

Kaggle 資料集實作

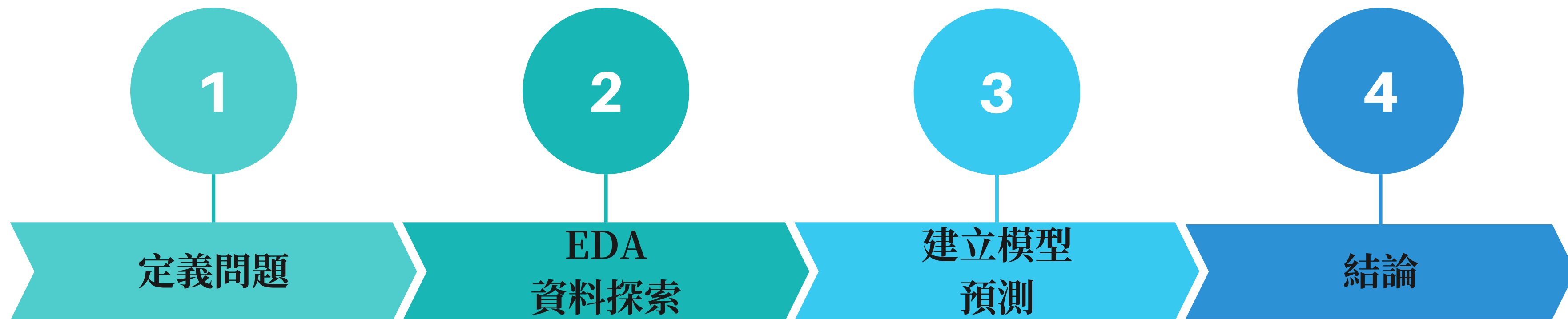
銀行客戶 流失分析報告

張智淵



<https://www.cleartouch.in/what-is-customer-churn-and-how-do-you-prevent-it/>

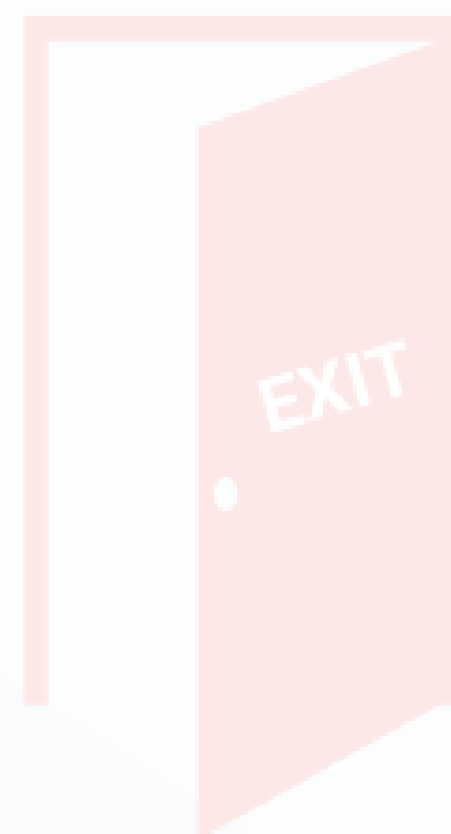
目錄



CUSTOMER
CHURN

1

定義問題

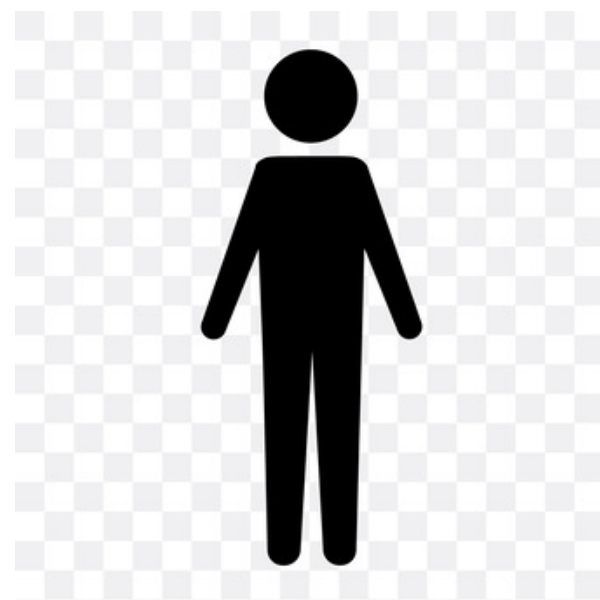


1

定義問題

影響客戶流失的因素？

(基本資料、交易行為、產品使用狀況)



客戶ID
性別
年齡
地理位置
薪資



信用評分
購買產品
活躍用戶
信用卡
存款
年數


預測目標：客戶是否流失？



CUSTOMER
CHURN

2

EDA / 資料探索



The background illustration depicts a business scenario of customer churn. On the left, five stylized human figures in business attire (blue, pink, grey, yellow, and teal) stand in a line. On the right, three similar figures are shown running away from the viewer towards a red door labeled 'EXIT'. The entire scene is set against a light grey background with a subtle grid pattern.

2

EDA / 資料探索

相關係數矩陣熱圖

- 顯示數據中各個特徵之間的相關性
- 和客戶流失(Exited)相關係數較高的特徵:

年齡(Age)

餘額(Balance)

是否為活躍用戶(IsActiveMember)

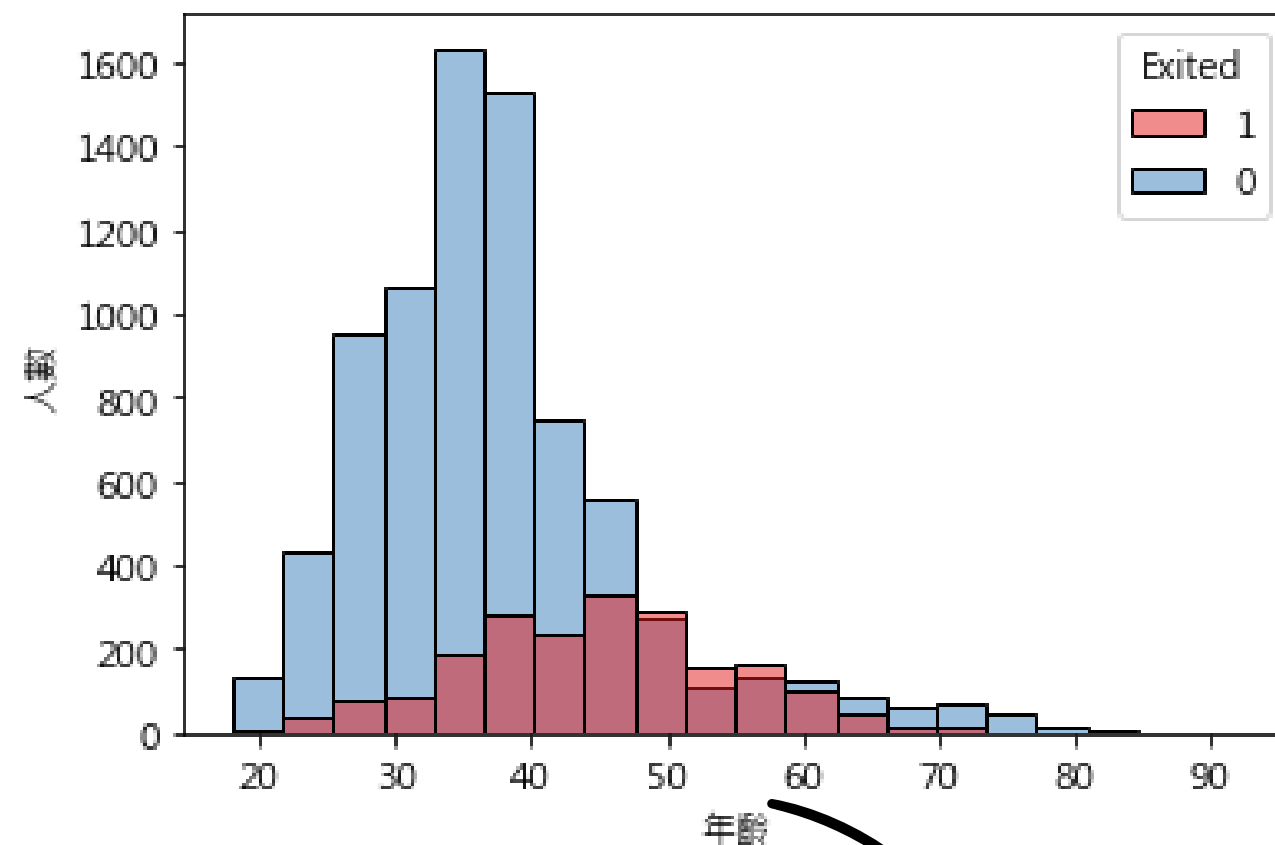


2

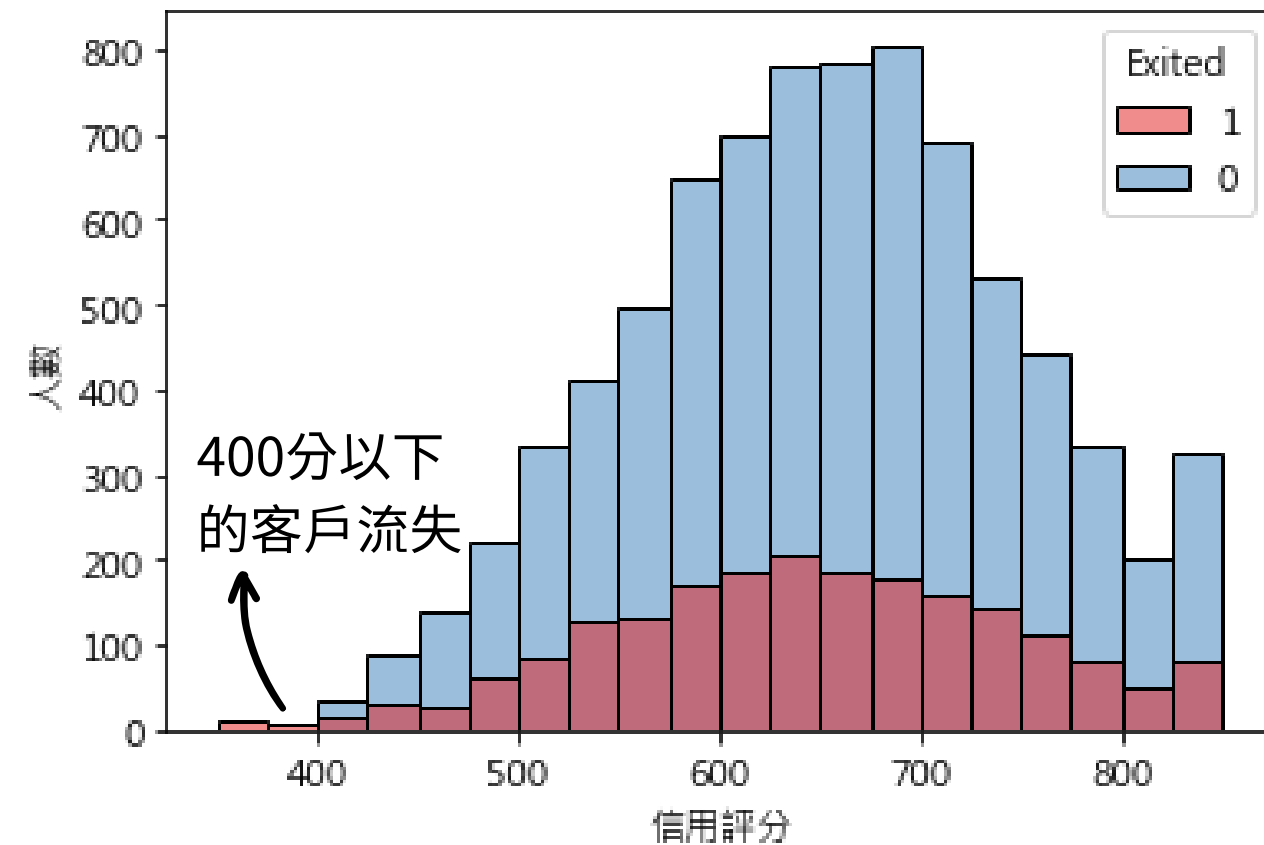
EDA / 資料探索

流失vs非流失客戶

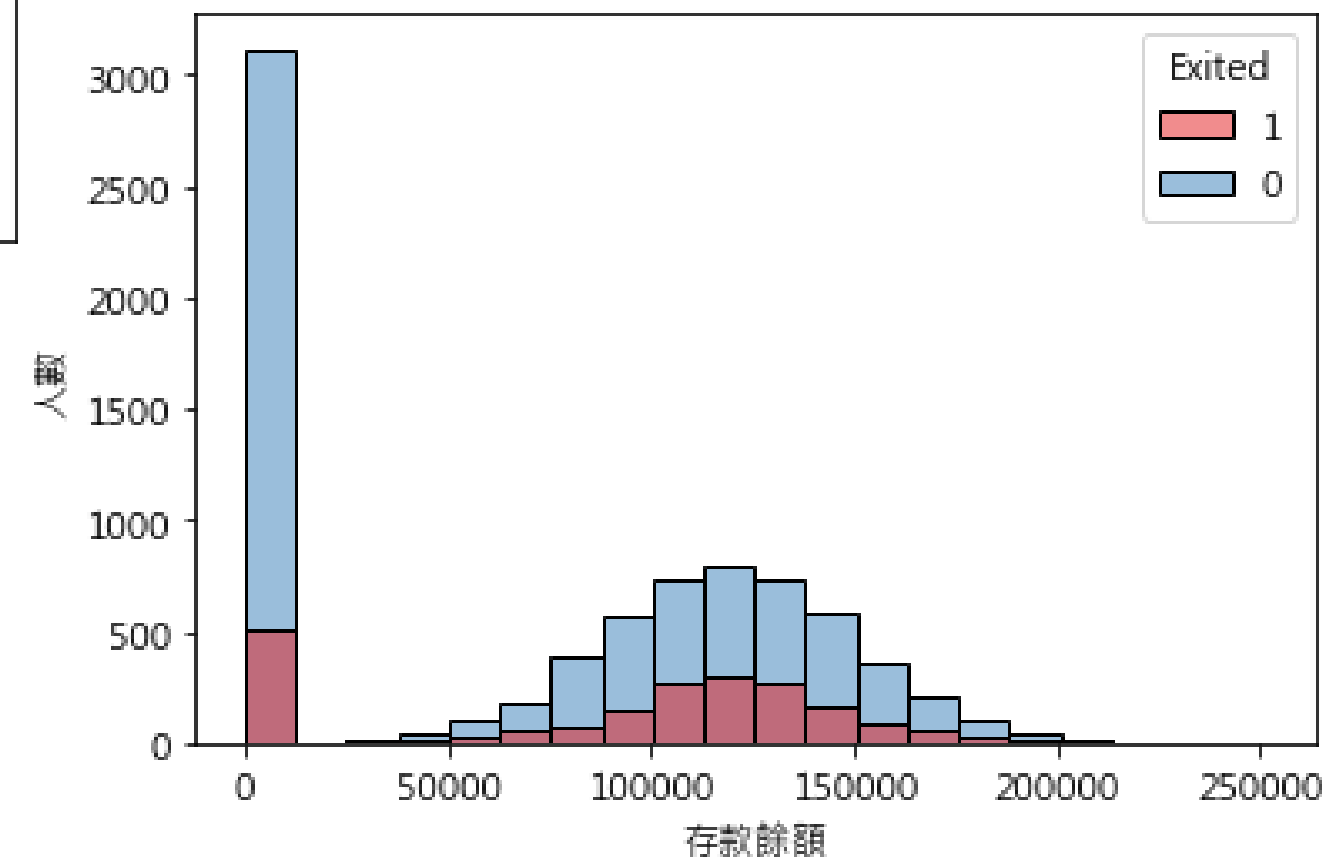
客戶年齡



年齡在接近50~60歲的客戶流失率高



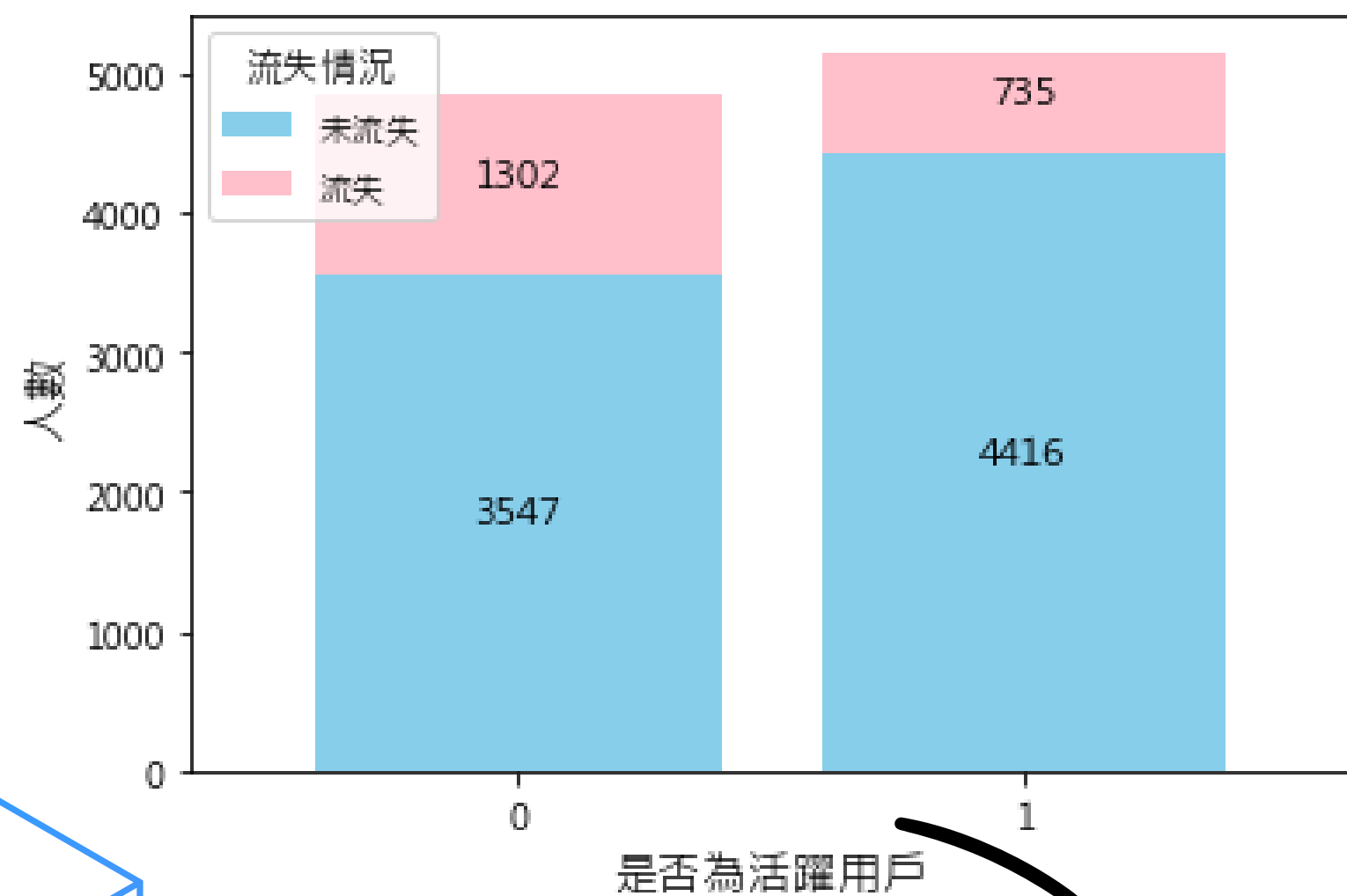
客戶信用評分



客戶帳戶餘額

2

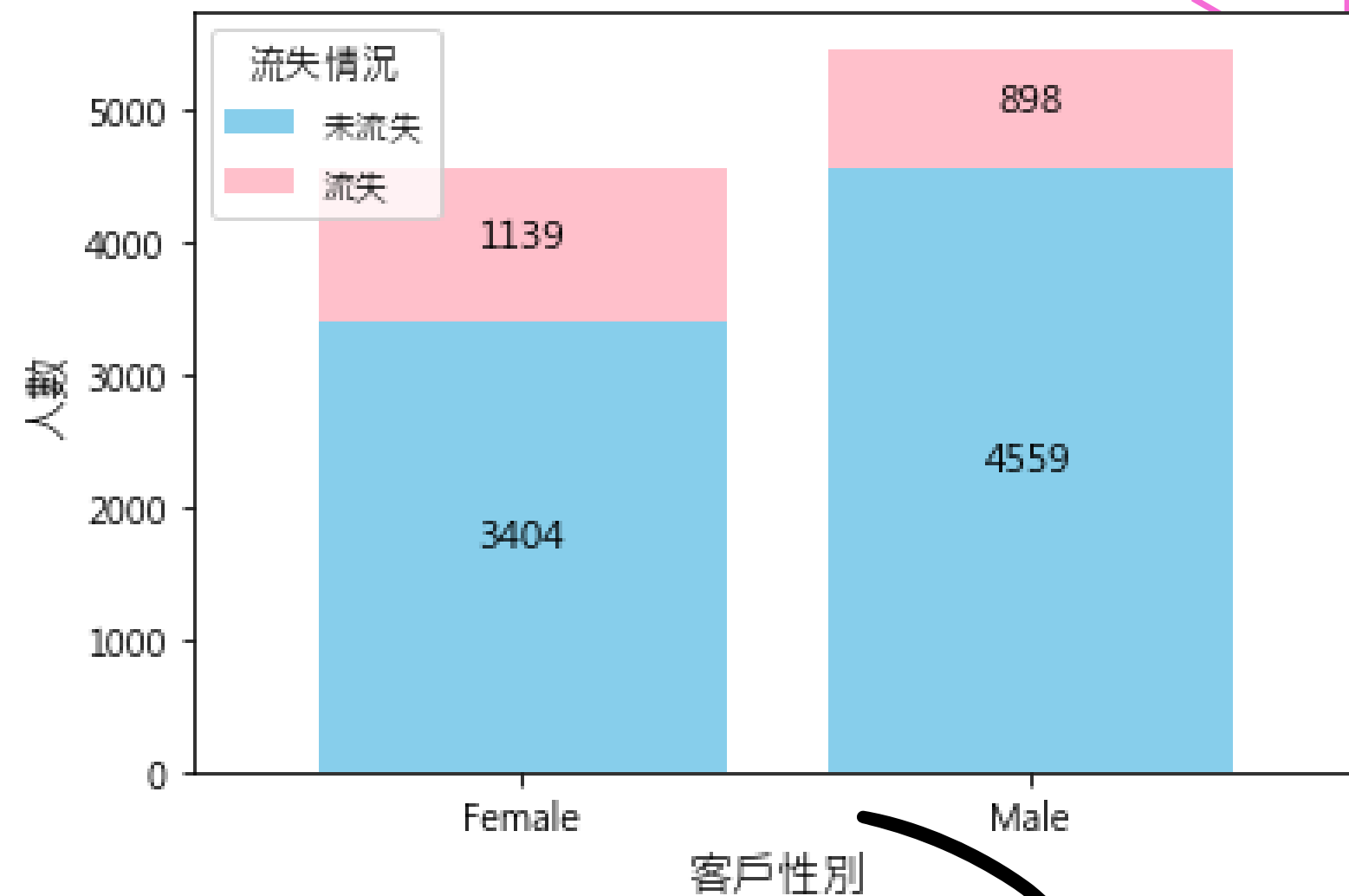
EDA / 資料探索



活躍客戶

非活躍用戶流失率約

27%



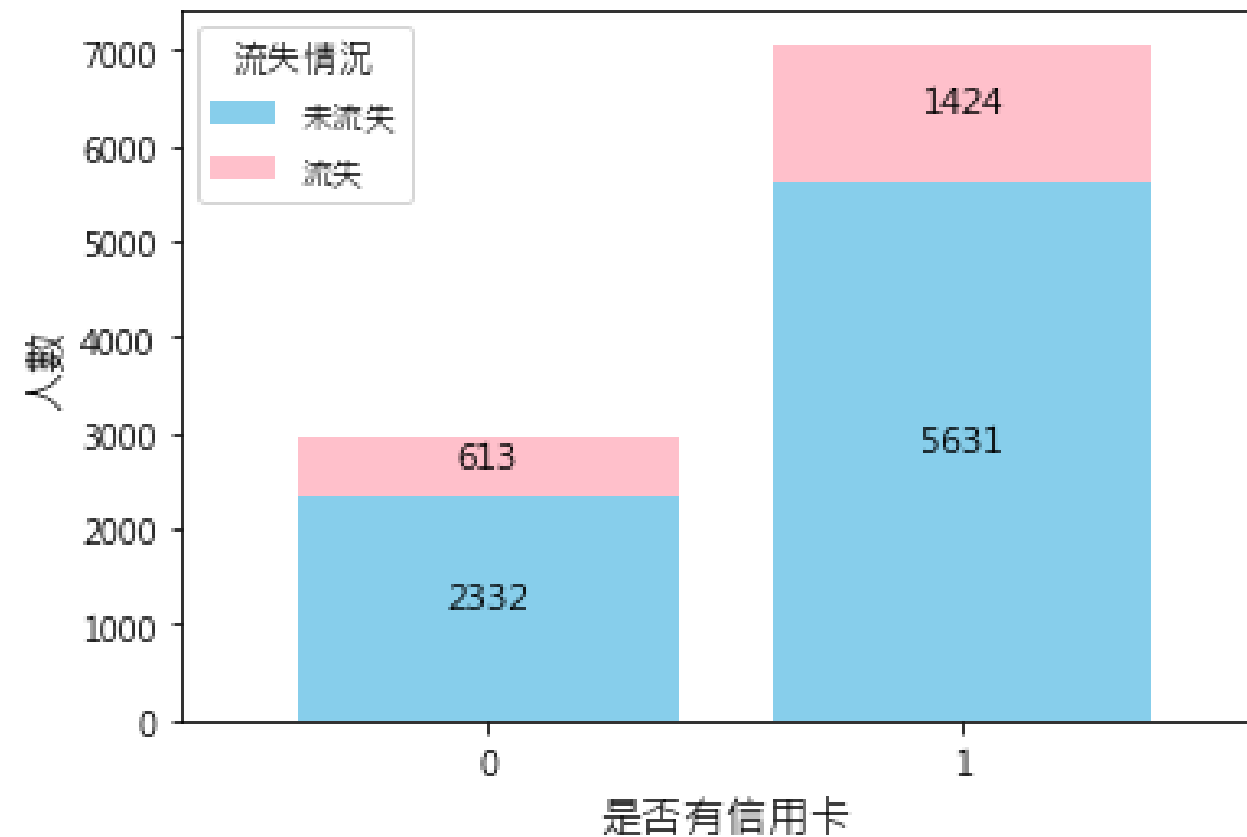
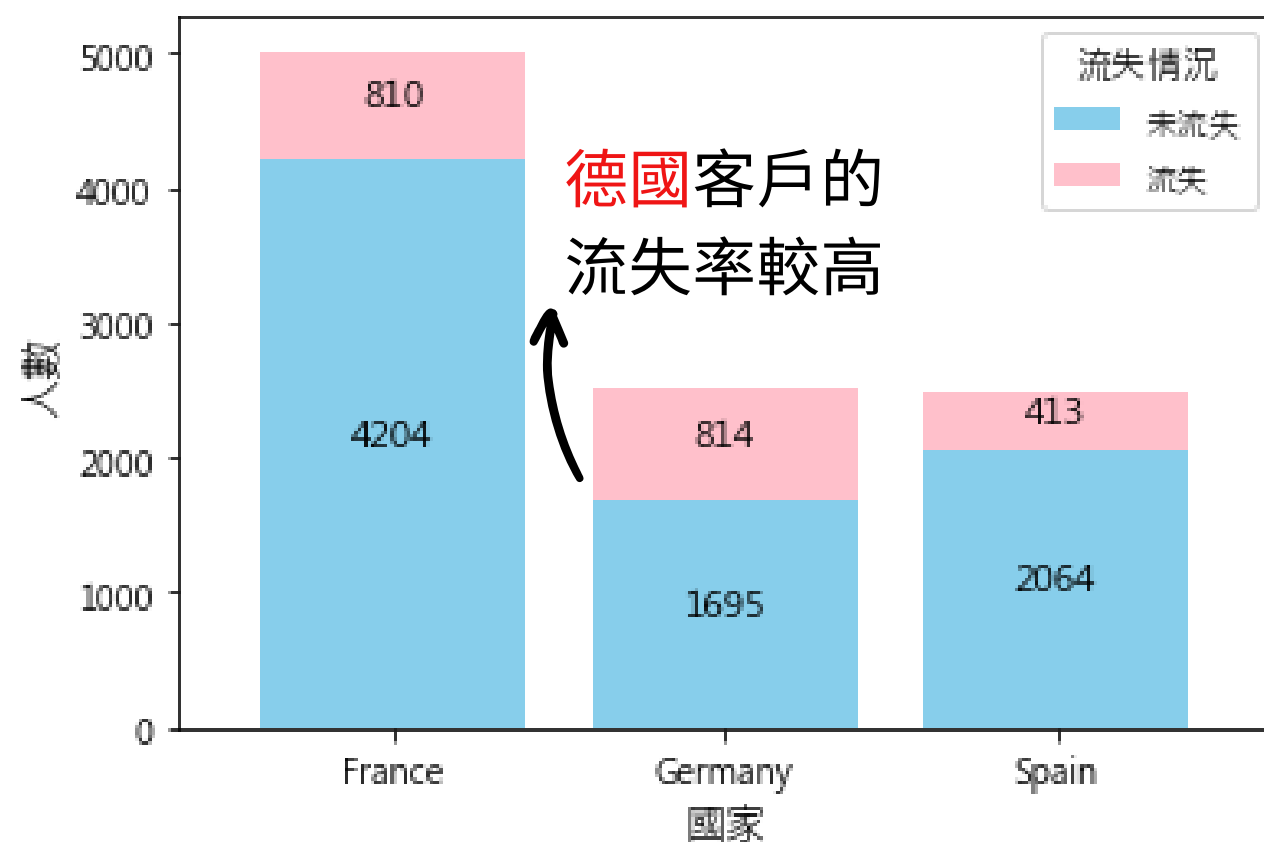
客戶性別

女性客戶流失率高

2

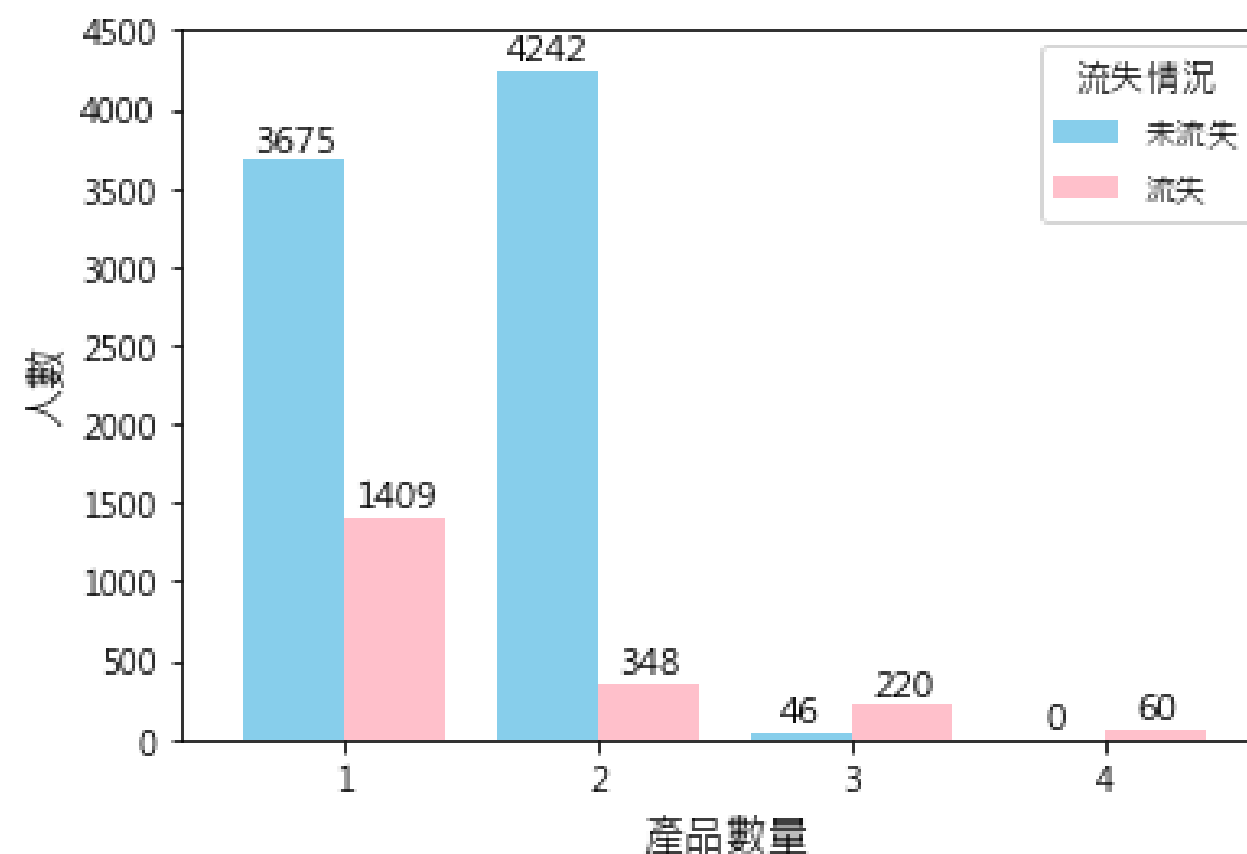
EDA / 資料探索

客戶地理位置



客戶是否有信用卡

客戶有無信用卡對流失沒有顯著差異

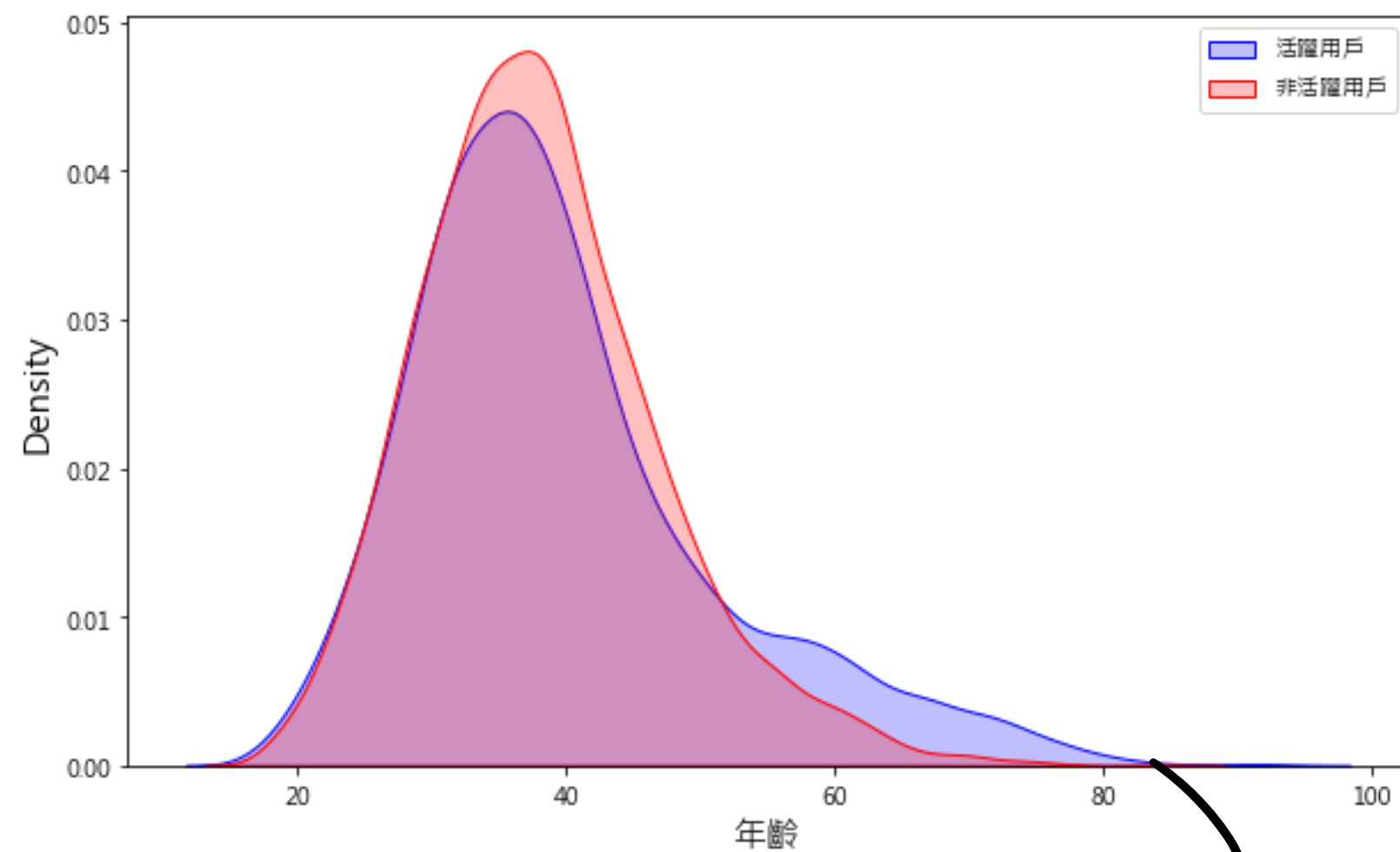


客戶購買銀行產品數量

購買3、4種產品的客戶流失比例很高

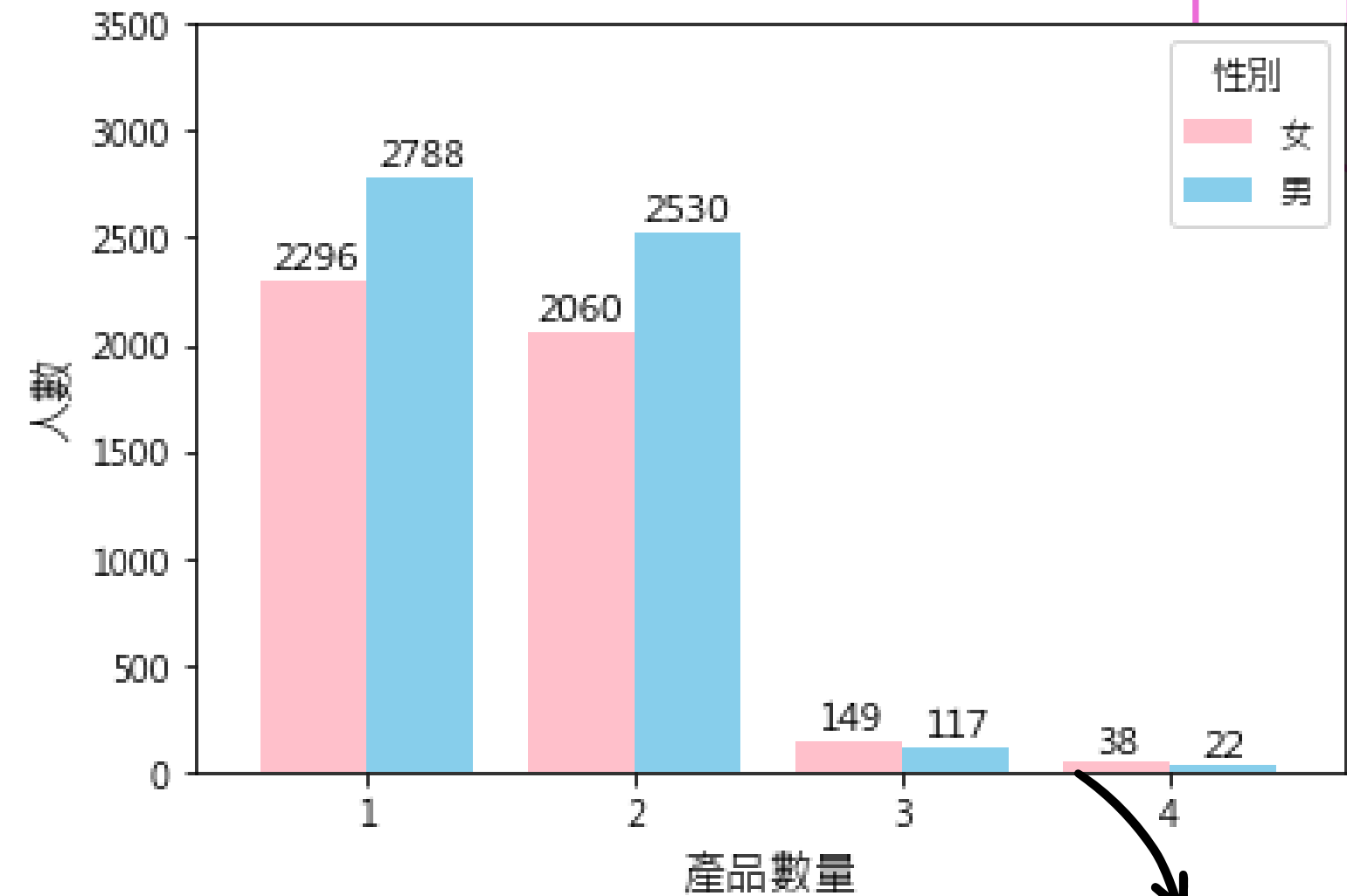
2

EDA / 資料探索



活躍客戶-年齡

高年齡層活躍客戶的比例高



產品數量-性別

女性購買3、4種銀行產品的比例更高

CUSTOMER
CHURN

3

建立預測模型

EXIT

3

建立預測模型

01

資料預處理

類別資料編碼、特徵標準化

02

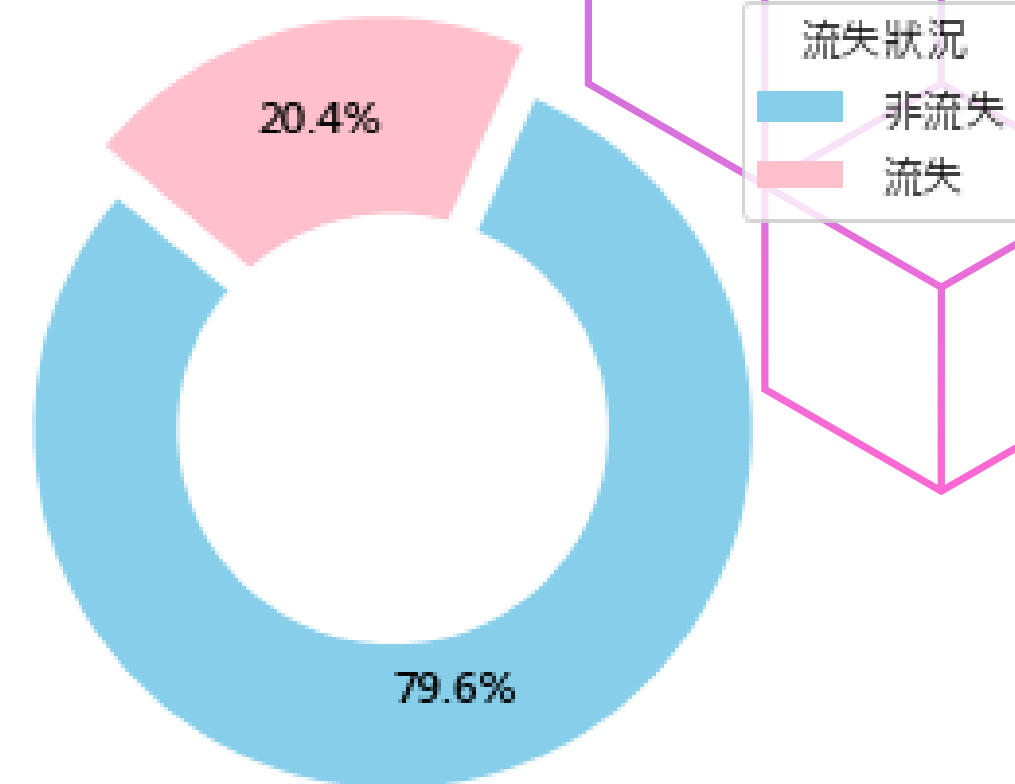
建立模型

邏輯回歸、隨機森林模型
Halving Grid Search 調整
參數

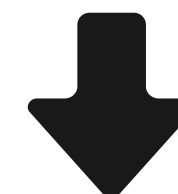
03

驗證
評估模型

交叉驗證(Cross validation)
比較模型表現(F1 Score)
特徵重要性



*不平衡資料(Imbalanced data)



- 對每個訓練集裡的類別給予
權重(class_weight)
- 評估指標選擇**F1 Score**比較

3

建立預測模型

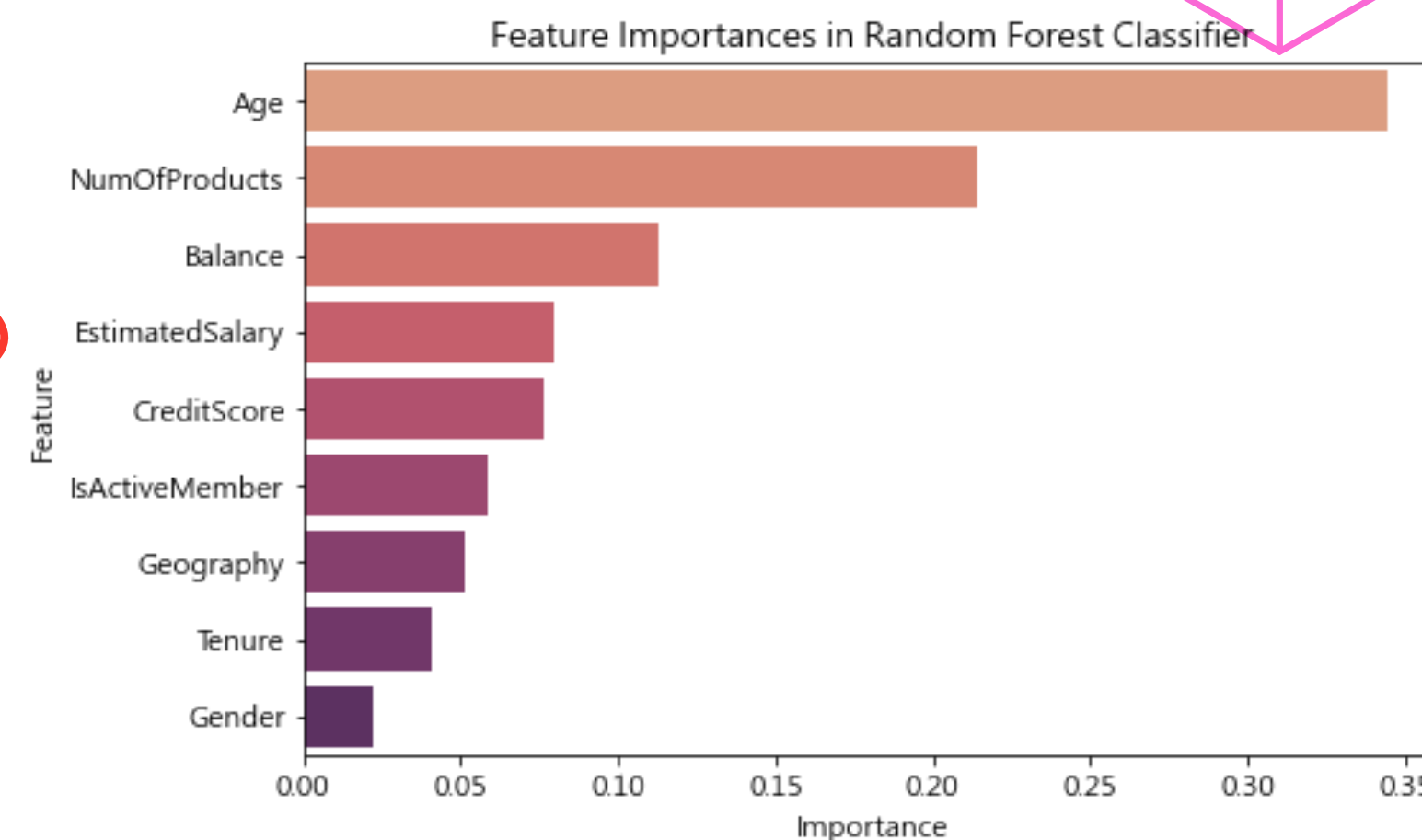
交叉驗證分數

	Accuracy	Precision	Recall	F1 Score
LogisticRegression	0.7900	0.3900	0.0569	0.0993
LogisticRegression (HalvingGridSearch)	0.8084	0.6007	0.1787	0.2749
RandomForest	0.8614	0.7553	0.4600	0.5738
RandomForest (HalvingGridSearch)	0.8215	0.5485	0.6937	0.6140

預測結果

	Accuracy	Precision	Recall	F1 Score
LogisticRegression	0.8013	0.6133	0.1743	0.2714
LogisticRegression (HalvingGridSearch)	0.8013	0.6133	0.1743	0.2714
RandomForest	0.8613	0.7978	0.4647	0.5873
RandomForest (HalvingGridSearch)	0.8260	0.5736	0.7033	0.6319

特徵重要性排名



CUSTOMER
CHURN

4

結論

EXIT

從資料探索和建立模型後得到的結果

- 非活躍客戶的流失率較高
- 年齡在接近50~60歲的客戶流失率高
- 購買三、四種產品的客戶流失率高，且女性購買的比例較高
- 高年齡層客戶能保持活躍的比例較高
- 年齡(Age)、購買銀行產品的種類(NumOfProducts)和餘額(Balance)等特徵在隨機森林模型的重要性較高