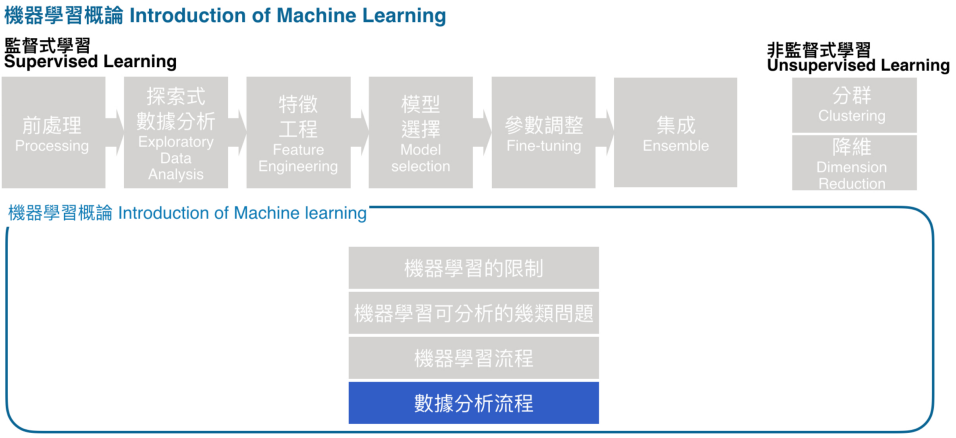
D4：EDA/讀取資料與分析流程

課程閱讀 本日作業 問題討論 學習心得(完成)

EDA/讀取資料與分析流程

知識地圖



課程講解

資料簡介：房貸風險預測

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.**

[資料來源]

> Home Credit Default Risk | Kaggle
> Can you predict how capable each applicant is of repaying a loan?
> www.kaggle.com

房貸風險預測問題與說明

問題一：為什麼這個問題重要？

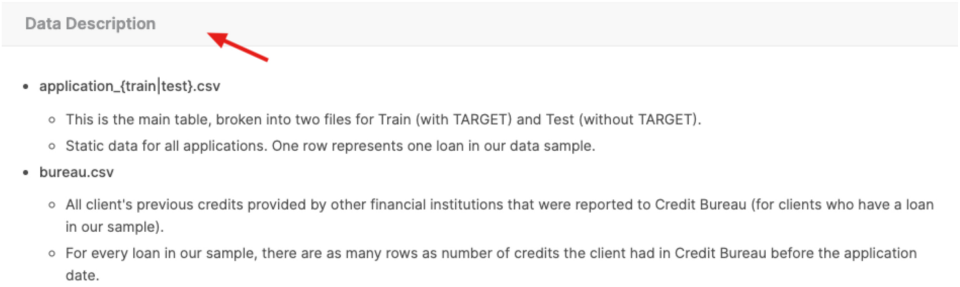
說明：許多人因為沒有信用歷史， 所以沒辦法申請貸款 → 這群人常會轉向風險較高的放款者 → 可能導致這群人的生活狀況更糟 → 如果這群人可以接受正向的幫助，他們將能步入良好正常生活。
Home Credit 想透過放寬貸款條件，提供給這群人可以有好的借貸經驗 → 但即使放寬貸款條件，公司仍不能接受嚴重呆帳 (未還款) 發生 → 預測還款能力，讓公司可以在放寬貸款條件下，仍不致有貸給無法還債者。

問題二：資料從何而來？

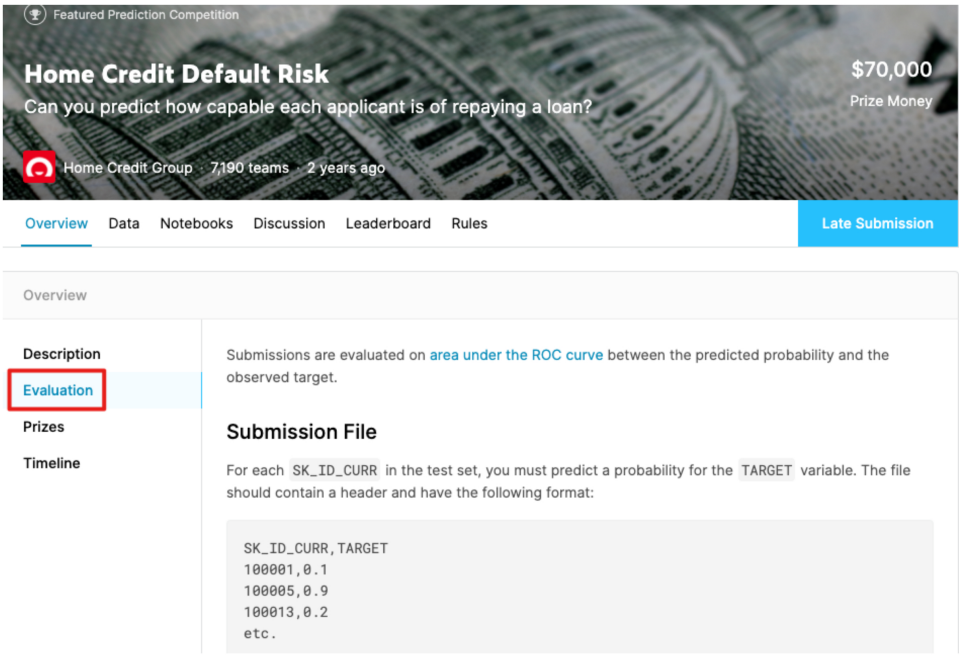
說明：信用局 (Credit Bureau) 調閱紀錄、 Home Credit 內部紀錄 (如過去借貸狀況、信用卡狀況)

問題三：資料的型態是什麼？

Data Description 皆為結構化資料：數值、類別資料

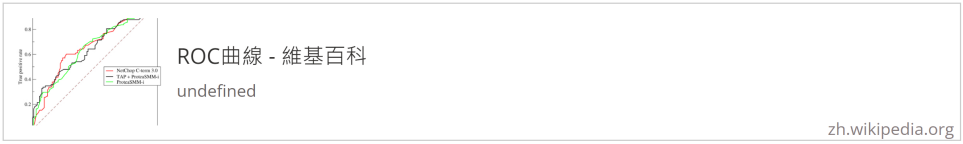


問題四：我們可以回答什麼問題 ？問題：指標



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]

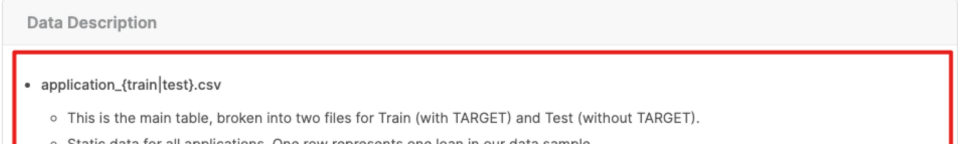
註1：在 AUROC, 0.5 代表隨機猜測, 越趨近於 1 則代表模型預測力越好



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



- bureau.csv
  - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
  - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- bureau_balance.csv
  - Monthly balances of previous credits in Credit Bureau.
  - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- POS_CASH_balance.csv
  - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.

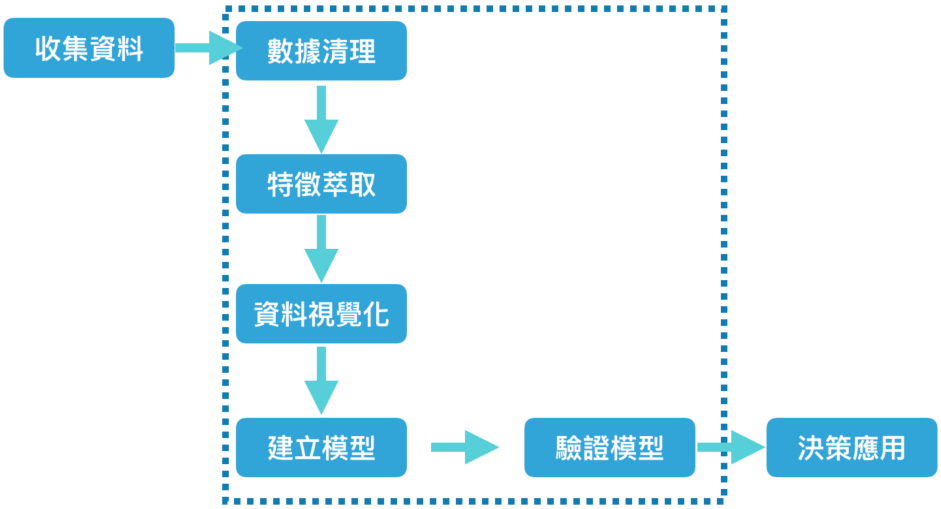
什麼是EDA？

- 初步透過**視覺化/統計工具**進行分析，達到三個主要目的

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

- 從 EDA 的過程中觀察現象，檢查資料是否符合分析前的假設

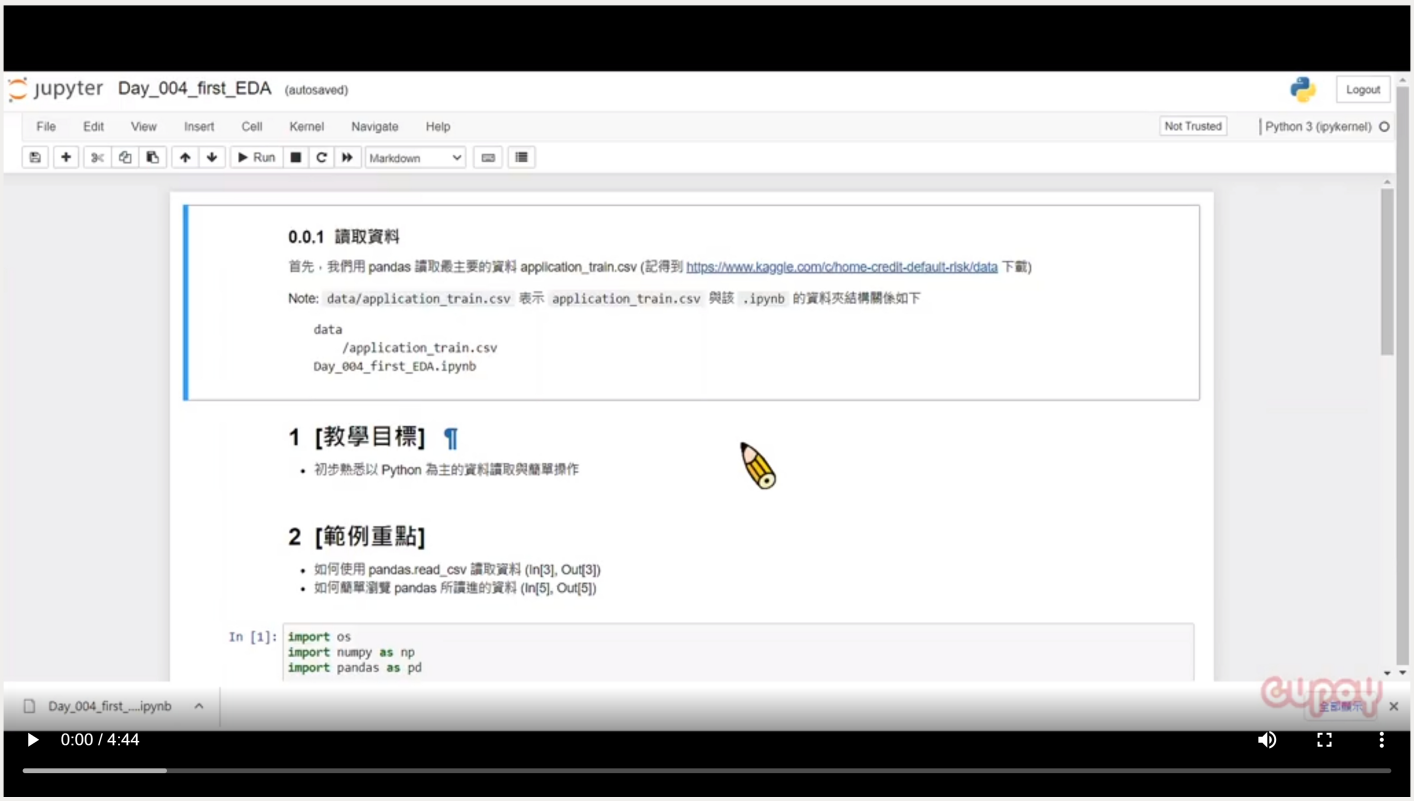
數據分析的流程



解題時間

Sample Code ＆作業
開始解題

範例影片



下一步：完成作業