



信用卡客戶流失預測

2023.01.10

統計碩二 黃文顯

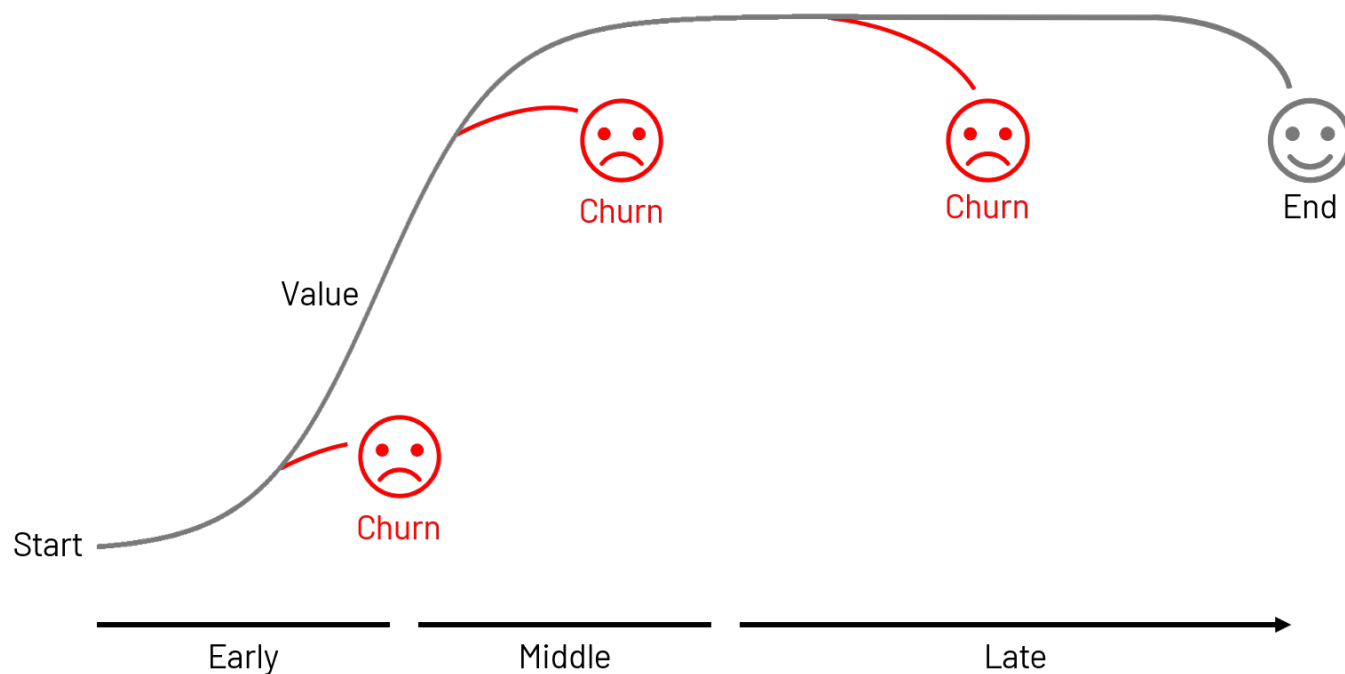
Agenda

- 分析目標 & 資料介紹
- 資料分析
 - 資料前處理 & EDA
 - 模型建置
- 分析總結 & Recommendation

分析目標

幫助銀行解決客戶流失問題

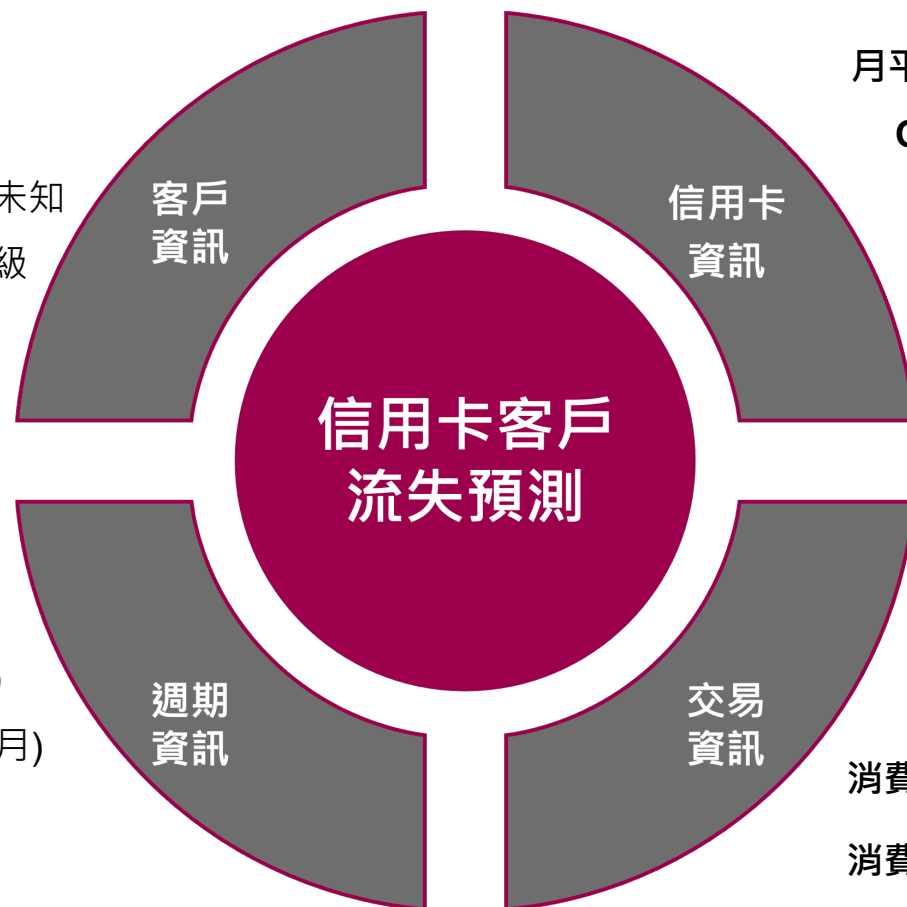
銀行對於客戶離開他們的信用卡服務感不到安，
希望能幫助銀行準確預測流失的客戶，這樣他們就可以
主動去找客戶，為他們提供更好的服務。



What is the reason causes customer churn?

年齡： 26 – 73歲
家庭人數： 0 – 4人
性別： 男性、女性
婚姻狀況： 已婚、離婚、單身、未知
收入類別： 由低至高分為六個等級
教育水準： 分為六個等級

不活躍月數(一年內)： 1 – 6 (月)
持卡時間： 13 – 56 (月)
持有其他產品數： 1 – 6
聯繫客服次數： 0 – 6



信用卡額度： \$1438 – 34516
月平均消費金額： \$0 – 2517
Open to buy： 信用卡月平均餘額
信用卡等級： Blue、Silver、Gold、Platinum
信用利用率： 0 – 1

總年度消費金額： \$510 – 18484
總年度消費次數： 10 – 139
消費金額變化率(Q4/Q1)： 0 – 3.397
消費次數變化率(Q4/Q1)： 0 – 3.714

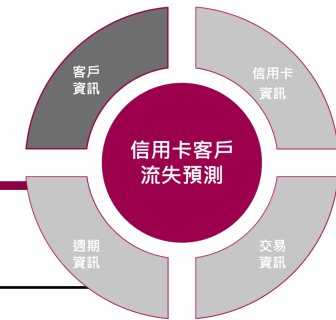
Agenda

- 分析目標 & 資料介紹
- 資料分析
- 資料前處理 & EDA
- 模型建置
- 分析總結 & Recommendation

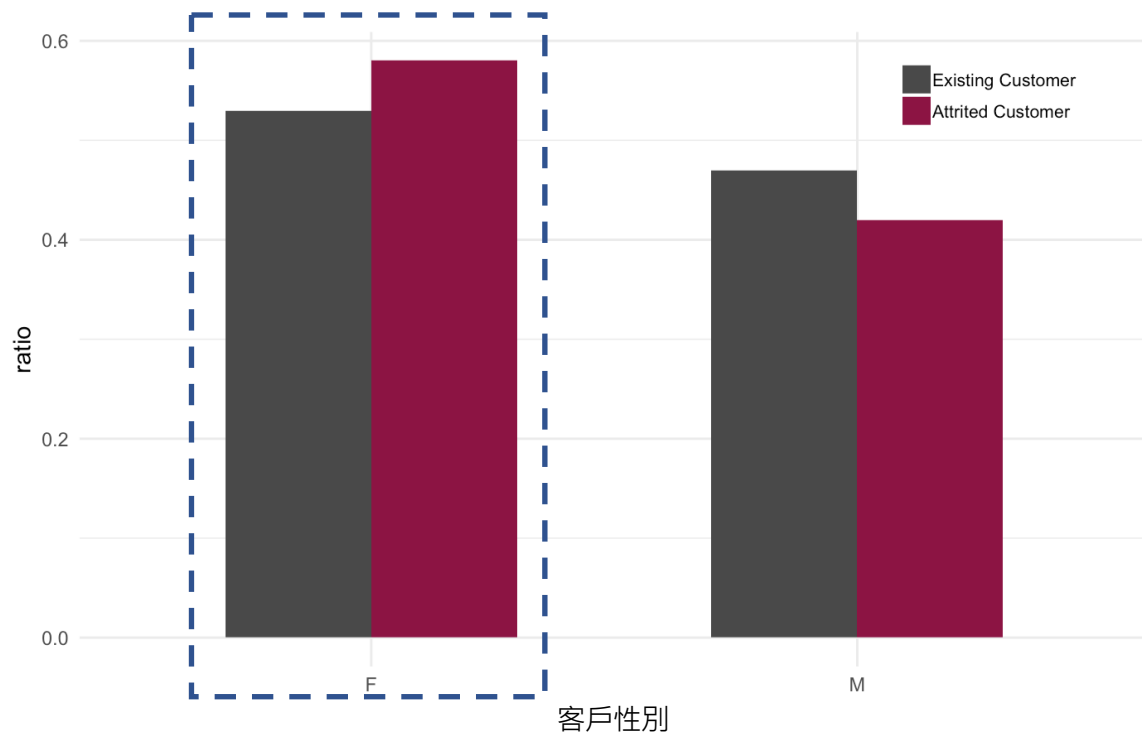
資料前處理，並拆成 Training 與 Testing data

資料說明與前處理		資料切分	
資料說明	資料含用戶特徵與在本行的信用卡消費資料	Dataset N=10127、p=19	
資料維度	樣本數N=10127、變數個數p=19 Response(Y)：客戶流失=1、客戶留存=0	7 Training N=7088、p=19	3 Testing N=3039、p=19
資料轉換	<ul style="list-style-type: none">在EDA過程中將年齡劃分為五個區間： (26,29]、[30,39]、[40,49]、[50,59]、[60,73]信用卡使用率從0-1等分： (0,0.2]、(0.2,0.4]、(0.4,0.6]、(0.6,0.8]、(0.8,1]	<ul style="list-style-type: none">EDA 100%使用Training進行建模皆以10 fold CV調整參數	<ul style="list-style-type: none">Evaluate在Testing在此資料上進行模型比較

「女性」與「中老年族群」有更高的流失率

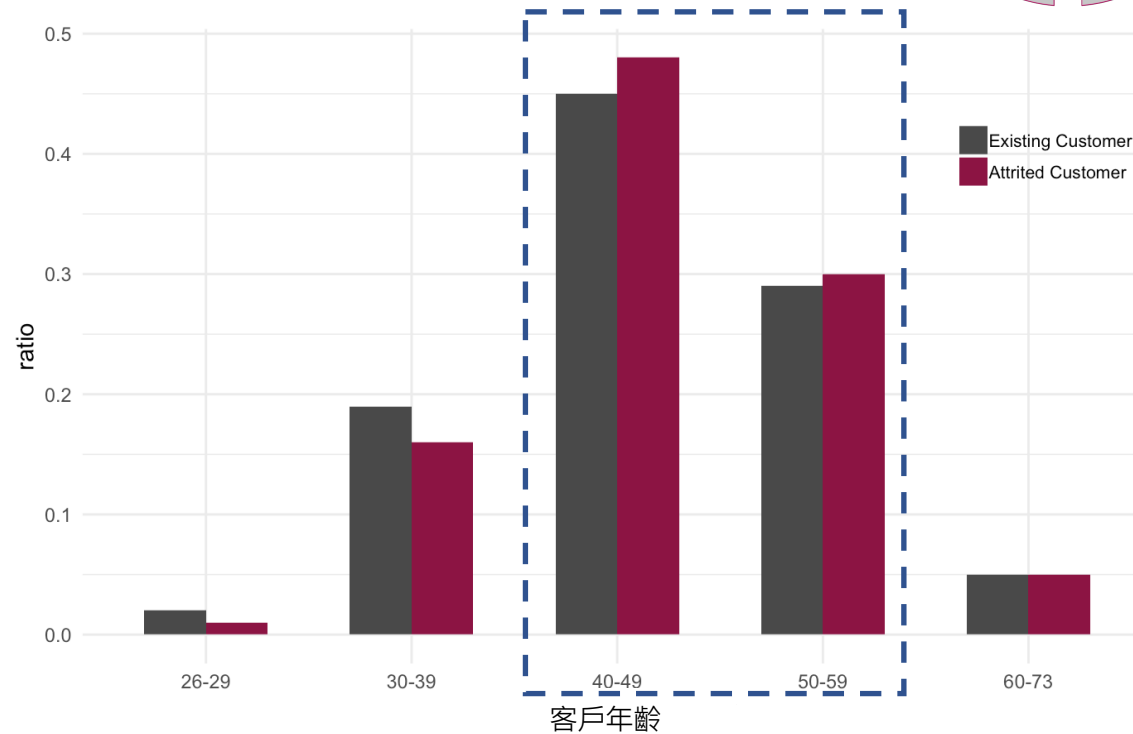


性別對流失率的影響



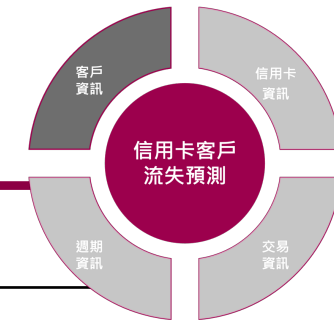
- 相較於留下的客戶，流失的客戶更多是女性用戶

隨年齡增長對流失率的影響



- 相較於留下的客戶，流失的客戶更多中年及老年人

「低薪族群」與「學士後、博士生」有更高的流失率

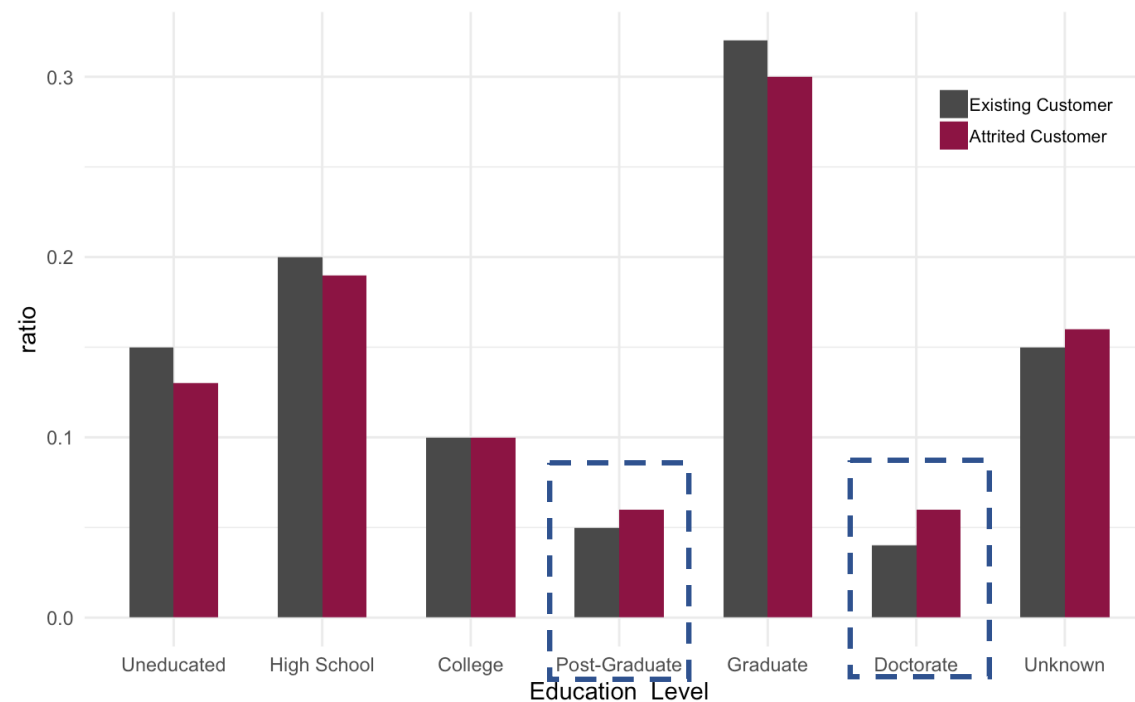


薪資對於流失率的影響



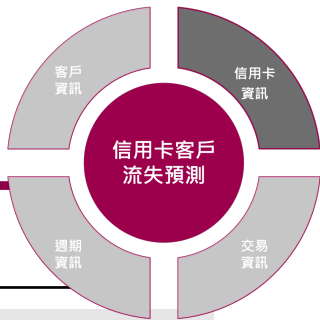
- 相較於留下的客戶，流失的客戶多數是年薪小於\$40K的客群

教育程度對於流失率的影響

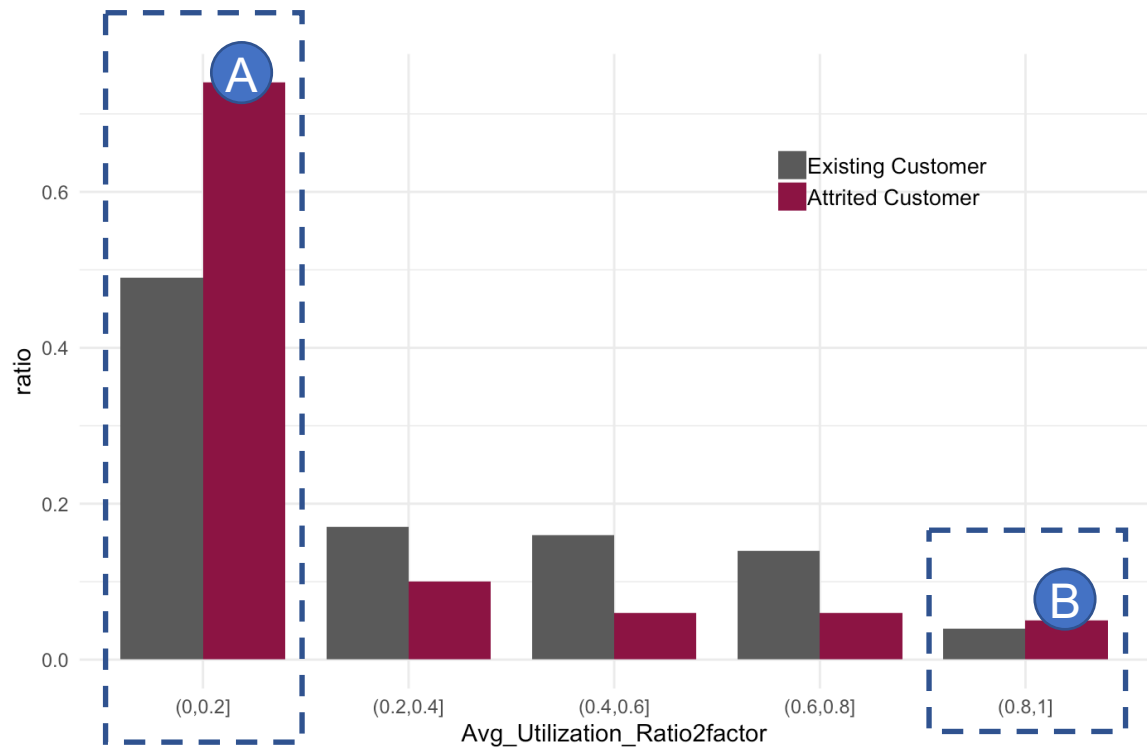


- 相較於留下的客戶，流失客戶的教育程度比較多是學士後與博士學位

信用卡使用率「極高」與「極低」的用戶有更高的流失率



信用卡使用率極高與極低的族群皆有較高的流失風險

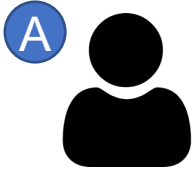


- 信用卡使用率 = 月平均消費金額 / 信用額度
- 信用額度為銀行對客戶的信任程度
- 相較於留下的客戶，位在信用卡使用率的兩端用戶有較高的流失率

全體用戶信用額度與月消費金額統計摘要

客戶人數	信用額度中位數	月平均消費金額中位數
7088	\$4534	\$1262

流失的低使用率用戶特徵



人數: 830
信用額度中位數: \$5134
月平均消費金額中位數: \$0

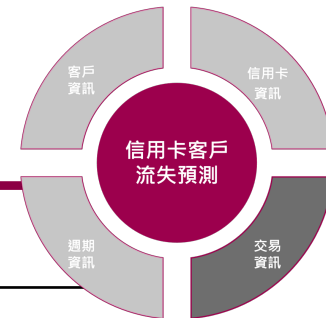
流失的高使用率用戶特徵



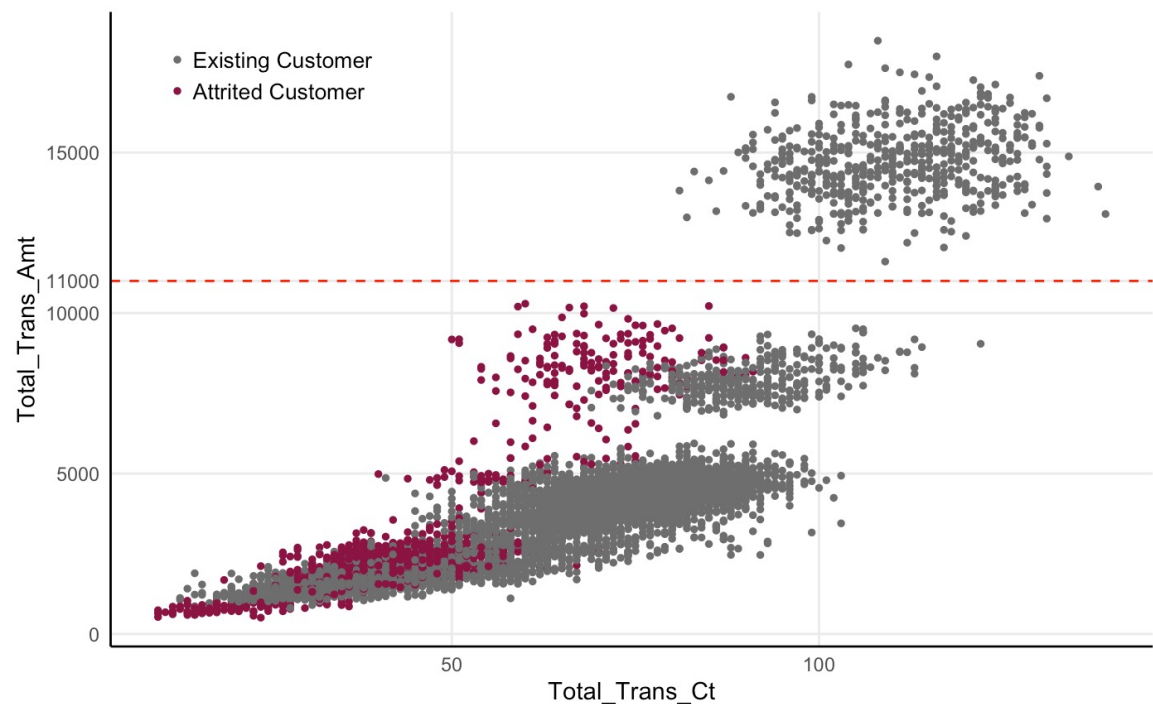
人數: 60
信用額度中位數: \$2474
月平均消費金額中位數: \$2177

- 低使用率的流失客戶為銀行更加信任的客戶，但客戶不買單產品，故銀行可以考慮對信任客戶提出更具吸引力的方案
- 高使用率的流失客戶為銀行較不信任的客戶，且客戶的月平均消費金額幾乎達信用額度上限，猜測客戶可能是在使用上受限故選擇離開

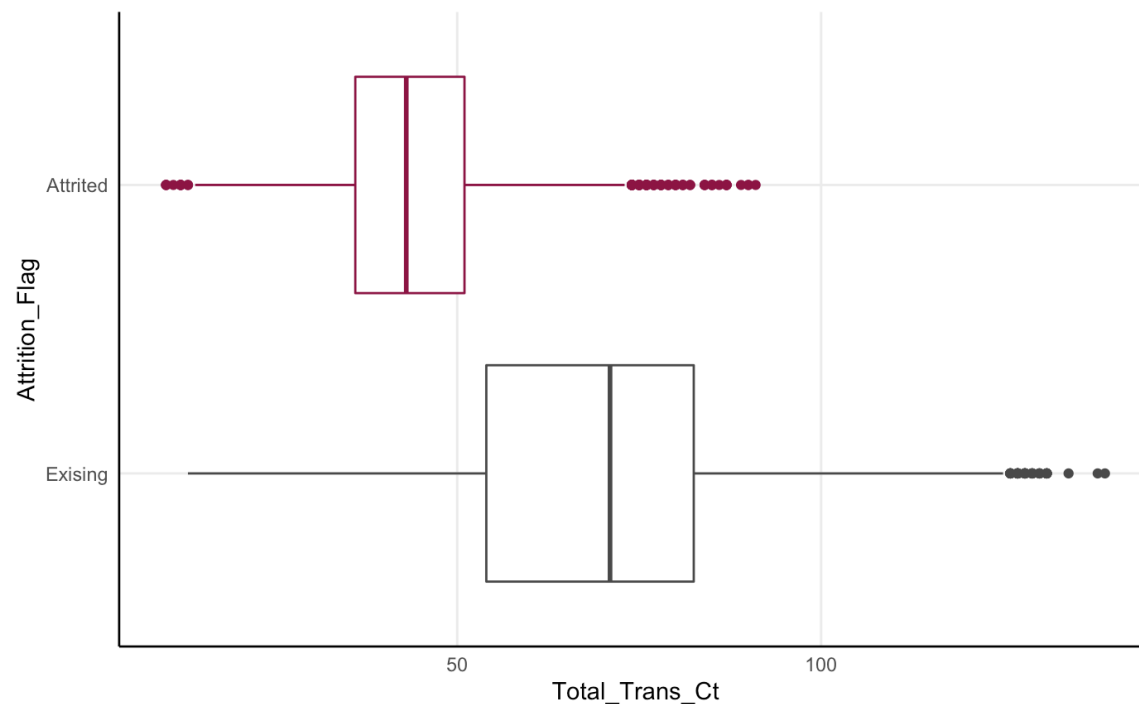
「年度總消費次數」與「年度總消費金額」對流失率的影響



流失客戶在年度消費次數與年度消費金額的分佈情形



全體用戶信用額度的統計摘要



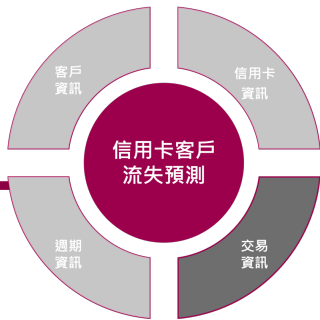
- 年度總消費次數與年度總消費金額將客戶劃分為三個不同族群
- 年度總消費金額超過\$11000的客戶留存率100%
- 年度總消費金額低於\$11000的客戶在年度總消費量的維度反映出高流失率的客戶有較低的消費次數的現象

年度消費金額低於\$11000的客戶：

- 流失的客戶年度總交易次數中位數為43
- 留存的客戶年度總交易次數中位數為69

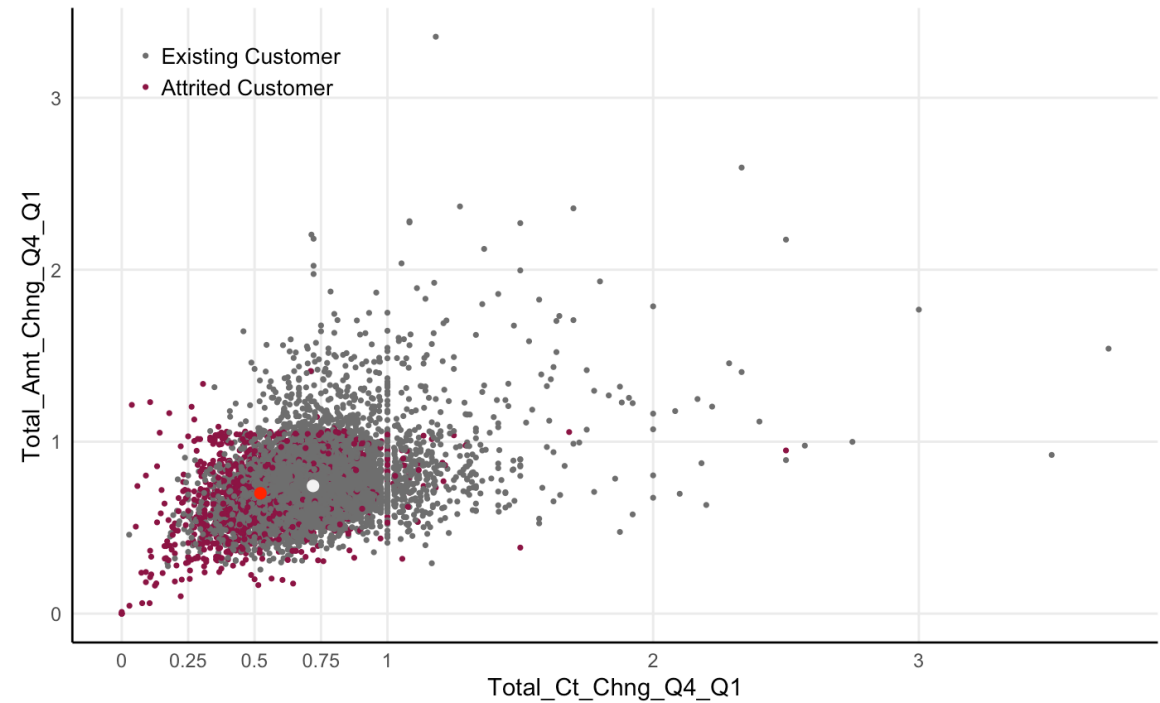
後續的預測上此二變數勢必對客戶分群起到重要影響

消費次數與消費金額的「變化率(Q4/Q1)」對流失率的影響



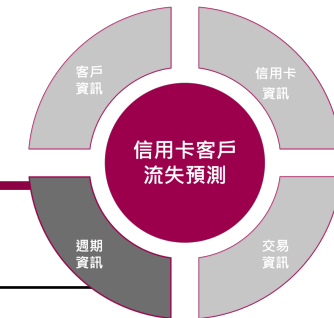
流失客戶於Q1到Q4消費習慣的變化

衡量指標	了解客戶在Q1到Q4消費次數與消費金額的轉變
分數計算	<ul style="list-style-type: none">消費次數變化率 = $\text{Q4消費次數} / \text{Q1消費次數}$消費金額變化率 = $\text{Q4消費金額} / \text{Q1消費金額}$

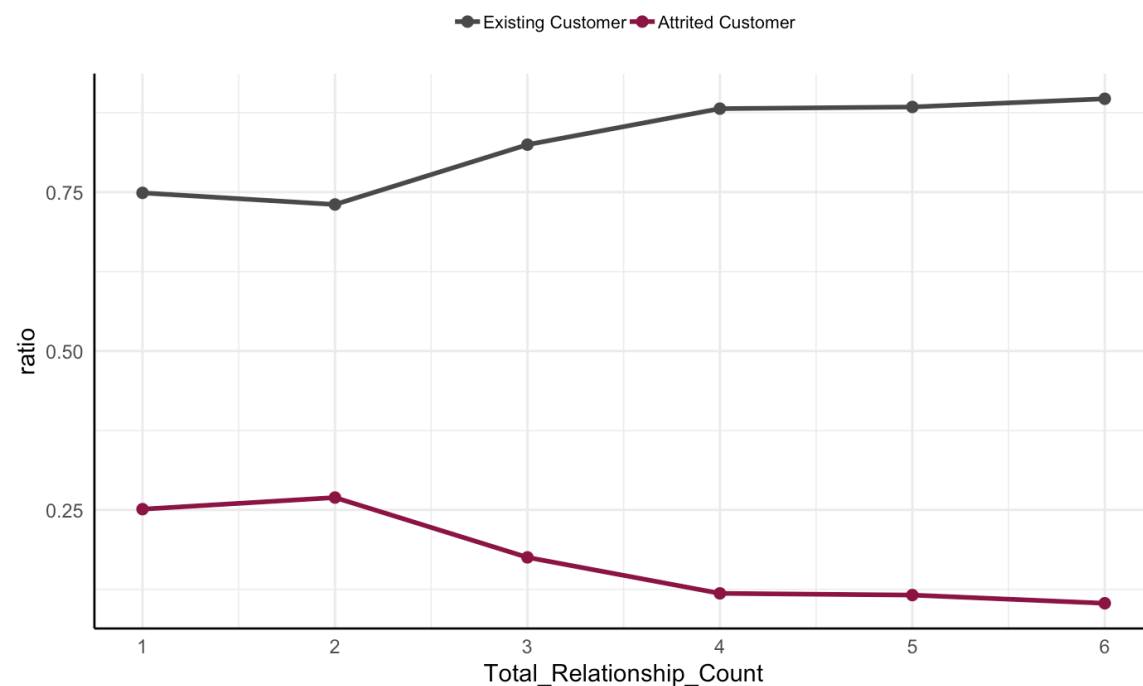


- 9成以上的客戶，Q4的消費次數與消費金額都低於Q1，不是銀行端需要過度反應的一個現象
- Q4的消費次數不到Q1的一半，是客戶流失的警訊**

客戶「持有產品數量」與「聯繫本銀行次數」對流失率的影響

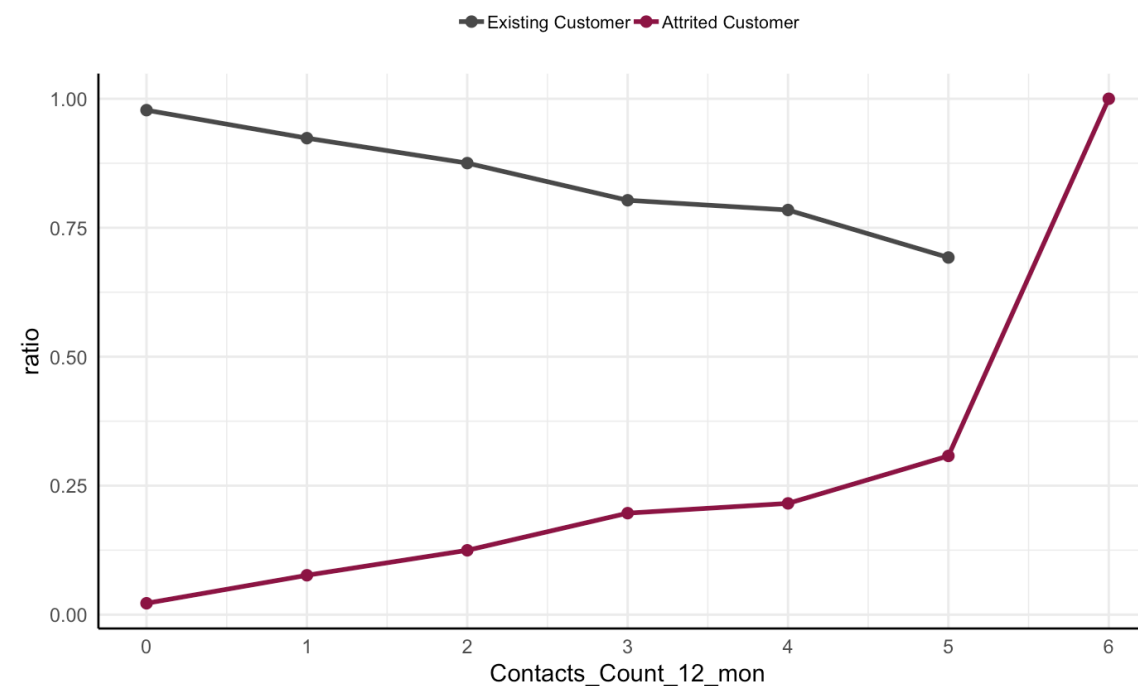


客戶持有的產品數量對流失率的影響



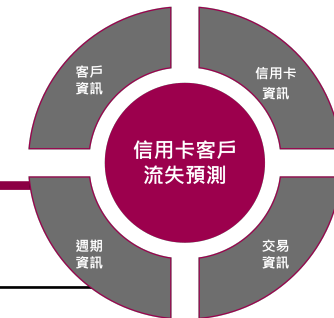
- 客戶流失率隨著持有商品數的增加而下降
- 因此推薦客戶相關金融商品可做為減少顧客流失的其一策略

客戶聯繫客服次數對對流失率的影響



- 客戶流失率隨著聯繫客服次數的增加而增長
- 因此在客戶聯繫客服的次數開始提升後，該客戶是銀行端更需要去留意的

What is the reason causes customer churn?



探討流失客戶的特徵與消費習慣

客戶特徵描述

相較於留下的客戶而言，流失客戶的使用者特徵更多是：

- 年齡為**40-59歲**的中老年與女性客戶
- 薪資水準較多是**低薪族群**
- 教育程度則分成「**學士後**」與「**博士學位**」兩族群
- 信用卡使用率位在**兩端(很少用/很常用)**的客戶

客戶交易行為描述

- **年度總消費金額**做為客戶分群的首要指標，**年度總消費次數**可做為客戶分群的次要指標
- **Q4的交易數量不到Q1一半**是客戶潛在的離開訊號
- 客戶流失率隨著持有商品數增加而下降，故**推薦金融商品**是減少客戶流失的其一策略
- 客戶流失率隨著聯繫客服次數的增加而增加，故**根本解決與本銀行互動頻繁的客戶問題**，是降低流失率的方式

Agenda

- 分析目標 & 資料介紹
- 資料分析
 - 資料前處理 & EDA
 - 模型建置
- 分析總結 & Recommendation

資料前處理與建模流程

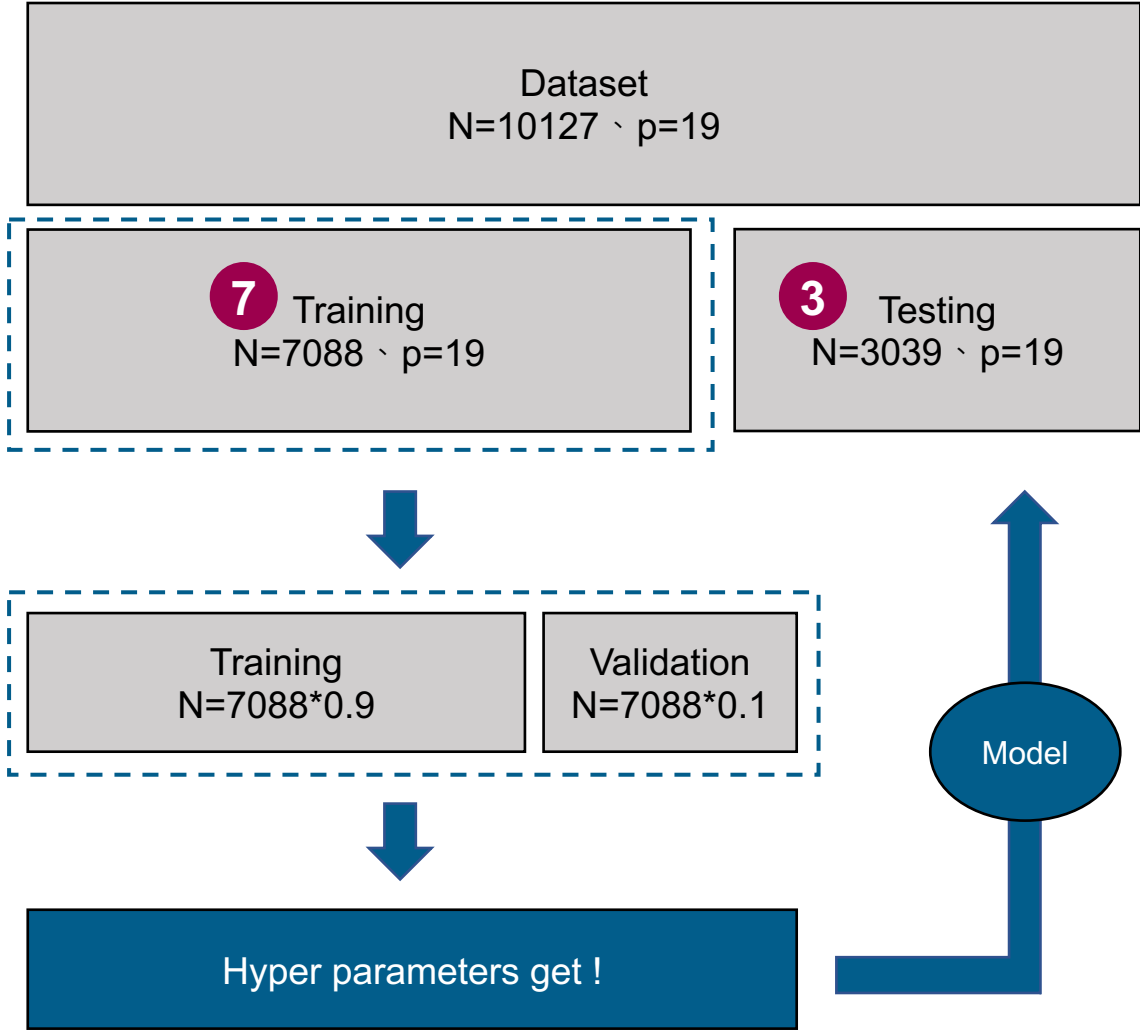
資料前處理、制定模型衡量標準

建模準備

- 對5個類別變數做**Dummy coding**：性別、教育水準、婚姻狀況、收入、信用卡等級
 - 解決**Imbalance**：透過Up-sampling
 - 移除完全共線性變數：Open to Buy
- 定義模型衡量標準：
- Accuracy
 - F1 Score
 - Recall(REC) = $TP / (TP+FN)$

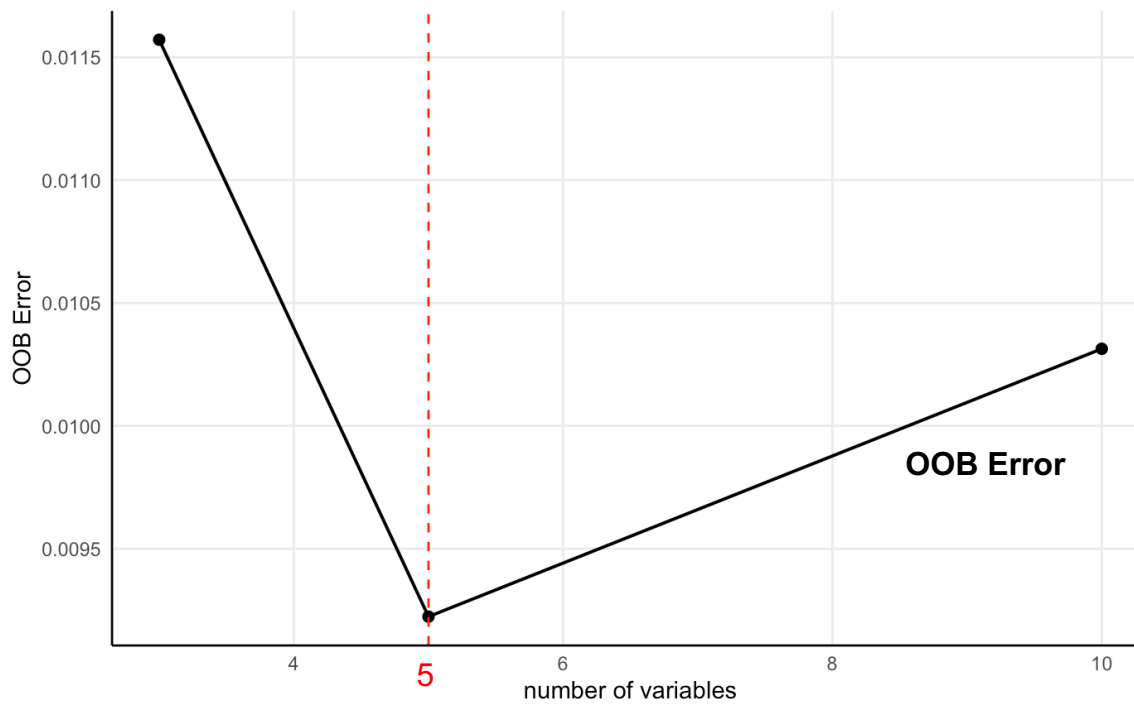
	Positive (流失)	Negative (留存)
Positive (流失)	TP	FN
Negative (留存)	FP	PN

資料切分、Hyper parameters 選取



Random Forest (1/2)

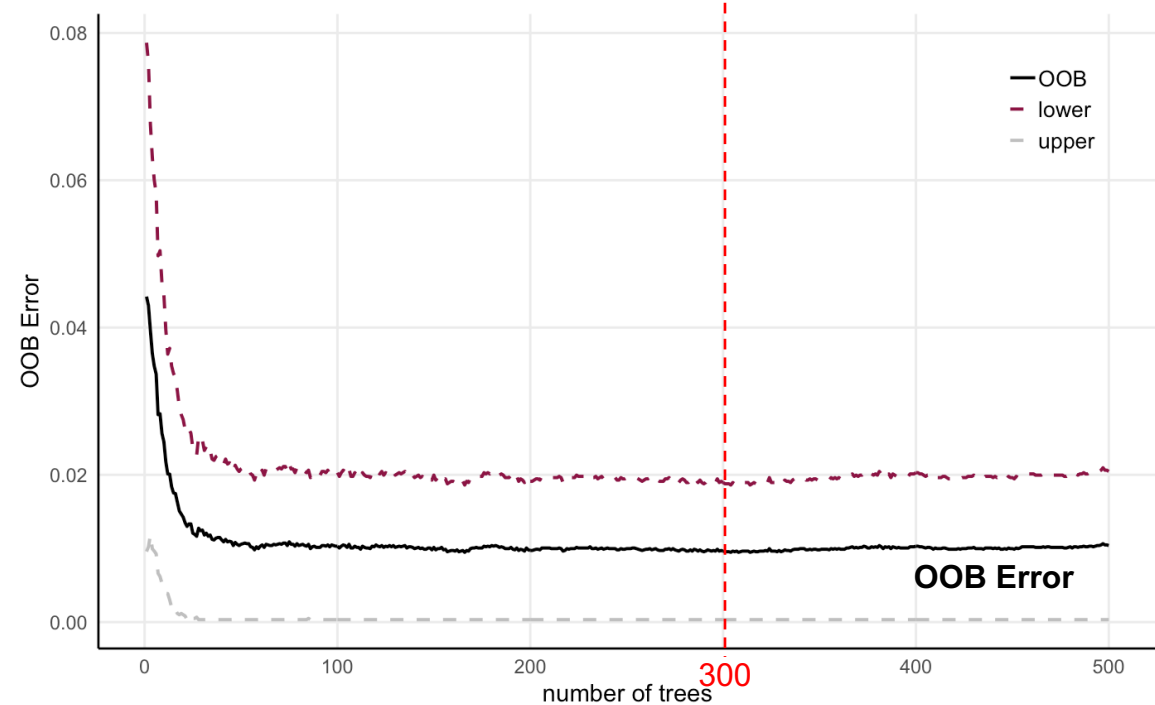
Tune Random forest 每個節點所需的變數個數



模型設定

- 給定300棵樹(T)的情況下，每個節點選出5個變數(M)為Random forest的最佳組合
- 並由右圖確認該組參數(T,M)=(300,5)確實可穩定收斂

檢查 Hyper parameters 是否收斂與模型 performance



	Positive (流失)	Negative (留存)
Positive (流失)	2502	35
Negative (留存)	79	423

➡

- Accuracy = 0.9624
- F1 Score = 0.8813
- Recall = 0.9236

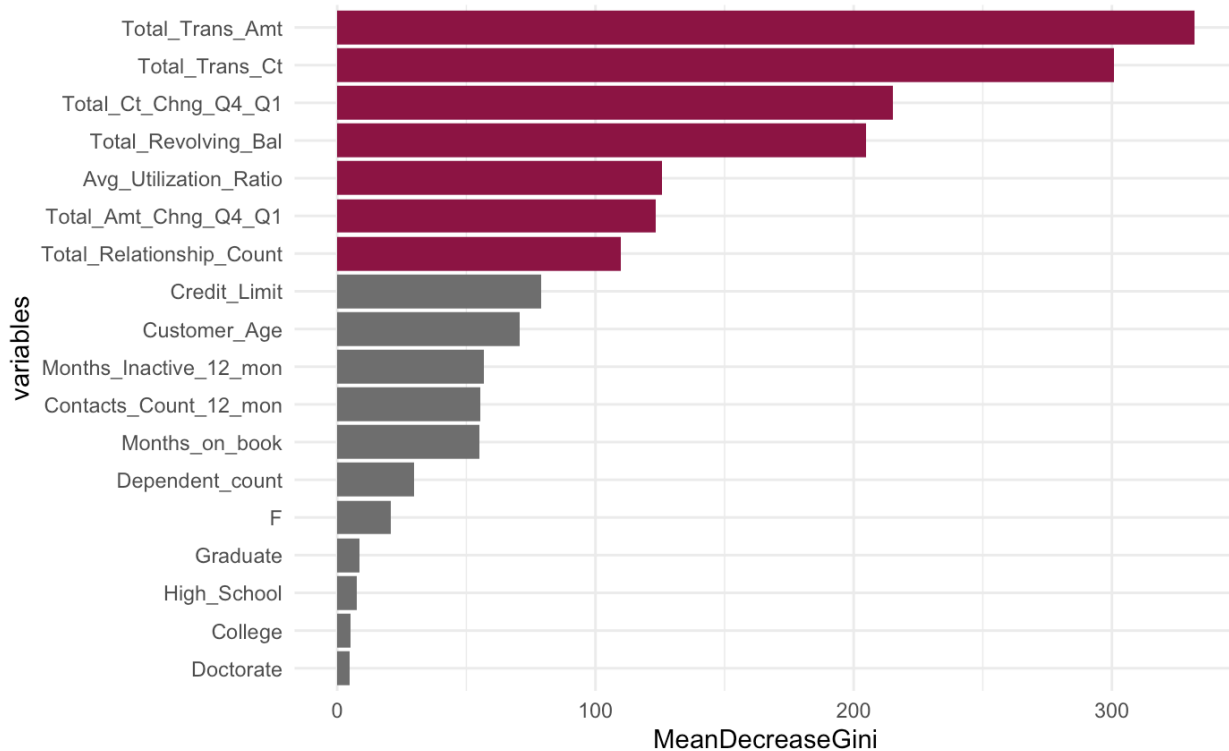
Random Forest (2/2)

Random forest 認為的重要變數與衡量標準(Gini)

Mean
Decrease
Gini

衡量變數如何影響每個節點與Terminal nodes的同質性

Variables importance plot



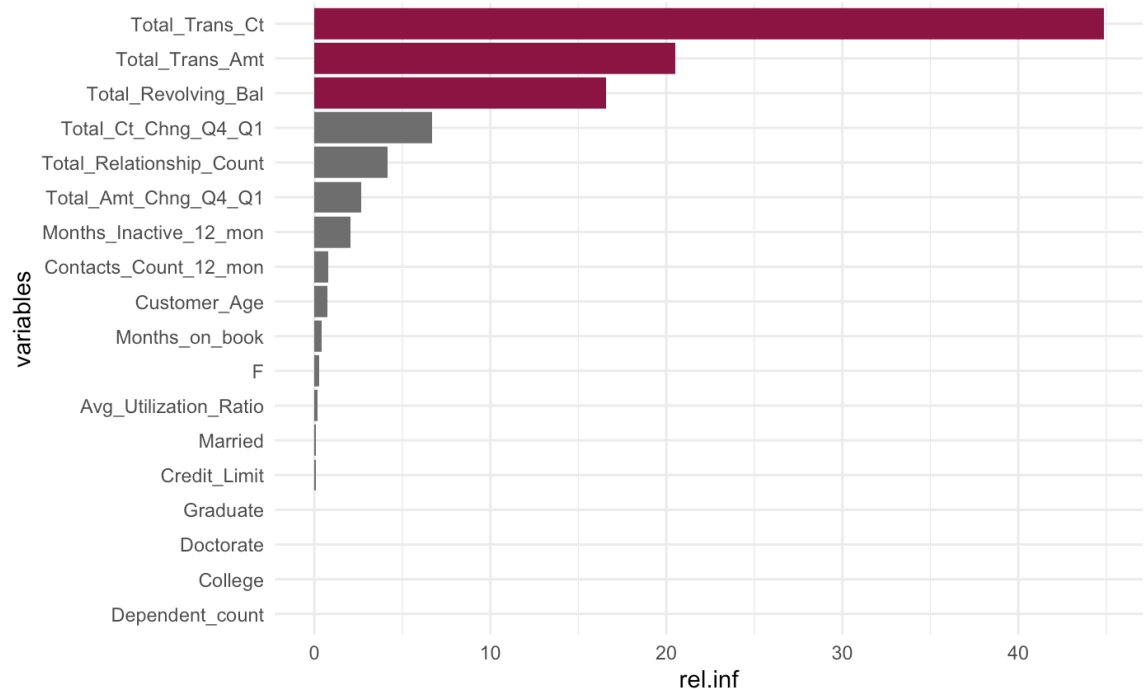
- 以變數能減少多大的Gini impurity為變數重要性指標
- 重要變數(Mean Decrease Gini > 100)
 - Total_Trans_Amt：總年度消費金額
 - Total_Trans_Ct：總年度消費次數
 - Total_Ct_Chng_Q4_Q1：消費次數變化率
 - Total_Revolving_Bal：月平均消費金額
 - Avg_Utilization_Ratio：信用卡使用率
 - Total_Amt_Chng_Q4_Q1：消費金額變化率
 - Total_Relationship_Count：持有產品數量

Gradient Boosting

Tune Gradient boosting

Learning rate	Depth of tree	Min obsinnode	Trees	Accuracy
0.1	3	10	150	0.9620469
0.1	2	10	150	0.9568283
0.1	3	10	100	0.9544287
0.1	2	10	100	0.9476582
0.1	3	10	50	0.9417328

Variables importance plot



Gradient boosting 認為的重要變數與模型 performance

- 以變數相對重要性為變數重要性指標
- 重要變數
 - Total_Trans_Ct：總年度消費次數
 - Total_Trans_Amt：總年度消費金額
 - Toal_Revolving_Bal：月平均消費金額

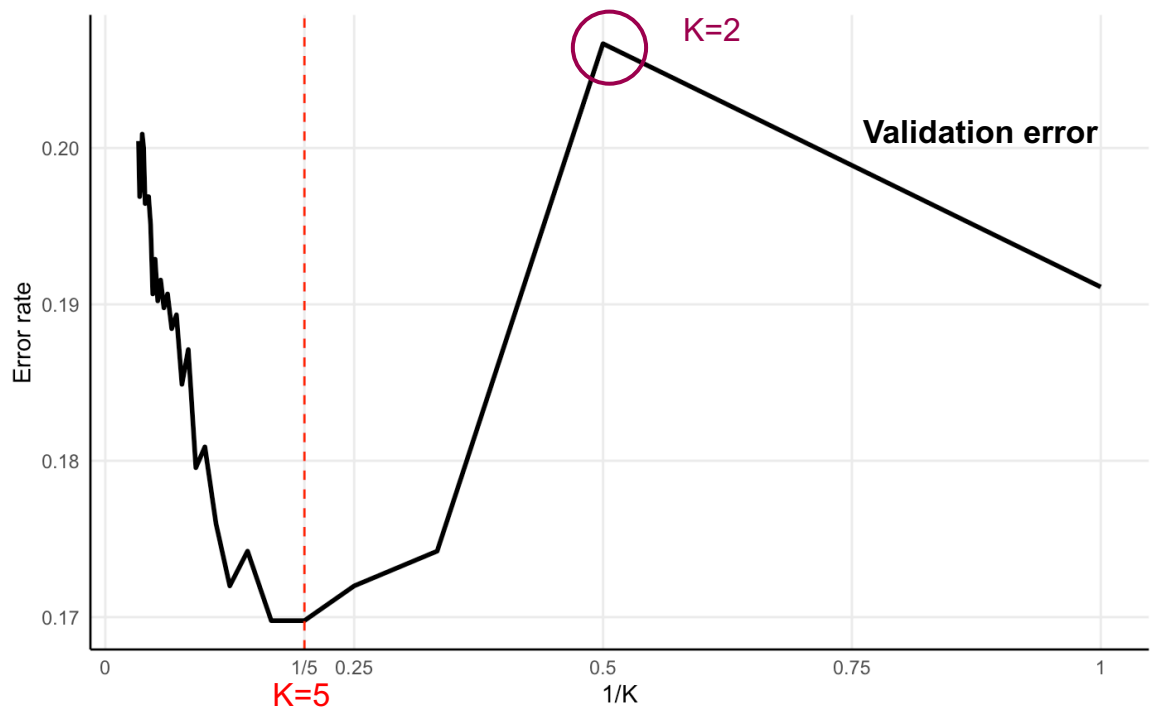
	Positive (流失)	Negative (留存)
Positive (流失)	2408	129
Negative (留存)	24	478



- Accuracy = 0.950
- F1 Score = 0.855
- Recall = 0.769

K Nearest Neighbor(KNN)

Tune KNN Optimal K



模型設定

- 10 fold cross validation → Optimal K=5

模型說明與模型 performance

- 隨著K從5到2的過程，Validation Error開始快速上升，代表出現overfitting的現象

	Positive (流失)	Negative (留存)
Positive (流失)	2136	401
Negative (留存)	82	420




- Accuracy = 0.841
- F1 Score = 0.635
- Recall = 0.512

Support Vector Machine(SVM)

Linear kernel

Kernel function : $K(x, z) = (1 + x'z)^1$


C	Error	<ul style="list-style-type: none">目標 : $\min(loss + C\sum\epsilon_i)$Cost(C) : 可容忍錯誤的大小Optimal C = 1e-7
1e-07	0.5213333	
1e-06	0.5213333	
1e-05	0.5213333	
1e-04	0.4257778	
1e-03	0.1973333	

	Positive (流失)	Negative (留存)	 <ul style="list-style-type: none">Accuracy = 0.833F1 Score = 0.614Recall = 0.500
Positive (流失)	2128	409	
Negative (留存)	98	404	

Gaussian kernel

Kernel function : $K(x, z) = \exp(-\gamma|x - z|^2)$

C	γ	Error	<ul style="list-style-type: none">目標 : $\min(loss + C\sum\epsilon_i)$Cost(C) : 可容忍錯誤的大小Optimal C = 1Optimal γ = 1e-5
1	1e-05	0.5280000	
2	1e-05	0.5280000	
3	1e-05	0.5280000	
4	1e-05	0.5186667	
1	1e-04	0.2217778	

	Positive (流失)	Negative (留存)	 <ul style="list-style-type: none">Accuracy = 0.847F1 Score = 0.645Recall = 0.524
Positive (流失)	2153	384	
Negative (留存)	80	422	

Agenda

- 分析目標 & 資料介紹
- 資料分析
 - 資料前處理 & EDA
 - 模型建置
- 分析總結 & Recommendation

多個模型的綜合比較與分析總結

Random forest 打敗所有模型，掌握了96%以上的客戶流向

	Random forest	Gradient boosting	Logistic	Forward selection	Backward selection	SVM linear	SVM RBF	KNN	LDA	QDA
acc	0.962	0.950	0.895	0.895	0.895	0.833	0.847	0.841	0.850	0.881
F1	0.881	0.855	0.633	0.633	0.630	0.614	0.645	0.635	0.651	0.686
REC	0.924	0.769	0.746	0.746	0.750	0.500	0.524	0.512	0.529	0.529

模型總結

- Random forest能有效解決overfitting的問題，故在Up-sampling後的資料能有優異的表現
- Gradient boosting雖然有不錯的表現，但從REC相較於Random forest大幅下降，代表可能開始出現overfitting的現象
- 其餘模型的則沒有太特殊的表現

重要變數總結

- 年度總消費額與消費次數能有效的切出三群不同的客戶特徵
- 消費情況產生的變化同時反映出客戶可能流失的跡象
- 極端的信用卡使用情形也反映出客戶流失的機率高低
- 持有產品數正比於客戶的留存率
- 月平均消費金額也有效的區分出客戶的留存與否

給銀行經理的策略建議

如今已能準確的預估**96%**以上的流失客戶，故銀行端能採取以下行動

- 銀行可以進一步了解**中老年客戶與女性客戶**在信用卡的使用需求，並推出符合對應客群的信用卡服務
- 針對**低使用率的高價值客戶**提供更具吸引力的方案，進而避免該群客戶流失
- 在客戶能力範圍內，適當地提升**高使用率但低信用額度客戶的信用額度**，進而滿足客戶的消費需求
- **盡可能的推薦金融商品於現有客戶**，即使利潤再低都可藉此增加客戶對銀行的黏著度
- 年末(Q4)時，分析客戶消費次數與消費金額是否接近年初(Q1)的一半，進而關心這些客戶的使用需求

A thick, solid maroon diagonal stripe runs from the top-left corner towards the bottom-right, separating the left side of the slide from the right side.

Thank you!

Reference

- Data resource: <https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction/code>
- Statistical learning: <https://www.statlearning.com/resources-second-edition>
- 分工: 黃文顯