Day 2

資料清理數據前處理

EDA-1/讀取資料

資料簡介：房貸風險預測

資料來源Source

[描述]
Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes **use of a variety of alternative data--including telco and transactional information**--**to predict their clients' repayment abilities**.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. **Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.**

資料簡介：房貸風險預測

Questions	Explanation
為什麼這個問題重要？	許多人因為沒有信用歷史， 所以沒辦法聲請貸款 → 這群人常會轉向風險較高的放款者 → 可能導致這群人的生活狀況更糟 → 如果這群人可以接受正向的幫助，他們將能步入良好正常生活 Home Credit 想透過放寬貸款條件，提供給這群人可以有好的借貸經驗 → 但即使放寬貸款條件，公司仍不能接受嚴重呆帳 (未還款) 發生 → 預測還款能力，讓公司可以在放寬貸款條件下，仍不致有貸給無法還債者。
資料從何而來？	信用局 (Credit Bureau) 調閱紀錄、 Home Credit 內部紀錄 (如過去借貸狀況、信用卡狀況)
資料的型態是什麼？	[Data] 皆為結構化資料：數值、類別資料
我們可以回答什麼問題 ？ 問題：指標	[Evaluation]分類問題, 預測各個客戶 ID 是否會還款，以還款機率 (0 ~ 1) 作為最終輸出 以 Area Under the ROC curve (ROC) 評估 [註1]

註1：在 AUROC, 0.5 代表隨機猜測, 越趨近於 1 則代表模型預測力越好

資料的樣子是什麼？

我們有多少資料

- 多少個資料來源？資料的格式是什麼？資料之間關係是什麼？
- 資料欄位的意義？每一 row 的意義？
- 仔細閱讀 Kaggle 上提供的資料說明

你會遇到很多具體的問題

- 怎麼讀資料？在 Python 做資料前處理，我們第一步就是引入常用的套件
  - pandas：用於讀取以及管理資料
  - numpy：用於數學函數的運算
- 有多少筆資料？有多少個欄位？
- 有沒有遺失值等等

這些問題的本質其實是在了解資料，我們稱為
「探索式資料分析」(Exploratory Data Analysis)

什麼是EDA？

01

初步透過**視覺化/統計工具**進行分析，達到三個主要目的

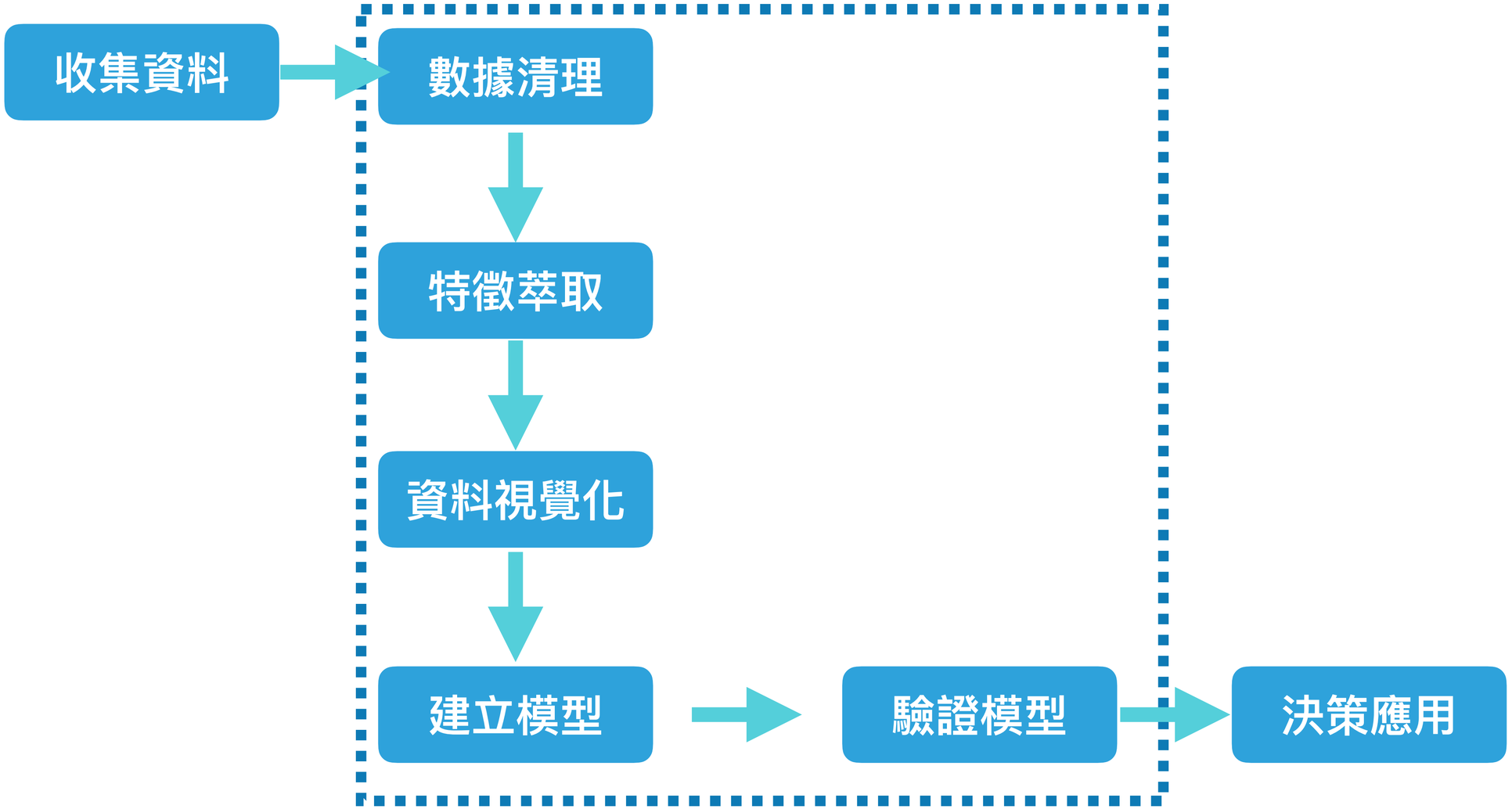
- **了解資料**
  - 獲取資料所包含的資訊、結構和特點
- **發現 outliers 或異常數值**
  - 檢查資料是否有誤
- **分析各變數間的關聯性**
  - 找出重要的變數

02

從 EDA 的過程中觀察現象，檢查資料是否符合分析前的假設

- 可以在模型建立之前，先發現潛在的錯誤
- 也可以根據 EDA 的結果來調整分析的方向

數據分析的流程



Day2 Homework 作業

- 請下載本次馬拉松建議的 Kaggle 資料或準備好自己的資料
- 如採用 Kaggle 資料，請透過HomeCredit_columns_description.csv，了解各個欄位的意義
- 參考 Day 2 的 ipynb 範例，了解如何讀取資料並可自行嘗試對資料進行操作 (操作的部分請參考該基礎教材)

補充資料 推薦閱讀文章

1. 探索式資料分析簡介 by 吳漢銘老師
2. What is Exploratory Data Analysis?
3. DataCamp: Pandas-foundations
   (需註冊登入，第一個 chapter 是免費的，建議可用來預習 pandas)