**資料科學導論競賽報告**

H34071039工資系111李若瑜

H34074087工資系111何彩綺

H34076039工資系111黃振維

**競賽敘述與目標**

本次競賽以銀行客戶資料做為資料集，分析其中欄位並預測銀行是否會流失該客戶，即未來不再於銀行進行交易。使用的資料集分別為：訓練數據集

(train.csv) 和測試數據集 (test.csv)，預測目標為Exited欄位。上傳upload.csv 檔案至競賽網站，最後評分以三個評估指標Accuracy (30%)、

Precision (30%) 和 F-Score (40%) 為標準，數值越高者為佳。

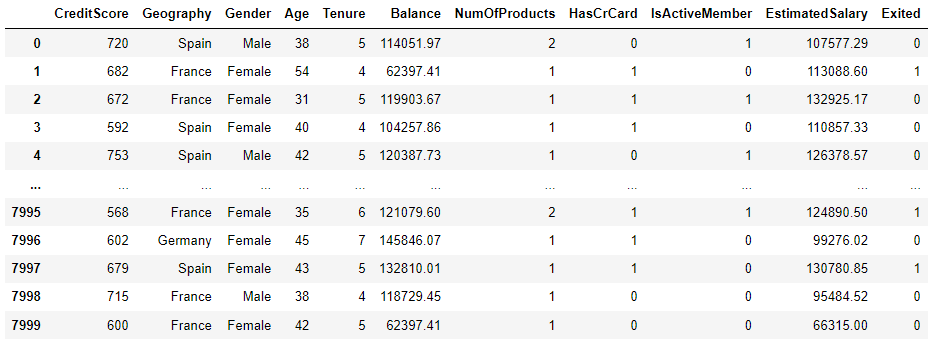
**資料前處理**

我們運用了sklearn套件中的MinMaxScaler以及OneHotEncoder

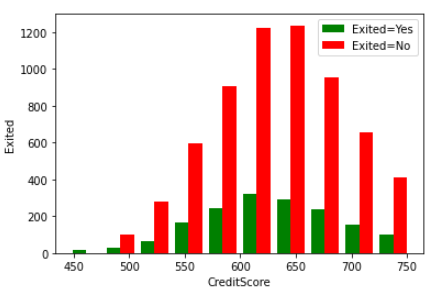
方法。使用MinMaxScaler給定一個明確的最大值與最小值，每個特徵中的最小值變成了0，最大值變成了1，數據會縮放到到[0,1]之間。OneHotEncoder的編碼邏輯為將類別拆成多個column，每個列中的數值由1、0替代，當某一列的資料存在的該行的類別則顯示1，反則顯示0。

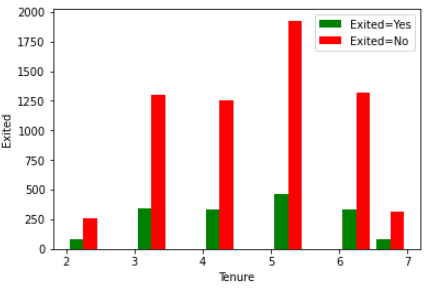
**特徵處理與分析**

(1) 刪除不影響Exited的三個欄位RowNumber、CustomerId、Surname。

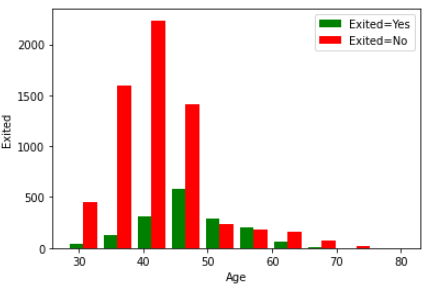


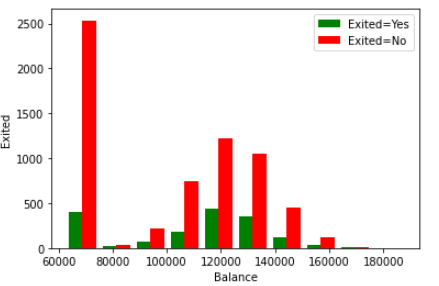
(2) 分析剩餘欄位與Exited的關係分佈，以直方圖呈現。

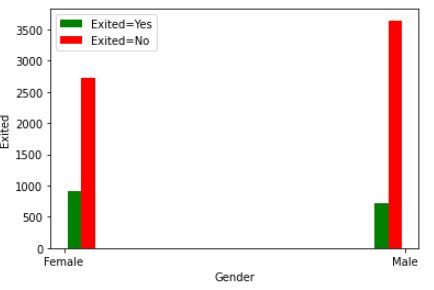


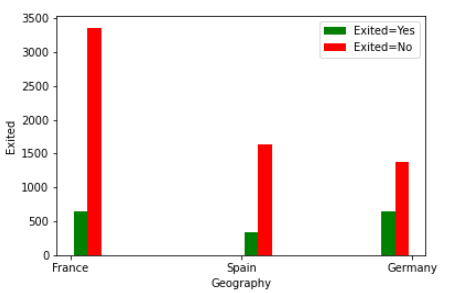


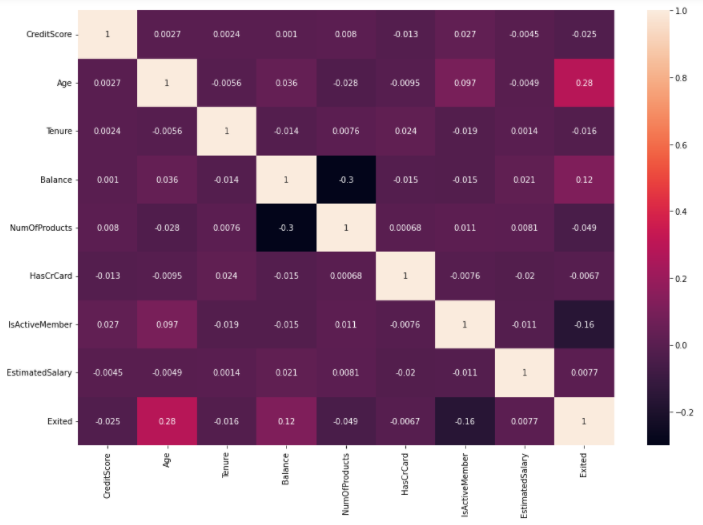












**預測訓練模型**

**預測結果分析**

**感想與心得**

剛開始的組內分工是分成sklearn和ANN，等撰寫完成後再討論預測的成效如何改善。我是負責sklearn的撰寫，但由於我先前並沒有實際使用過預測模型，所以花了一些時間研究前處理、引入package等等，並參考網路上的Youtube教學。但是sklearn比起ANN的預測結果差，所以我們後來改成用ANN的方式。

我們在建模時遇到最大的瓶頸決定刪除哪個欄位，因為刪除掉多餘的欄位，才會讓預測結果更準確，但關鍵是要選擇哪些欄位。我們上網找了很多分析方法，例如共變異數分析，資料視覺化找出關鍵特徵，最後保留適合欄位。

經過一學期的課程，幸運能在學期末接觸到資料科學競賽，讓我多了實務運用的經驗，和組員討論與自學的過程中獲益良多，也希望老師能在課堂中有更多的code講解和案例分析。最後，期許自己能在寒假自主參加Kaggle的競賽，複習本學期教學內容，也謝謝老師的精闢講解！