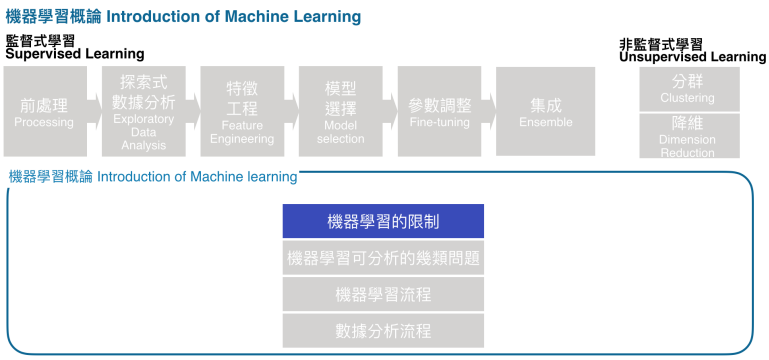
D1：資料介紹與評估資料

簡報閱讀 範例與作業 問題討論

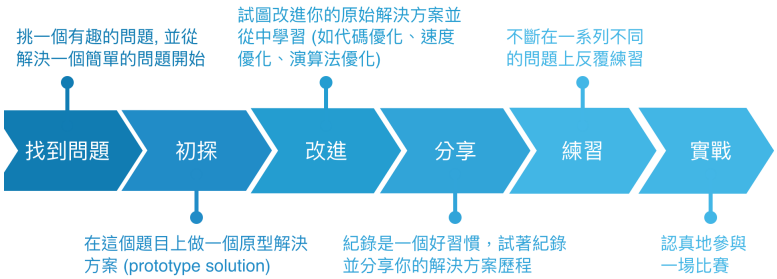
資料介紹與評估資料

知識地圖



- **準備進入資料科學領域的概念與流程與關鍵**

學習路徑



首次面對資料，我們應該思考哪些問題？

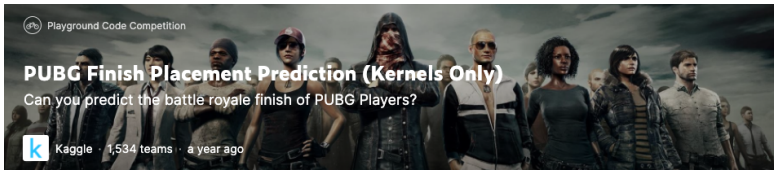
以下介紹四個問題

問題一：為什麼這個問題重要？(Why it is important?)

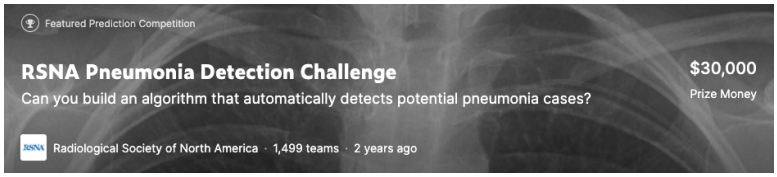
說明：

1. 好玩
2. 企業的核心問題
3. 公眾利益 / 影響政策方向
4. 對世界很有貢獻

舉例：



1. 預測生存 (吃雞) 遊戲誰可以活得久, PUBG
2. 用戶廣告投放, ADPC



問題二：資料從何而來？(Where do data come from?)

說明：

1. 來源與品質息息相關
2. 根據不同資料源，我們可以合理的推測/懷疑異常資料異常的理由與頻率

舉例：

資料來源如網站流量、購物車紀錄、網路爬蟲、格式化表單、Crowdsourcing、紙本轉電子檔

> Crowdsourcing - Wikipedia
> Crowdsourcing is a sourcing model in which individuals or
> en.wikipedia.org

問題三：資料的型態是什麼？(What are they?)

說明：

1. 結構化資料需要檢視欄位意義以及名稱
2. 非結構化資料需要思考資料轉換與標準化方式

舉例：

1. 結構化：數值、表格...等
2. 非結構化：圖像、影片、文字、音訊...等

問題四：我們可以回答什麼問題 ？問題：指標 (What is our goal?)

說明：每個問題都應該要可以被驗證 → 有一個可供衡量的數學評估指標 (Evaluation Metrics)

舉例：常見的衡量指標如下

1. 分類問題：正確率、AUC、MAP...等
2. 迴歸問題：MAE、RMSE...等

範例一：我們應該要 / 可以回答什麼問題？

生存 (吃雞) 遊戲

- 玩家排名：平均絕對誤差 (Mean Absolute Error, MAE)

- 怎麼樣的人通常活得久/不久 (如加入遊戲的時間、開始地點、單位時間內取得的資源量, ...) → 玩家在一場遊戲中的存活時間：迴歸 (Mean Squared Error, MSE)

範例二：我們應該要 / 可以回答什麼問題？

廣告投放

Receiver Operating Curve, ROC

- 哪些素材很好/不好 → 廣告點擊預測 → 預測在版面上的哪個廣告會被點擊：ROC / MAP@N (eg. MAP@5, MAP@12)

重要知識點複習

- 初入資料科學的探索流程
- 找到問題 → 初探 → 改進 → 分享 → 練習 → 實戰
- 面對問題需要思考的關鍵點
  a. 為什麼這個問題重要
  b. 資料從何而來
  c. 資料的型態是什麼
  d. 回答問題的關鍵指標是什麼

解題時間

Sample Code &作業
開始解題

下一步：閱讀範例與完成作業