HW6

(1) 小組成員

物理系 陳昱丞 C24080024
物理系 李國成 C24085016
統計系 劉臣洋 H24045010

(2) 競賽敘述與目標

將 8000 筆銀行資料以 13 個特徵做為參考並將其訓練，再用另外 2000 筆資料做為測試，目標是盡可能的預測銀行客戶最終是否會流失。

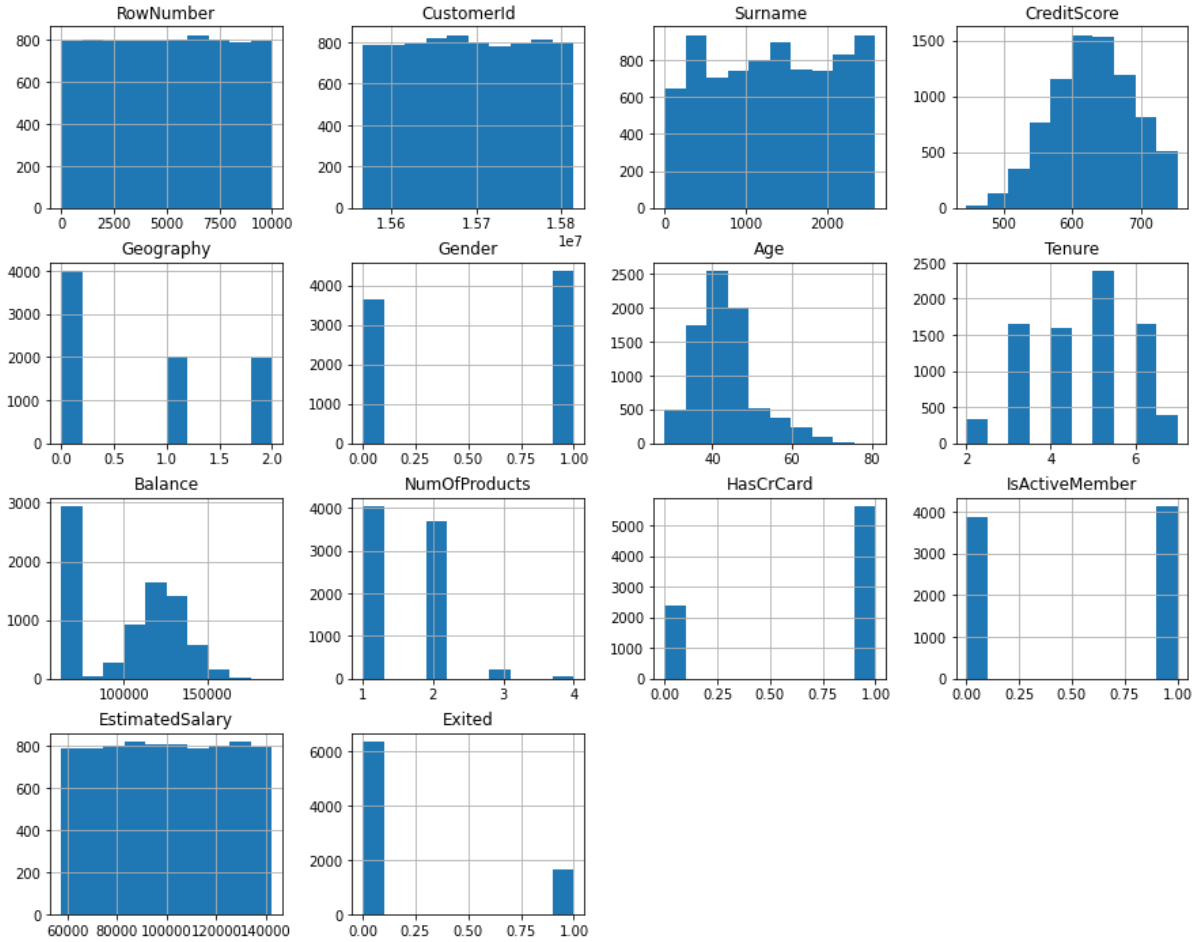
(3) 資料前處理

我們主要做的前處理有以下幾點：

1. check NA：檢查資料中是否有出現 NA 的數據。
2. label encoding：由於數據中有類別資料及數值資料，因此一開始先以 LabelEncoder 把標籤資料轉換為數值，方便劃出長條圖觀察各特徵的分布情形。並且對連續型特徵和類別型特徵
3. onehot：由於為無序的離散值，用 1, 2, 3 的順序在空間中會造成離原點距離較遠，因此我們用 onehot encoding 的方法，並作後續分析。而 onehot encoding 無法直接對字串進行編碼，因此也必須先透過上衣步驟 Label encoding 將字串以數字取代後再進行 onehot encoding 處理。
4. scaling：以利於電腦運算。
5. 先做 train dataset 和 test dataset split，避免和 resampling 次序弄錯導致 test dataset 的污染。
6. resampling：由於數據中我們發現 Exited=1 太少了，以至於造成 imbalance，因此我們隨機對 Exited=1 的數據用升採樣(upsampling)，設法將數據更加平衡。

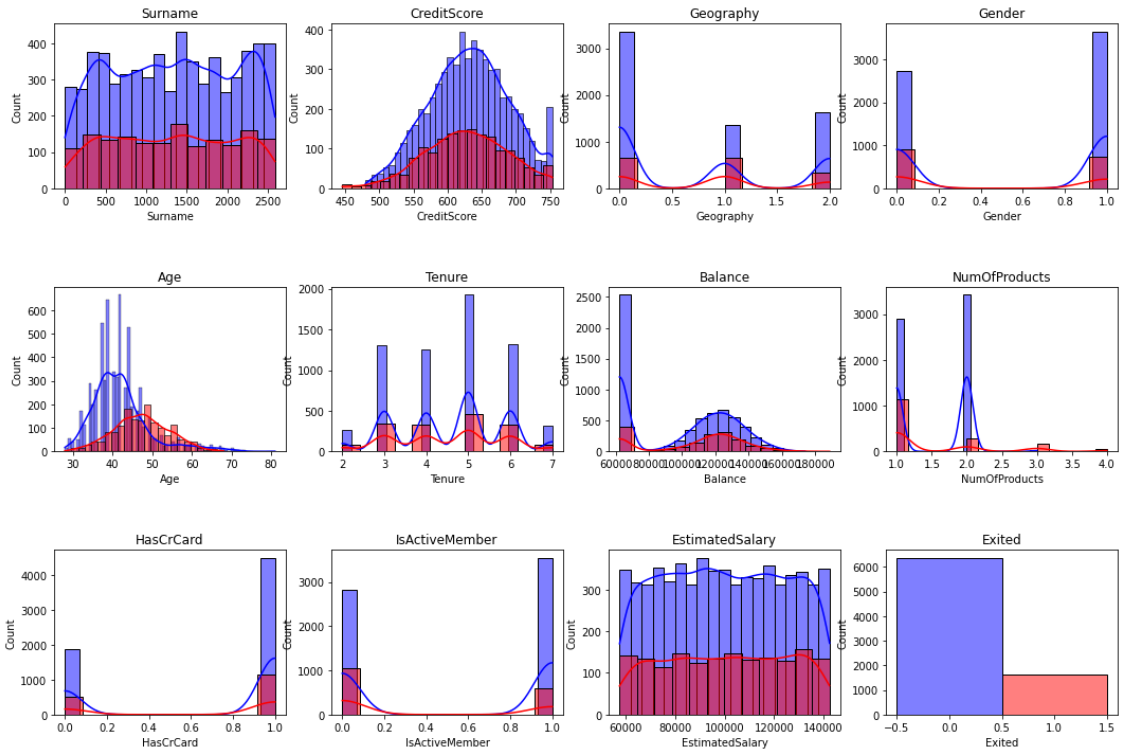
(4) 特徵處理與分析

特徵分配圖：



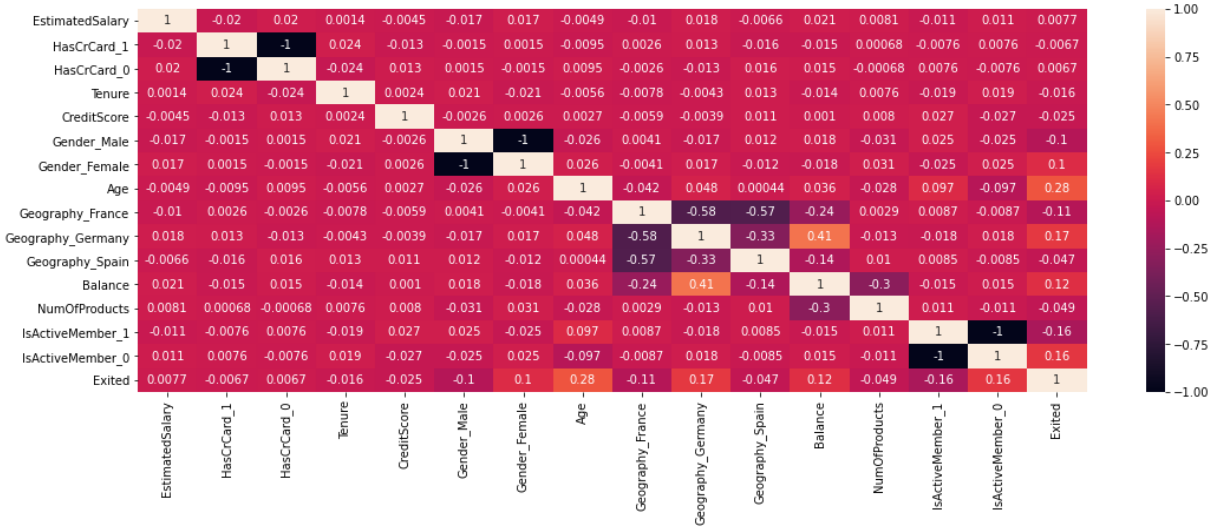
看特徵分配圖會發現 預測目標 Exited 是 imbalance data，需要對它進行 resampling 的動作。 此外 CreditScore, age, Tenure, Balance, NumOfProduct 也有 imbalance data.

特徵分配圖（藍色是 Exited=0，紅色是 Exited=1）：



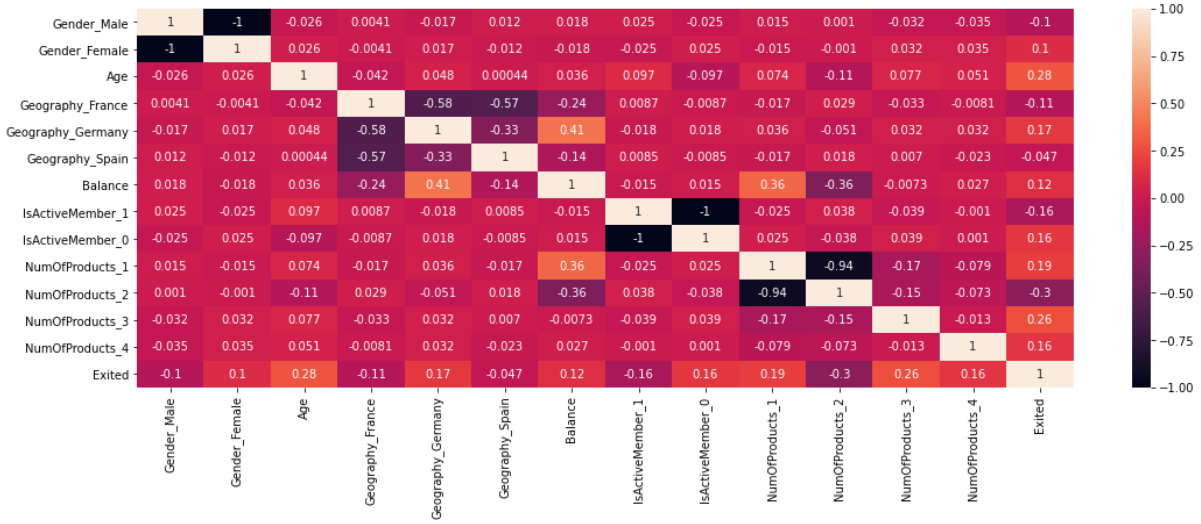
可以發現 Gender, Age, Geography, Balance, NumOfProducts, IsActiveMember 的分配都都不同。

Correlation heatmap：



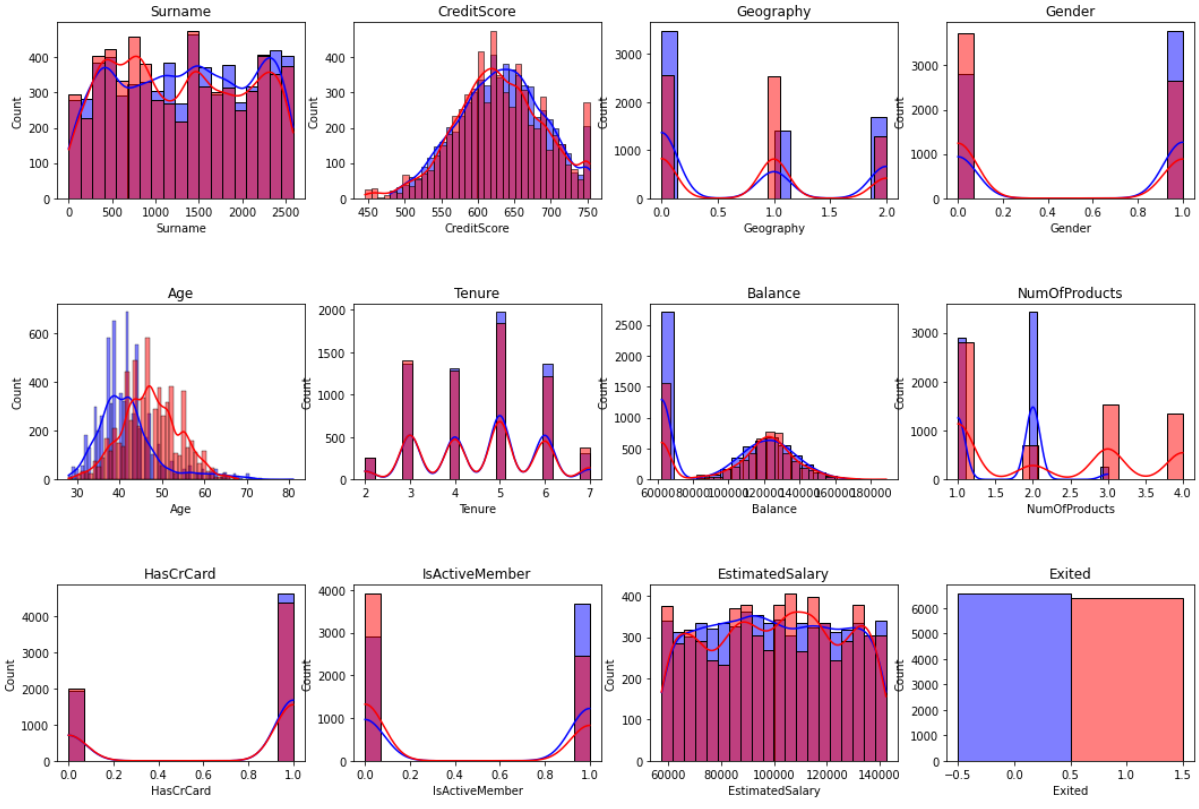
檢查 Correlation heatmap，會發現分配圖中明顯有相關性的 NumOfProducts ，在 Correlation 下幾乎被忽略。推測對 NumOfProduct 做類別型資料的 OneHot Encoding 會比較好。並且 EstimatedSalary, HasCrCard, Tenure, CreditScore 關聯性不大，可以考慮移除這些特徵。

重做一次 Correlation heatmap：



會發現 NumOfProducts 的關聯性增加很多。但這個部分我們並沒有在比賽當中用，因為隊友國成初期對資料前處理和分析完全沒有頭緒，而之後稍微掌握了一些方法也處於期末噩夢，比賽截止後再重新分析這筆資料時才發現這個細節。

Resampling 後的特徵分配：



用 Cross Validation 初步預測的結果：

RandomForrest

Precision	Accuracy	Recall	F1	Score
0.75	0.86688	0.49211	0.59429	0.72278
0.723	0.85	0.4597	0.56204	0.69672
0.71548	0.8525	0.50442	0.5917	0.70707
0.71782	0.85562	0.45455	0.55662	0.69468
0.7277	0.85938	0.48137	0.57944	0.7079
0.7268	0.85688	0.47843	0.57682	0.70583 (AVERAGE)

可以看到調整 NumOfProducts 會導致 Recall 降低但 Precision 會增加，需要再對資料比例做調整，讓一些 Exited 有的特徵更明顯。但因為時間關係就沒做了。

Surname 探討：

另外對也對一部分 Surname 畫了 Histogram，發現有些 Surname 和 Exited 的關聯性很大。推測是因為再西方 Surname 就是代表所屬的家族，應該也會是有效幫助預測的特徵值，不過因為它是類別型資料且樹木龐大。也礙於時間關係沒有進一步擴展這塊。

(5) 預測訓練模型

自動化不調參數大致跑過一些模型：
Random Forest, GradientBoost, Logistic Regression, KNeighbors, DecisionTree,SVM 等。

(6) 預測結果分析

(Accuracy、Precision 與 F Score)的預測結果之圖形或表格呈現；若有記錄上傳到競賽網站的預測結果，可整理呈現在報告中；若有自行切自己的訓練與測試資料，也可以回報自行測試的預測結果。 此外，也可以分析討論哪些模型與哪些特徵會產生較高與較低的效果，最終歸納出對於特徵、分類模型選擇的建議。得明確指出最終上傳的結果是哪一組分類模型與其參數。

大致跑過一些模型（沒什麼調整模型參數）

Model	Precision	Accuracy	Recall	F1	Score
RandomForrest	0.73786	0.85312	0.44667	0.55644	0.69987 (AVERAGE)
GradientBoost	0.75	0.855	0.44545	0.55894	0.70507 (AVERAGE)
GradientBoost2	0.76263	0.85875	0.45758	0.57197	0.7152 (AVERAGE)
LogisticReg	0.68421	0.83313	0.35455	0.46707	0.64203 (AVERAGE)
KNeighbors	0.60204	0.81875	0.35758	0.44867	0.6057 (AVERAGE)
DesisionTree	0.48936	0.78938	0.48788	0.48862	0.57907 (AVERAGE)
SVM	0.7963	0.81375	0.1303	0.22396	0.5726 (AVERAGE)

以下是不對 NumOfProduct resampling，只對 Exited resampling，並且調試模型的結果：RandomForestClassifier(n_estimators = 300, max_depth = 5)

	precision	recall	f1-score	support
0	0.91	0.85	0.88	1095
1	0.68	0.80	0.74	444
accuracy			0.83	1539
macro avg	0.80	0.82	0.81	1539
weighted avg	0.85	0.83	0.84	1539

後來我們丟掉部分不重要的值，並再執行一次，其目的有二；

1. 想提升運算效率
2. 想看那些丟掉的特徵有多不重要

	precision	recall	f1-score	support
0	0.91	0.84	0.88	1096
1	0.67	0.80	0.73	443
accuracy			0.83	1539
macro avg	0.79	0.82	0.80	1539
weighted avg	0.84	0.83	0.83	1539

不過結果較稍微不準確，但可忽略不計。

(7) 感想與心得

陳昱丞：

這次的資料比賽，是我第一次關於資料科學的比賽，沒想到比我想像中困難許多，從一開始選題目時不只要選一個自己覺得有趣的，還需要評估其可行性，像我們本來有想要嘗試爬國外留學資料，或是判斷假新聞等等，不過由於機乎都是類別型資料，因此難度高出許多，最後選擇一個我們覺得有趣也算是可行的預測成績題目，不過過程中的挑戰更多，像一開始前處理資料不平衡或是到後來模型試了好幾次卻沒有進步，有時真的會做到灰心喪志。資料科學真的比我想像中的挑戰多出許多，就像老師在最後一堂課說的，表面上看起來很美好，自己去嘗試了就需要足夠的恆毅力，這次的過程也讓我了解到自己找資料及自學的重要性！

最後，真的要謝謝老師還有助教，真的感受的到你們滿滿的用心。

李國成：

高中就在理論上有接觸過幾個機械學習算法的。但這次是我第一次系統性的學習資料科學。課程初期只是簡單的分類器算法，並沒有很大的課業壓力。但期中開始有的實作（專題和比賽）讓我覺得很困難。

首先就是身為系外選修這門課而且又是物理系的同學，對資料分析和前處理和完全沒有概念，只會把重心放在模型和理論上面，連資料集都會先假設是“理想的資料”。老師在課上並沒有教太多資料分析和前處理，導致一頭霧水也繞了很多彎路。知道資料不平衡會導致模型預測偏差，但不知道怎麼平衡數據。

後來發現到資料分析和前處理是更重要的，因此後期開始專注投入學習這些內容而暫時放棄學習更難的模型適配。實作讓我明白老師講的 meme 笑點何在，也知道什麼是“弄髒雙手”。期中開始的學習過程雖然辛苦，陷在迴圈也很痛苦，但整體上是還是值的和享受的。

最後就是雖然老師講課有點愛睡，但有感受到老師的用心，課程規劃也很棒。並且也有幸聽到老師的對資料科學的洞見。

劉臣洋：

這次的競賽真的讓我獲益良多。首先，在處理資料的過程中，它讓我回想起自己以往曾經學過的課程知識，像機械學習，多變量分析，巨量資料分析等。其次，因爲這次與我同組的組員是非本科系的學生，因此要學習如何與他們溝通及共事。這可以說是我這次競賽最大的收穫之一。

然而，儘管有修過或看過資料科學相關書籍，自己在修課過程中還是覺得很辛苦。我發現這是因爲我的程式語言能力不足所產生的問題。雖然瞭解模型的原理(像 RandomForest, Gradient Boosting 等)，但是當自己手刻模型時，才理解到要把原理轉換成程式碼的難度超乎自己的想象。這也提醒我，如果未來要繼續往這行業邁進，自己需要更多加強程式語言相關的能力。

最後，就這次的競賽而言，我個人的感想是做的有點沮喪，因爲無論我用什麼樣的方法，都無法提升預測結果的準確率。可是，自己仍然有學到一些東西，像 model stacking，error analysis 等。然而，因爲忙於期末專題的緣故，沒有把這些方法的預測結果上傳到網站，我對此感到非常遺憾。

總而言之，無論是與組員共事，或是課堂知識等，這門課都讓我獲益良多。就像老師剛開始分享的，資料科學是一門跨領域的學科。期待自己未來可以在自己不足之處有更多精進自己的機會，讓我不斷學習與成長，持續往資料科學家這標杆前進。