資料科學導論
HW6 競賽報告書

Team 13 PUIPUI

組員:

統計系 111 H24074019 黃科銓
統計系 111 H24076142 邱瑞麒
經濟系 111 C34061200 莊詠鈞

目錄

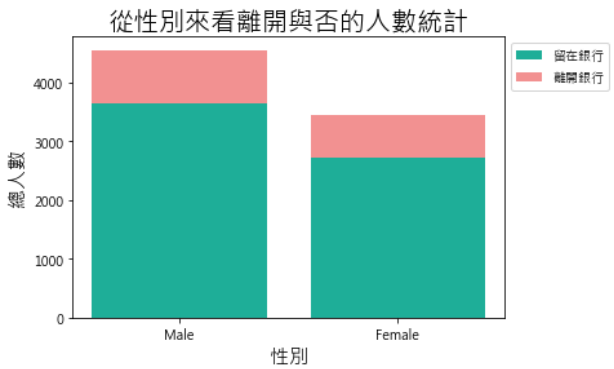
(1)競賽敘述與目標

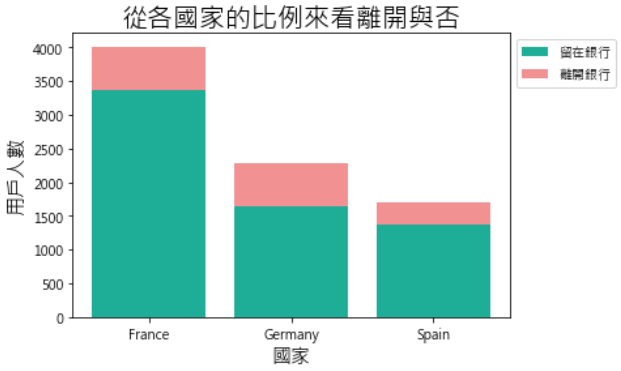
這次競賽所使用的資料是某個銀行的顧客資料，訓練資料 8000 筆，測試資料 2000 筆，一共 10000 筆，我們要利用這些資料去預測顧客會離開還是留下。競賽的評分著重於 F-score，因此我們的目標不僅是準確預測 exited 的 0、1，預測的結果最好盡可能包含到所有答案為 1 的顧客。

(2)資料分析

首先，我們觀察訓練資料及測試資料的變數比例，發現不管是年齡、性別、國籍…等等，都是差不多的，因此我們暫時不會對資料進行過濾，盡量以相同的比例去訓練、預測，等到後續有更多發現再決定是否進行過濾。

接著，我們決定觀察各個變數中，離開與留下的人數的比例，作為挑選訓練變數的參考。離開的人數占全部的比例是 0.1632。

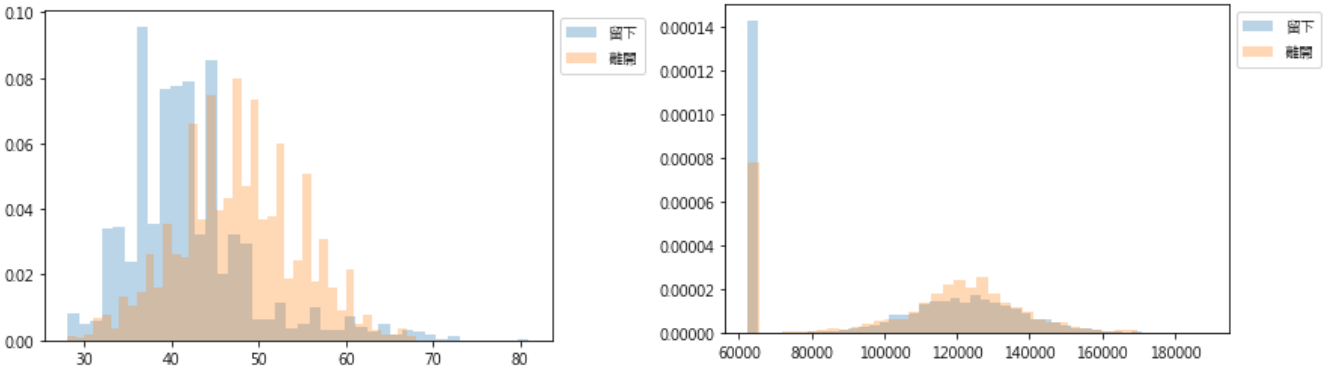




男生	女生
0.165637	0.250207

法國	德國	西班牙
0.161008	0.321606	0.171053

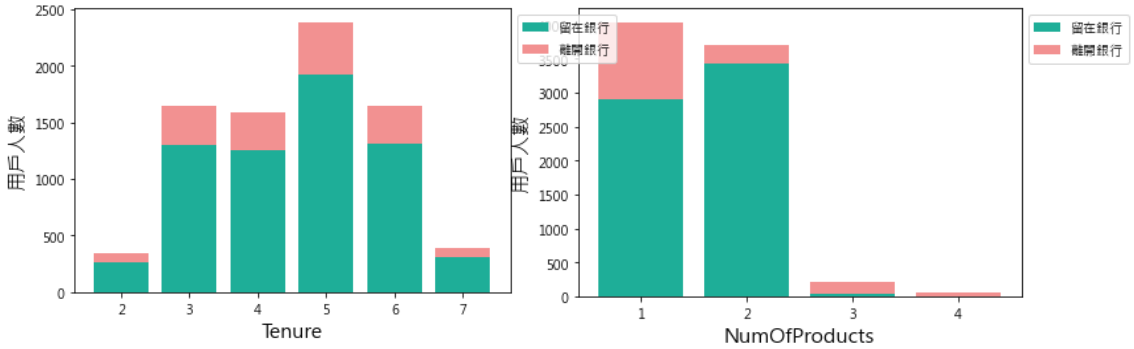
從上面的表格看來，女生離開的比例比男生高一些，德國人離開的比例比其他兩國高一些，挑選訓練資料時，性別及國家或許是不錯的選擇。



再來是年齡的分布(左圖)以及 balance 的分布(右圖)，單純看年齡與是否離開的相關性是所有變數中最高的，從圖中可以看出離開與留下的年齡分布明顯不同。Balance 在最小值的地方人數特別多，且留下的比例較高，若不看最小值，

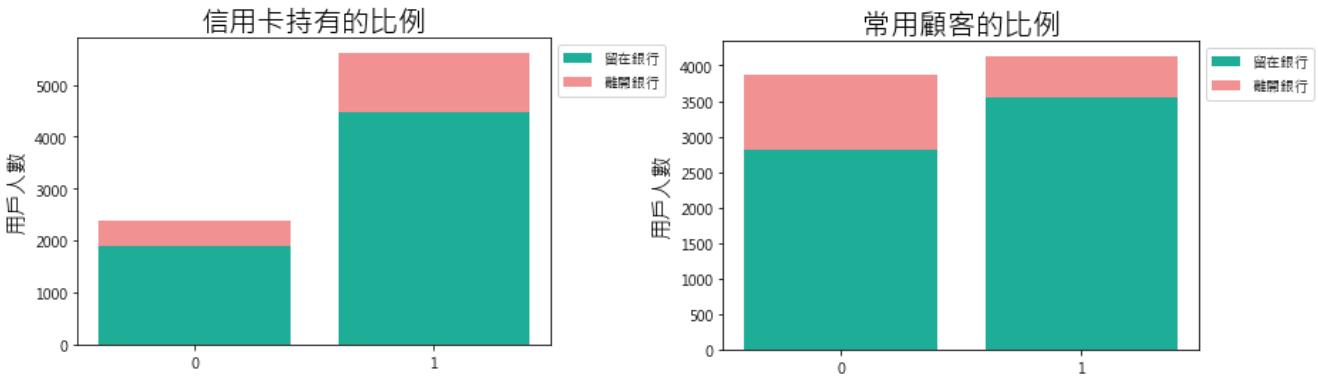
balance 大致呈現常態分佈，且離開的比例整體看起來較高。

因此挑選訓練的變數時，我們可能會優先考慮年齡和 balance。



左圖是 tenure，右圖是 NumOfProducts，雖然 tenure 的比例看起來都差不多，可能不會優先考慮，但是主要還是取決於加進模型訓練是否能更準確。

NumOfProducts 在 3、4 的部分雖然特別少，但是幾乎都是離開，而 1 離開的比例



也比 2 離開的比例高，特徵明顯，會優先考慮加進模型訓練。

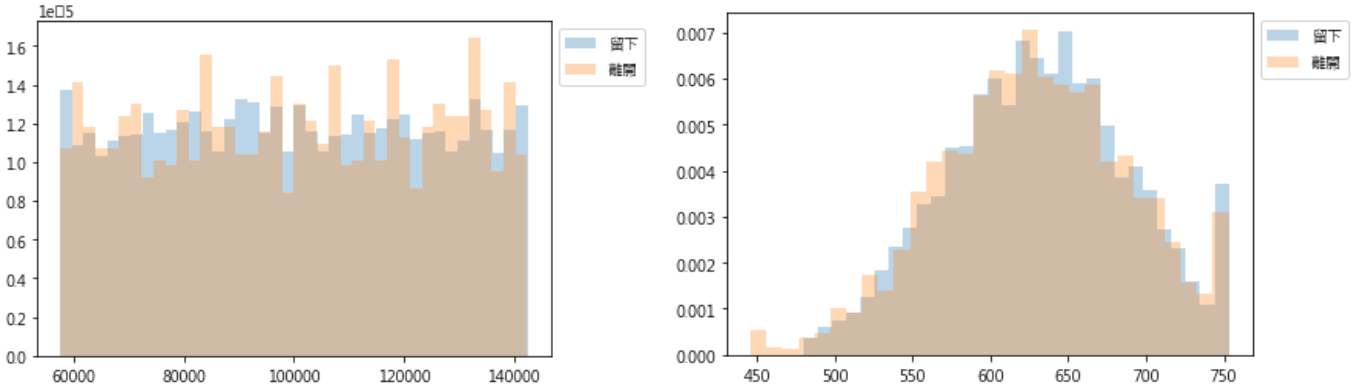
0	1
0.208	0.202

0	1
0.270	0.142

是否持有信用卡對於是否離開銀行的影響看似不大，暫時不考慮加進模型。

常客離開的比率比非常客離開的比率低，符合常理，我們會優先考慮將其加進模型。

左圖是估計薪資，右圖是信用分數，在估計薪資的部分，從圖上看不出甚麼特徵。



再看信用分數的部分雖然留下與離開的分布看起來差不多，但是在信用分數最低和最高的地方都能觀察到一些特徵，在信用分數最低的地方幾乎沒人留下，在信用分

數最高的地方，留下的人數比離開的還多，因此我們會優先考慮將它加進模型訓練。

最後我們檢查了各變數間的相關性，相關度都非常低，其中最高的是NumOfProducts 和 Balance，相關性有大約-0.3，不確定對訓練模型的影響大不大，不過可以確定大部分的變數彼此間都是獨立的。

(3)資料處理

1. id，CustomerID 取前四個數字
2. is_4，將財務商品整合成兩類，大於 2 個=1、少於等於 2 個=0
3. 性別數值化(男生 0、女生 1)
4. 增加變數 greater than 70000，Balance 大於 70000=1，否則=0
5. NAME_ID，用 preprocessing.LabelEncoder()把顧客名字數值化
6. 增加變數 money，對"Balance","EstimatedSalary"進行 pca 轉換，第一主成分解釋約 0.62，第二主成分解釋約 0.38
7. 增加變數 is_less_35000，money 小於-35000=1，否則=0
8. 將國家轉成 dummy variables

以上是我們對資料進行過的處理，除了將類別變數轉換成 0、1 以外，我們也對一些變數進行 pca 轉換，希望能夠提取出它們的特徵。

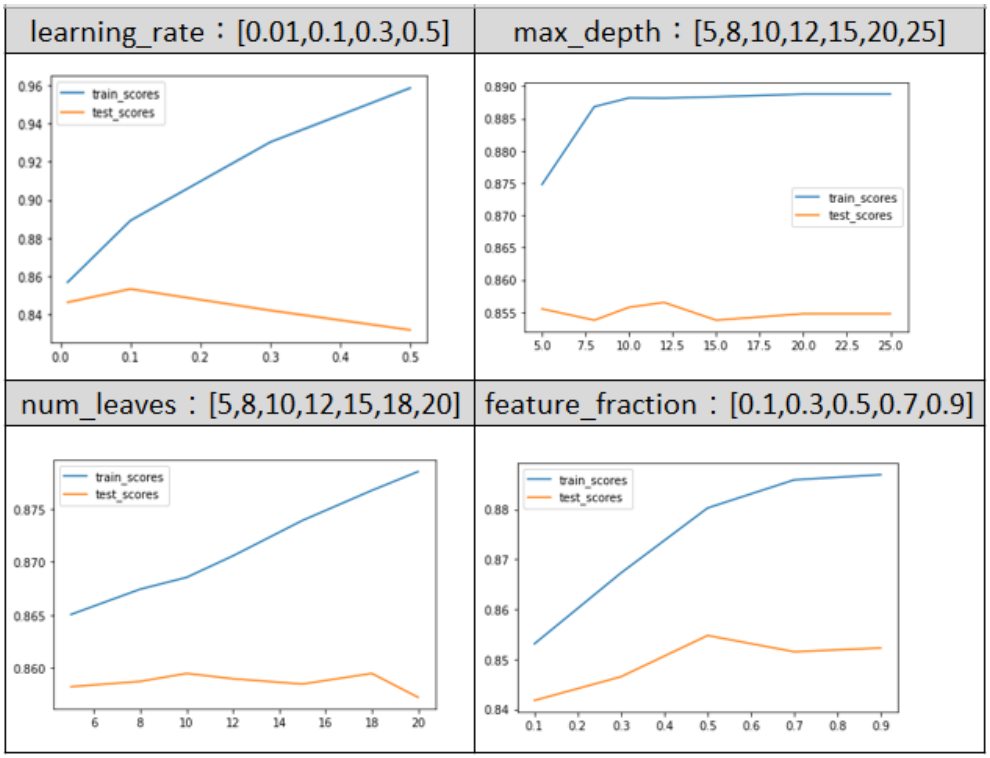
當然，不同的訓練模型適用於不同的變數組合，雖然對上面的資料進行處理，但不代表一定會套用到它們，像是將名字數值化套用在羅吉斯回歸的效果不錯，但套用在其他模型就沒甚麼效果。

(4&5) 預測訓練模型&結果分析

在此次預測比賽中，其目標是要分類銀行使用者是否離開的二元分類問題，我們總共用7種監督式分類模型，其中4種為我們在課堂中所學之機器學習模型，分別為「KNN」、「LogisticRegression」、「Decision Tree」、「Random Forest」，以及3種進階boosting機器學習模型，分別為「LightGBM」、「XGBoost」、「Gradient BoostingClassifier」，來深入分析每種分類模型的好壞。

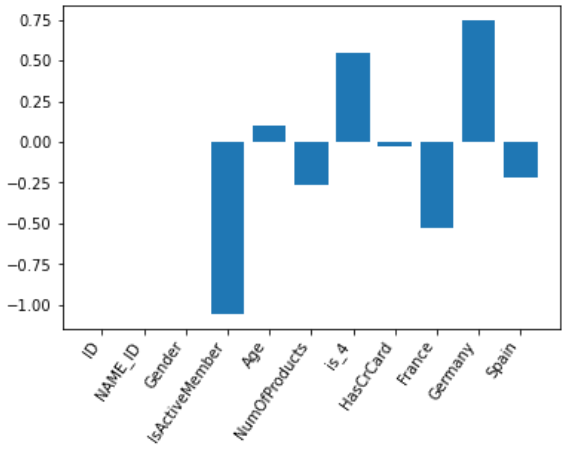
首先，我們特別針對「LightGBM」做超參數調整以及「Random Forest」、「LogisticRegression」來做特徵權重了解。

- LightGBM超參數調整
  - 固定參數：objective、metric、is_unbalance、boosting、num_boost_round、early_stopping_rounds。
  - 調整參數：
    learning_rate、max_depth 、num_leaves、feature_fraction



  - 調參數結果：
    原本是希望透過調超參數，可以得到更好的結果，然而實際利用test預測下來，並沒有得到更加的預測結果，反而降低效果，因此我們後來便對超參數不做任何調整。

- LogisticRegression之特徵解讀



  - 結論討論

    在羅吉斯回歸下，我們特別在資料集之特徵選擇上，與其他模型不同，經過不斷嘗試下，發現以「活躍使用者」、「德國人」、「是否有超過3個財務產品」、「法國人」之特徵係數權重影響較大，也提供我們洞悉了解影響離開銀行的潛在原因。

再者，為了在同基準下比較，我們透過交叉驗證的方式，以StratifiedKFold的分式，並切成10等分，分別討論各模型的「accuracy」、「recall」、「precision」及「roc_auc」之比較，由表一可知，以下結論：

- 準確率（Accuracy）

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

  - 代表意義：模型預測正確數量所佔整體的比例。
  - 白話文：我們預測出實際離開與留下銀行的使用者占全使用者之比例。
  - 最好的三個模型：GBC > LGBM > forest

- 召回率（Recall）

$$Recall = \frac{TP}{TP + FN}$$

  - 代表意義：在原本Positive的資料中被預測出多少。
  - 白話文：實際離開銀行的使用者中，我們預測多少人之比例。
  - 平均來看，最好模型：forest > GBC > LGBM

- 精確率（Precision）

$$Precision = \frac{TP}{TP + FP}$$

- 代表意義：被預測為Positive的資料中，有多少是真的Positive。
- 白話文：被我們預測出離開銀行的使用者中，占實際離開銀行之比例。
- 平均來看，最好模型：forest > LGBM > xgb

- ROC/AUC 曲線
  - 代表意義：ROC是以「Sensitivity」與「Specificity」間的變化圖，其中以曲線下的面積視為AUC，當AUC越高越好。
  - 平均來看，最好模型： GBC > forest > LGBM

	accuracy	recall	precision	roc_auc
knn	0.763000	0.087034	0.260868	0.530228
DT	0.799500	0.508570	0.498451	0.693965
forest	0.861500	0.441796	0.785673	0.854920
Logistic	0.786875	0.056371	0.360175	0.695562
LightGBM	0.863000	0.488987	0.753795	0.854145
gbc	0.863375	0.463258	0.777740	0.864155
xgb	0.854000	0.483466	0.709191	0.847016

- 綜觀結論

  整體來說，GBC、LGBM、forest和xgb等模型都得到較好的預測結果，由此可知，Ensemble learning(集成學習）可以增加預測效果，而在test資料上預測，在不斷的隨機嘗試下，以「LGBM」模型得到最好的預測結果。

最後，我們透過「模型投票」的方式，將以上所提的四種最好模型進行投票，分別用3組LGBM+2組forest+1組GBC+1組xgb之組合，每一組模型皆有1票，0表示使用者留下銀行，1代表使用者離開銀行，以「total>=6」為投票基準，來做最後投票加權的預測，而結果在public set中，得到最佳預測：acc為0.8875、precision為0.8409、fScore為0.6218。

(7) 感想與心得

H24074019黃科銓

對於此次比賽過程中，我最印象深刻的是，我特別用心花許多心思在對資料集的預先分析，進而了解哪些特徵與最後的結果是有很大的相關，還有某些特徵有特別的分布，像是給之後套各種模型奠定一些基礎。也因為此次比賽，不管是對於團隊討論及分工，抑或是對於模型的最佳選擇及模型評估方式，有了更深一層的了解，未來有機會的話，會透過參加kaggle的方式或是找有興趣的開放資料，來磨練自己的專業技能，以及處理資料的看法，進而增進我對數據的實戰經驗。

H24076142邱瑞麒

這個競賽非常刺激，一開始就有一些組別拿到很高的分數，給了後面的組別不小的壓力。因為每天都有30次機會，所以剛開始有試過一下用for迴圈去找能讓預測分數比較高的seed，不過這真的是沒效果沒效率的辦法，還是著重於不同的模型，和找出適合該模型的參數比較有用。對於統計大三的學生來說，這個競賽是一個很好的機會，讓自己活用過去的課程學過的方法、工具，雖然還不太熟練，但有了這次的經驗，下一次要用機器學習對資料進行分析、預測時，能夠更輕鬆的踏出第一步，並接著一步步穩妥地走下去。

C34061200 莊詠鈞

這次的比賽蠻有挑戰性的，每天的排名都會重新洗牌，當發現自己被別人超越時，心中難免有些失落，但當再次超越時，心中又會百般喜悅，或許這就是程式的魔力，永遠沒有最正確的答案或模型，只有透過不斷嘗試以後才可以更加精進自己的結果。希望可以在接下來成大的最後一年，多學幾種程式語言搭配使用，寫出一個開放的資料庫給學弟妹使用，也會繼續參加kaggle上的比賽增進自己資料處理的能力。