統計 111 楊正毅 H24071265
統計 111 蕭云雅 H24071299

一、競賽敘述與目標

我們會拿到數百筆有關銀行客戶的數據行為數據(如：在銀行存入的資產、是否有信用卡…等)，目標是預測客戶最終是否會退出銀行，即將來不再與銀行進行交易。

二、資料前處理

在確認資料中沒有缺失值後，將我們認為不需要的特徵直接刪除，所以會先將三行資料刪除，那三行分別為 RowNumber、CustemerId、Surname。之後我們把只含兩類的變數轉換為 0 和 1、超過三類的類別型變數：如 Geography 做資理處理 One hot encoding。在觀察資料時我們發現 balance 的資料有些問題，有大約 2500 左右筆資料的 balance 都是 62397.41，因為如果要將那些 balance 有問題的資料都刪掉會刪掉過多的資料，所以我們決定不使用 balance 的資料進行分析。

做完上述的處理後，最終的我們使用的 columns name 為以下這些：

CreditScore	Gender	Age	Tenure
NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
Exited	Geography_France	Geography_Germany	Geography_Spain

之後也將資料標準化（減去平均後除以標準差）以去除尺度差異。’

三、特徵處理與分析

使用相關係數矩陣觀察，發現變數之間的相關都不高，且我們認為變數的維度不算高，所以沒有使用非監督式學習降維。我們使用隨機森林內建的功能(feature_importance)來選擇重要的特徵，在之後配合模型選擇特徵的過程可以用來參考，下圖為 model 給出的特徵重要程度。

Age	0.233354
EstimatedSalary	0.154793
CreditScore	0.149711
Balance	0.147074
NumOfProducts	0.132223
Tenure	0.062792
IsActiveMember	0.040957
Geography_Germany	0.020134
Gender	0.019702
HasCrCard	0.019527
Geography_France	0.010423
Geography_Spain	0.009309

四、預測訓練模型

我們選了決策樹、隨機森林、SVC 還有 xgboost 和 lightgbm 作為候選的模型，在全部使用預測參數的狀況下使用 CrossValidation（Scoring = （F1+accuracy+precision）/3）觀察哪個模型表現較好，選擇表現最好的 lightGBM，之後同樣使用 CV 去調整參數，最終選用 n_estimators=160、class_weight = {1:1,0:2}、metric= 'binary_logloss' 、colsample_bytree = 0.8、reg_alpha = 0.01、max_depth = 2、min_child_samples = 6 , objective = 'binary'，沒有更動的則是預設值。

模型在 K=5 fold 的 Crossvalidation 上的表現為 0.7599、0.7675、0.7259、0.7071、0.7518，

五、預測結果分析

最後使用在上面提到的 lightGBM 和對應的參數訓練所有 train.csv 內的資料，預測 test.csv 的資料上傳的結果為

Acc：0.87 、Precision：0.7451、F-score：0.5938、final-score：0.7363

六、心得與感想

楊正毅

在這次的競賽應用了很多課程中和課程外學到的東西，我覺得整個過程很有趣，也學到了很多課程中沒學到的東西，只可惜上傳一次之後因為其他課程都有期中考跟報告所以就沒辦法 tune 出一個很滿意的模型 QQ。能夠完整的跑過一個競賽的流程也讓我學到了蠻多的，也很謝謝老師舉辦了這次的競賽！

蕭云雅

其實從高中就一直很排斥寫程式，上大學發現要寫程式，心情很錯愕……，但一想到快要畢業了，就想給自己機會嘗試看看，結果心得就是，真的很難☹，但也從這次競賽中學習到很，我覺得競多想法，謝謝，我的組員總是很耐心向我解釋，跟傾聽我的想法賽的舉辦很有意義，讓我們在這堂課學習到的東西加以應用，謝謝老師！

七、Github

楊正毅：https://github.com/yang890813/IDS_HW

蕭云雅：https://github.com/yunyahsiao/data-science-hw