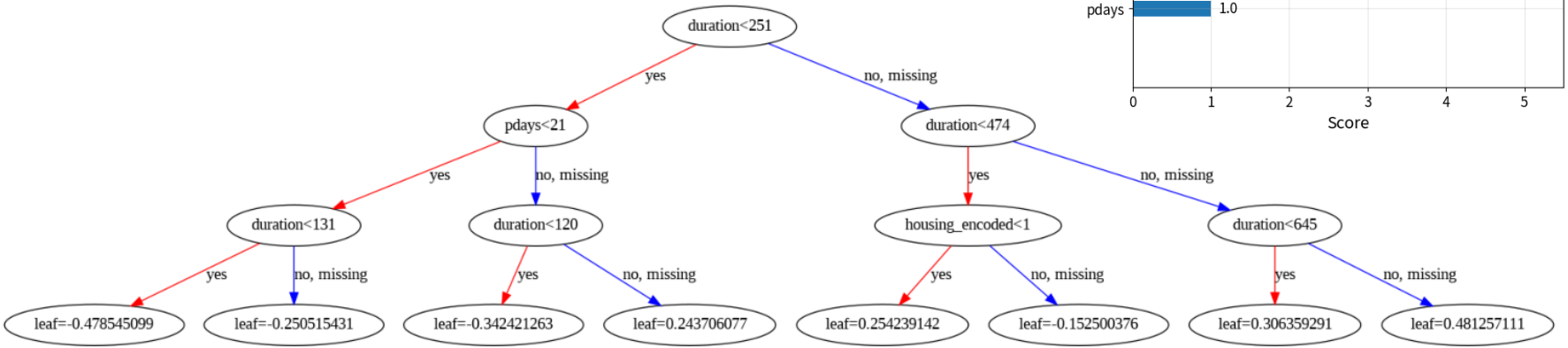
$$P(y=1|\mathbf{x}) = \frac{1}{1+e^{-\text{leaf}(\mathbf{x})}}$$

Question: 有沒有自動化的方法計算每個特徵被用作分裂的次數

特徵重要性 (Feature Importance)

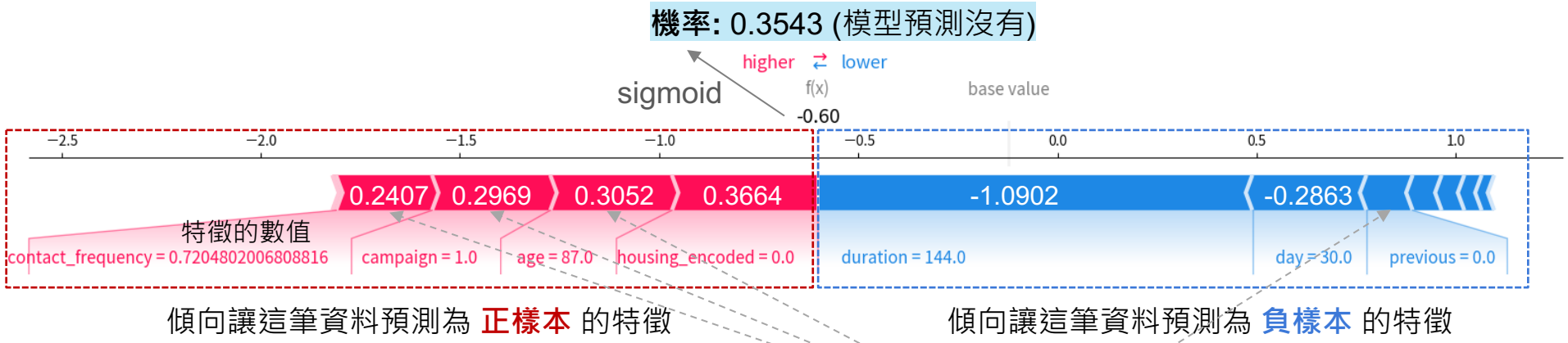
特徵在所有樹中被用來作為分裂節點的 **總次數**





Question: 有沒有辦法知道每個特徵對預測結果的貢獻?

SHAP (SHapley Additive exPlanations): 單筆資料



傾向讓這筆資料預測為 **正樣本** 的特徵

傾向讓這筆資料預測為 **負樣本** 的特徵

(箭頭的長度表示每個特徵的 SHAP 值)

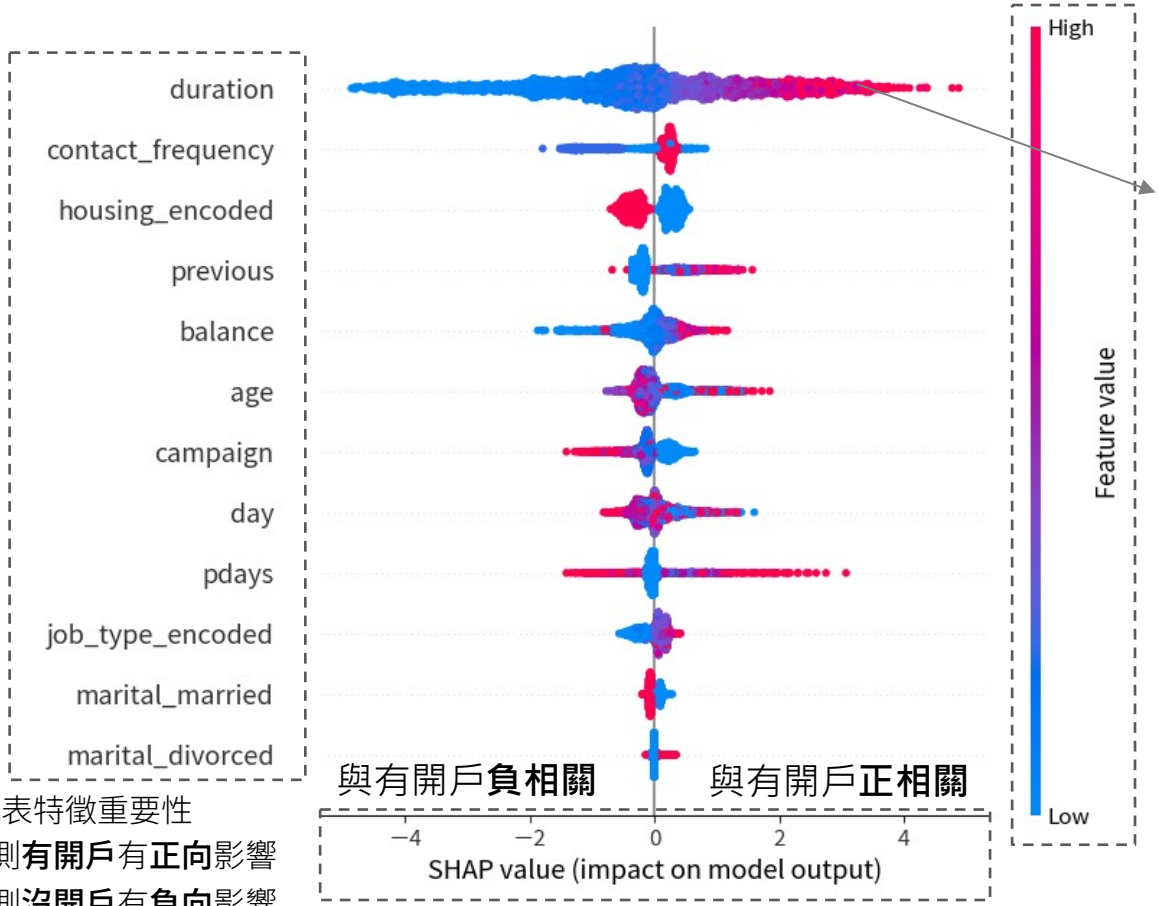
$$f(x) = (\text{base value}) + \sum_{\text{feature } i} \text{SHAP}_i$$

起點 (每筆資料的值相同)　第 *i* 個特徵對預測的貢獻

原始論文: A Unified Approach to Interpreting Model Predictions (2017 NIPS)

Medium 推薦好文: 可解釋 AI (XAI) 系列 — SHAP [張家銘]

SHAP (SHapley Additive exPlanations): 多筆資料



由上到下為
11 大特徵排序

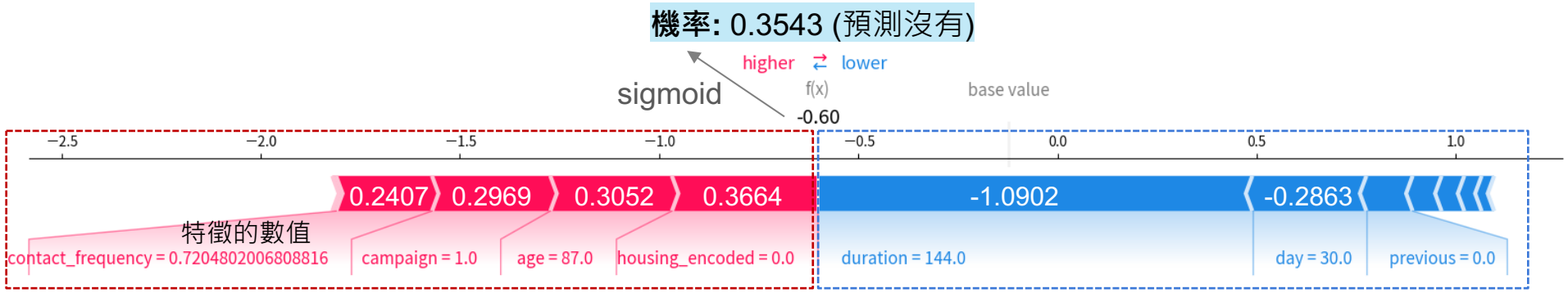
圖上的每一個點代表這筆資料
於這個特徵的 SHAP 值

資料點的顏色代表特徵的數值大小
越紅: 特徵的數值 **越大**
越藍: 特徵的數值 **越小**

SHAP 值代表特徵重要性
正值: 對預測**有開戶**有**正向**影響
負值: 對預測**沒開戶**有**負向**影響

利用 SHAP 做分析: 看單筆資料

用 SHAP 分析 false negatives (預測沒有，實際卻有)



傾向讓這筆資料預測為 **正樣本** 的特徵

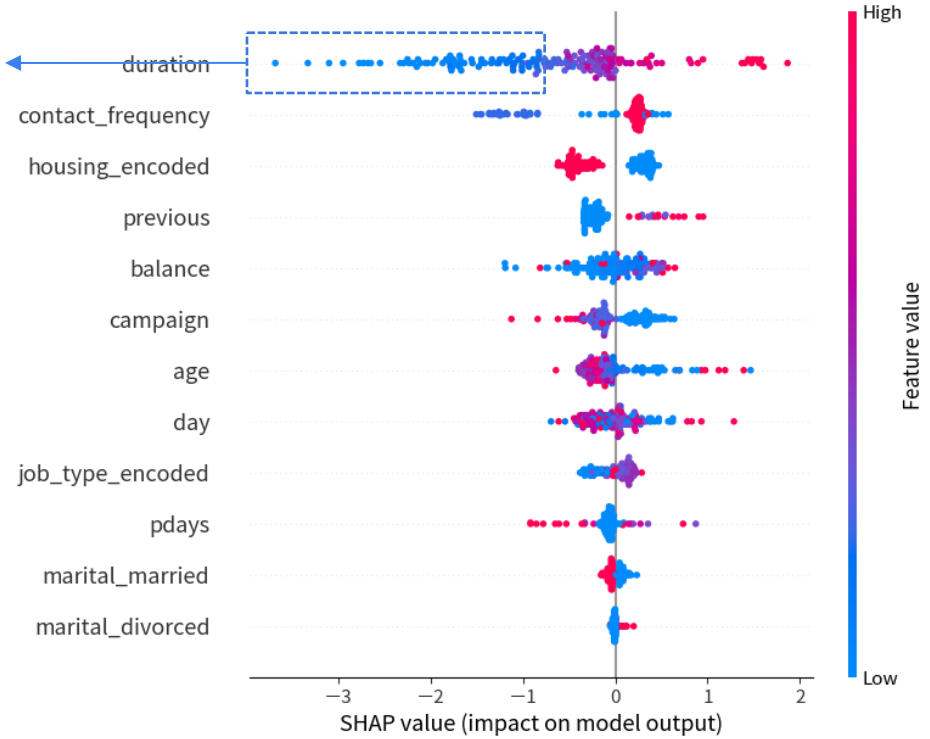
傾向讓這筆資料預測為 **負樣本** 的特徵

觀察: duration 特徵讓這筆資料有強烈的負樣本趨勢

利用 SHAP 做分析: 看單筆資料

用 SHAP 分析 false negatives (預測沒有，實際卻有)

duration 數值小的區域
對模型預測有負類的傾向
(因為 SHAP 小)



額外拆分驗證集

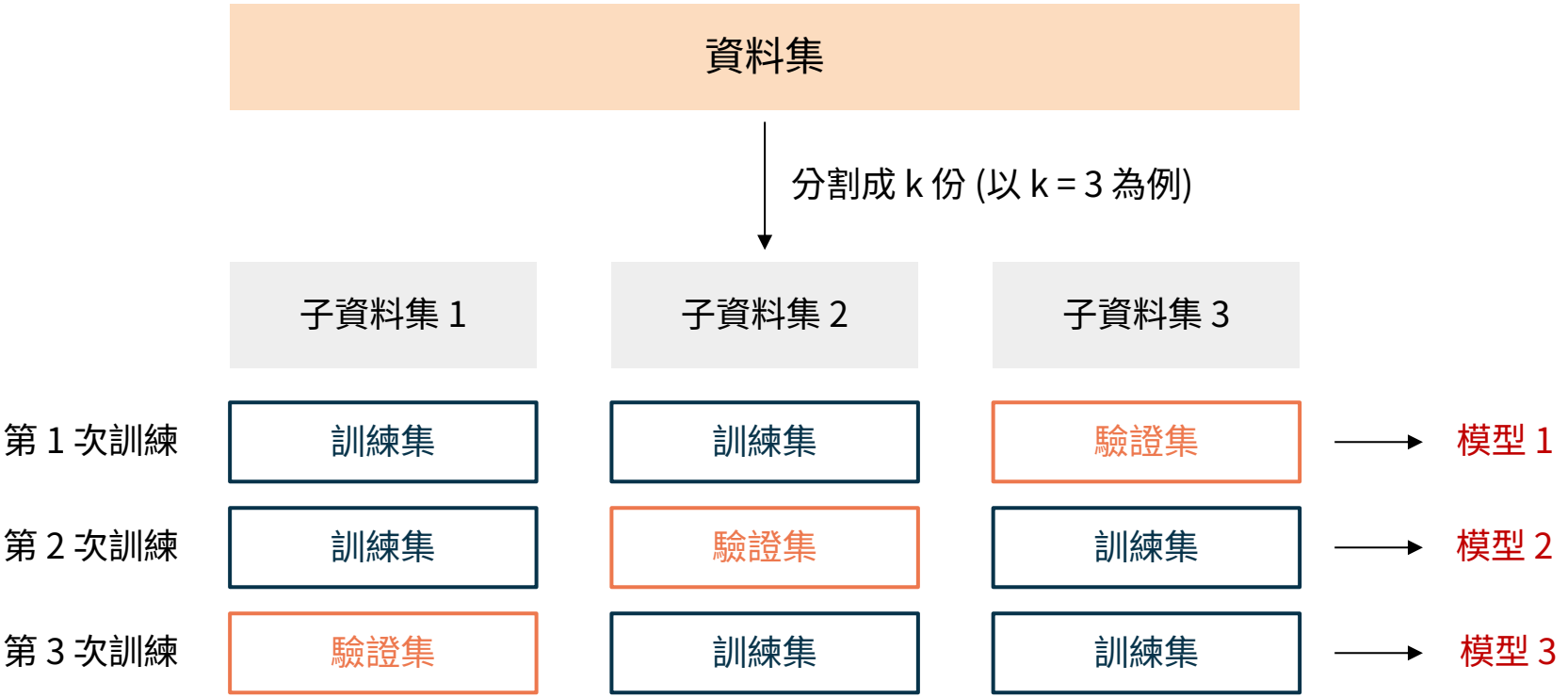
考大考前先來場模擬考壓壓驚吧

拆分資料集、驗證集、測試集

	資料集	功能	類比
訓練集	60-80% X y	建立模型	小考、平時考
驗證集	10-20%	尋找模型最佳參數	模擬考
測試集	10-20%	評估模型成效	學測、指考

k 折疊交叉驗證 (*k* Fold Cross Validation)

`KFold`



如何找到準確率最高的模型參數?

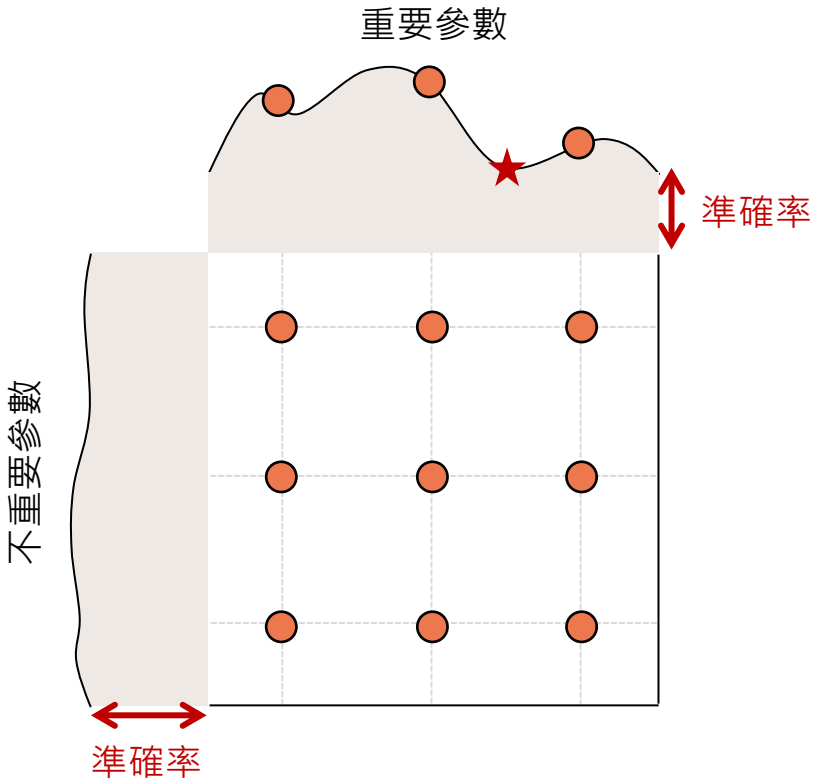
超參數搜索

讓模型自動找到最喜歡的參數發揮最大值

超參數搜索方法

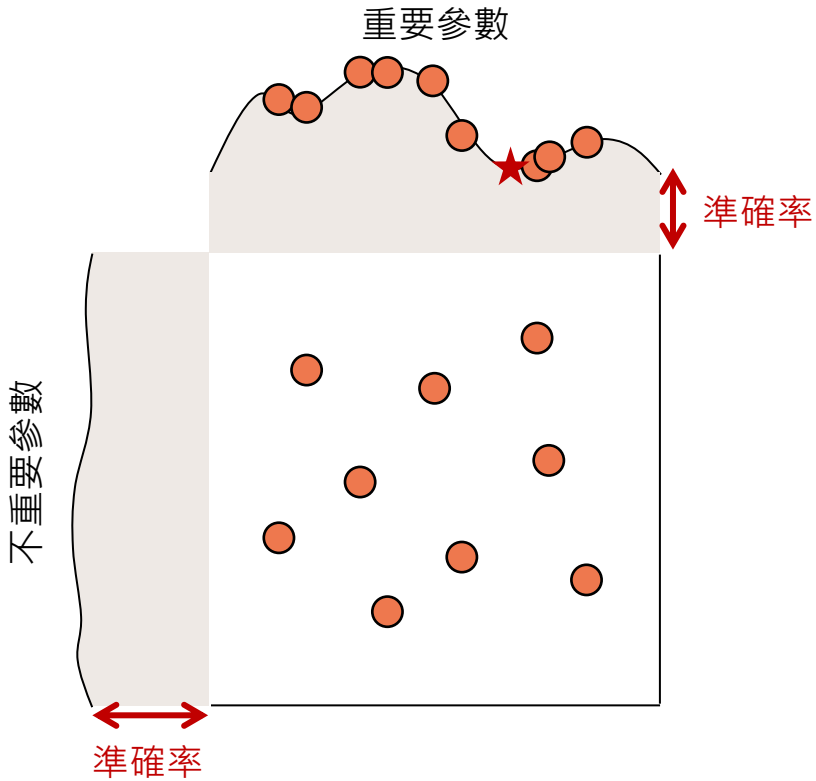
網格搜索 (Grid Search)

`GridSearchCV`



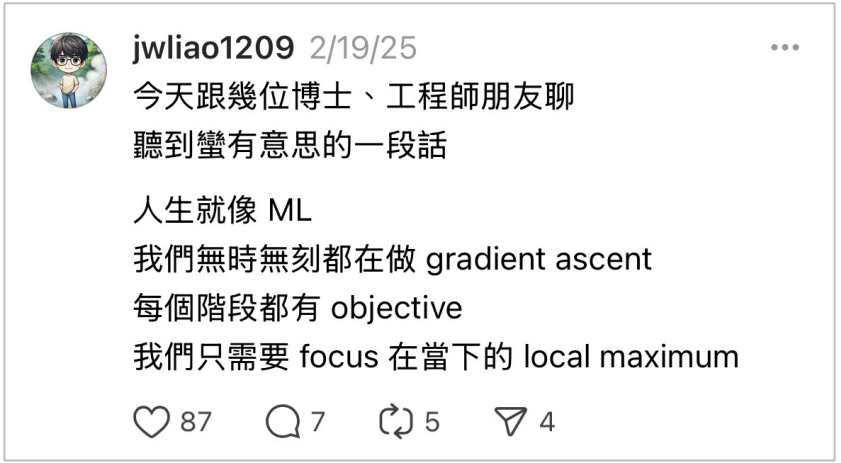
隨機搜索 (Random Search)

`RandomizedSearchCV`



小結論

- 隨機搜索比網格搜索找最佳解更有效率
- 不過實務上還是較常使用**網格搜索**
- 如果可以對目標的指標求取微分，可以使用梯度下降法找最佳解

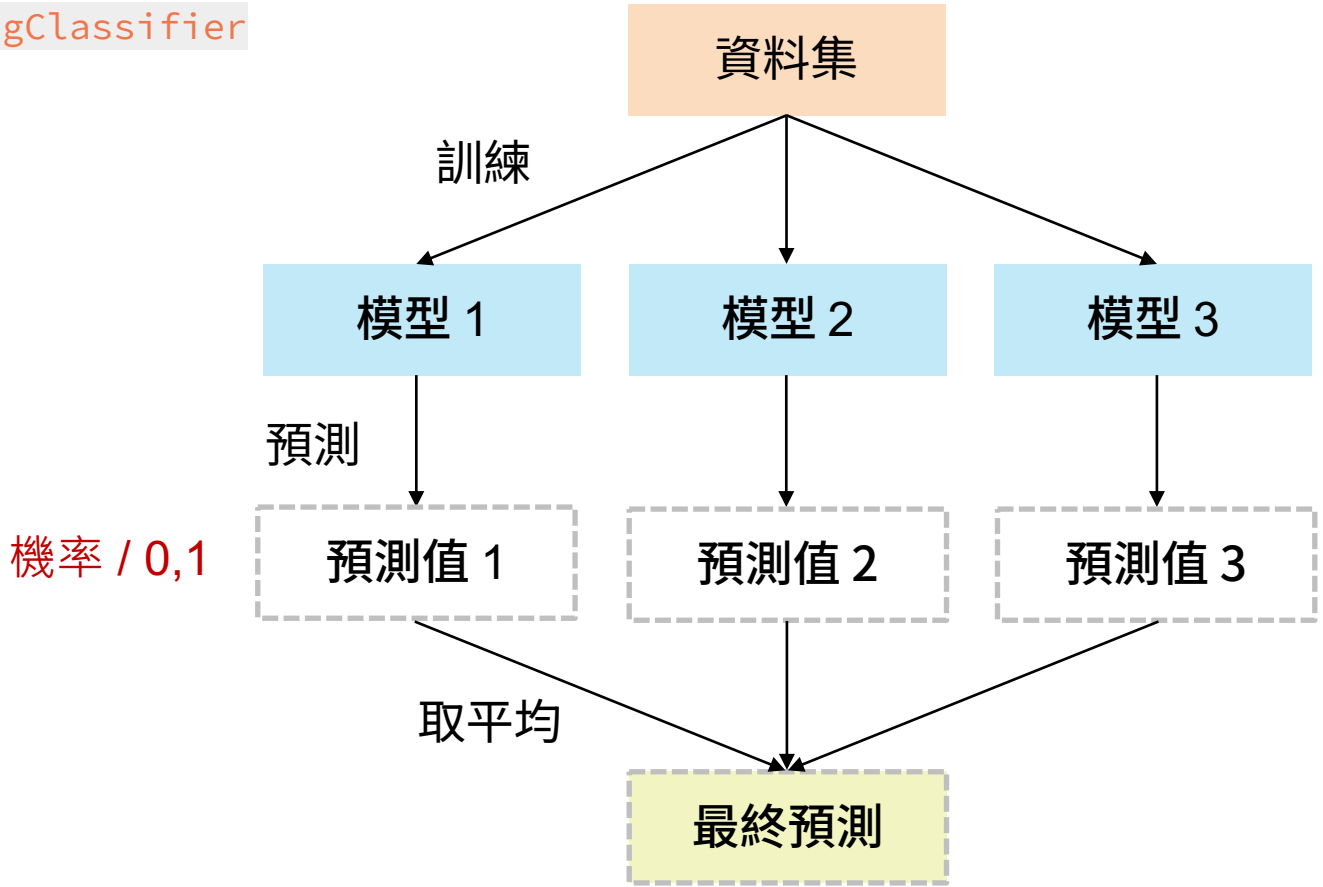


集成學習

三個臭皮匠勝過一個諸葛亮

投票 (Voting)

`VotingClassifier`



分析報告

製作一個讓人感動的專案簡報

簡報呈現

- 問題介紹
  - 定義明確的量化指標
  - 為何選選擇這個指標，其對應到的商業意義為何?
- 闡述想法
  - 從 EDA 中觀察到什麼現象
  - 因為 … 所以使用某某特徵
  - 因為 … 問題所以做某某特徵轉換
  - 因為 … 特性所以使用某某模型
  - 設計的方法如何對應定義的問題
- 不要出現過多的數學式子，盡量以圖呈現
- 不要放程式碼，可以簡述模型概念
- 提供分析結果與洞察

回家作業 1: 特徵工程練習

請用第一次上課所教的特徵工程與 SVM 模型，提高 test dataset 的準確率，並將你的方法整理成 5 頁投影片，與程式碼一起繳交至雲端資料夾，前 3 名的同學將可獲得小禮物。

作業繳交連結

命名規定: ml01_廖家緯.pdf / ml01_廖家緯.ipynb

回家作業 2: 可解釋性 AI 練習

請用上課所教的特徵工程建立 XGBoost 模型，使用 SHAP 值進行分析，並改進模型。將你的方法整理成 10 頁內的投影片，與程式碼一起繳交至雲端資料夾，表現優良者會頒發特製獎狀。

作業繳交連結

命名規定: ml02_廖家緯.pdf / ml02_廖家緯.ipynb

自學資源

這堂可結束後如果你還意猶未盡，可以透過以下資源自學更多機器學習

- [機器學習] 林軒田 機器學習基石 / 機器學習技法
- [深度學習] 李宏毅 機器學習 / 生成式 AI (LLM)
- [深度學習與自然語言處理] 陳縕儂 深度學習之應用
- [機器學習] 吳恩達 Machine Learning (coursera)
- [統計與機器學習] StatQuest

致謝

- 感謝社長瑾叡與我在 5th NTU DAC 設計 Workshop 課程
- 感謝 6th NTU DAC 課程長 (Benson, 佳妤, Maggie, 謙文) 與我一同討論本課程
- 感謝 Benson 協助設計初版簡報模板
- 感謝昕怡給我簡報設計許多建議
- 感謝佳妤給我許多課程建議
- 感謝 6th NTU DAC 社員給我很多鼓勵，順利完成這 3 堂課程

Thank you