信用卡流失客戶預測

107302003 林書亞

110753208 蔣明憲

110352011 柯騰達

106703056 張晏瑄

Table of contents

01

資料集/目標介紹

研究目的

金融科技應用:信用卡客戶流失預測

台灣金融業在信用卡市場面臨激烈的競爭，銀行為了確保經營的規模及利益，對於**留住客戶變得越來越重視**，因此不斷的投入資金，進行活動促銷來刺激客戶刷卡，透過關懷外撥來**挽留客戶、預防流失**

有限的預算內，企業應如何配置才可發揮最大效益，即須有效的方法來**分析客戶特性及行為**，以提供企業決策參考，訂定合適的客戶關係經營策略

資料簡介

樣本總數:10127筆

目標欄位:

流失用戶標記(流失/留存)

數值型欄位:

年齡、家人數量、成為會員月數、目前擁有產品數、近一年不活躍月份數、近一年簽約數、信用額度、總循環信用、近一年月平均額度、總交易次數、總交易金額、第四季與第一季交易變化數、額度使用率

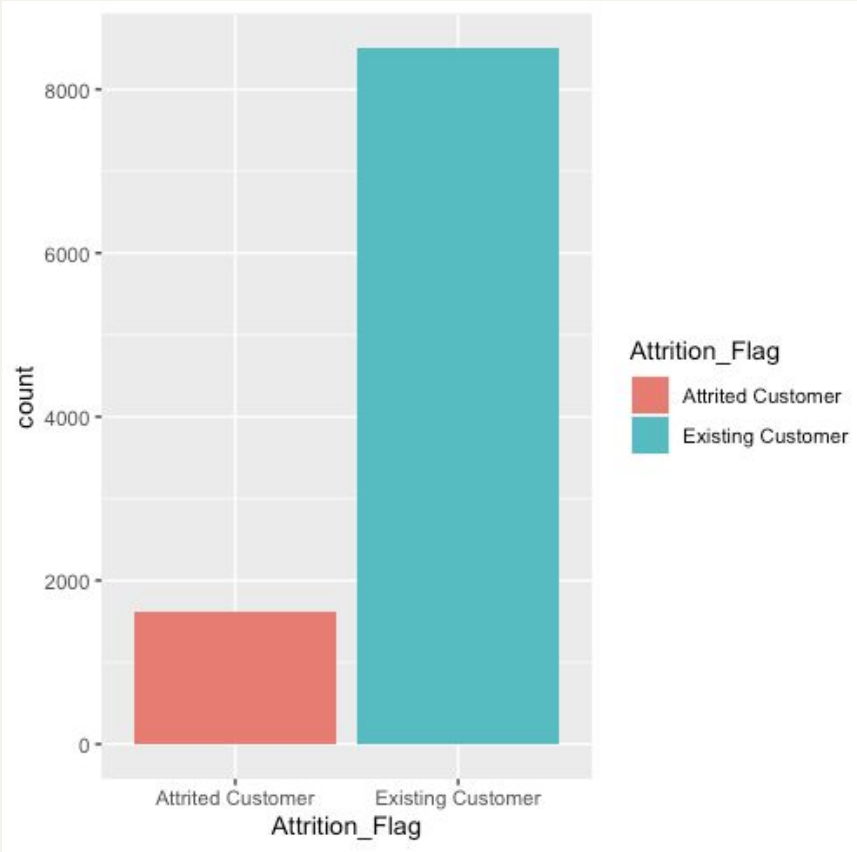
類別型欄位:

性別、教育程度、婚姻狀態、收入類別、信用卡等級

EDA

用戶流失標籤

- 流失客戶少，僅佔總體客戶16%
- 在樣本不平均情形下，可能需要做抽樣來平衡樣本



EDA

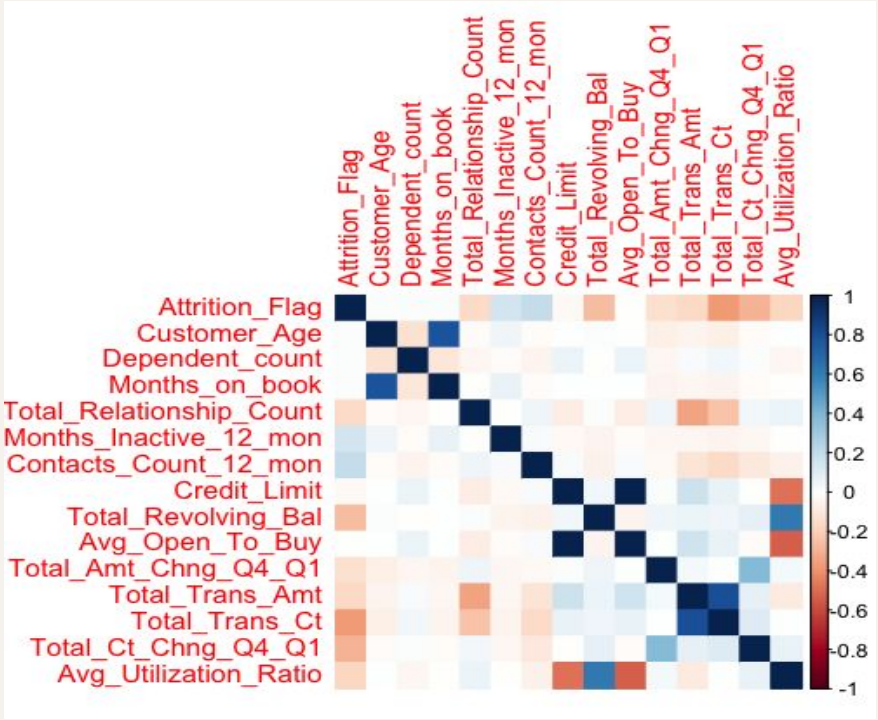
正相關

- 用戶年齡和在成為銀行用戶時間
- **信用額度**和**近12個月平均額度**
- 總交易次數和總交易金額

  (總交易金額改為平均交易金額)
- 信用額度使用率和總循環信用

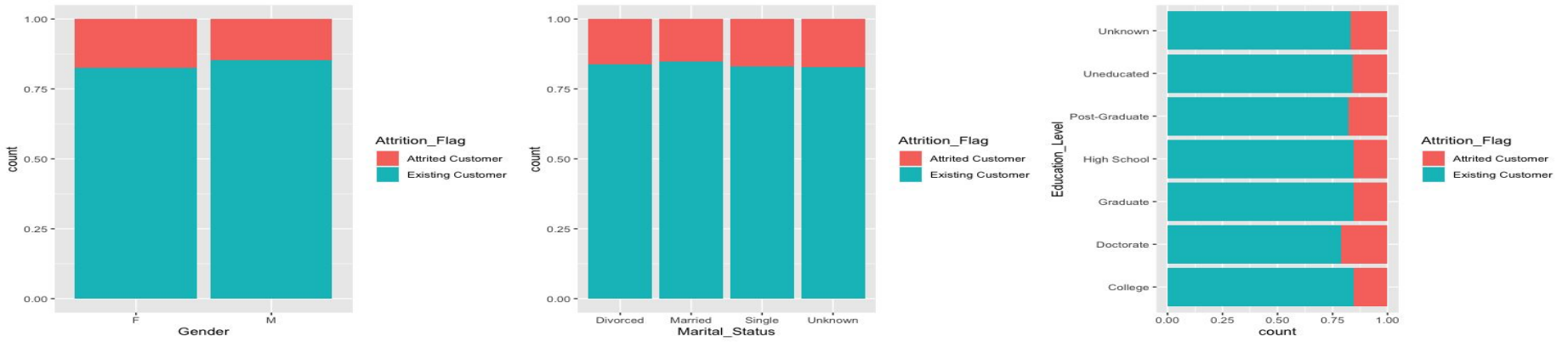
負相關

- 信用額度和信用額度使用率
- 近12個月平均額度和信用額度使用率



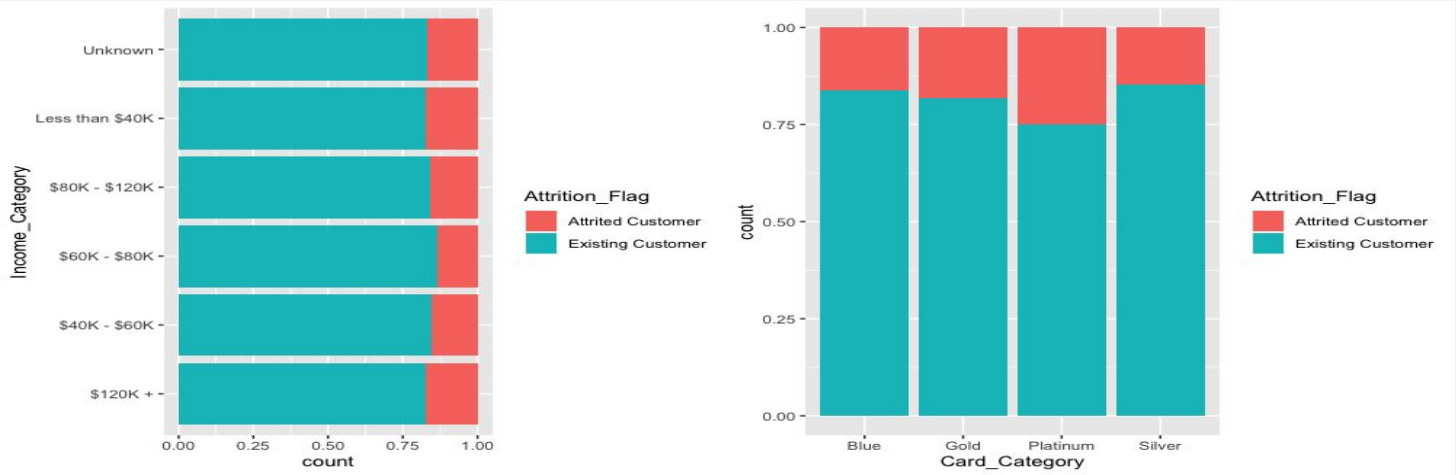
EDA

- 女性的流失客戶比例較男性略高，但差異不大
- 婚姻狀態與流失無明顯關係
- 擁有博士學位的客戶流失比例較其他教育程度顧客高



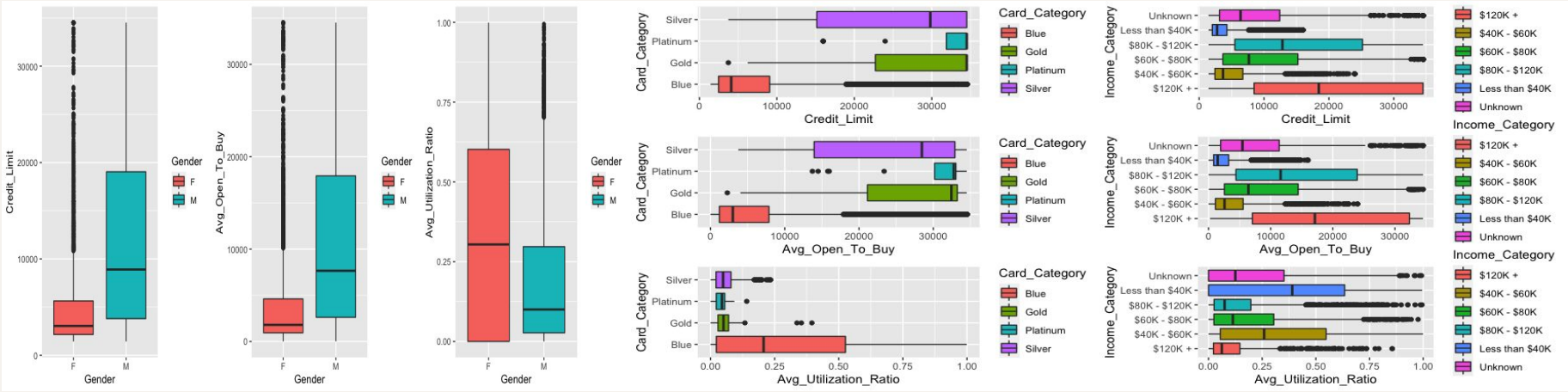
EDA

- 收入不同與流失與否無明顯關係
- 等級最高的白金會員流失比例較高



EDA

- 男性用戶信用卡額度較女性高，而女性額度使用率較男性高
- 藍等級會員在信用額度上明顯低於其他等級會員，但額度使用率較高
- 收入未知會員額度分佈與 $60-80K 會員相似



02

資料處理

資料處理

欄位刪減

- 將相關性高的信用額度和近12個月平均額度擇其一留下(選擇保留12個月平均額度)

數值調整

- 將總交易金額調整為平均交易金額

類別型資料的Unknown

- 收入類別資料中Unknown類別的處理:以$60-80K收入水準取代
- 教育程度類別資料中Unknown類別的處理:以高中程度教育水準取代

變數轉換

- 年齡變數(數值型)轉類別型:26 ~ 35、36 ~ 45、46 ~ 55、56 ~ 65、>65
- 成為會員月數變數(數值型)轉類別型:0 ~ 10、11 ~ 20、21 ~ 30、31 ~ 40、>40

資料處理-樣本不平衡

合成少數類別過取樣技術：

- SMOTE
- MWMOTE

樣本不平衡-SMOTE

- 是一種綜合取樣人工合成資料演算法
- 同時進行 Oversampling 與 Undersampling

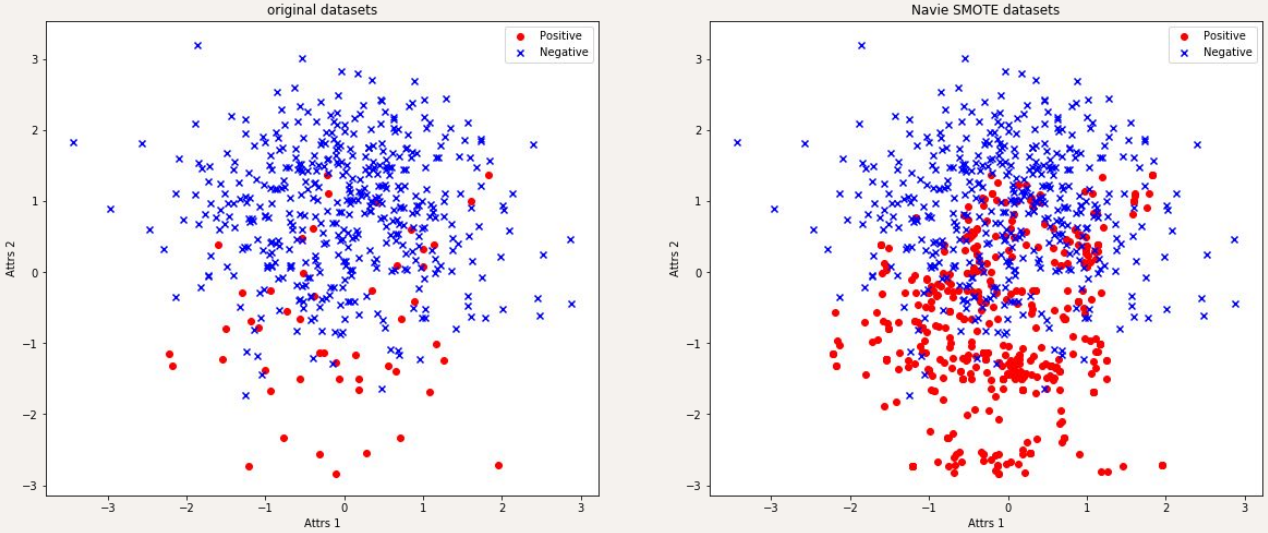
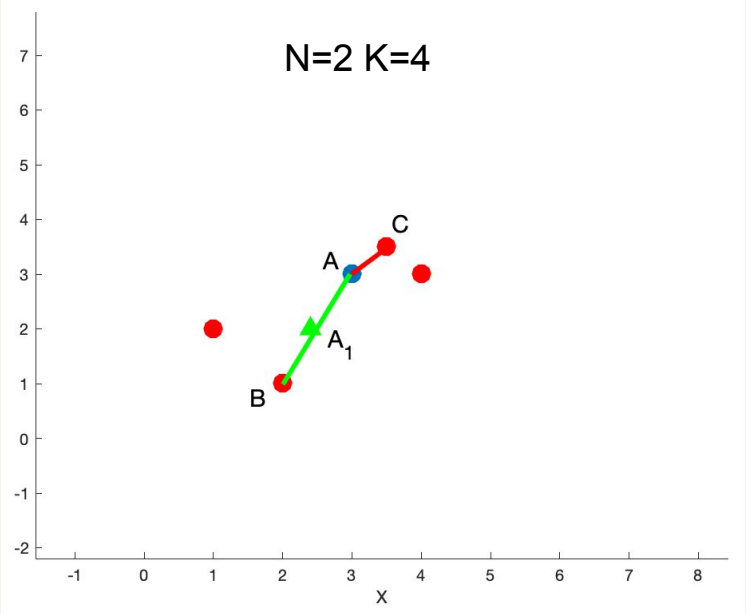


示意圖: 探索SMOTE演算法

樣本不平衡-SMOTE

- Synthesized Minority Oversampling Technique
- 重要參數：N、K
  - N 代表採樣倍率
  - K 代表鄰近值

- 從K個鄰近值中選取N個樣本來進行生成
- A到B間：乘上[0, 1]之間的隨機因子(A1)



示意圖：探索SMOTE演算法

樣本不平衡-SMOTE

- 使用 DMwR package 中的 SMOTE
- 設定 perc.over、perc.under、k
  - perc.over = xx：表示少樣本變成原來的（1+xx/100）倍
  - perc.under = yy：表示多樣本變成少樣本的 yy/100 * (xx/100) 倍
  - k：鄰近值 (default = 5)
- perc.over = 200, perc.under = 150 即少數過採樣成三倍，多數亦降採樣為少數的三倍

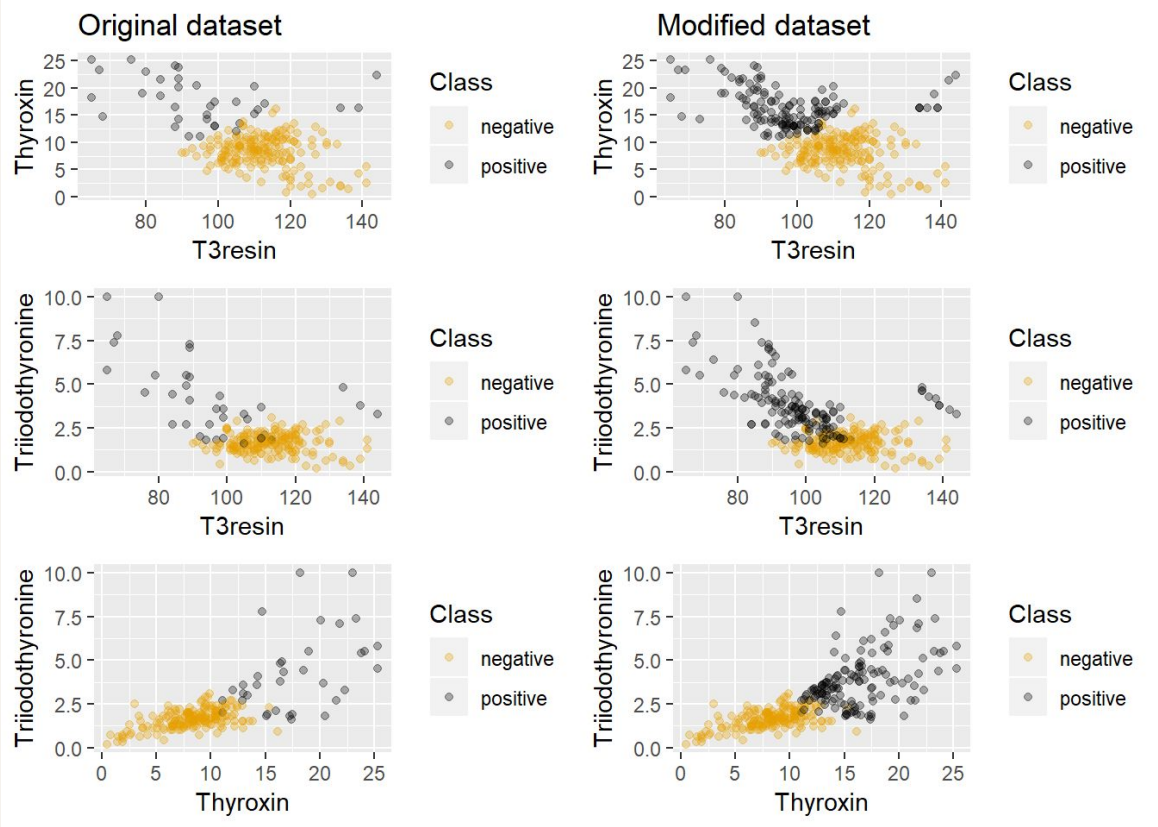
樣本不平衡-MWMOTE

- Majority Weighted Minority Oversampling TEchnique
- 利用與多數類樣本的距離來分配權重給少數類樣本，利用權重合成少數類樣本
  - 利用距離除去少數類中的噪點樣本(根據Euclidean Distance)
  - 根據到少數類樣本的距離，找出多數類樣本邊界
  - 對每個邊界附近的少數類樣本進行權重的計算(跟多數類樣本的距離及密度，越近越高)
  - 用權重來當成抽樣合成樣本的機率，進行SMOTE的插值合成

樣本不平衡-MWMOTE

- 使用 imbalance package 中的 mwmote
- 設定 numInstances、classAttr
  - numInstances：要生成的總數
  - classAttr：判斷多數少數類別的欄位名稱
- numInstances = x, classAttr = 'Attrition_Flag'

	TrainRecall	ValidRecall	TrainAccuracy	ValidAccuracy
MWMOTE(+3000)	0.9986	0.8677	0.9995	**0.9556**
MWMOTE(balance)	0.9997	0.8621	0.9996	**0.9505**
MWMOTE(+10000)	0.9995	0.8761	0.9995	**0.9475**



示意圖：imbalance: Oversampling Algorithms for Imbalanced Classification in R

03

模型選擇

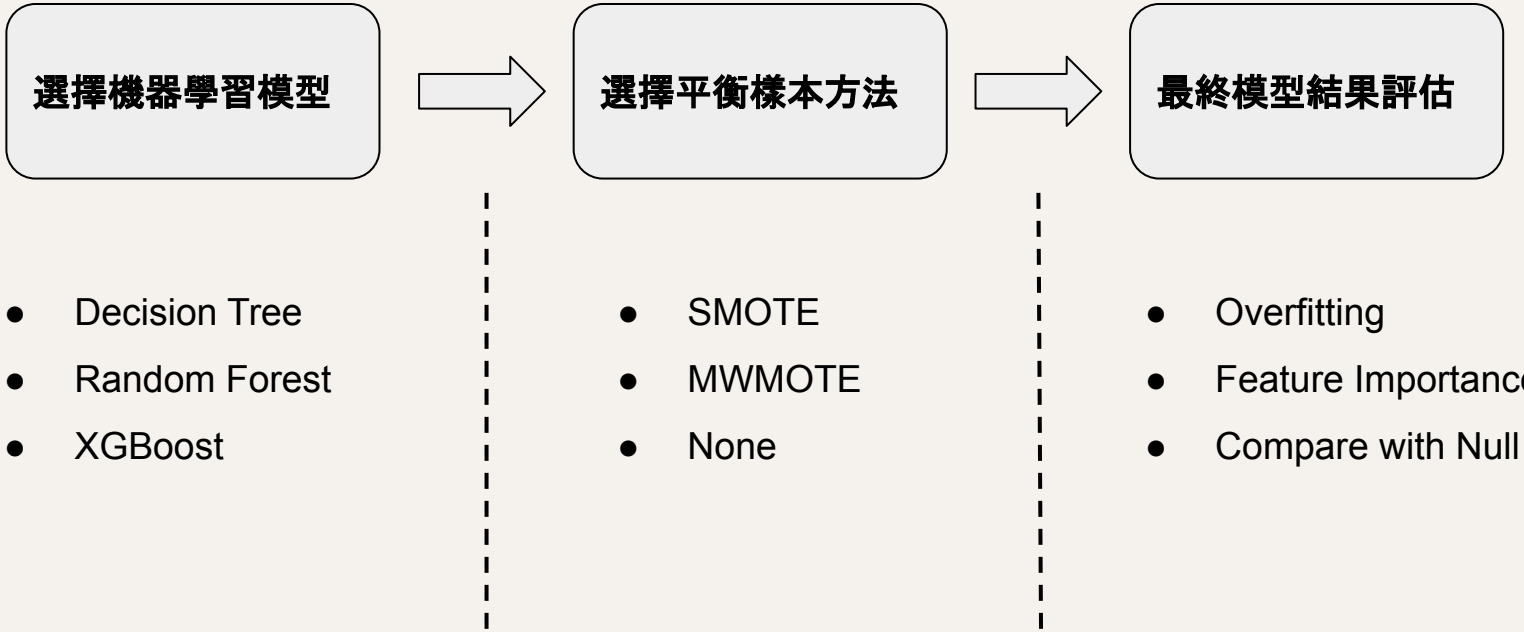
XGBoost 優點

- **正則化(控制模型複雜度，防止過擬合)**
- **兼具Boosting跟Bagging的性質**
- **較不容易落入局部最佳解**
- **並行處理(運算速度快)**
- **模型預測能力高**

模型預測能力預期:

1. XGBoost
2. Random Forest
3. Decision Tree

模型選擇流程


選擇機器學習模型
選擇平衡樣本方法
最終模型結果評估
Decision Tree
Random Forest
XGBoost
SMOTE
MWMOTE
None
Overfitting
Feature Importance
Compare with Null Model

模型選擇

fold	training	validation	test	auc	model
5	0.9321	0.9241	0.9352	0.9160	Decision Tree
5	0.9077	0.9342	0.9346	0.9600	Random Forest
5	0.9820	0.9563	**0.9520**	**0.9828**	XGBoost

XGBoost 在 Testing Data 的 Accuaracy 以及 AUC 表現最好

選用 XGBoost 模型進行後續建模

模型選擇

比較MWMOTE, SMOTE, 與不做樣本平衡的模型表現

- 不做樣本平衡的precision和AUC表現較好
- 透過SMOTE平衡樣本在Recall上表現較好

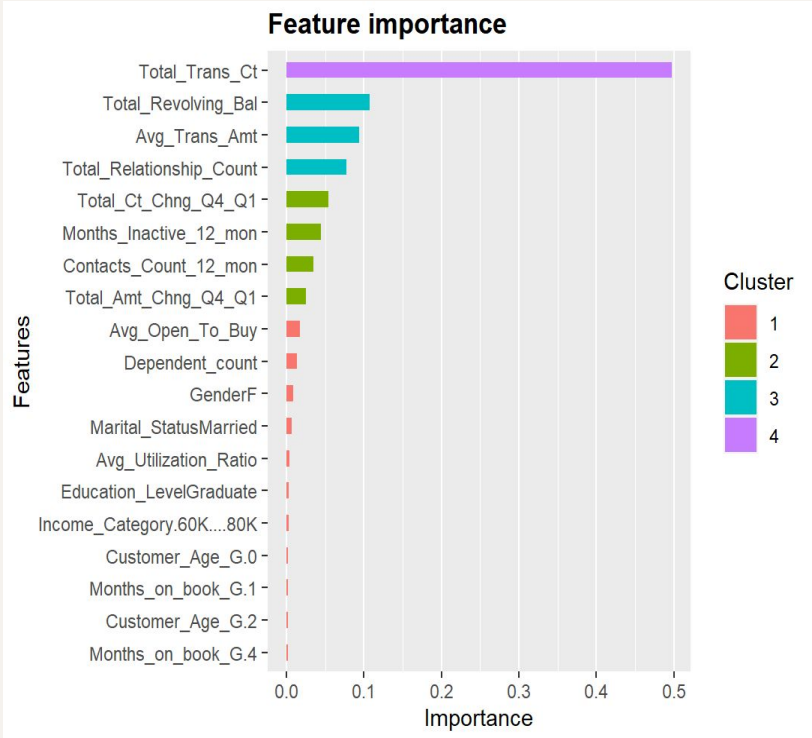
☞ 以預測流失客戶為目標下，選擇Recall表現好的SMOTE

fold	training	validation	test	auc	recall	sample
5	0.976	0.9468	0.95	0.9782	0.8213	MWMOTE
5	0.9827	0.9408	0.9434	0.9813	**0.9103**	SMOTE
5	0.9808	0.9521	**0.9555**	**0.9827**	0.8007	None

模型選擇

- 多數特徵重要性低，交易相關欄位重要性高
- 選擇前8個重要特徵建模，模型表現較使用全部特徵來得優異

fold	training	validation	test	auc	recall	fs_label
5	0.9833	0.9396	0.9331	0.9796	0.8799	ALL
5	0.9808	0.9375	0.9319	**0.9818**	**0.9081**	Feature Importance Top 8



Null Model之比較

Null Model 設計：

- 資料的流失客戶比例作為其**流失機率（15.7%）**

Deviance	Model
1760.631	Null
478.4788	Final Model

pseudo R-Squared：0.7282

最終模型優於Null Model

04

結論

結論

- 重要性低欄位多為**基本個人資料相關的類別型變數**，像是教育程度、收入類別、成為會員的時間區間等
- 特徵重要性高的欄位多為**信用卡使用情形相關的變數**，如：交易金額、交易次數、使用產品數量等欄位

☞ 用戶型態和成為會員時間較不影響客戶流失與否，影響客戶流失主因為**產品面**，當客戶開始減少交易次數、少辦理新卡，代表該客戶很有可為潛在會流失的客戶