金融實務專案

預測什麼樣的銀行客戶流失機率比較高

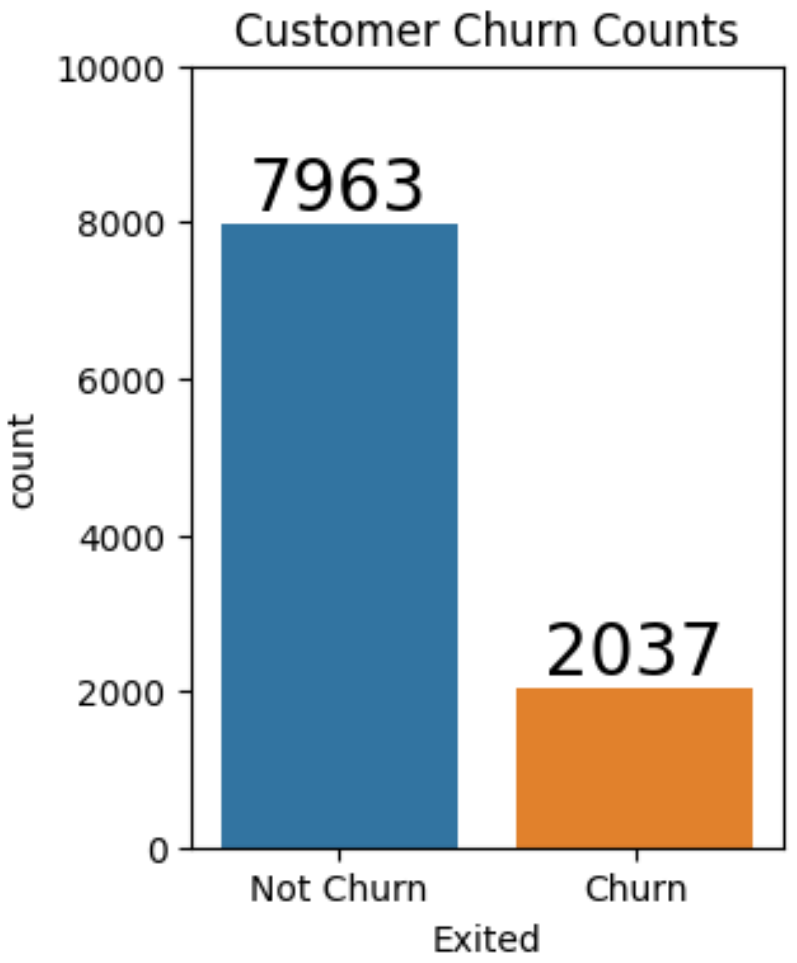
2023/7/15 凌銘陽

資料集統計描述

10000筆客戶資料，14個欄位，無缺失值

ColumnName	Description	Type
RowNumber	資料編號	連續型資料
CustomerId	客戶ID	連續型資料
Surname	客戶姓氏	字串資料
Geography	居住國家	離散型資料
Gender	性別	二元型資料
Age	年齡	連續型資料
CreditScore	信用分數	連續型資料
Tenure	往來期間	離散型資料
Balance	帳務餘額	連續型資料
NumOfProducts	持有產品數	離散型資料
HasCrCard	是否持卡	二元型資料
IsActiveMember	是否有效會員	二元型資料
EstimatedSalary	預估收入	連續型資料
Exited	是否流失 (0:未流失, 1:流失)	二元型資料

約20%的人流失



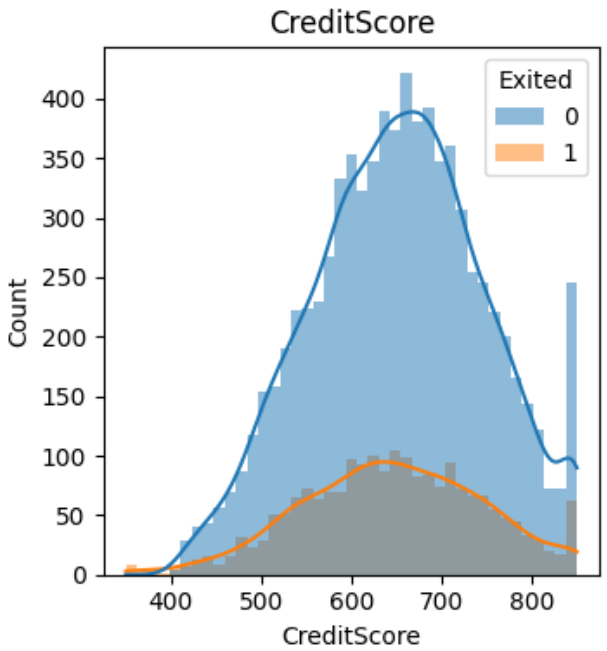
一、 確認你的分析範圍

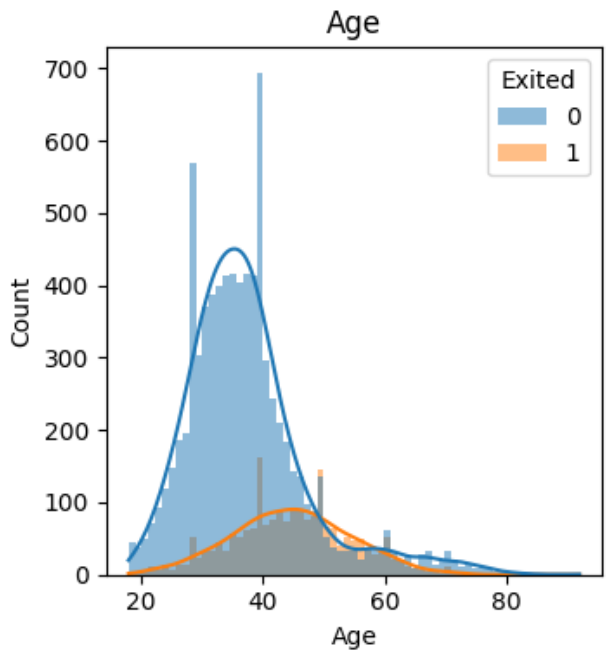
提出假設	與問題、假設相關的欄位、變數可能有哪些？
客戶持有產品數量多，相對較為忠實客群，流失率較低	客戶資料、持有產品數、是否流失
客戶帳務餘額少的, 越可能會流失	客戶資料、帳務餘額、是否流失
客戶年齡低可能有較低忠誠度，流失率高	客戶資料、年齡、是否流失
信用分數低，是否影響到借款成功率，導致有換行的趨勢，進而影響流失率	客戶資料、信用分數、是否流失
地區會不會影響流失率	客戶資料、居住國家、是否流失
客戶來往時間長短影響忠誠度，用越久越不容易流失	客戶資料、來往期間、是否流失
客戶是會員，可能流失率較低	客戶資料、是否會員、是否流失
客戶有持卡，可能流失率較低	客戶資料、是否持卡、是否流失
客戶性別，可能流失率會有落差	客戶資料、性別、是否流失

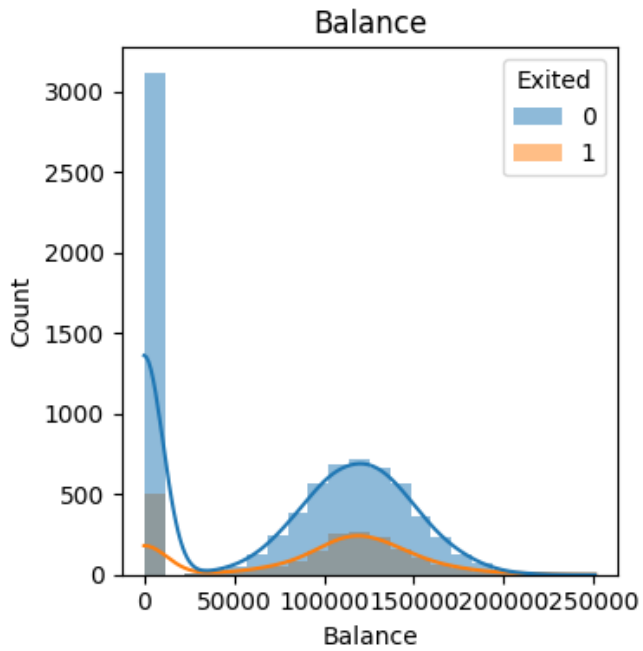
二、 開始你的分析 (連續型資料)

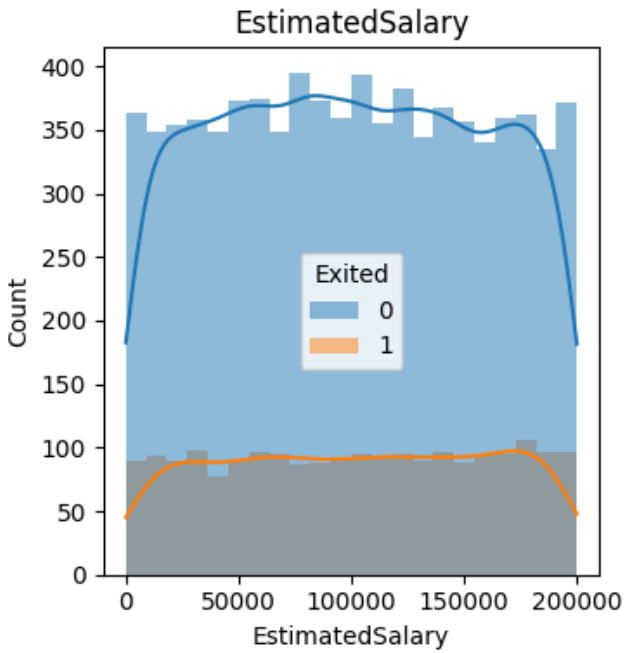
- CreditScore、Age 呈現常態分佈，在平均值有較高的流失數
- 40 ~ 50歲有較多的流失量(橘色)
- Balance 呈現雙峰分佈，在5萬以下及約12萬左右有較高的流失量
- EstimatedSalary 則是與流失率無太大關聯

提出假設	與問題、假設相關的欄位、變數可能有哪些？	驗證結果
客戶年齡低可能有較低忠誠度，流失率高	客戶資料、年齡、是否流失	否，流失數量與年齡成常態分佈
信用分數低，是否影響到借款成功率，導致有換行的趨勢，進而影響流失率	客戶資料、信用分數、是否流失	否，流失數量與信用分數成常態分佈
客戶帳務餘額少的, 越可能會流失	客戶資料、帳務餘額、是否流失	是，呈現雙峰分佈，餘額5萬以下的客戶與12萬左右有較高流失率



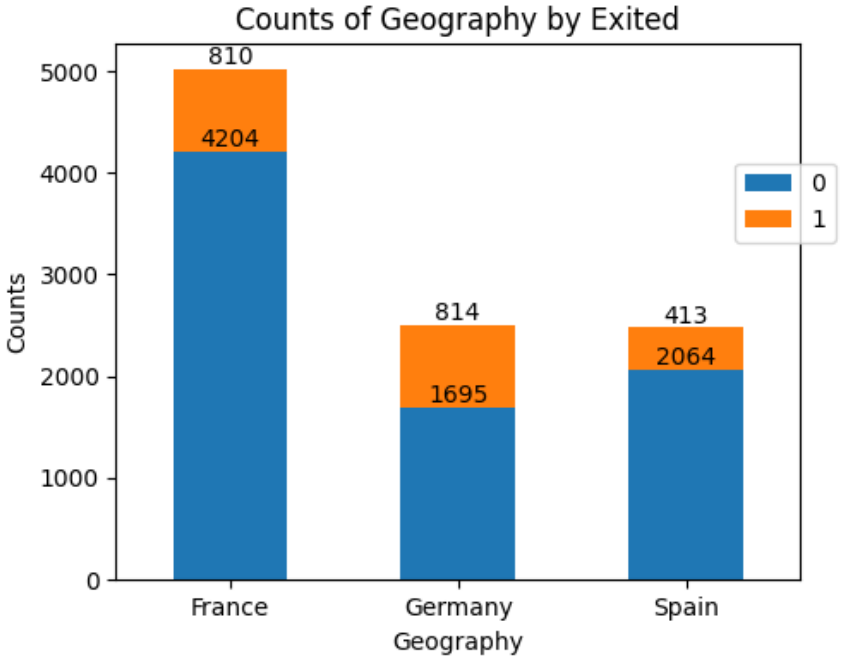


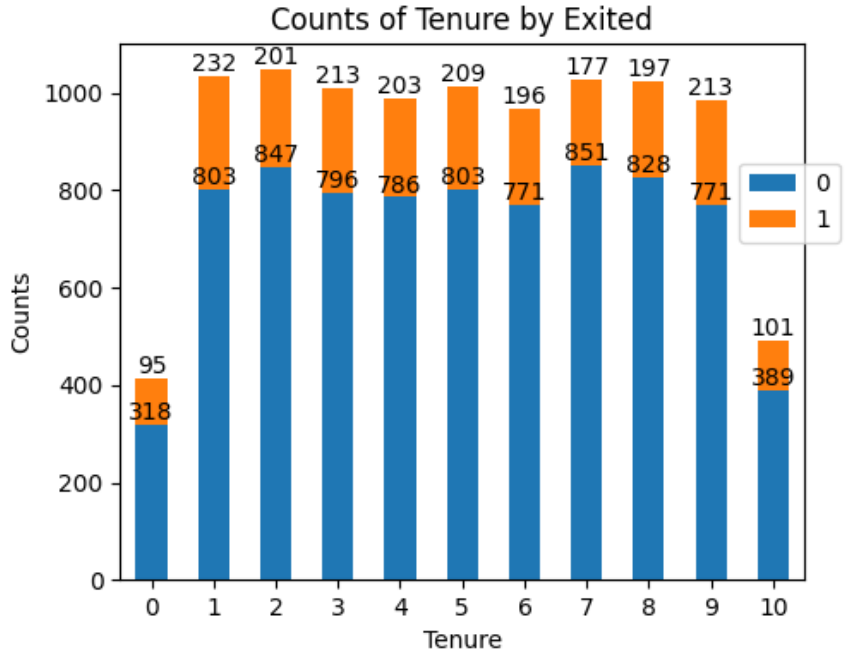


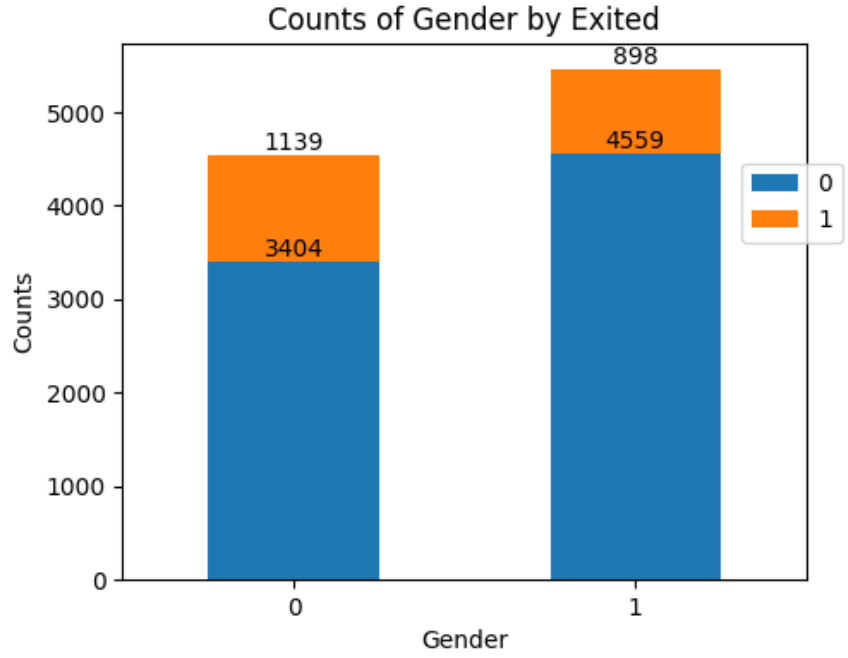


二、 開始你的分析 (離散型資料)

提出假設	與問題、假設相關的欄位、變數可能有哪些？	驗證結果
地區會不會影響流失率	客戶資料、居住國家、是否流失	是，德國地區流失率為西班牙與法國的2倍(32、16%)
客戶來往時間長短影響忠誠度，用越久越不容易流失	客戶資料、來往期間、是否流失	否，無論來往多久流失占比幾乎相同(~20%)
客戶性別，可能流失率會有落差	客戶資料、性別、是否流失	是，女性流失率高於男性(25% 、16%)

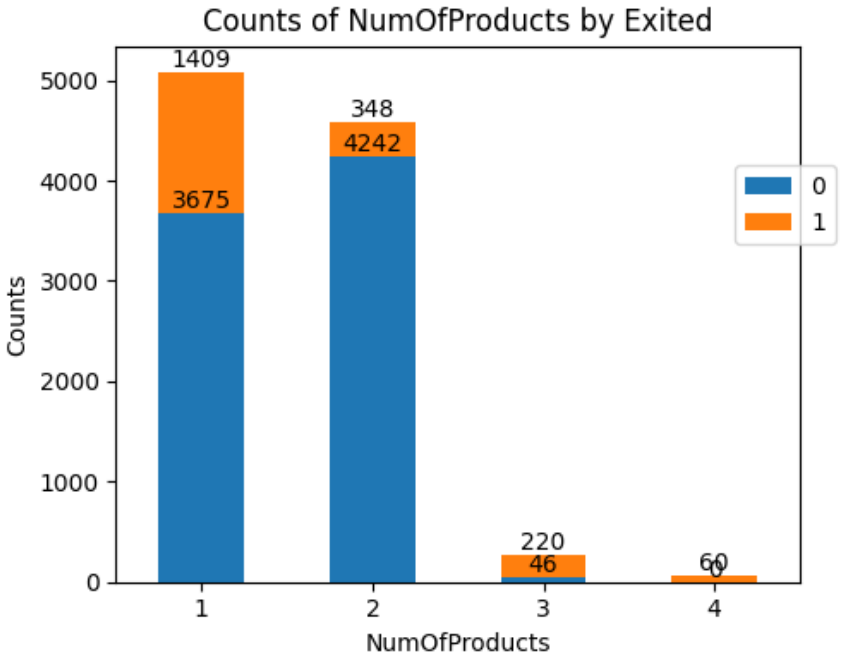


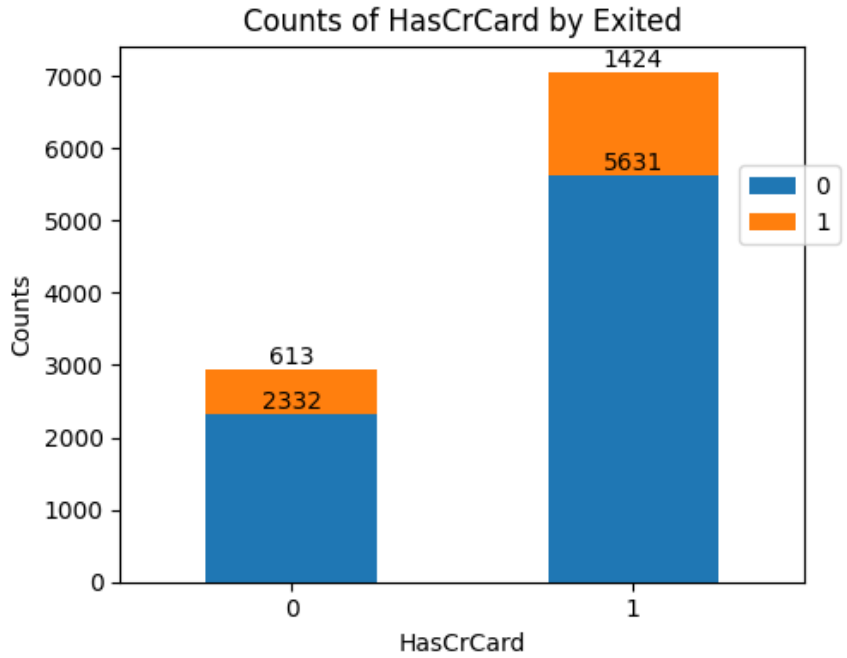


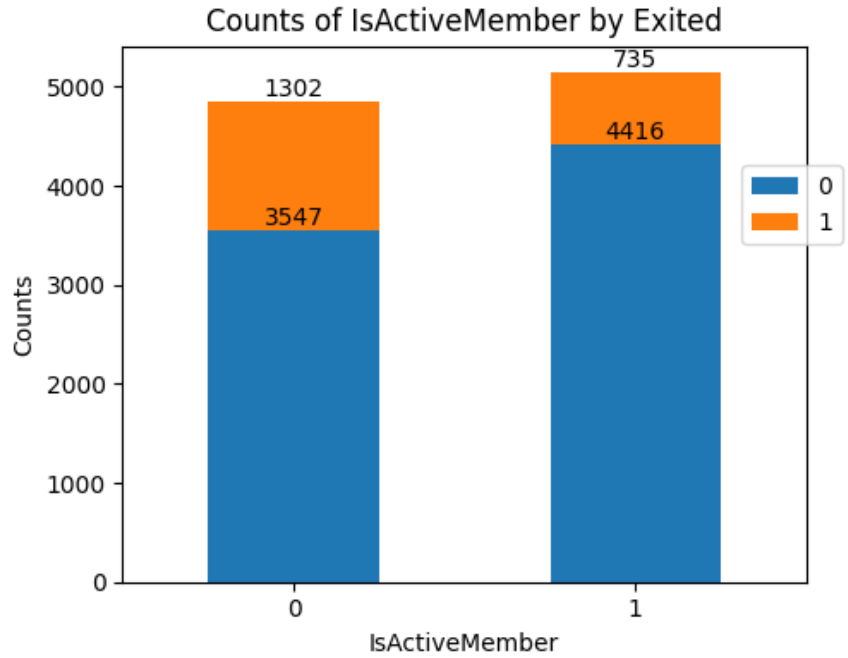


二、 開始你的分析 (離散型資料)

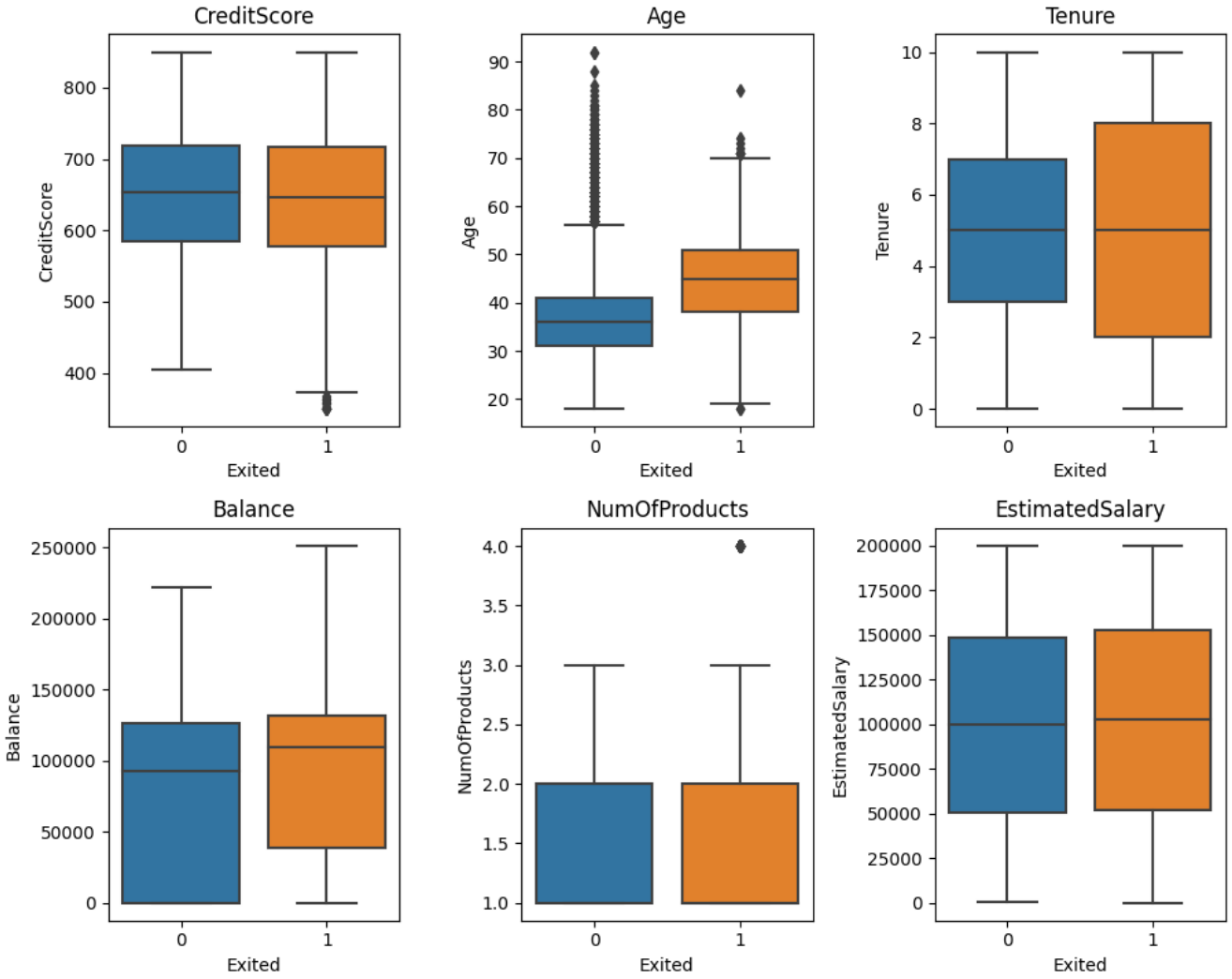
提出假設	與問題、假設相關的欄位、變數可能有哪些？	驗證結果
客戶持有產品數量多，相對較為忠實客群，流失率較低	客戶資料、持有產品數、是否流失	是，因為持有複數產品者流失率較低，因此假設為真(由於持有3個以上的基數少，固有較大的流失率，選擇將2個以上是為一類)
客戶有持卡，可能流失率較低	客戶資料、是否持卡、是否流失	否，無倫有無持卡皆有相同的流失率(~20%)
客戶是會員，可能流失率較低	客戶資料、是否會員、是否流失	是，無會員流失率為27%，反之則14%



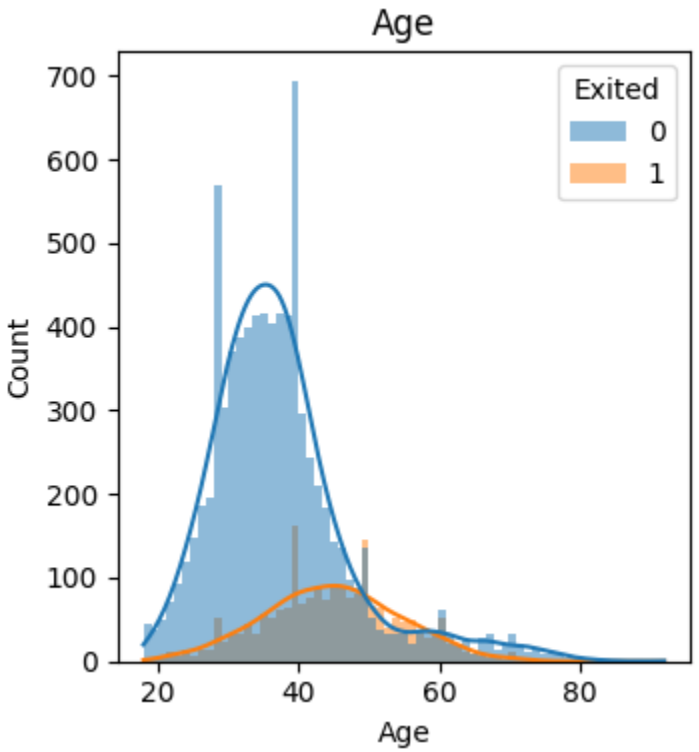




二、 開始你的分析(檢查離群值)

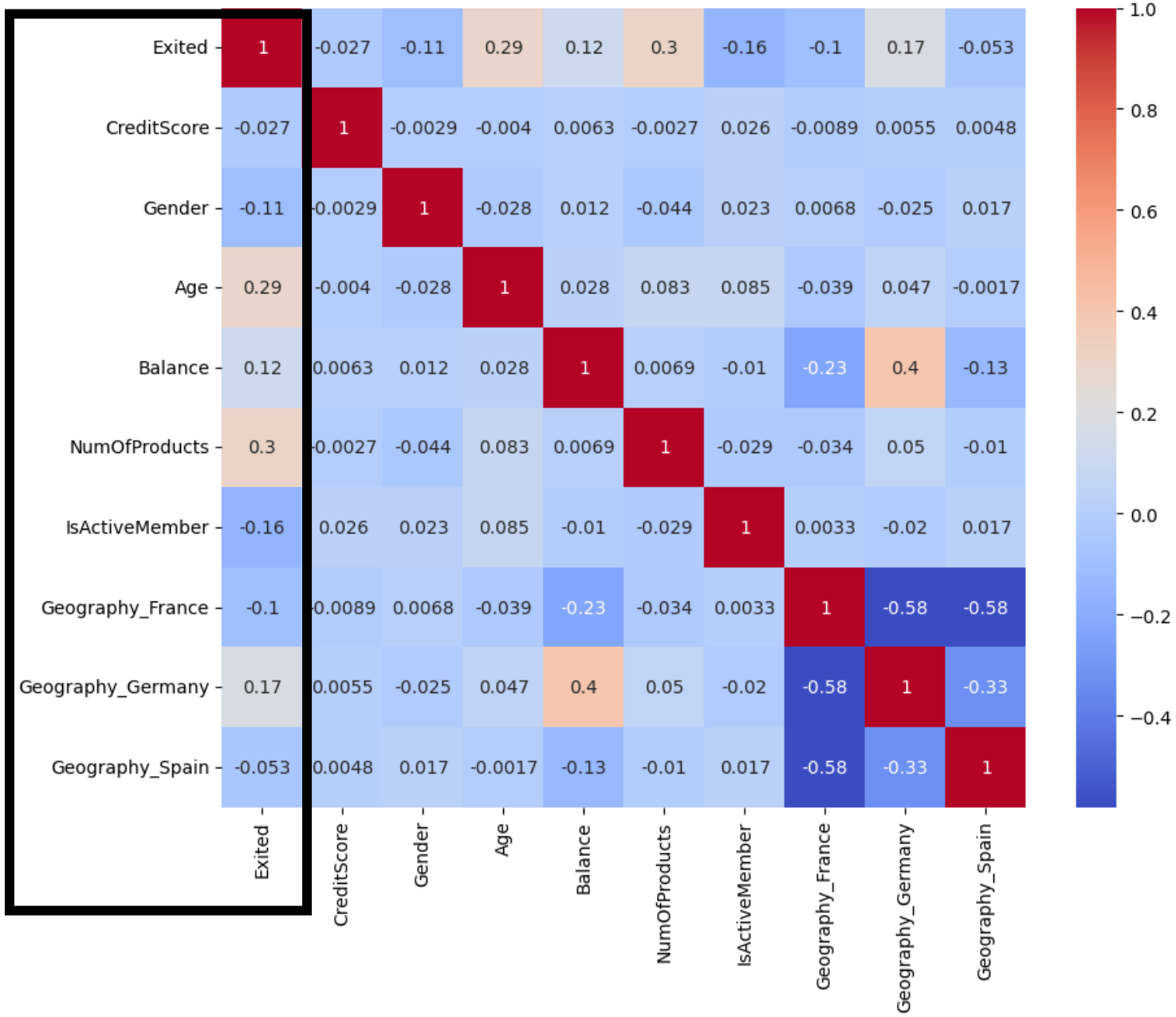


年齡有較大離群值(約為57歲以上)



二、 開始你的分析(特徵工程)

1. 保留會影響流失的特徵
   CreditScore、Geography 、 Gender 、 Age 、 Balance 、 NumOfProducts 、 IsActiveMember
2. 處理NumOfProducts，因3、4基數少，合併到2，轉為二元型資料
3. 離散型資料轉換為二元欄位(是否德國、是否法國、是否西班牙)
4. 連續型資料標準化
   1. CreditScore、 Age 常態分佈 > z轉換
   2. Balance雙峰分佈 > 最大最小轉換
5. 相關係數分析
   - 1 : 完全正相關
   - -1 : 完全負相關
   - 0 : 無關聯



三、提交你的分析報告

總結論

- 對流失較有影響力的特徵有(相關係數)
  1. 持有產品數 (0.29)
  2. 年齡 (0.28)
  3. 地區 (德國0.17)
  4. 是否會員 (0.15)
  5. 帳戶餘額 (0.12)
  6. 性別 (0.11)
- 可控制的建議
  - 提高客戶購買產品數
  - 維持中年齡層(35~50)的客戶數量
  - 使客戶成為有效會員
- 建立有效模型進行預測與評估

四、模型與預測

1. Baseline model
   - 邏輯回歸 LogisticRegression
   - 支持向量機 SVC
   - 決策樹 DecisionTreeClassifier
   - 隨機森林 RandomForestClassifier
   - 梯度提升樹 GradientBoostingClassifier
   - KNN KNeighborsClassifier
   - 貝葉司分類器 Naive_bayes
   - XGBooting XGB
2. 模型驗證(Area Under the Receiver Operating Characteristic Curve, AUC)
   - 選擇適合的模型(邏輯回歸、梯度提升樹)
3. 模型優化
   - 調整權重(流失客戶通常較稀少，會有不平衡的問題，此資料集流失比例為1:5
   - 超參數優化

四、模型與預測 - Baseline model

因為我們在意的是Exited為1的預測(流失)

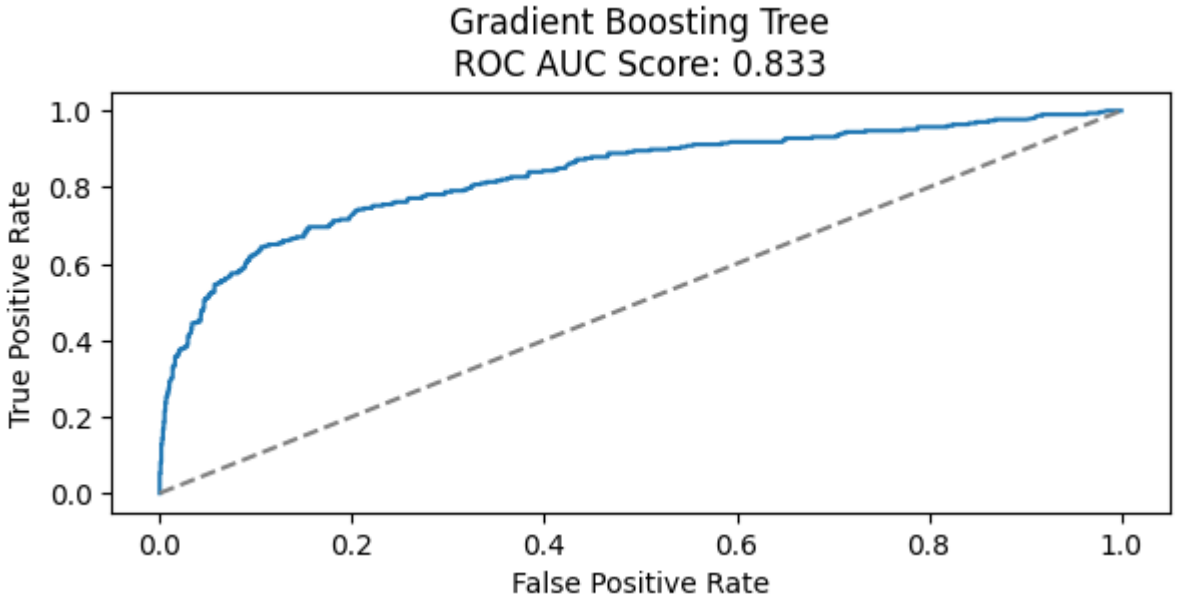
- **ROC_AUC : 衡量二元分類模型性能的一個常用指標**
  ROC : 該曲線代表真陽率（True Positive Rate）vs 假陽率（False Positive Rate）
  AUC : 代表該曲線下的面積，該值越接近1表示模型的性能越好

	Model	AUC Average Score	ROC_AUC_Score	ROC_AUC_Score_1
0	Gradient Boosting Tree	0.826762	0.704593	0.833426
1	Logistic Regression	0.792484	0.644177	0.811966
2	XGB	0.804581	0.695257	0.801698
3	NBC	0.776863	0.582373	0.792129
4	Random Forest	0.802822	0.705735	0.791129
5	SVC	0.758981	0.672915	0.782499
6	KNN	0.745890	0.679213	0.760071
7	Decision Tree	0.659379	0.657577	0.657364

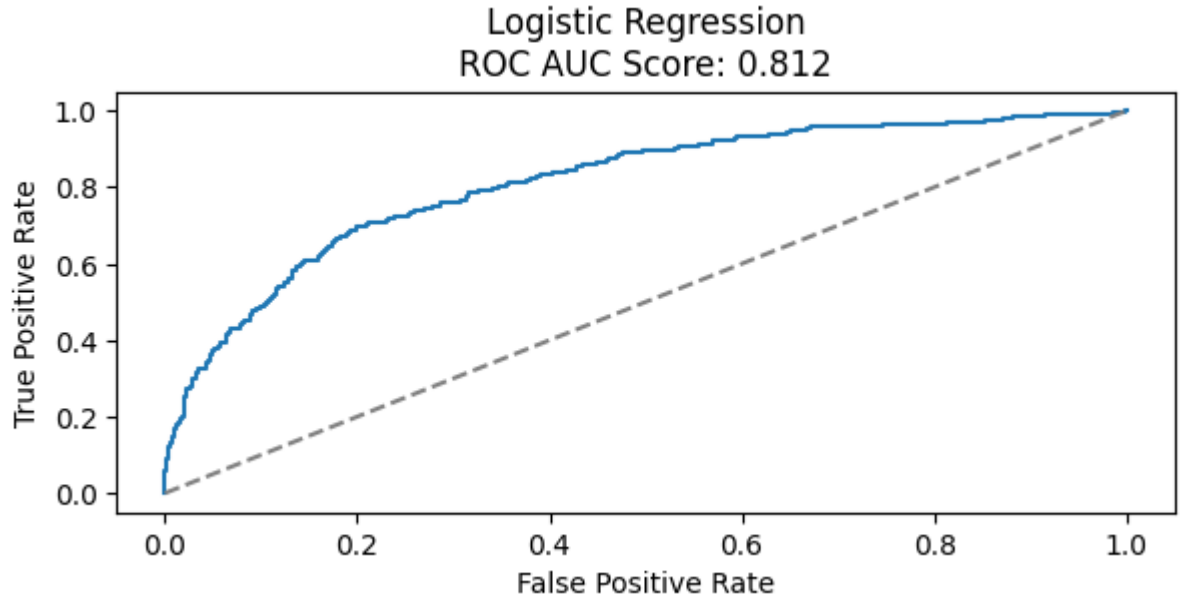
四、模型與預測 – 選擇模型(線下面積較高的2個)

曲線下面積：

- 1：完美預測分類
- 0.5：隨機55分類(灰線以下)
- 0：模型性能差，幾乎無能力預測



	precision	recall	f1-score	support
0	0.87	0.96	0.91	1194
1	0.75	0.45	0.56	306
accuracy			0.86	1500
macro avg	0.81	0.70	0.74	1500
weighted avg	0.85	0.86	0.84	1500



	precision	recall	f1-score	support
0	0.85	0.96	0.90	1194
1	0.70	0.32	0.44	306
accuracy			0.83	1500
macro avg	0.77	0.64	0.67	1500
weighted avg	0.82	0.83	0.81	1500

四、模型與預測 – 優化模型

1. 調整流失比例權重 (0 : 1 = 1 : 5)
2. 超參數調整、驗證曲線微調
3. 這兩個模型AUC略為上升

	Model	ROC_AUC_Score_1
0	LOGR_opt	0.812491
1	LOGR	0.811966
2	GDBT_opt	0.834521
3	GDBT	0.833426

LOGR_opt

	precision	recall	f1-score	support
0	0.93	0.66	0.77	1194
1	0.38	0.80	0.51	306
accuracy			0.69	1500
macro avg	0.65	0.73	0.64	1500
weighted avg	0.82	0.69	0.72	1500

GDBT_opt

	precision	recall	f1-score	support
0	0.87	0.97	0.91	1194
1	0.78	0.42	0.54	306
accuracy			0.86	1500
macro avg	0.82	0.69	0.73	1500
weighted avg	0.85	0.86	0.84	1500

五、BackUp

- EDA : Python
- Feature Engineer : Python
- ML : Python
- 參考資料 :
  - 自己筆記
  - Google
  - ChatGPT
  - Link