信用卡盜刷偵測
模型分享

Tim Wang

2020/9/24

流程

- 資料集概況
- 建模規劃
- 資料探索分析（Exploratory Data Analysis）
- 特徵工程（Feature Engineering）
- 建模（Modelling）
- 問題與嘗試
- 後續方向

資料集概況

- 90天內，1,521,787筆信用卡交易紀錄， 23欄位

交易序號

歸戶帳號（95,214）
交易卡號（129,413）

交易金額
授權日期
授權時間
分期註記（2）
分期期數

消費國別（103）
消費城市（5,698）
消費幣別（72）
MCC_CODE（434）
特店代號（89,316）
收單行代碼（6,051）

交易類別（7）
交易型態（11）
支付型態（9）

網路交易註記（2）
超額註記（2）
Fallback註記（2+1）
3DS註記（2+1）

狀態碼（5）

盜刷註記（2）
20,355筆盜刷紀錄
盜刷率約1.3%

建模規劃

變數

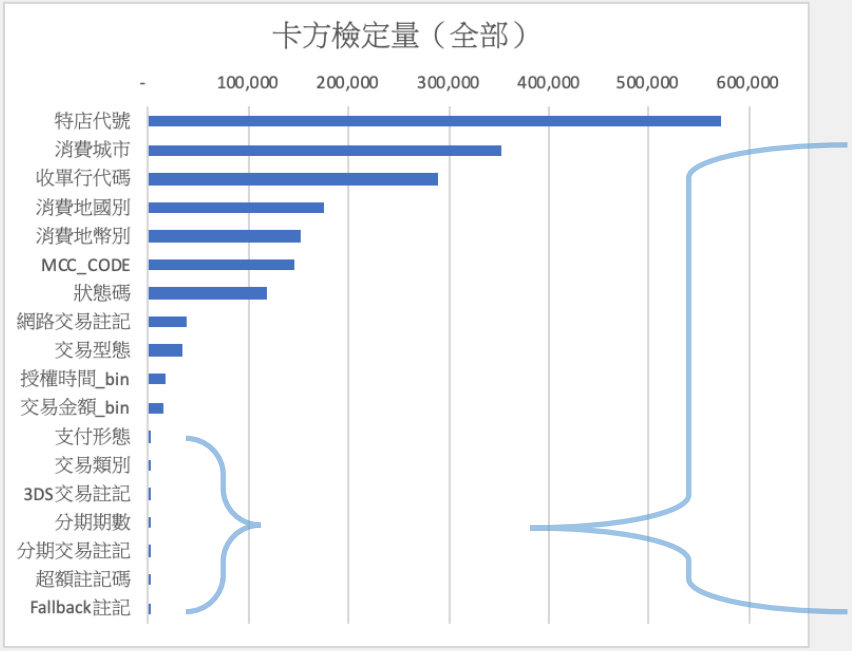
1. 該筆交易資料
2. 該帳號/卡號歷史「正常」交易資料

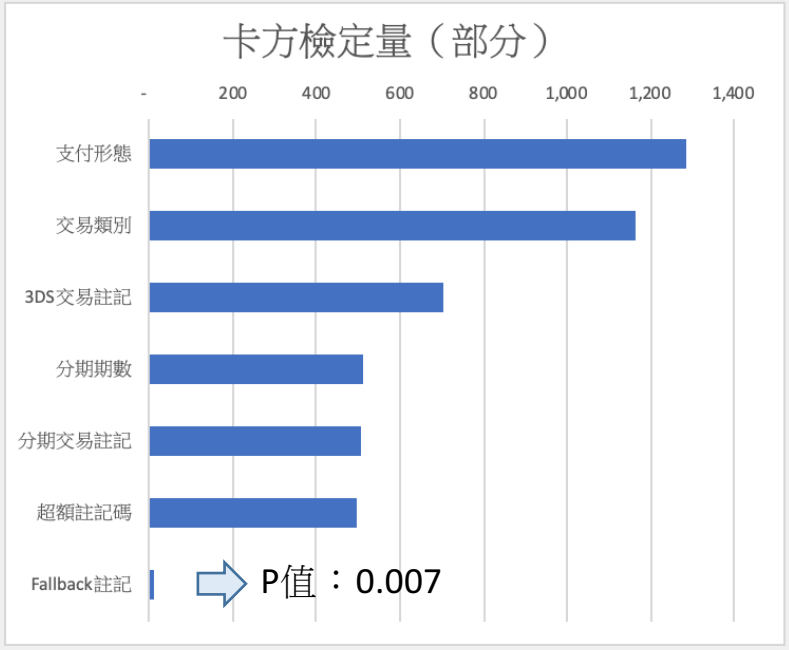
演算法

1. 監督式
   Random Forest Classifier
   Gradient Boosting Classifier
   Ada Boost Classifier
   Neural Network
2. 非監督式（isolation forest）

資料探索分析（一）

主要變數卡方檢定

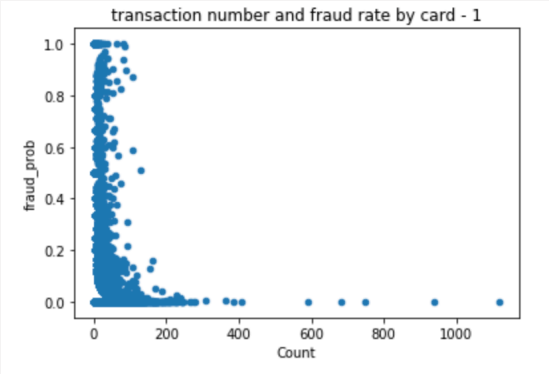




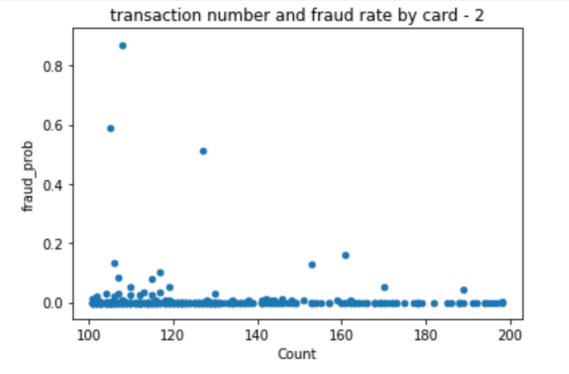
資料探索分析（二）
帳號/卡號盜刷風險分析

- 被盜刷的機率因【帳號/卡號】而異（75%以上帳號僅有一個卡號）

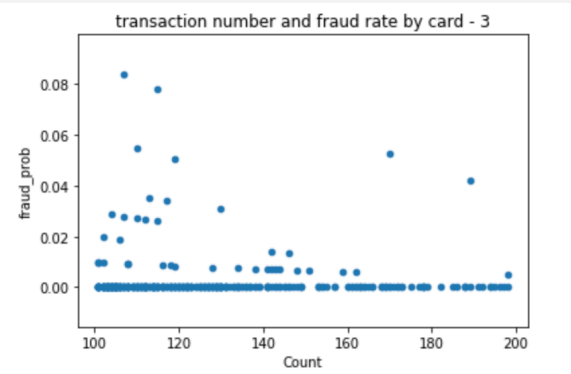
全數卡號



100 < 刷卡次數 < 200



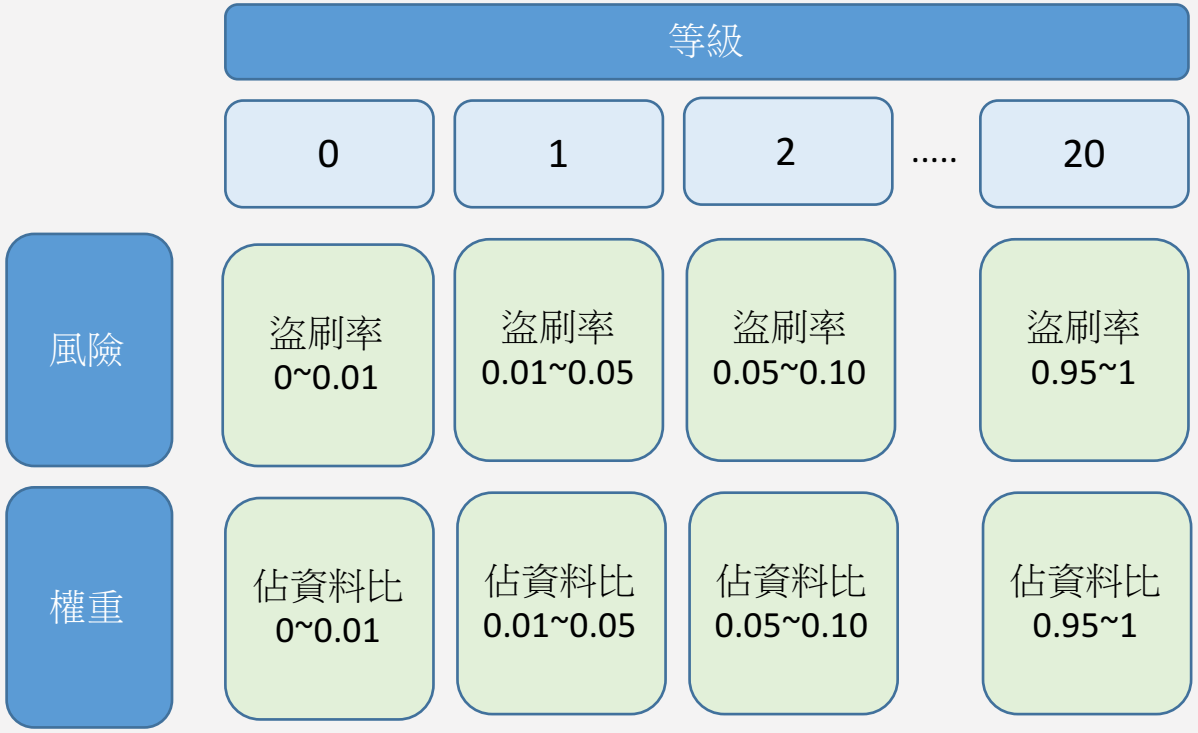
100 < 刷卡次數 < 200
盜刷率 < 0.1



特徵工程（一）

三個類別以上的名目變數，以風險等級與權重等級重新定義

Fallback註記（2+1）
3DS交易註記（2+1）
狀態碼（5）
交易類別（7）
交易型態（11）
支付型態（9）
消費國別（103）
消費城市（5,698）
消費幣別（72）
MCC_CODE（434）
特店代號（89,316）
收單行代碼（6,051）



特徵工程（二）

回顧每筆交易所屬帳號、卡號的歷史資訊

主要資訊

- 過去盜刷機率
- 平均消費/此次消費金額、最大金額/此次消費金額
- 距上次消費天數、最接近消費時段
- 交易類別、交易型態、支付形態、分期期數、狀態碼、網路交易註記、 Fallback註記、 3DS交易註記、分期交易註記、超額註記碼、消費地國別、消費城市、消費地幣別、 MCC_CODE 、特店代號、收單行代碼

輔助資訊

- 過去全部刷卡次數
- 過去正常刷卡次數
- 過去正常刷卡次數分佈型態：
  - 距今天數平均、距今天數標準差（較佳）
  - 距今天數【0、25、50、75、100】百分位數

建模（一）

- 依【授權日期】先後拆分資料：

	訓練集	驗證集	測試集
授權日期	1 ... 70	71 ... 80	81 ... 90
交易筆數	1,184,411（78%）	169,529（11%）	167,847（11%）
盜刷率	1.4%	1.0%	0.9%

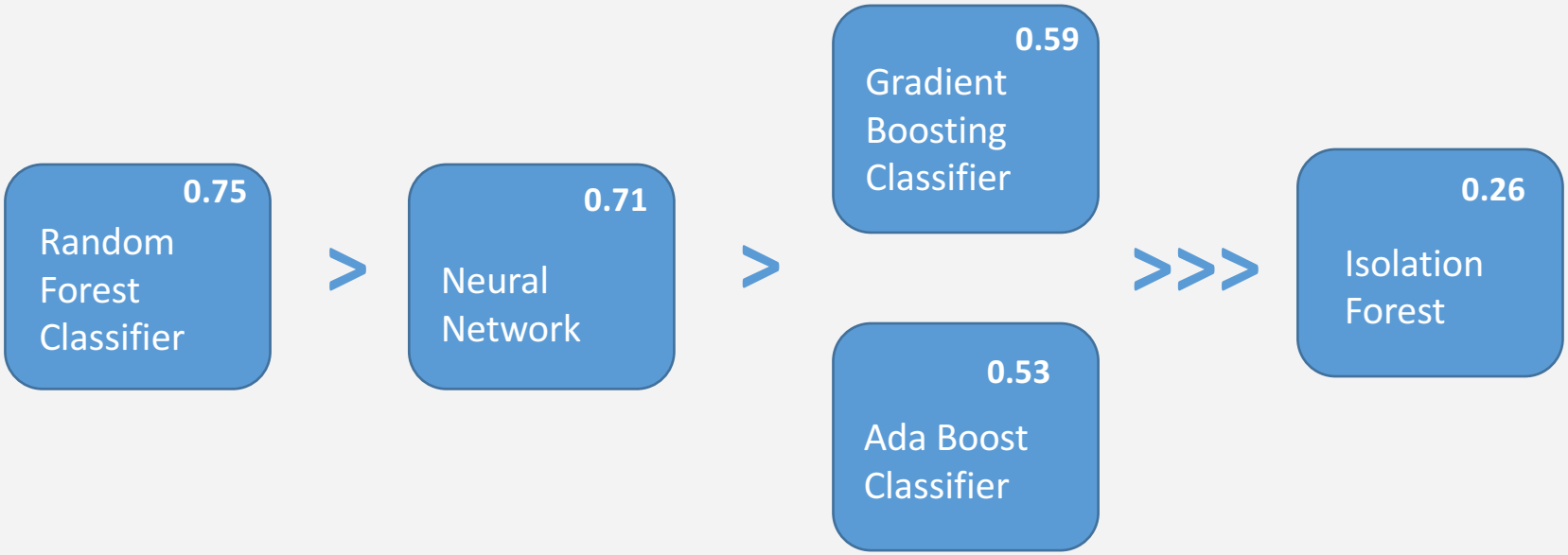
建模（二）

- 共採用變數81個

當前交易資料	歷史交易紀錄
• 一戶多卡、歸戶卡數 • 交易金額、授權時間、分期期數、分期交易註記、超額註記碼、網路交易註記 • 風險等級＋權重等級變數	• 帳號、卡號 • 主要資訊、輔助資訊

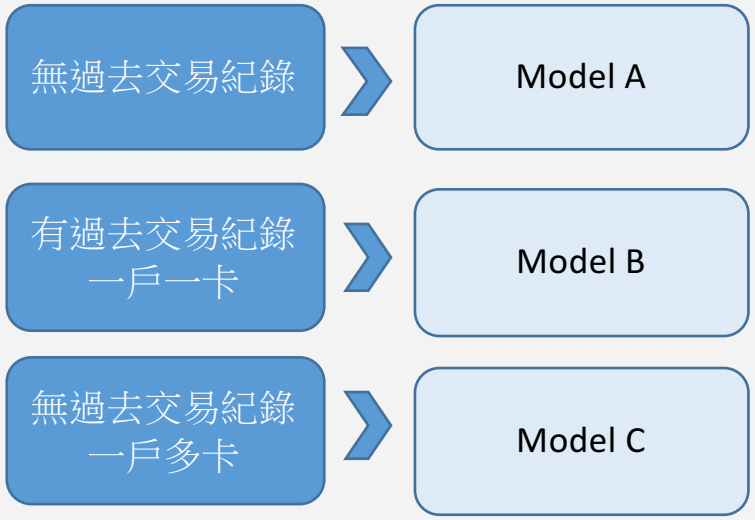
建模（三）

- 模型篩選（以驗證集 F1 Score 衡量）：



建模（四）

組合模型



驗證集
F1 Score
0.72

單一模型

驗證集
F1 Score
0.75

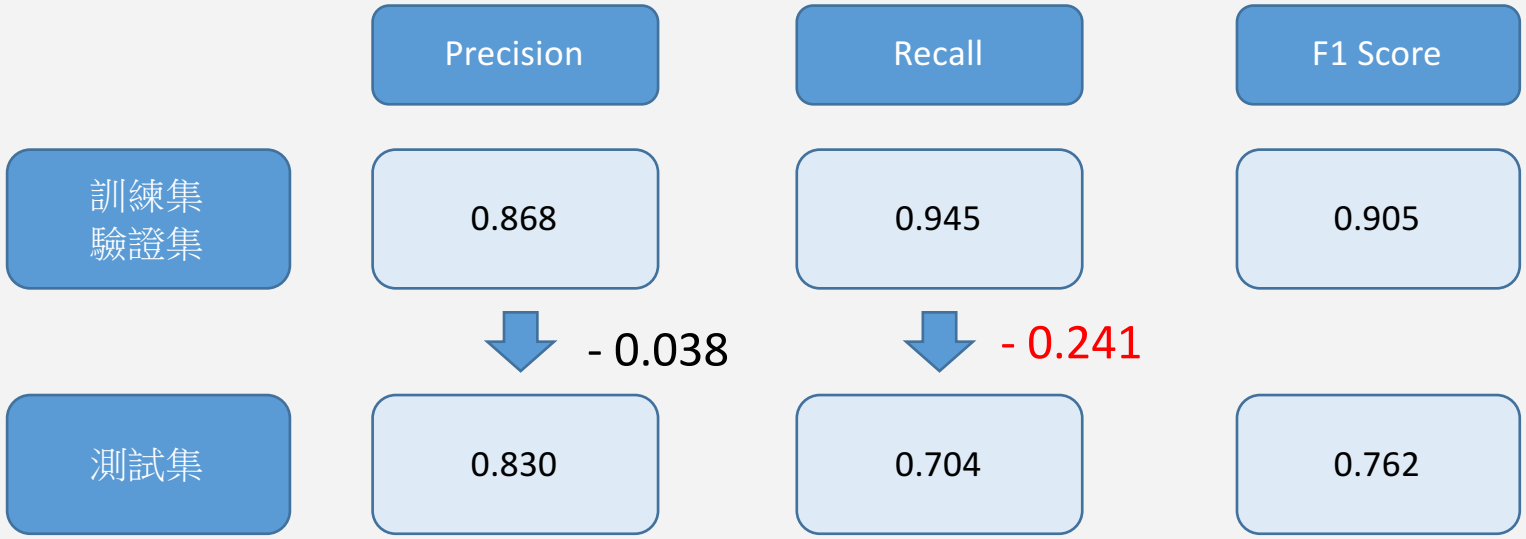
建模（五）- 最終結果

- 用原有設定，以【訓練集】＋【驗證集】的資料重新訓練模型，得到的【測試集】F1 Score為 0.762
- 81個解釋變數中，以該卡號過去遭到盜刷的機率所佔的權重最大（右表：前十大解釋變數）

排名	變數	權重
1	past_100_交易卡號_過去盜刷機率	14.0%
2	收單行代碼_FRAUD	8.7%
3	特店代號_FRAUD	8.3%
4	消費城市_FRAUD	7.7%
5	past_100_歸戶帳號_過去盜刷機率	4.6%
6	消費地國別_SHARE	4.4%
7	past_100_交易卡號_消費地國別	4.0%
8	消費地國別_FRAUD	3.6%
9	past_100_歸戶帳號_消費地國別	2.9%
10	MCC_CODE_FRAUD	2.6%
11-81	其他	41.6%
	總計	100.0%

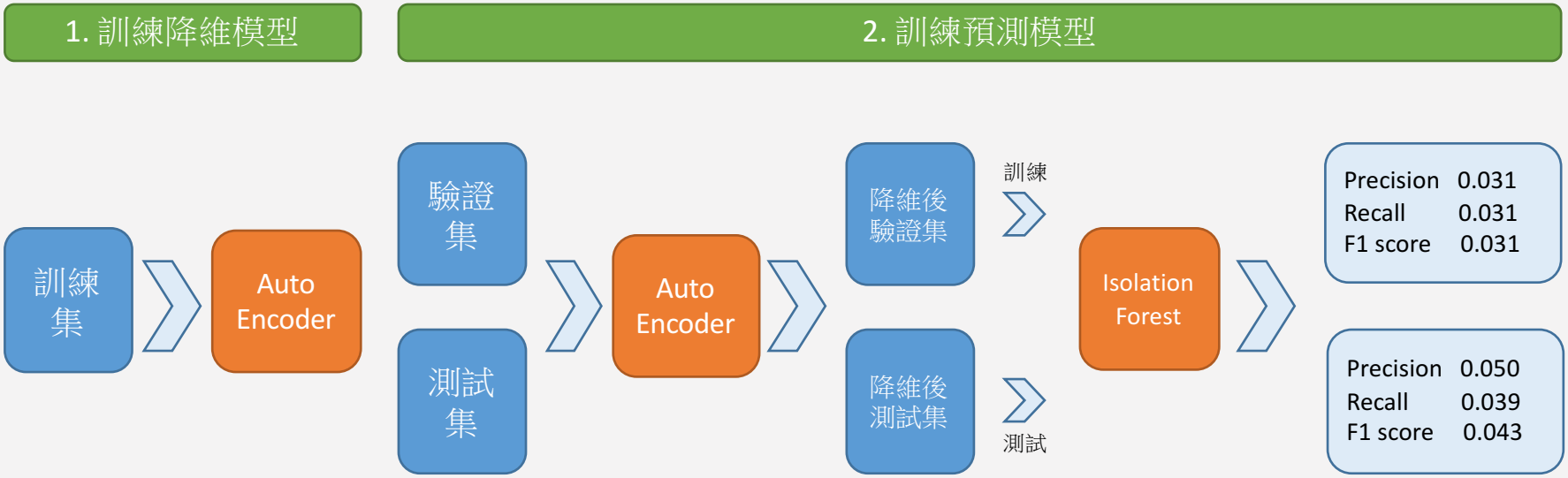
問題與嘗試（一）

- 盜刷型態不斷變化，用監督式學習建立的模型容易有漏網之魚



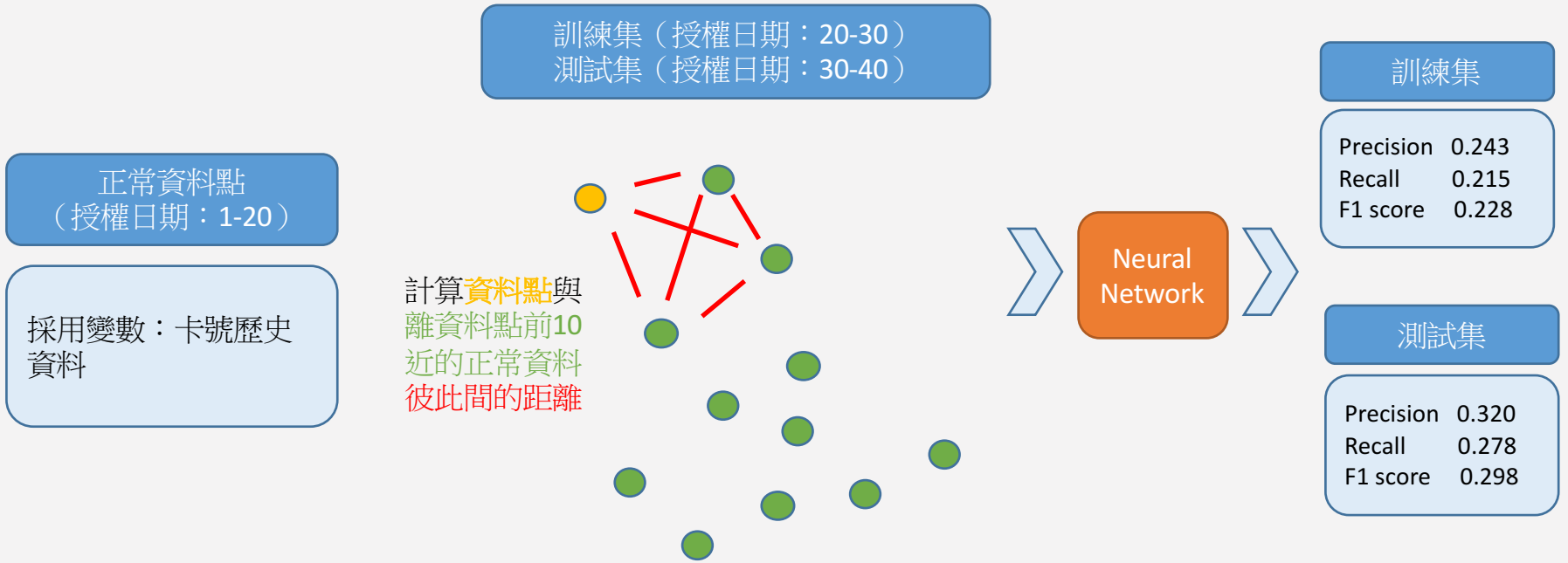
問題與嘗試（二）

- 降維後（81->10）再訓練非監督式學習模型



問題與嘗試（三）

- 以離正常資料點的距離當作變數，訓練預測模型



後續方向

- 變數
  - 加入卡主個人資訊（年齡、職業、收入、住所等）
  - 時間序列的歷史資料
- 方法
  - 非監督學習模型
  - 神經網路
  - 組合不同模型（ensemble）