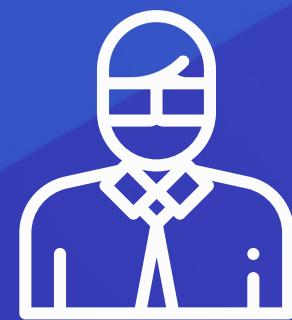


Day 22 特徵工程

特徵工程簡介



出題教練

陳明佑

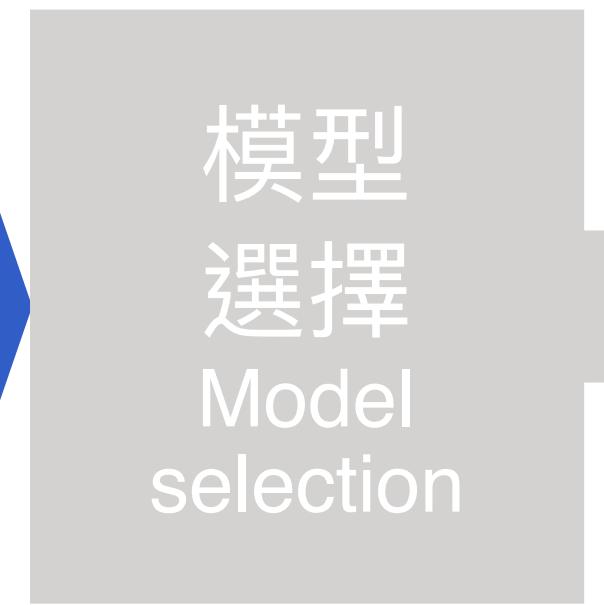


知識地圖 特徵工程 特徵工程簡介

特徵工程

監督式學習

Supervised Learning



非監督式學習

Unsupervised Learning



特徵工程 Feature Engineering

概論

數值型特徵

類別型特徵

時間型特徵

填補缺值

去離群值

類別型特徵處理

時間型特徵處理

去偏態

特徵縮放

特徵組合

特徵篩選

特徵評估

本日知識點目標

- 初步理解特徵工程的概念
- 能從程式中辨識特徵工程的區塊與意義
- 知道特徵工程至少需要那些部分

什麼是特徵工程? (1 / 2)

通常我們會這樣考慮事情...



什麼是特徵工程? (2 / 2)

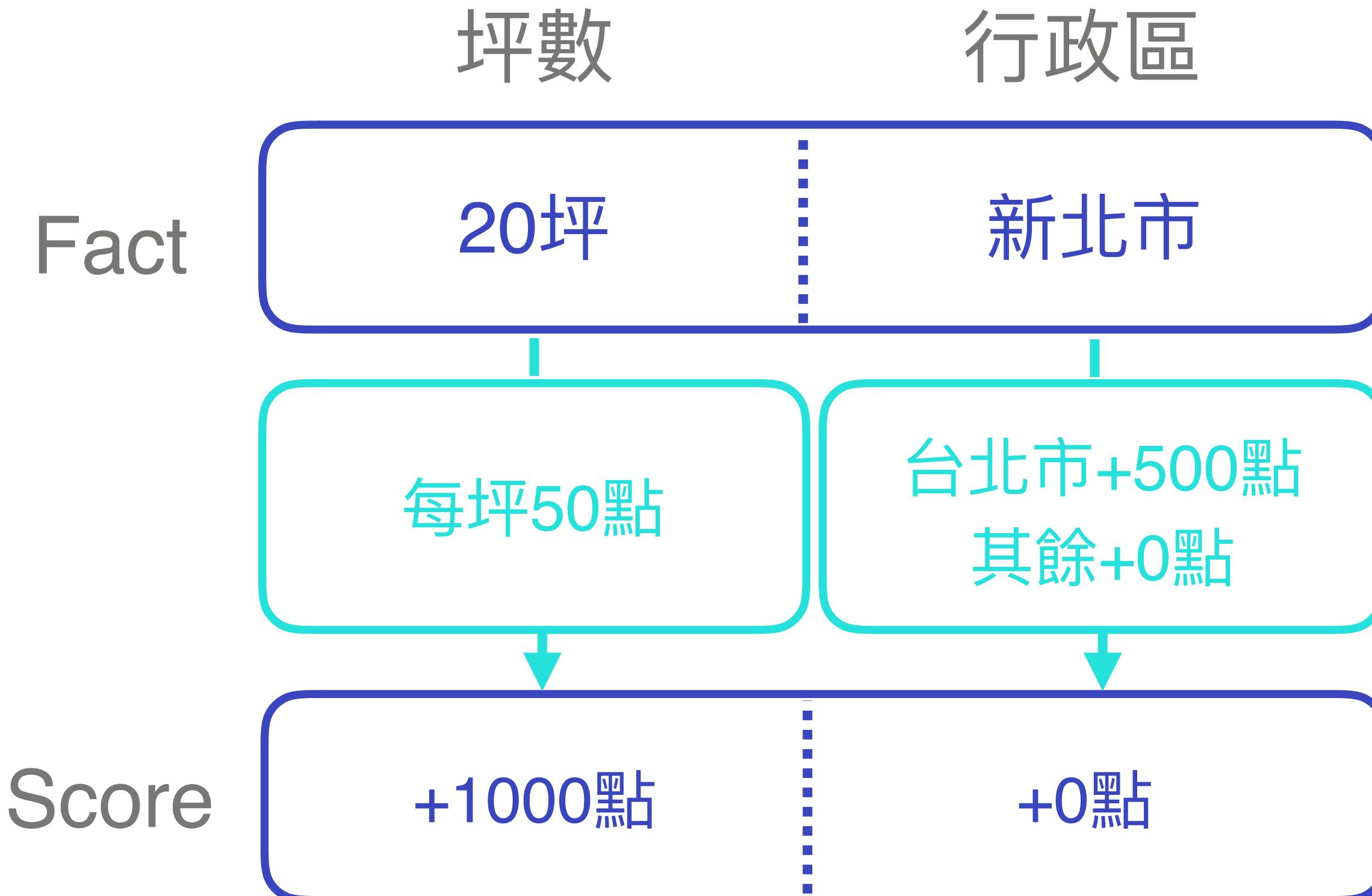
但其實考慮的關鍵是...



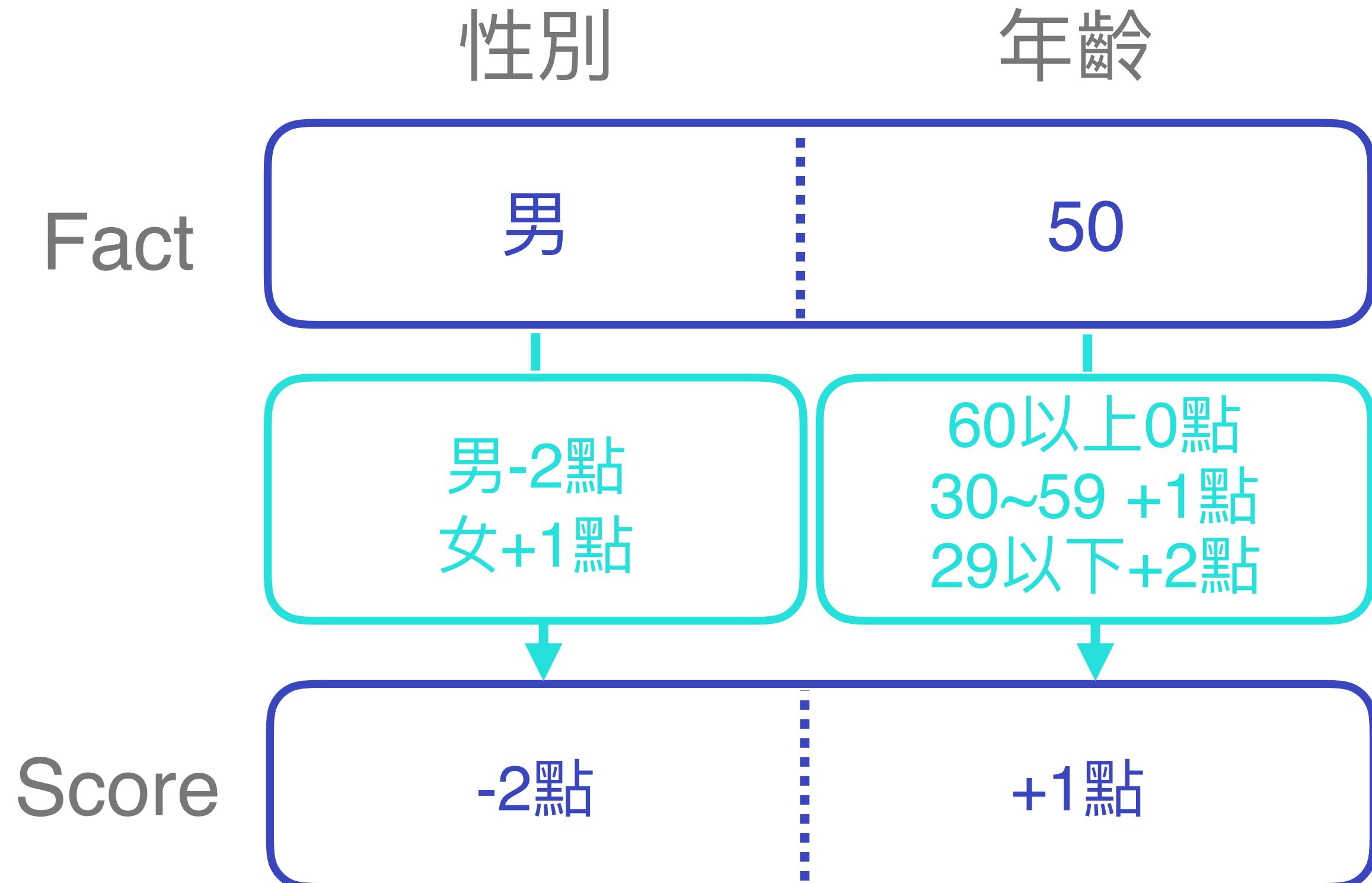
從事實到對應分數的轉換，我們稱為特徵工程

特徵工程舉例

房價預測

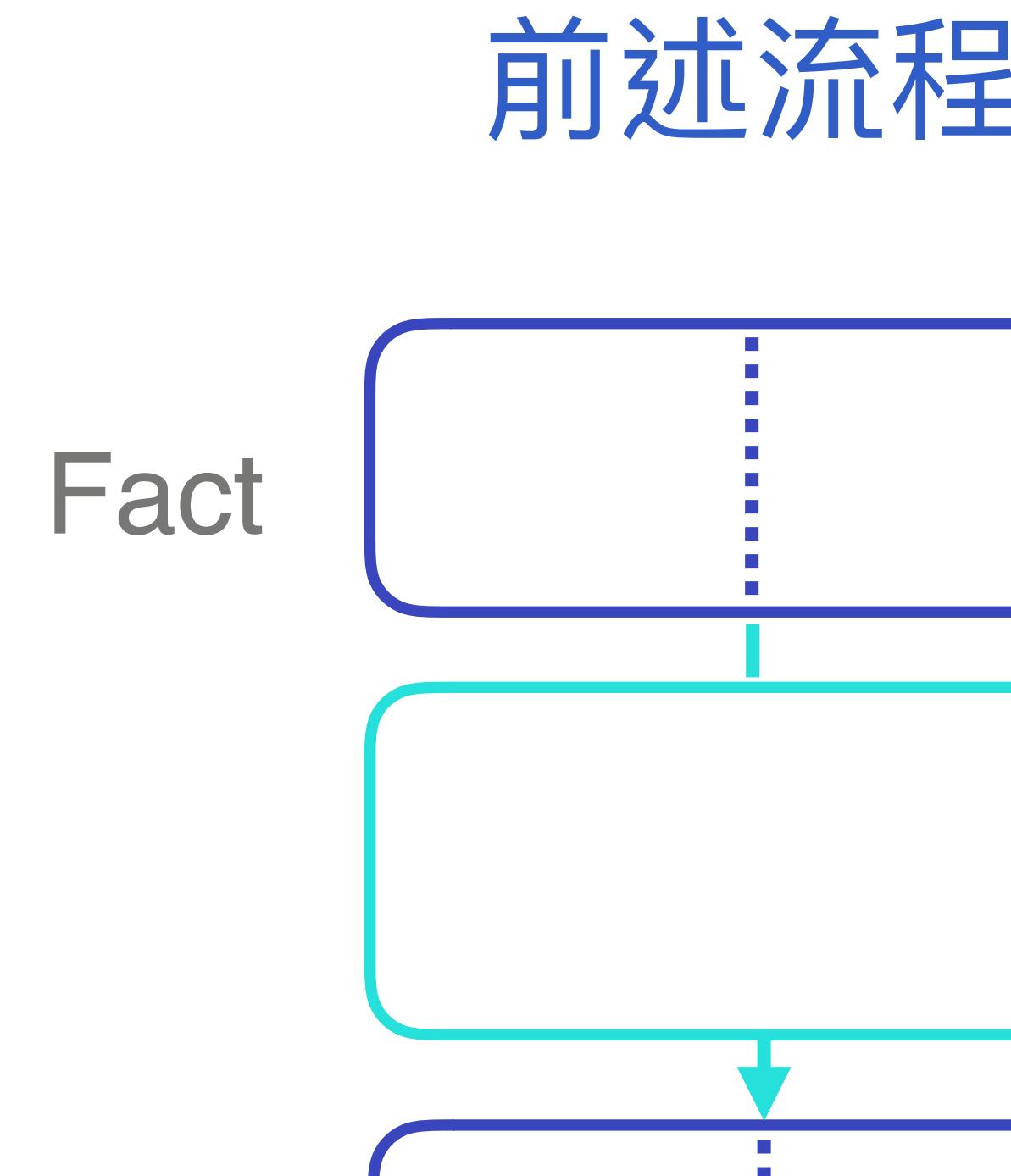


鐵達尼號生存預測

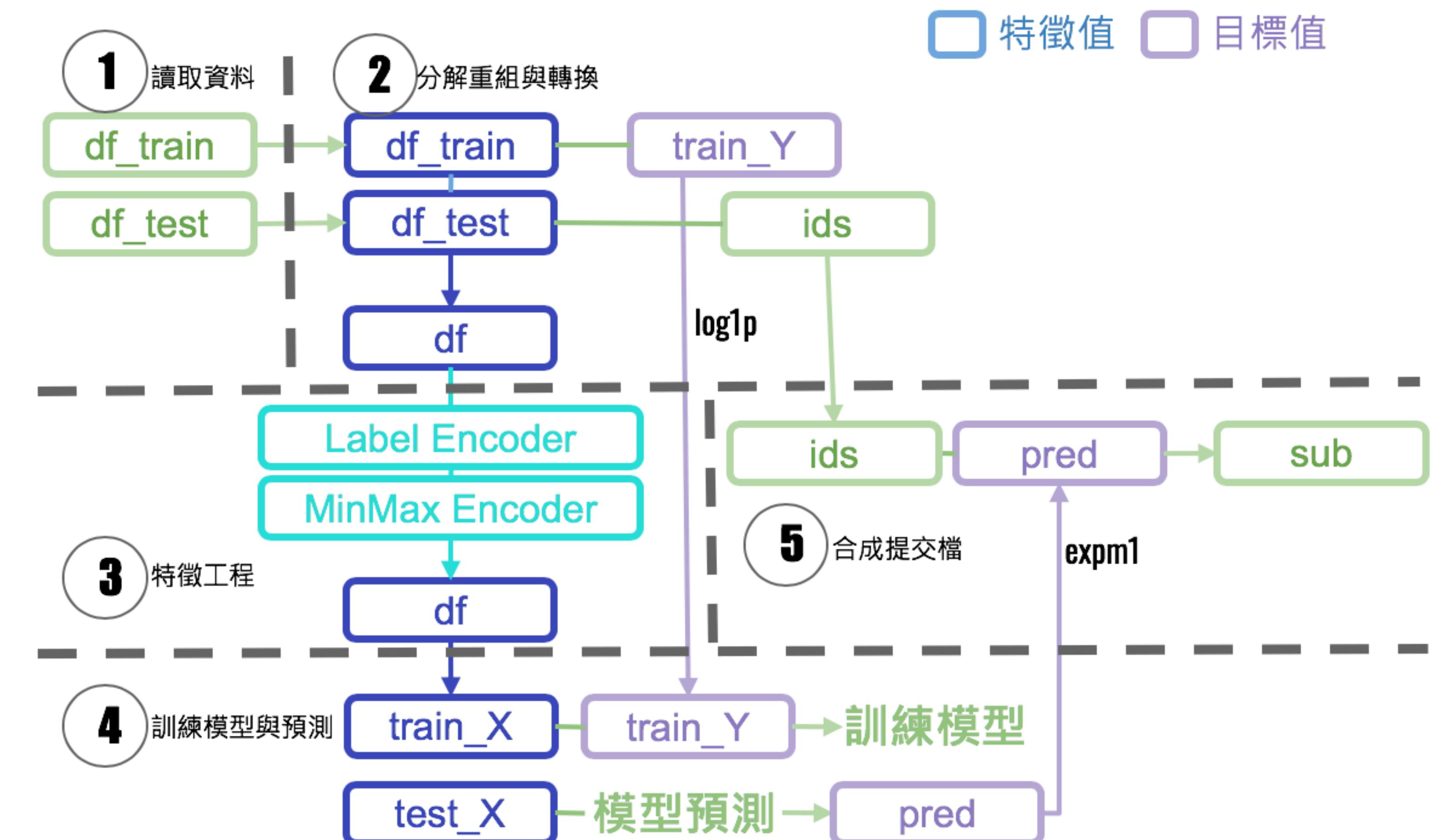


點數未必直接對應到總價或機率，但會是後續評估的依據

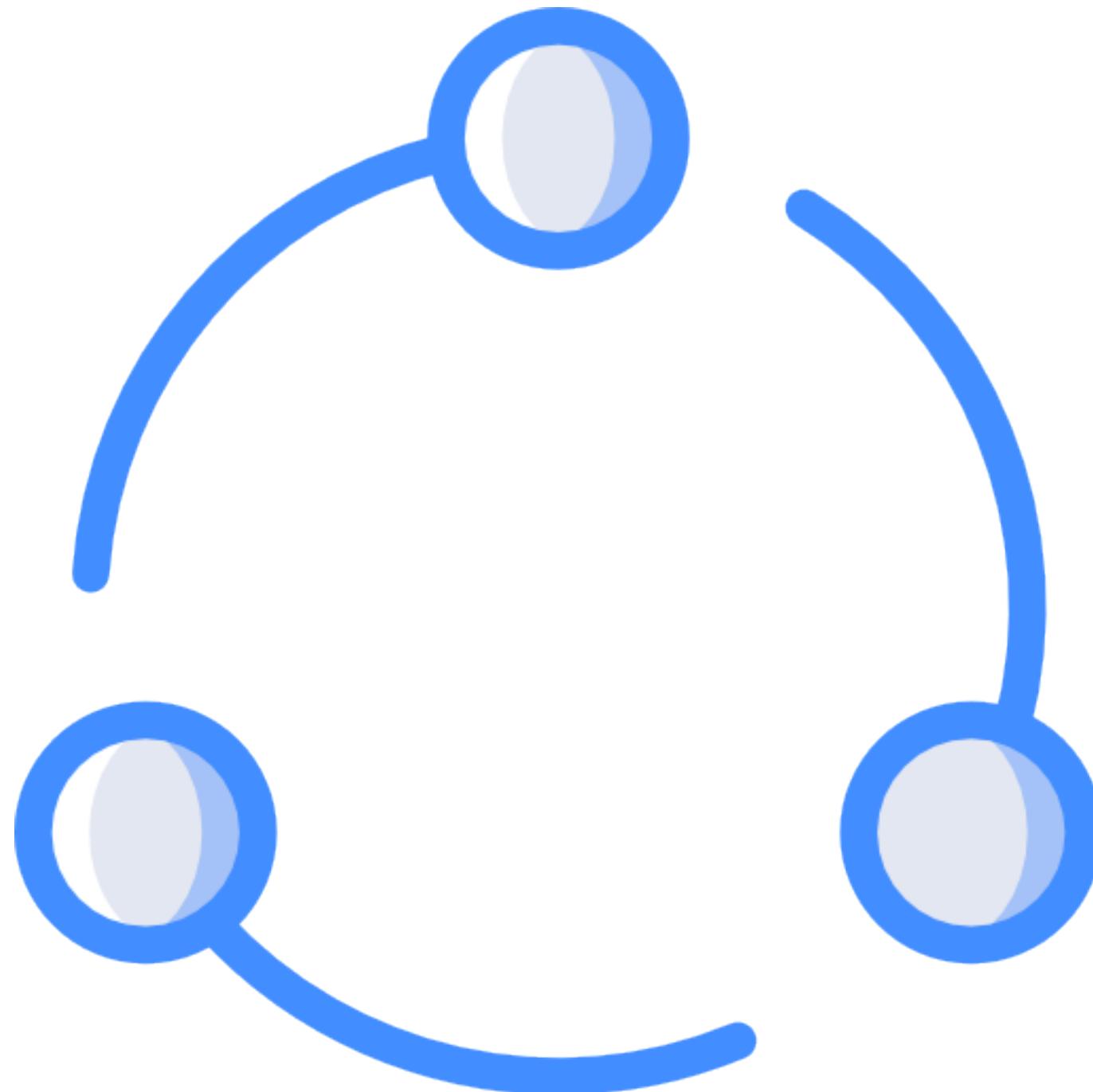
前述流程 / python程式 對照(範例：房價預測)



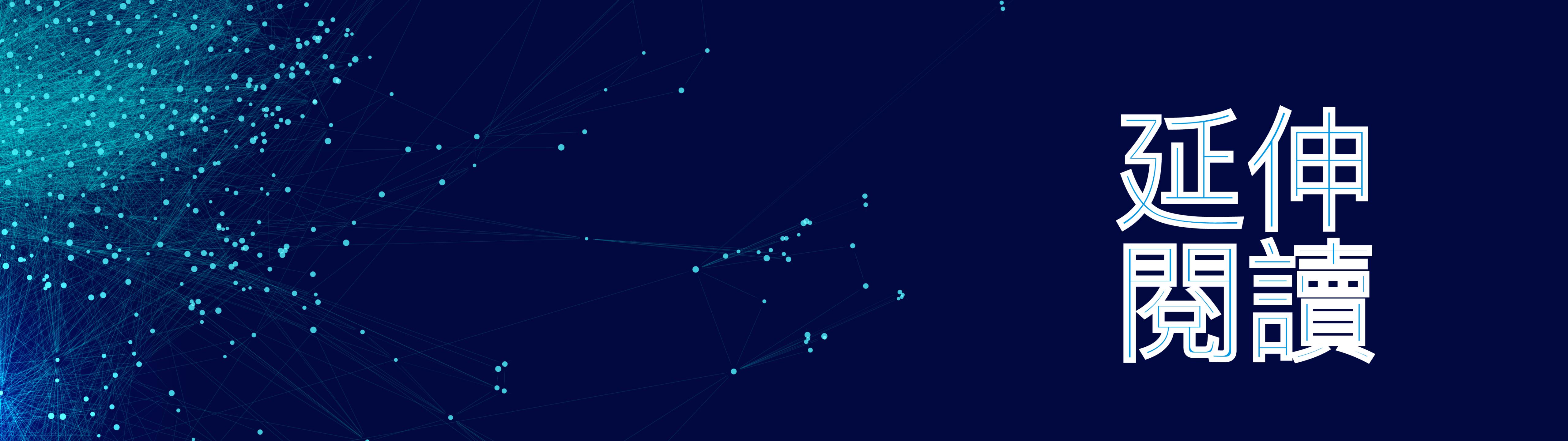
python程式 (請參閱今日範例)



重要知識點複習



- 特徵工程是**事實對應**到後續評估分數的**轉換**
- 在程式語法中，特徵工程位於**資料彙整之後**，以及**擬合模型之前**
- 由於資料包含類別型(文字型)特徵以及數值型特徵，所以最小的特徵工程至少要包含一種**類別編碼**(範例使用**標籤編碼**)，以及一種**特徵縮放方法**(範例使用**最小最大化**)



延伸 閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有
多餘時間，可再補充延伸閱讀文章內容。

推薦延伸閱讀

知乎-特徵工程到底是什麼

網頁連結

- 本文重點為右圖，主要是希望同學大致知道特徵工程大致包含哪些部分，若對細節有興趣，還可以從這篇中了解一些概念其中一部分的內容，會在後面的課程中說明並練習，詳情請參閱百日馬拉松課綱。

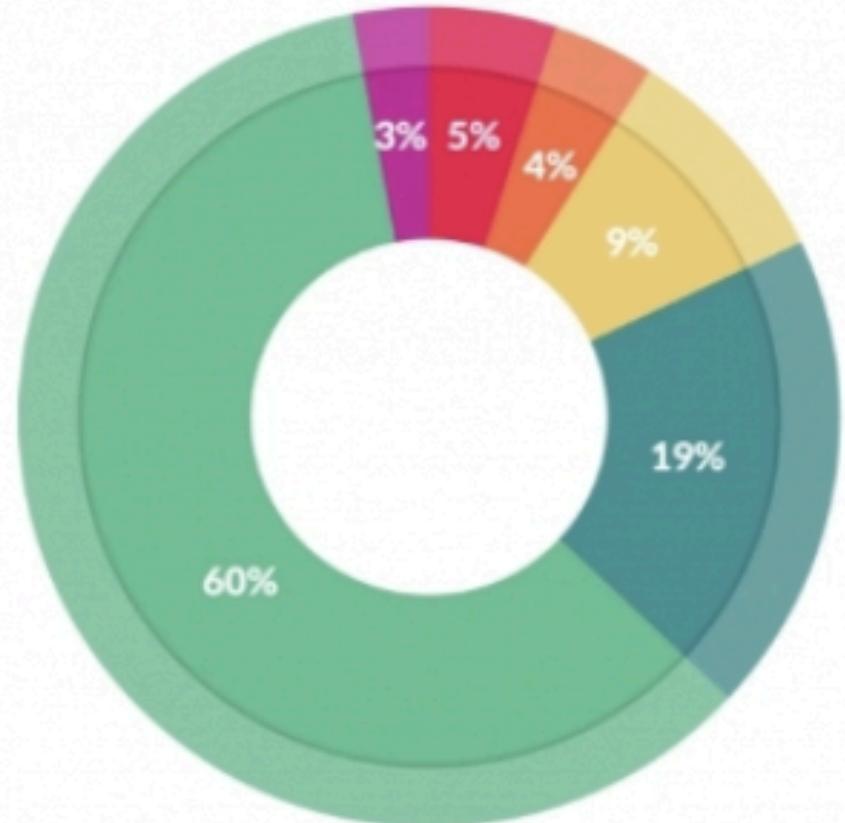


推薦延伸閱讀

痞客邦-iT邦2019鐵人賽：為什麼特徵工程很重要
網頁連結

- 本文主要在描述現實中資料科學工作的時間比重(右圖)，其中大部分的時間在於資料清理，少部分為特徵探勘，雖然這兩部份都是特徵工程，但在學習階段與實務階段，比重卻有著天壤之別。

資料科學家所花費工作項目時間佔比



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

- 創造訓練資料集: 3%
- 清理並組織資料: 60%
- 搜集資料: 19%
- 特徵探勘: 9%
- 優化機器學習演算法: 5%



解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

