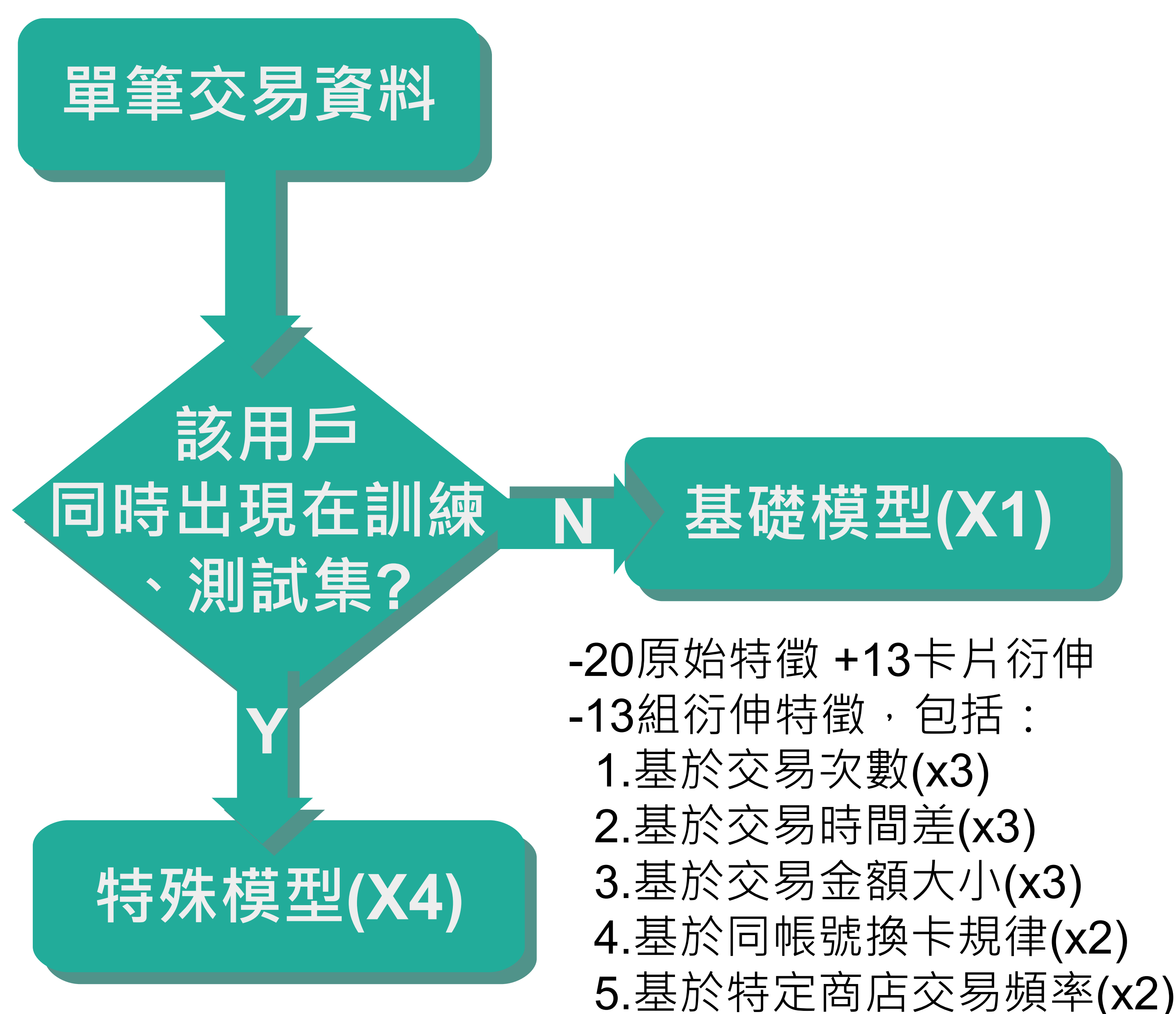


阿嬤我要吃糖果小隊-信用卡盜刷預測模型建構

簡介：

本團隊僅使用少量特徵(33項)建構模型，模型特色為簡單，且可有效防止過度擬合，以利企業實際能快速導入部屬環境，並降低維護模型的人力成本。模型本身由5個子模型組成，其中包括1個基礎模型(20原始特徵+13基於卡片(cano)衍伸)，及4個根據盜刷紀錄衍伸的特殊模型。當單筆交易資料進入模型，首先判斷該用戶(bacno)是否同時出現在訓練和測試集，若沒有則採用基礎模型，若有則依卡片(cano)在特定商店(mchno)、特定金額(conam)、盜刷時間差(locdt)作為是否盜刷的判斷。

模型預測流程



- 基礎模型+4特殊特徵
- 1.同卡同店盜刷紀錄x2
- 2.特定刷卡金額盜刷x1
- 3.盜刷紀錄時間差x1

注意：特殊特徵在現實生活中未必有用，因為我們在預測時未必知道該用戶以往的盜刷標注

特徵組成

	原始特徵X20	參見官網資料
基礎模型特徵	卡片衍伸X13	同卡總交易次數 (cano_num)
		同卡交易天數 (cano_locdt)
		同帳戶在特定店交易次數 (bacno_mchno)
		交易次數X3
		上/下次交易時間差 (locdt_div)
		同卡和交易日標準差 (cano_locdt loctm sigma)
		同卡和交易時間標準差 (cano_locdt locdt sigma)
		交易時間差X3
		同卡同交易日，交易金額上限 (cano_locdt conam_max)
		同卡同交易日，交易金額下限 (cano_locdt conam_min)
特殊模型特徵	如果該帳戶同時存在於測試/訓練集	同卡同交易日，交易金額為0時差 (cano_locdt conam_0_div)
		同卡該交易與下一筆時差 (cano_locdt_next_div)
		同戶不同卡交易時間差 (bacno_cano_A/B_locdt_div)
		同戶同店最後和第一筆交易時差 (bacno_mchno_1st_last_locdt_div)
		同卡同店的第幾筆交易 (bacno_mchno_num)
		特定店交易頻率X2
		同卡同店之前有盜刷紀錄 (cano_mchno_fraud_ind_Y)
		同卡同店之前無盜刷紀錄 (cano_mchno_fraud_ind_N)
		同卡同交易金額前是否為盜刷 (cano_conam_fraud_ind_Y)
		同卡第一次盜刷與該交易時差 (cano_1st_fraud_ind_locdt_div)

模型特色與結論

- 1.特徵少模型簡單 → 可有效防止過擬合(Overfitting)
- 2.只出現在訓練集的特徵轉為N/A → 減少模型學到無法用在測試集的規則
- 3.以用戶(bacno)分割訓練集，使用Early Stopping 和 GroupKfold減少訓練的迭代次數 → 防止過擬合
- 4.使用10 folds，以1個標準差去排除一些極端的fold，再以平均值估算 → 確保模型無受極端資料引導。

實際操作請見：https://github.com/aarontong95/TBrain_Credit_Card

