

Credit Card Fraud Detection

First Group Presentation

2024/04/09

Outline

01

組員分工

02

專案背景說明

03

資料說明

04

預期目標

05

執行計畫

06

資料前處理

07

甘特圖

01. 組員分工



PM

林貫原

專案進度管理
決策管理
文件管理

PM

許政揚

協助進度管理
網頁雛形
小組報告

TS

易祐辰

資料視覺化
資料搜集

小組組員

職位分工表

DS

楊廷紳

程式管理
資料分析

DS

周昱宏

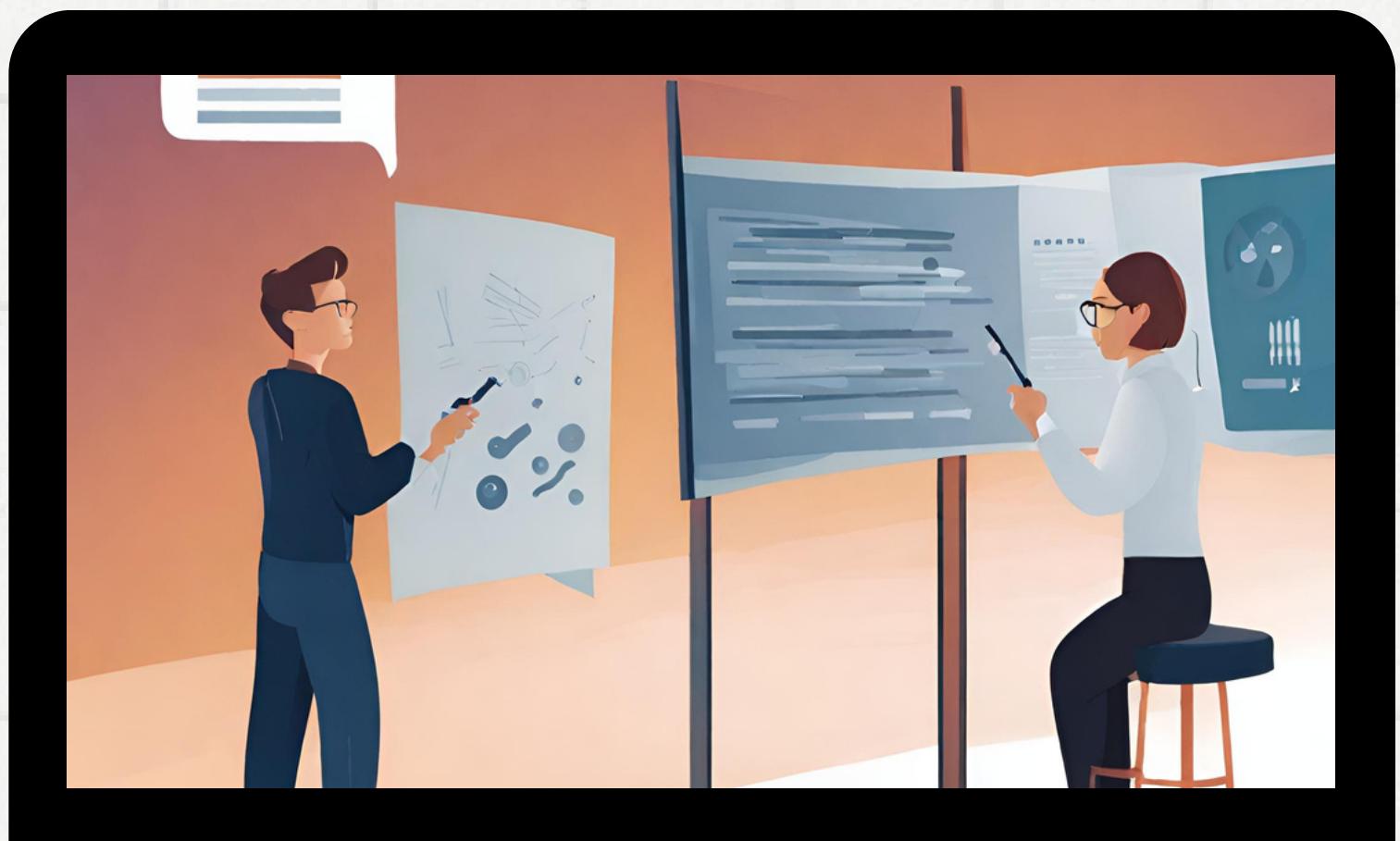
資料清洗
資料分析

SD

留筠雅

系統架構
介面架構
使用說明書撰寫

02. 背景說明



前言

- 研究信用卡違約的統計資料，有助於了解違約行為的特徵、趨勢和影響因素，進而提供預測和管理違約風險的依據。
- 對於**金融機構**評估信用風險、制定信用政策以及發展適當的風險管理策略至關重要。
- 對於**消費者**而言，了解違約的可能原因和風險因素，可以引導其更加理性地使用信用卡，避免陷入財務困境。



研究背景

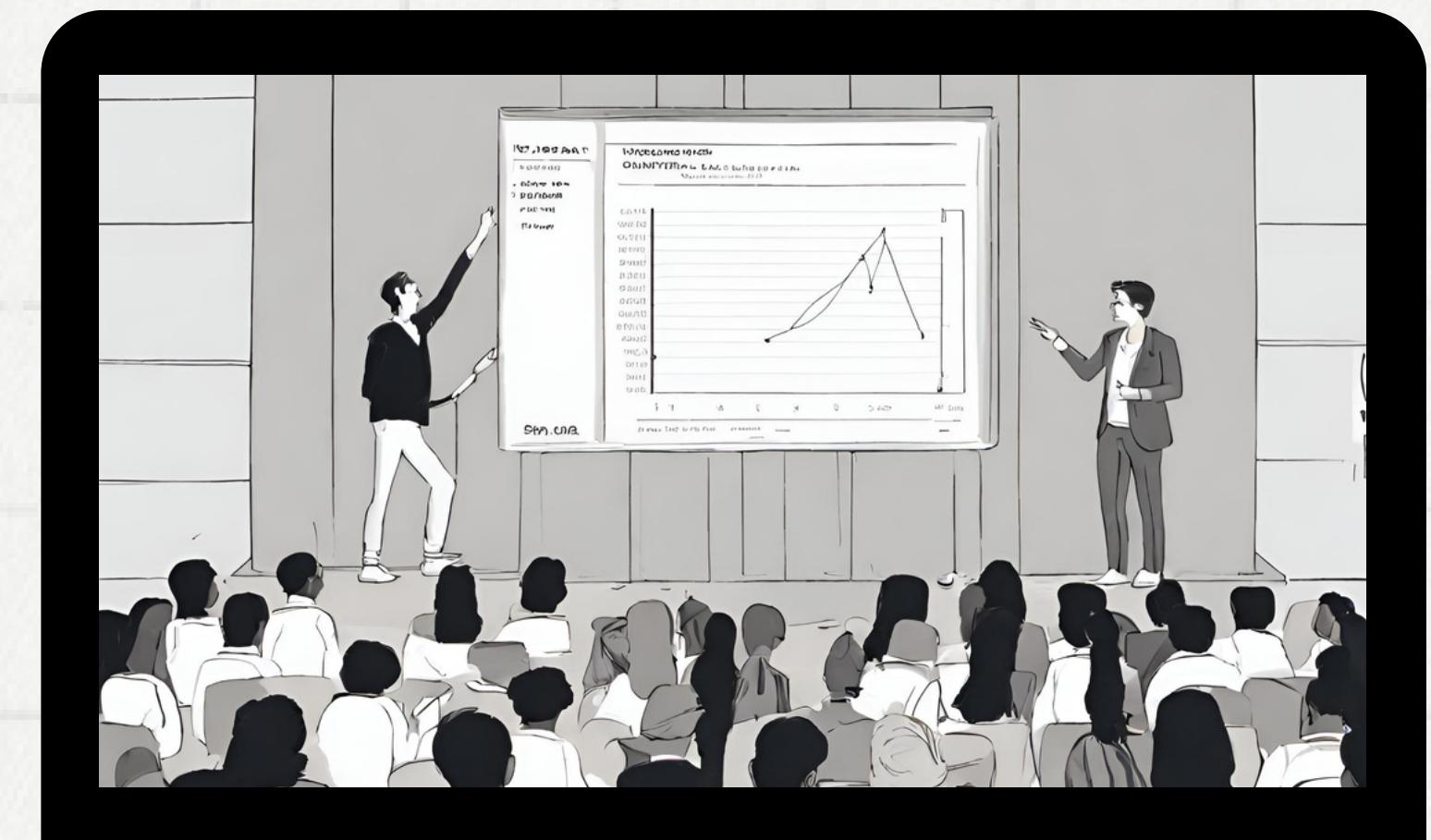
1. 經濟環境的影響
2. 個人特徵和行為模式
3. 信用卡產品本身的特性、使用方式以及支付習慣
4. 法規和監管政策對信用卡市場的規範



研究目的

1. 挖掘資料中有用的變數
2. 找出違約客戶的分群特徵
3. 預測未來客戶是否會有信用卡違約

03. 資料說明

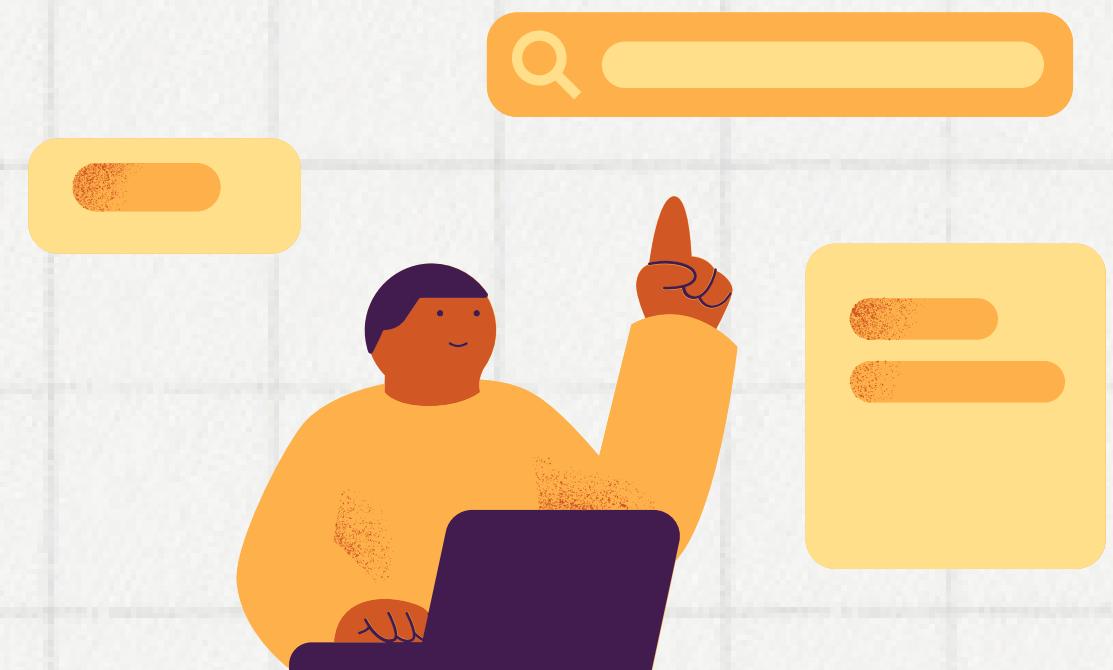


資料來源

資料來自[kaggle](#)上的信用卡違約預測 (Credit Card Fraud Detection)

且資料集是由[班加羅爾國際資訊科技學院](#)收集而來

(International Institute of Information Technology Bangalore)



資料檔案總覽

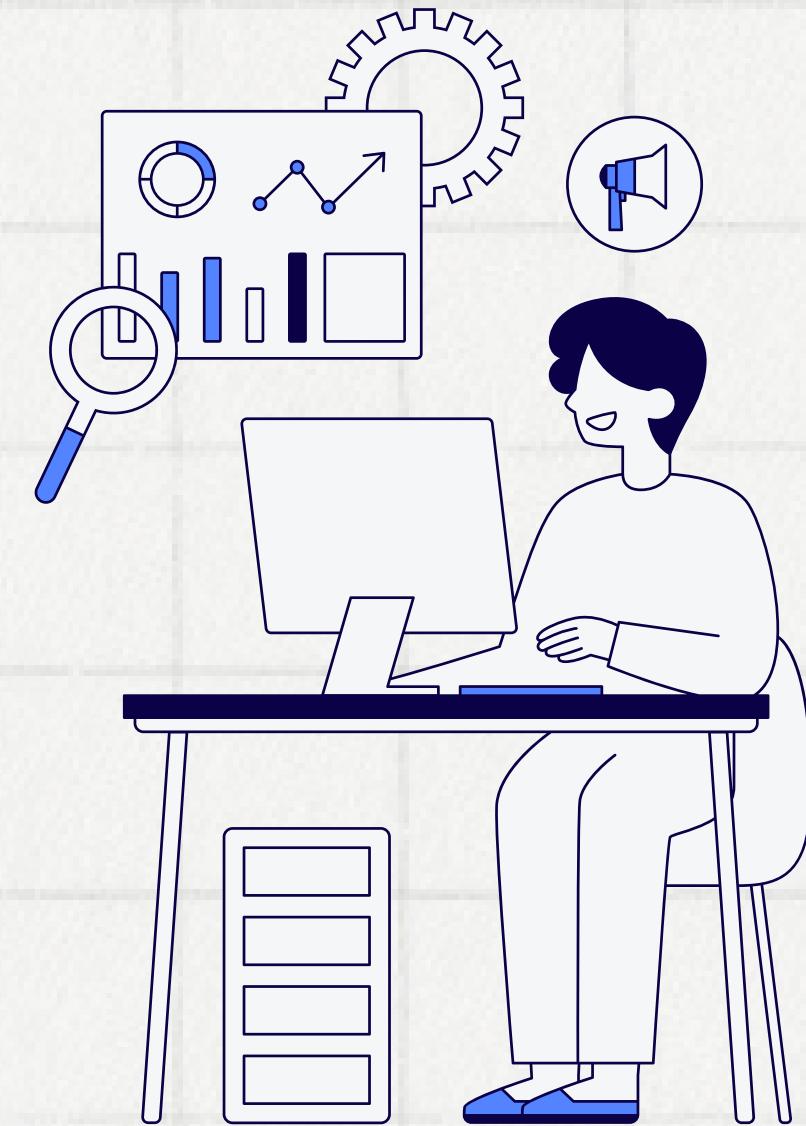
資料檔案	資料目的
creditcard_train	主要使用之訓練集 包含客戶個人資料
creditcard_test	主要使用之測試集
columns_description	變數名稱解釋
previous_application	客戶申請貸款前的資料
creditcard_test_true	用來確認答案

資料分群

總共122 個變數。分為訓練及與測試集

訓練集資料共有 306611 筆

測試集資料共有 900 筆。



資料展示

變數

類別型	Nominal	合約類型、收入狀態、家庭狀態、居住地等
	Ordinal	客戶居住地評分、客戶申請貸款之星期
數值型	Interval	客戶之子女數、客戶擁有車輛數、家庭成員數等
	Ratio	客戶收入、貸款信用額度、貸款年金等

資料展示

不平衡資料

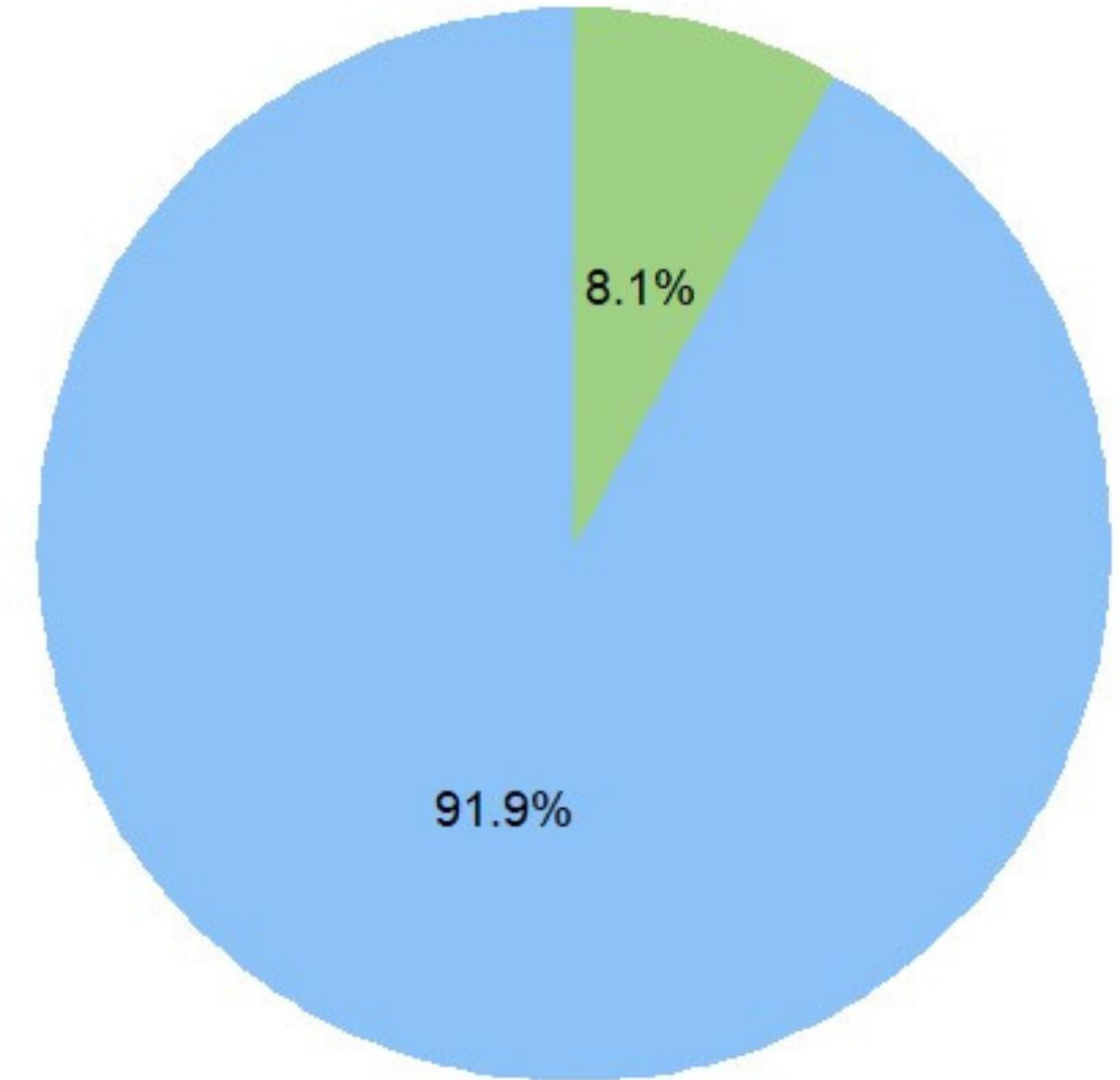
變數 ‘Target’ 的值為0、1
代表客戶是否違約

在不平衡的分類問題上若達到90%

甚至是99%的分類準確率

一般而言，準確率在不平衡資料中幫助不大

Distribution of Target Variable



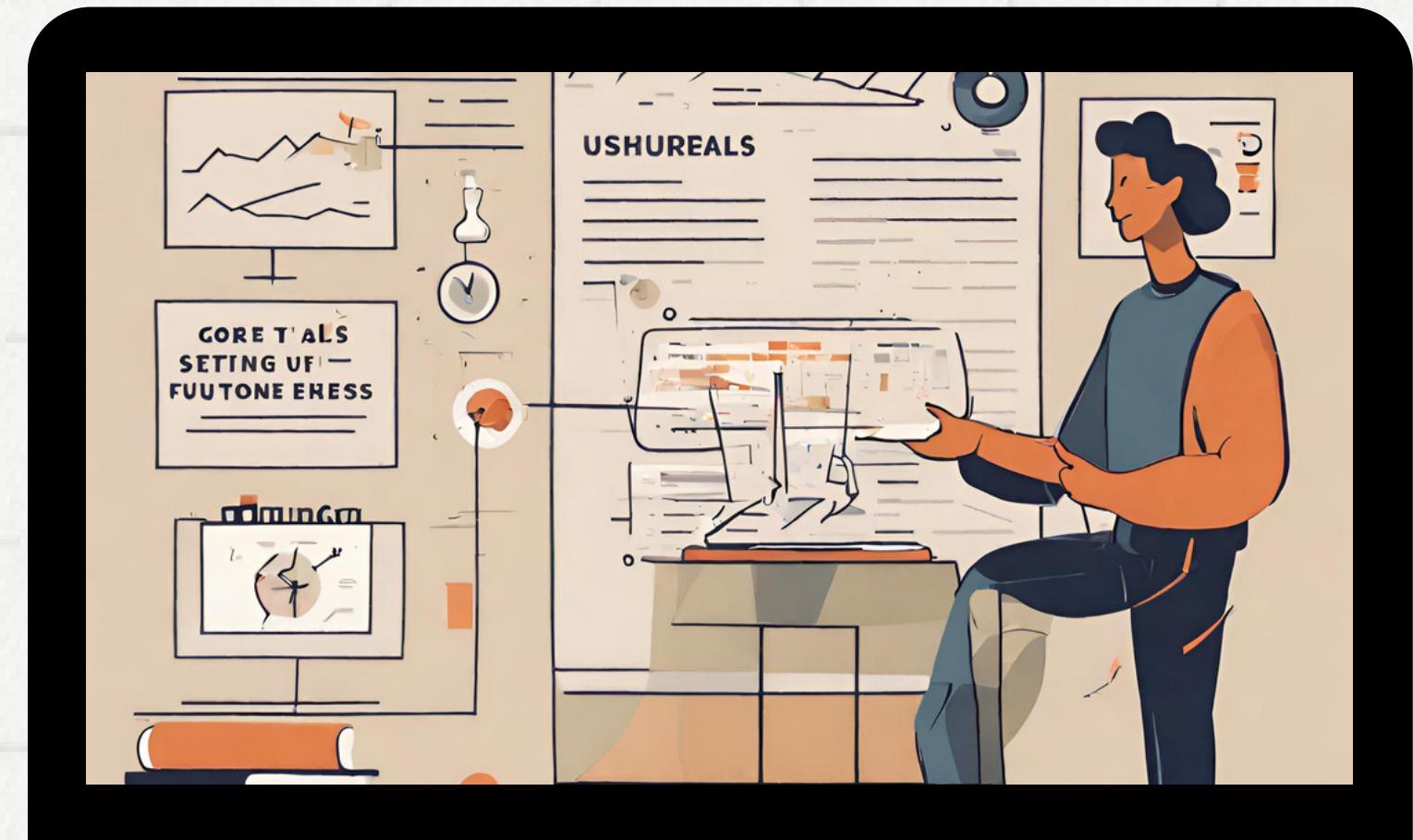
Target No Yes

資料展示

遺失值

遺失值類型	遺失值產生之原因
完全隨機遺失	遺失值的產生皆為隨機的
隨機遺失	遺失值的產生並非完全隨機，但可觀察資料得到有些變數與資料缺失是有相關的。因此當控制這些變數，缺失資料的情形就可視為隨機。
非隨機遺失	遺失值的產生並不隨機。當資料不為上述兩者時，不能忽視此處的資料遺失，必須進一步確認研究資料為何出現遺失值。

04. 預期目標



本研究旨在以客戶基本資料與交易資料為基礎，探討信用卡違約風險的判斷方法

①

利用資料探勘的方法，
找出影響違約與否之變數

②

獲得客戶分群以及是否違約的特徵之視覺化

③

建立網頁系統，提供使用者輸入客戶資料後
利用系統判斷該客戶違約的風險為何

05. 執行計畫



過去的研究

來源為Kaggle中同一筆資料集，他人所做的研究

<https://www.kaggle.com/code/mnhcngbi/fraud-detection>

步驟1： 資料前處理-利用平均數、中位數、眾數等對遺失值直接進行插補

步驟2： 產生與目標變數交互訊息的評分函數做特徵選取

步驟3： 挑選前十名的變數

步驟4： 使用四種不同的模型觀察其準確率

羅吉斯回歸

k近鄰分群

高斯貝氏分群

決策樹分群

執行計畫

資料前處理

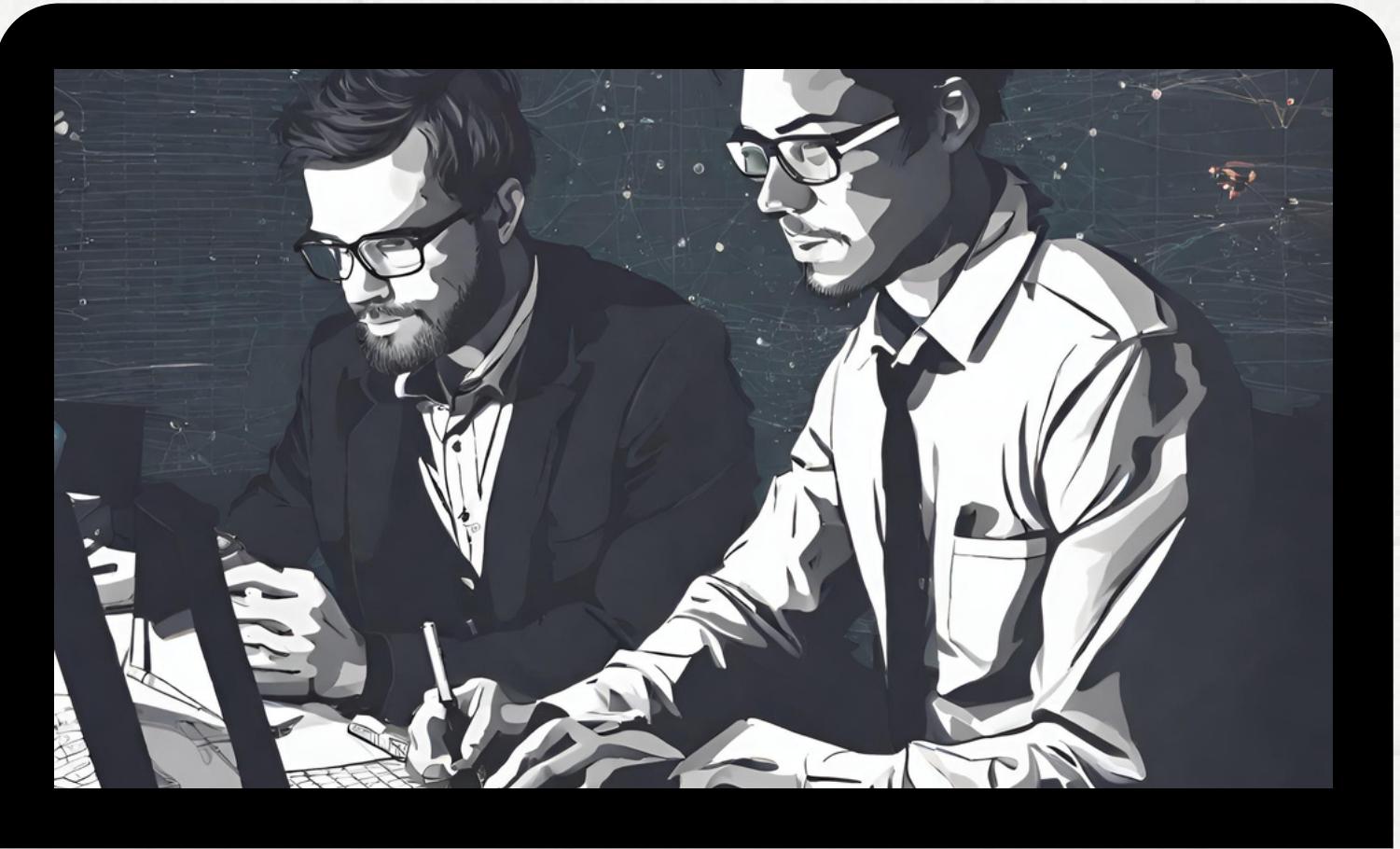


參考Kaggle以及網路文獻對於信用卡違約的探討



根據專案得到的資料檔與變數解釋

06. 資料前處理



原始資料

訓練集

creditcard_train



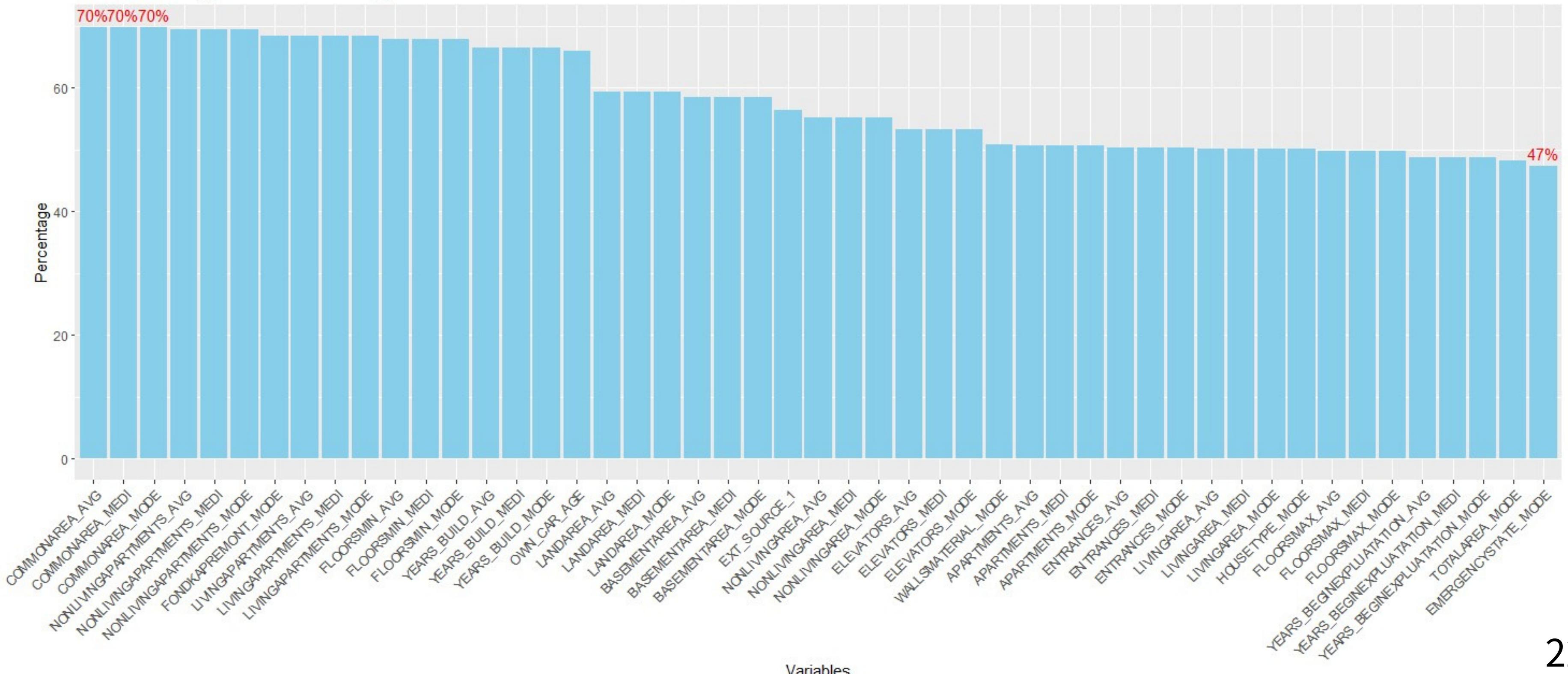
306611筆資料

122個變數

去除遺失值超過32%的變數

篩選出49個變數
保留： $122 - 49 = 73$ 個變數

Percentage of missing values in variables > 32%



將FLAG_Document1~20
變數
合併成1個變數



共有20個FLAG_Document
保留： $73-20+1=54$ 個變數

將AMT_REQ_CREDIT_...變
數合併成1個變數



共6個變數
保留： $54-6+1=49$ 個變數

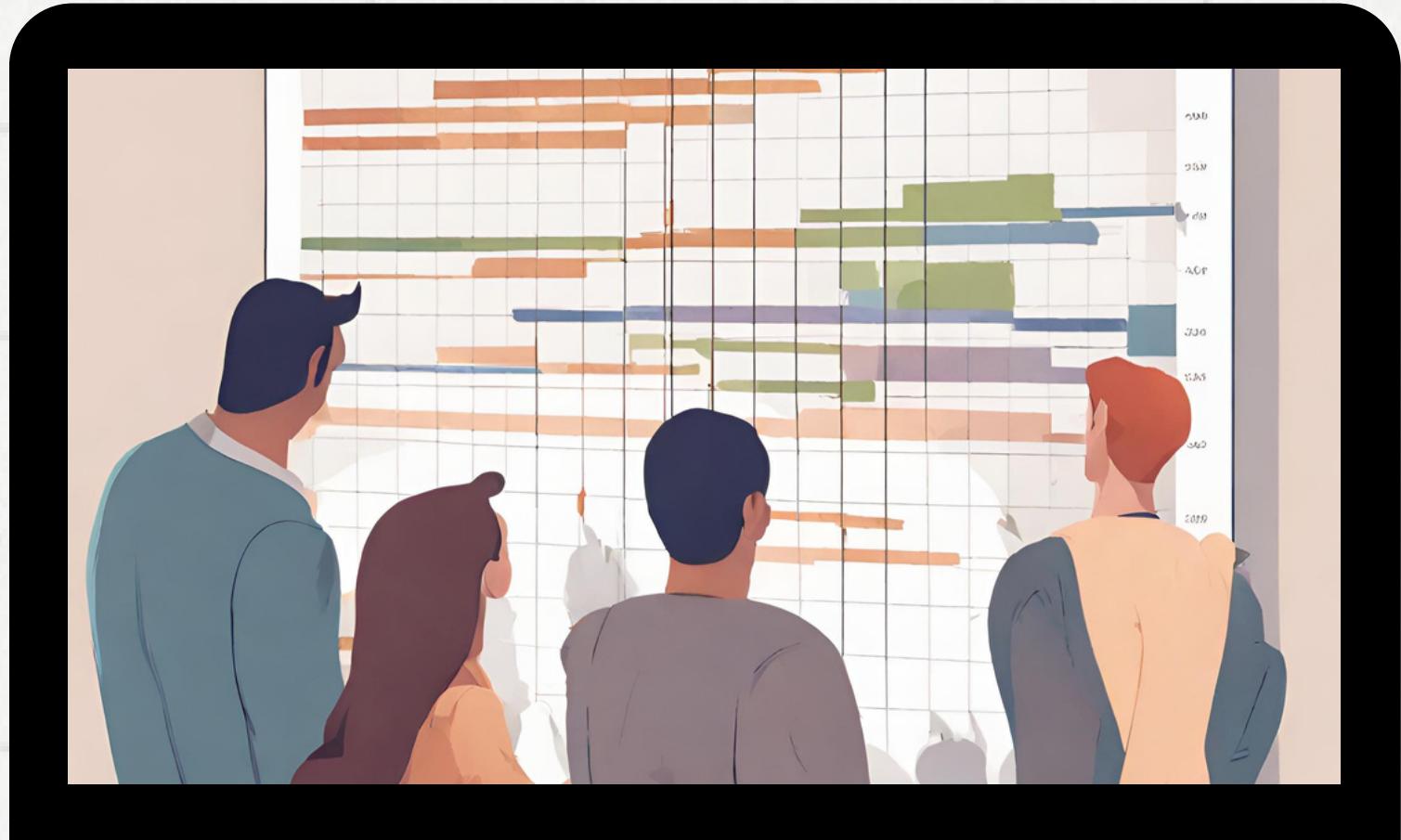
增加一項變數
missing_ratio



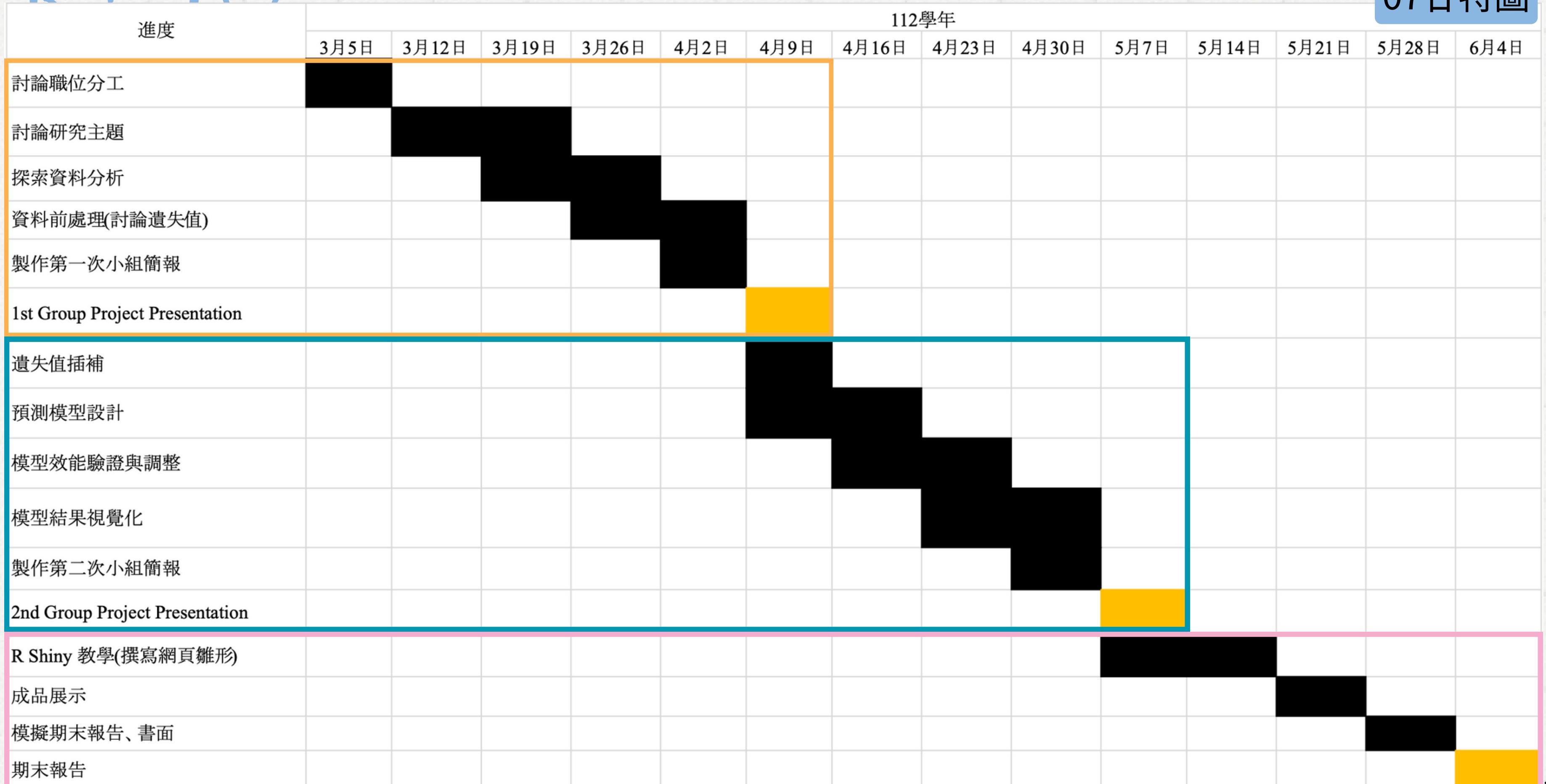
保留： $49+1=50$ 個變數

其中missing_ratio為遺失值比例

07. 甘特圖



07甘特圖





Thank you