

客戶流失預測

羅健華



目標



1. 透過資料視覺化，對客戶特質進行初步了解
2. 建立一個分類模型，預測銀行客戶是否即將流失



資料來源

Kaggle是一個知名的數據分析競賽平台，此次練習以以下連結的資料集作為資料來源

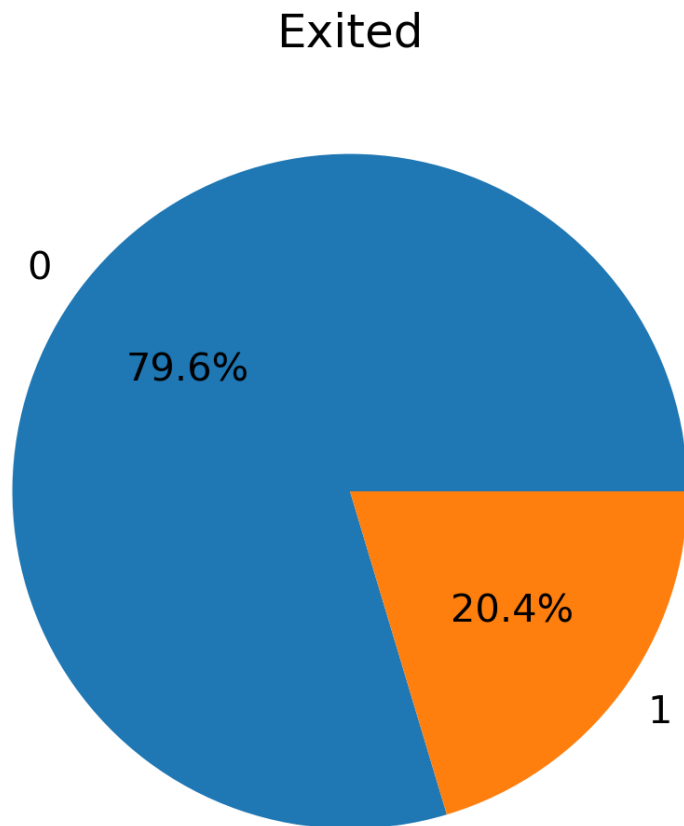
<https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling>



變數

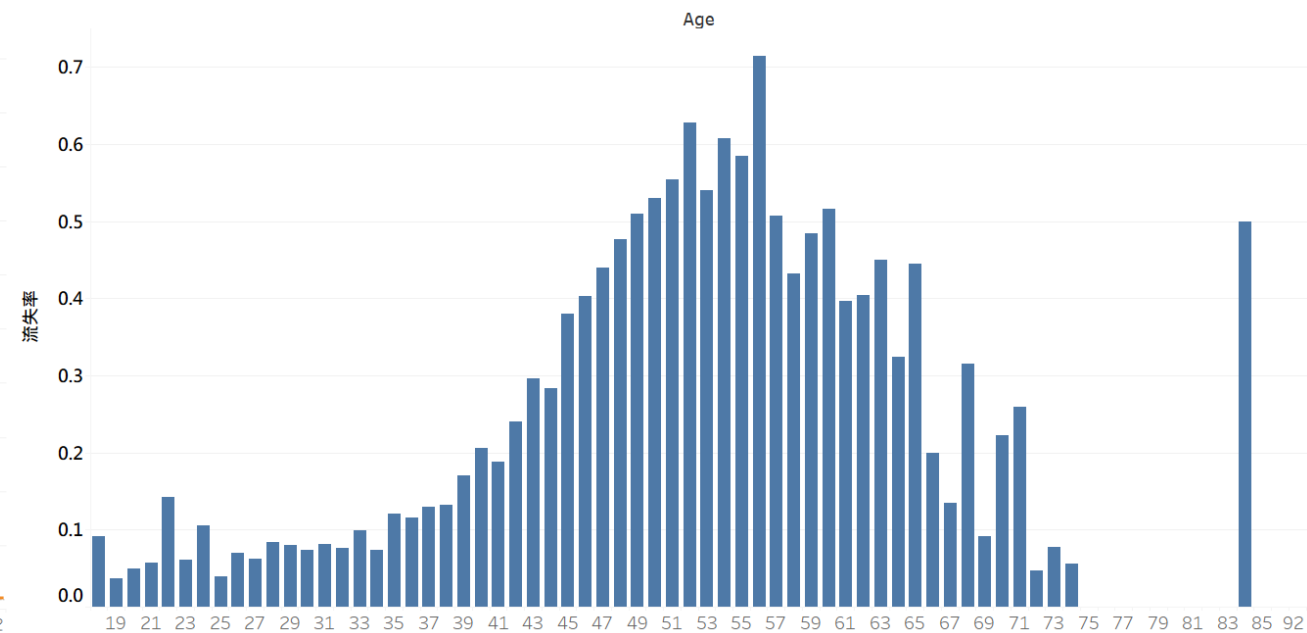
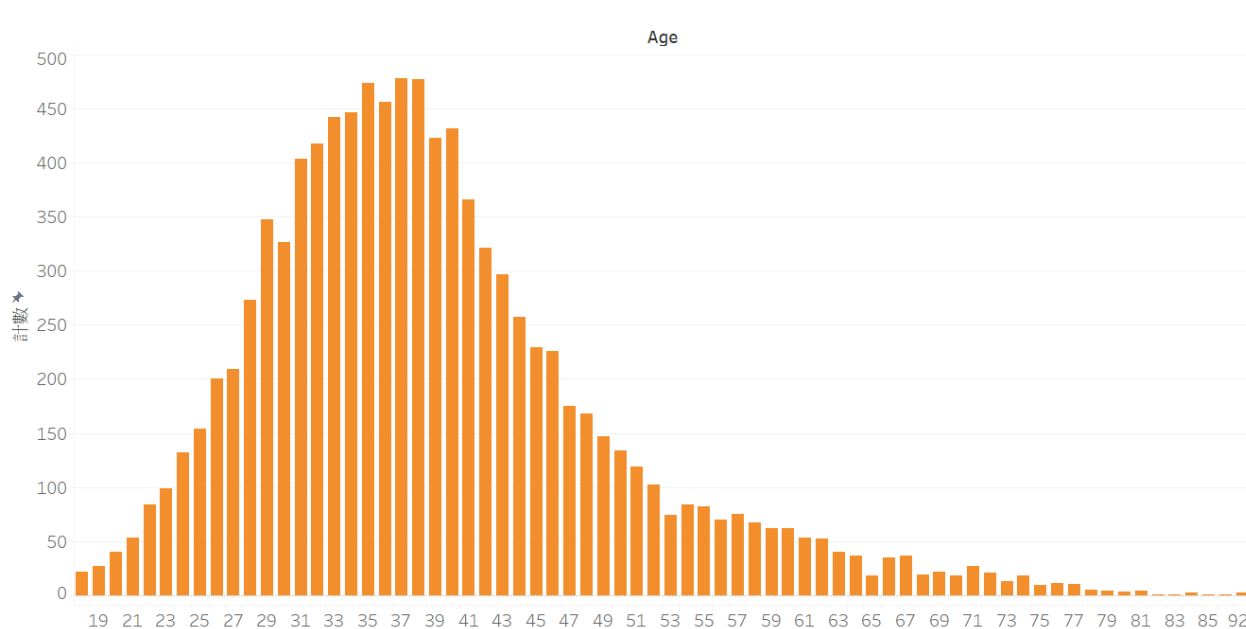
dependent variable
Exited
independent variable
Age
Gender
Geography (France 、 Germany 、 Spain)
Estimated Salary
Tenure
Balance
Credit Score
Number Of Products
Has Credit Card
Is Active Member

流失率



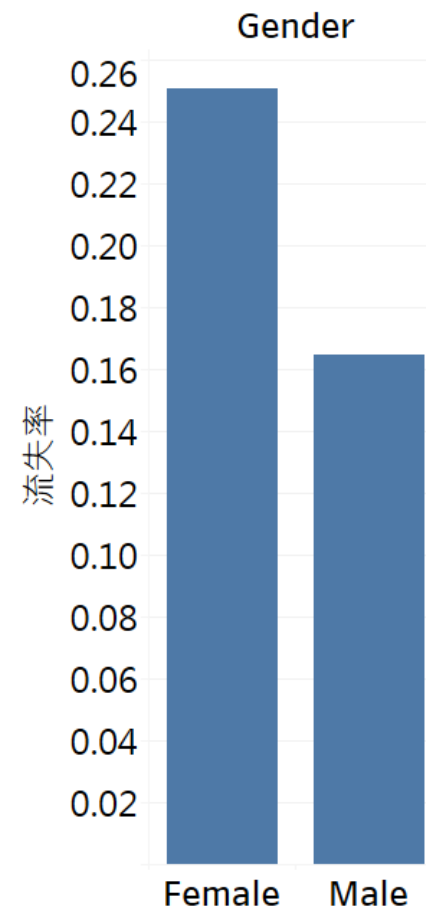
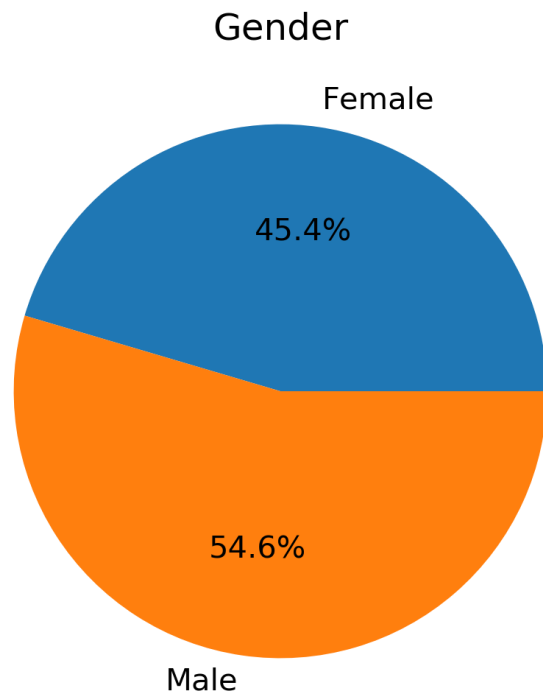
在資料集中，約有20%的客戶流失，此數字可當作一個比較的基準，後續模型的正確率應至少大於80%，才有比較的意義。

年齡



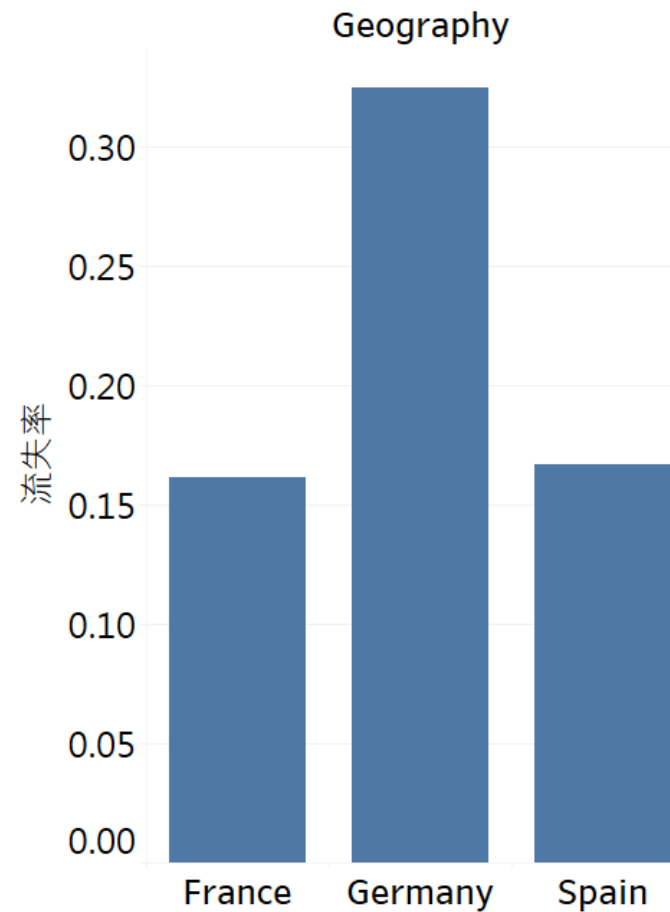
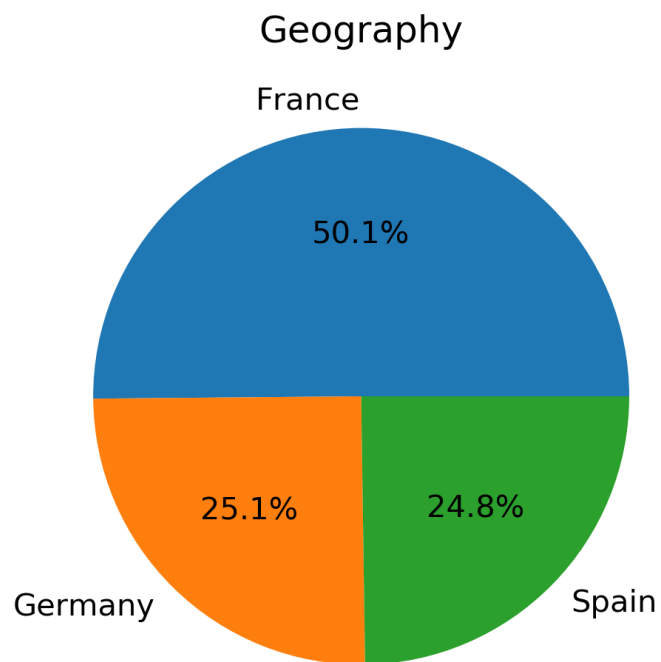
1. 隨著年齡的增加，客戶流失率逐漸升高，並在56歲達到最大值，之後又開始降低

性別



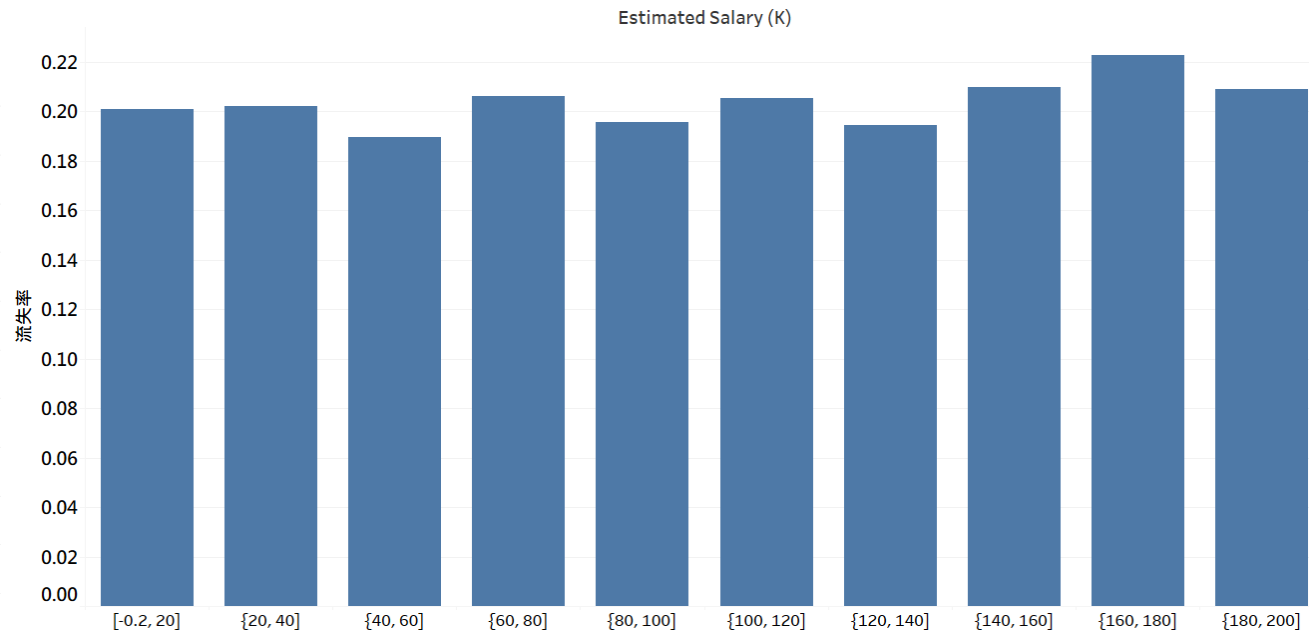
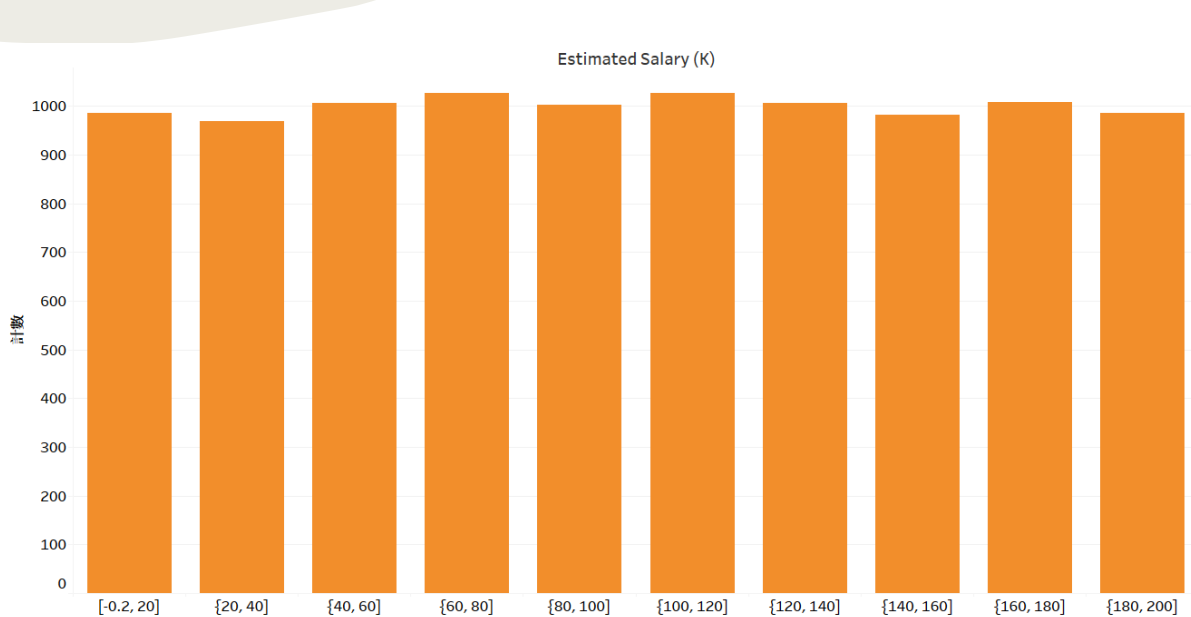
女性客戶的比例低於男性，但流失率卻高於男性

地區



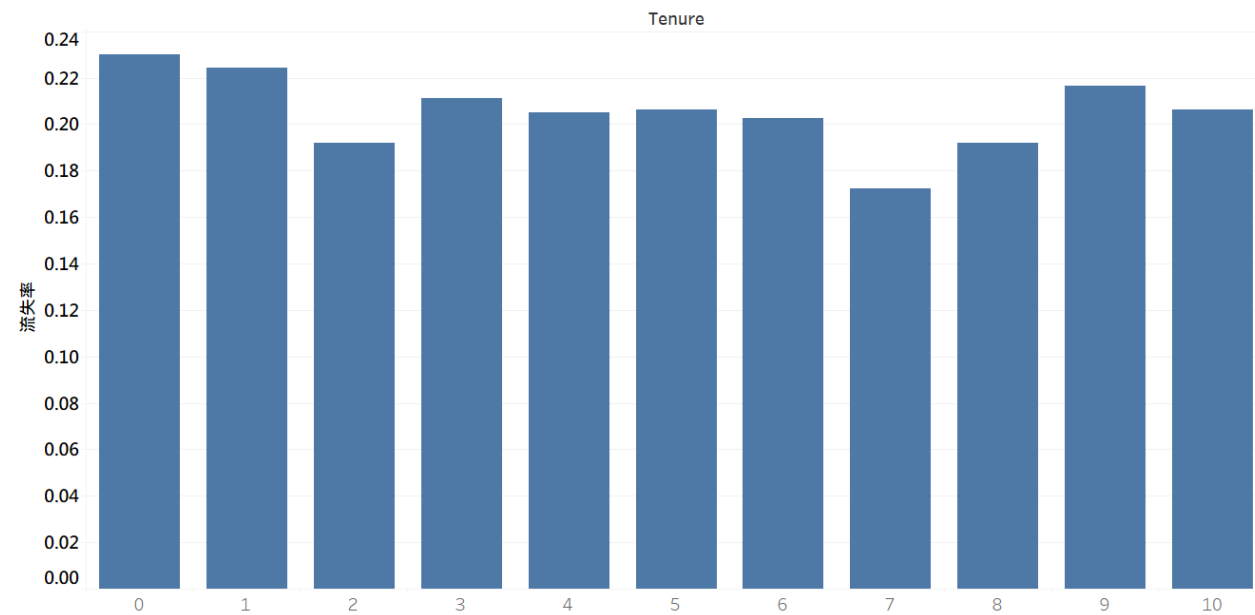
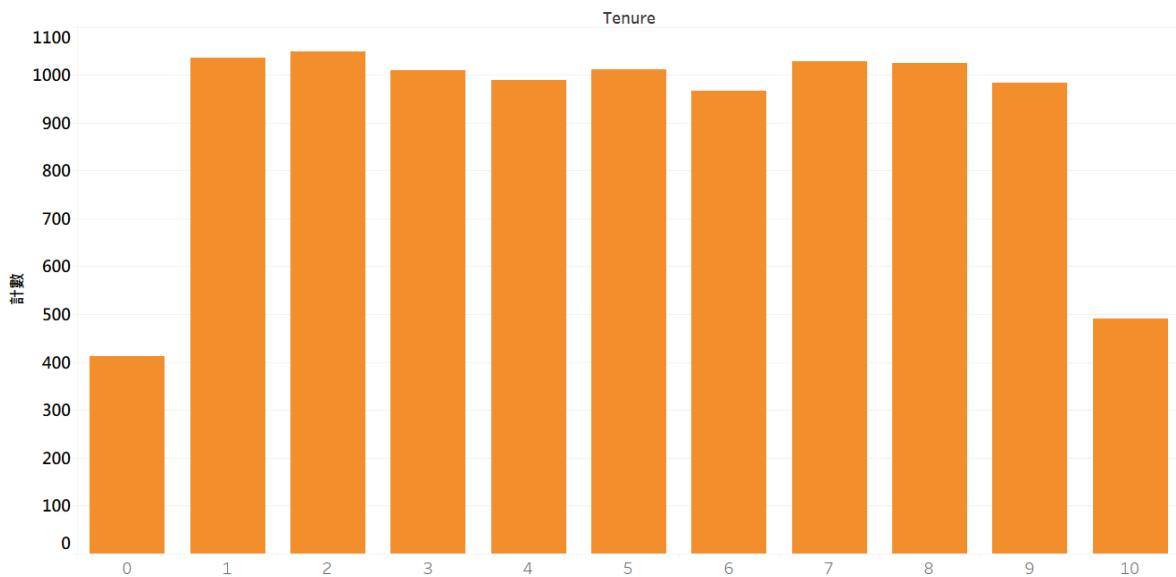
1. 大部分的客戶來自法國
2. 德國的客戶流失率異常偏高，顯示在該地區可能存在問題

估計薪水



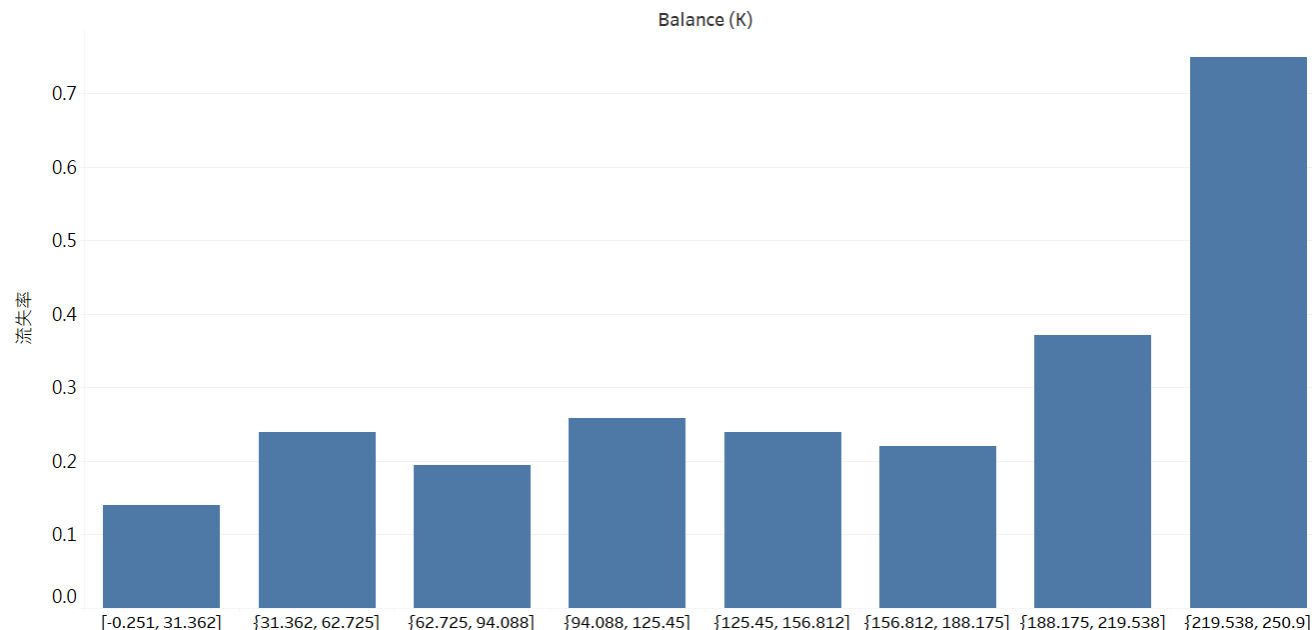
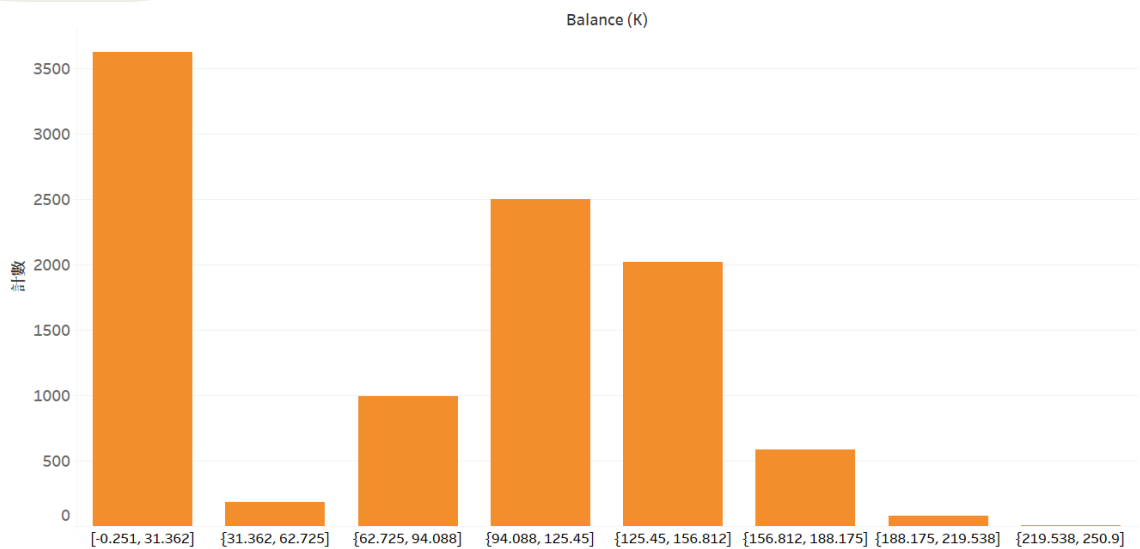
在薪水方面，不管是分布狀況，或是流失率，都沒有明顯的差異

成為客戶 時間長短



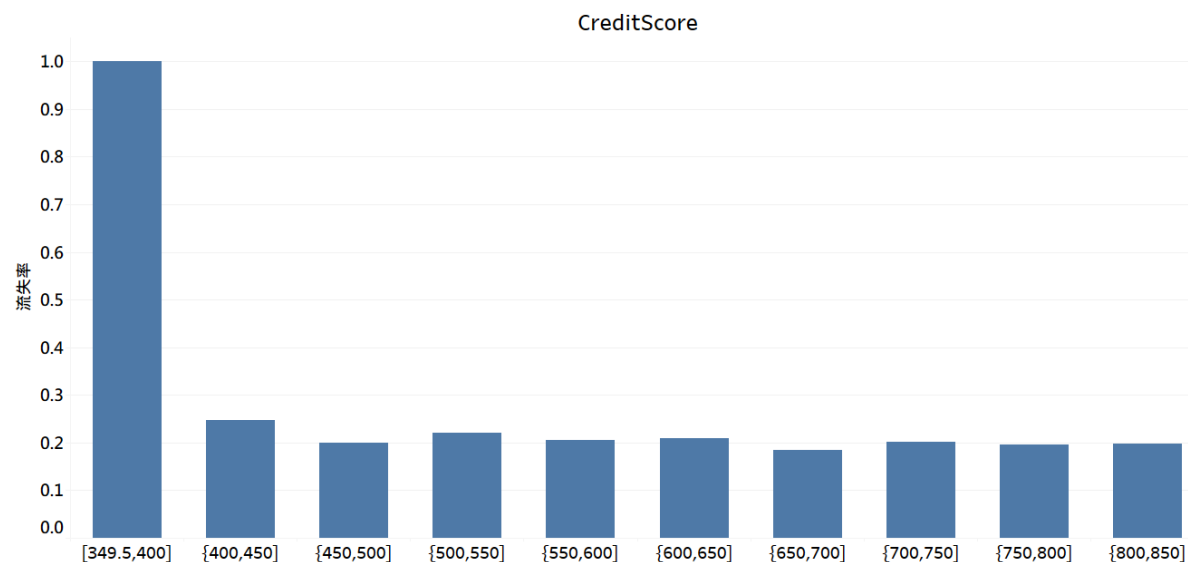
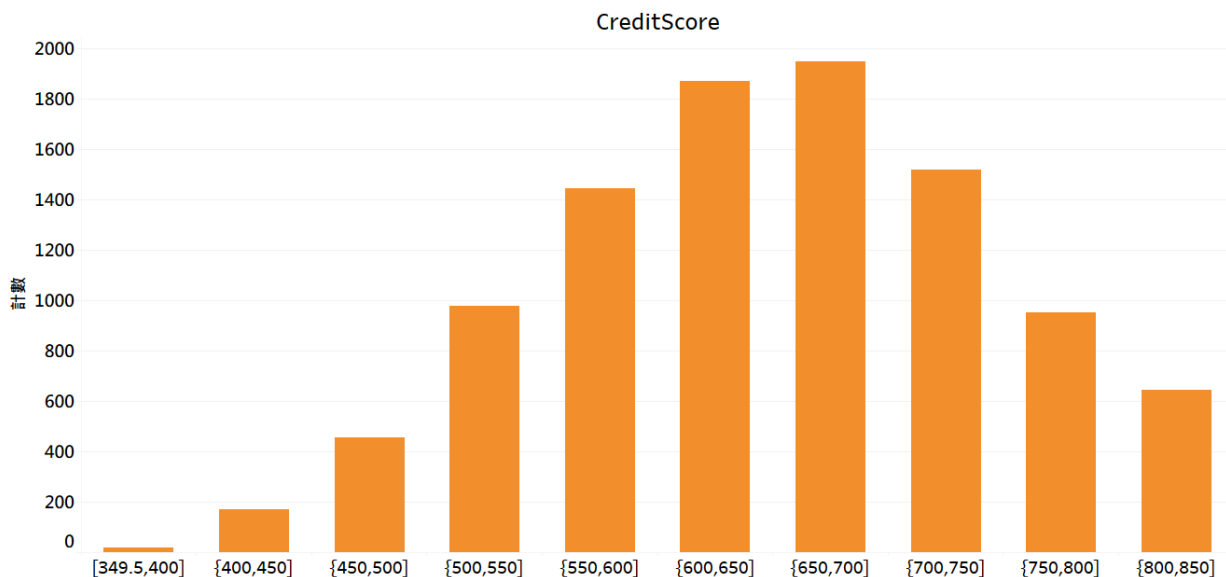
- 1.今年成為客戶的數量比往年較少，值得注意是否有獲取新客戶上的問題
- 2.在流失率上，沒有明顯的差異

戶頭餘額



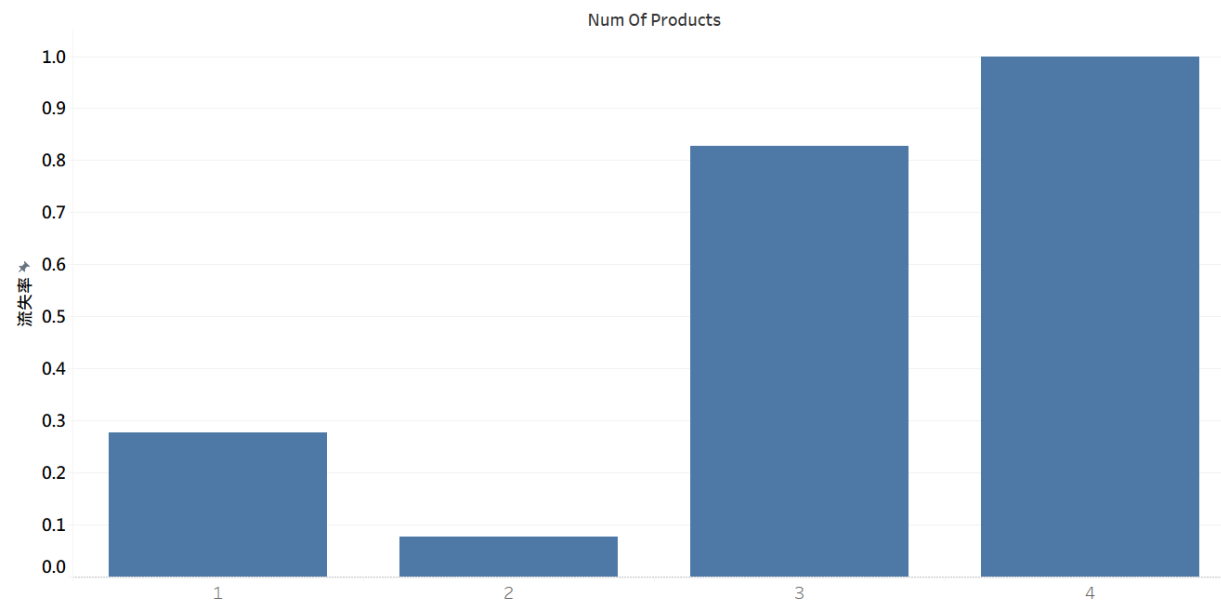
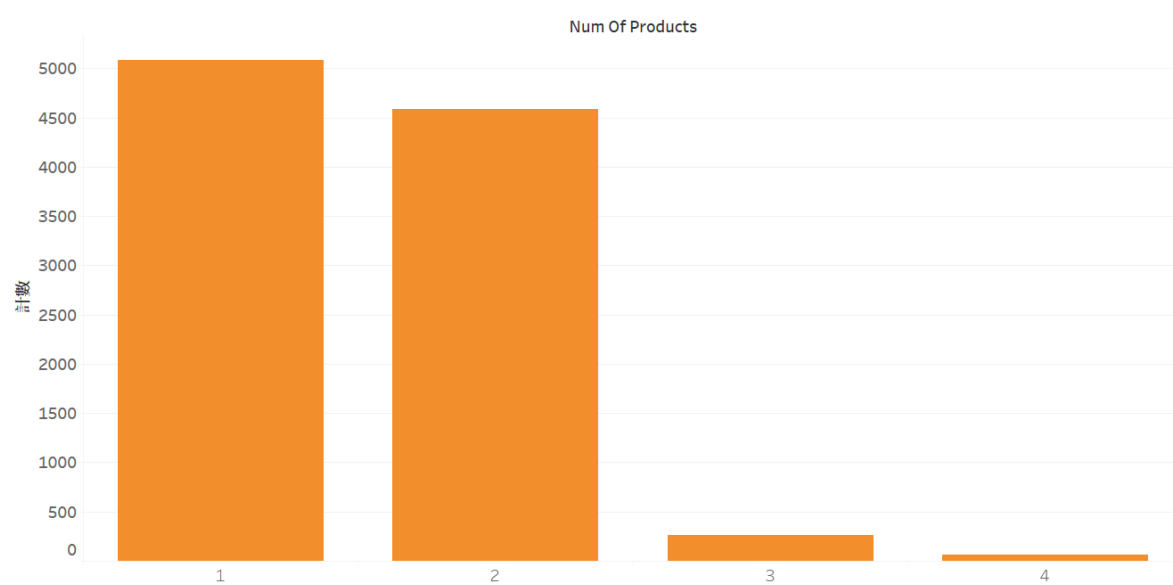
1. 36%的客戶，戶頭餘額少於3萬元(事實上大部分都是戶頭餘額為0)，而戶頭餘額超過3萬元的用戶則約呈現常態分佈
2. 戶頭餘額超過19萬元的客戶，流失率會提高

信貸評分



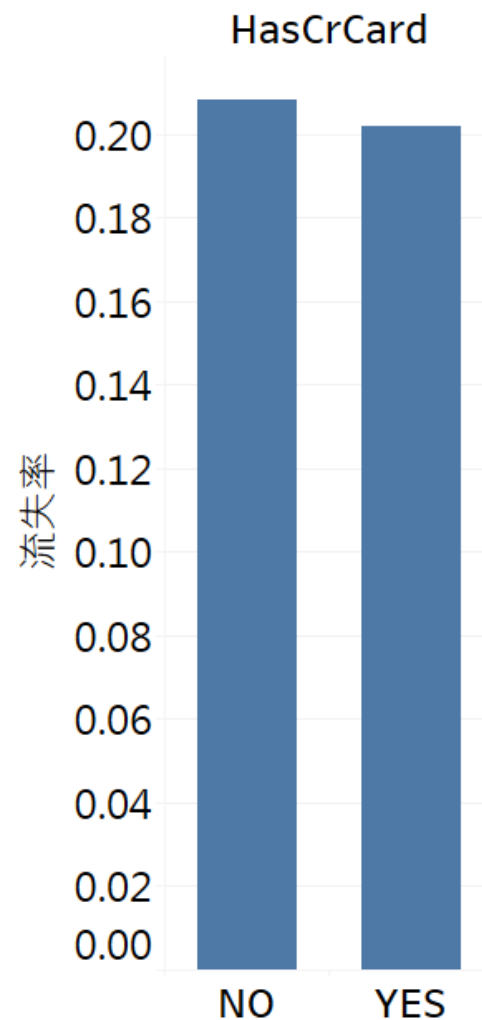
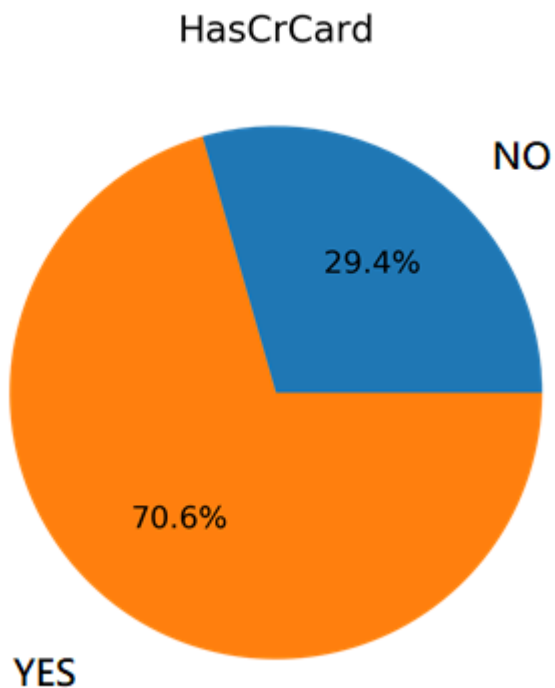
1. 信貸評分在400以下的客戶，流失的機率相當高，以上則沒有顯著的差別
2. 然而400分以下的客戶只占極小的比例

商品數



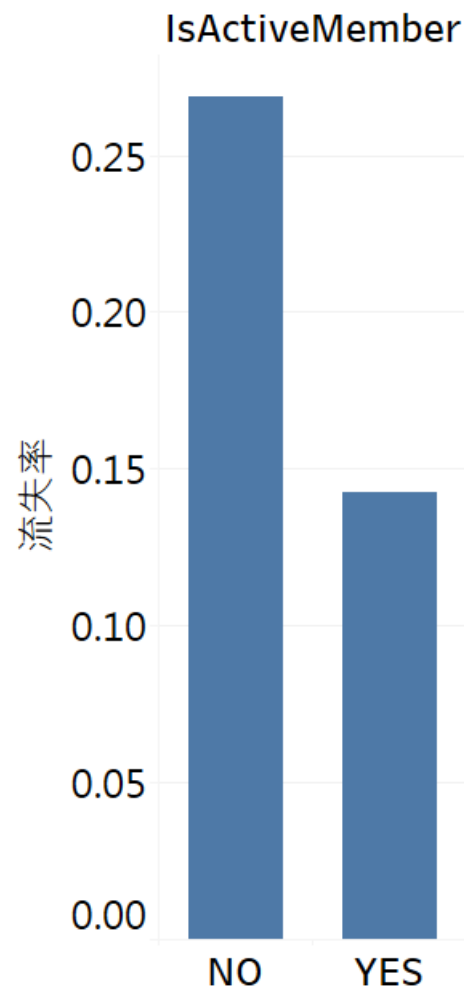
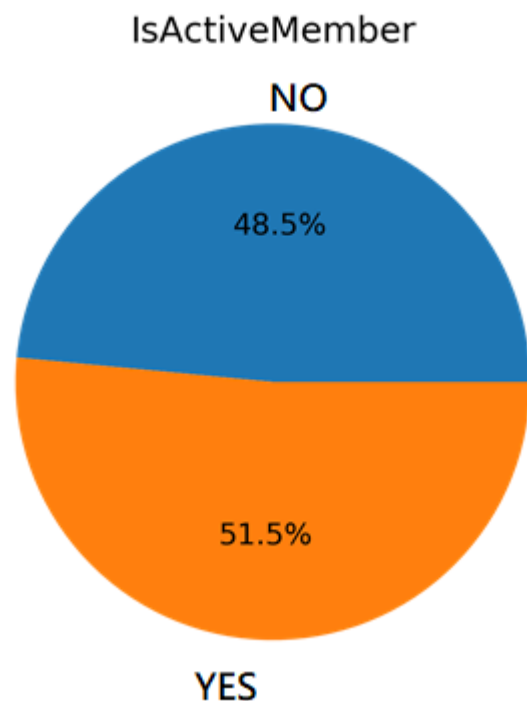
1. 大部分的客戶在該銀行都只擁有1~2樣商品
2. 擁有3~4樣商品的客戶，流失率較高

是否擁有 信用卡



是否擁有信用卡對於客戶是否流失似乎沒有影響

是否活躍



1. 理所當然地，非活躍客戶的流失率較高
2. 值得注意的是，非活躍客戶的占比將近一半，銀行應採取積極措施，將這群人轉為活躍客戶，有機會大幅降低流失率

建模流程

資料前處理

- 遺漏值補值 (可嘗試用平均數、中位數、眾數)
- 對類別變數做 One hot encoding
- 對數值變數做標準化或取對數


數據分割

- 將資料隨機拆分成訓練集與測試集

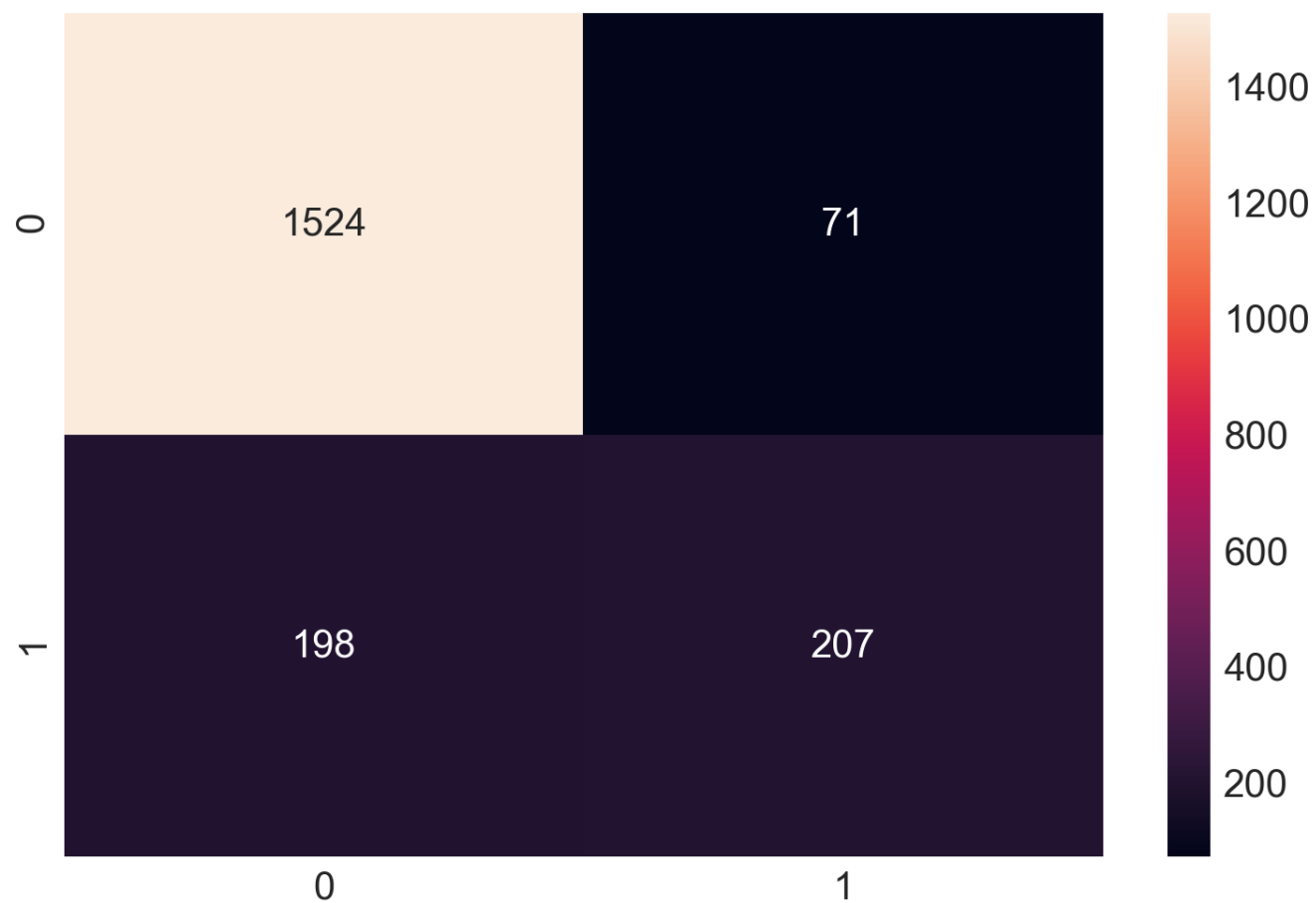
建立模型

- 運用訓練集調整模型中的參數，並用測試集驗證最後的結果

模型預測

Model	Accuracy
Logistic Regression	81.20%
KNN (k=7)	82.60%
SVM	84.70%
Naive Bayes	80.35%
Decision Tree	81.15%
Random Forest	86.55% 

混淆矩陣 - 隨機森林



結論

- 什麼樣的客戶比較有可能流失？
 - 女性
 - 德國地區
 - 非活躍客戶
 - 介於40~60歲之間
 - 擁有的商品數 > 2
 - 戶頭餘額 > 190000