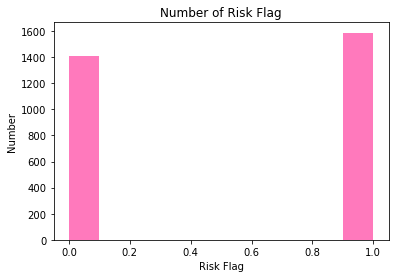
# Lab 1

**Programming, Due 21:00, Thursday, October 7th , 2021**

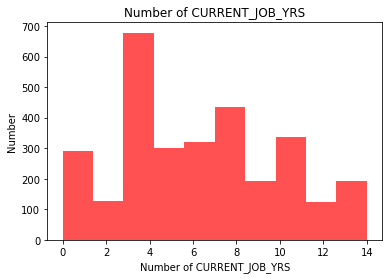
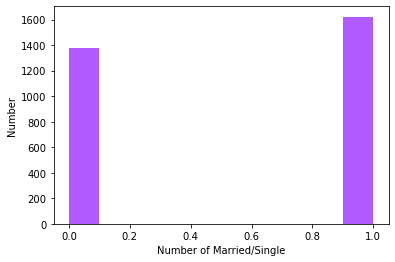
**Student: 廖沁旋 Chin-Hsuan, Liao**

**Student ID: N96104080**

1. **Bank of LAiMM wants to predict who are possible defaulters for the consumer loans product. They have data about historic customer behavior based on what they have observed. Hence when they acquire new customers, they want to predict who is riskier and who is not. You are required to use the training dataset to identify patterns that predict “potential” defaulters. Name your Jupyter notebook SVM and report SVM\_report.docx.**
   1. **(20%) Plot the statics of training dataset. You may need to process training data before feeding them into model.**



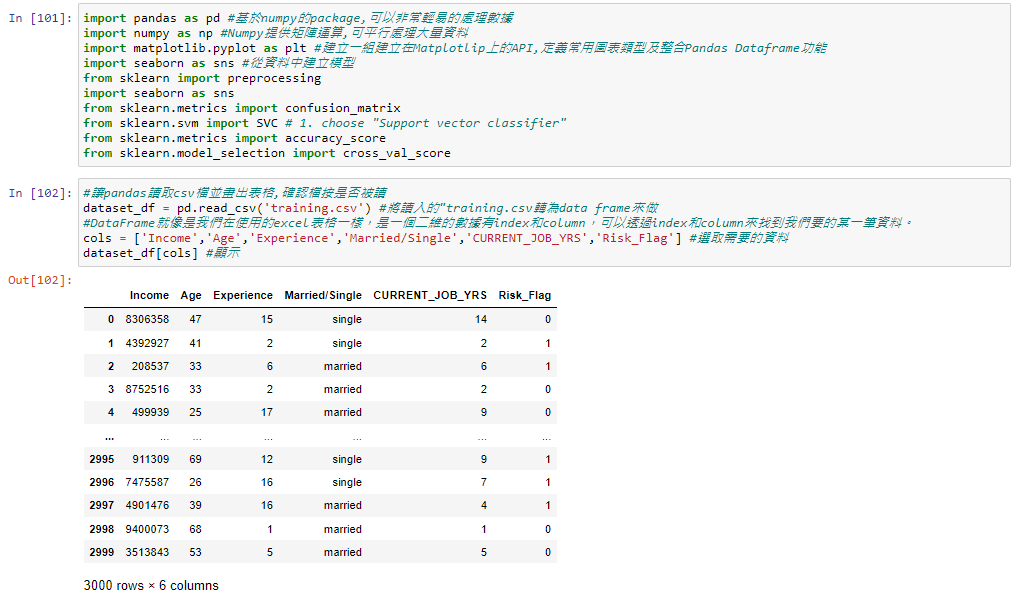




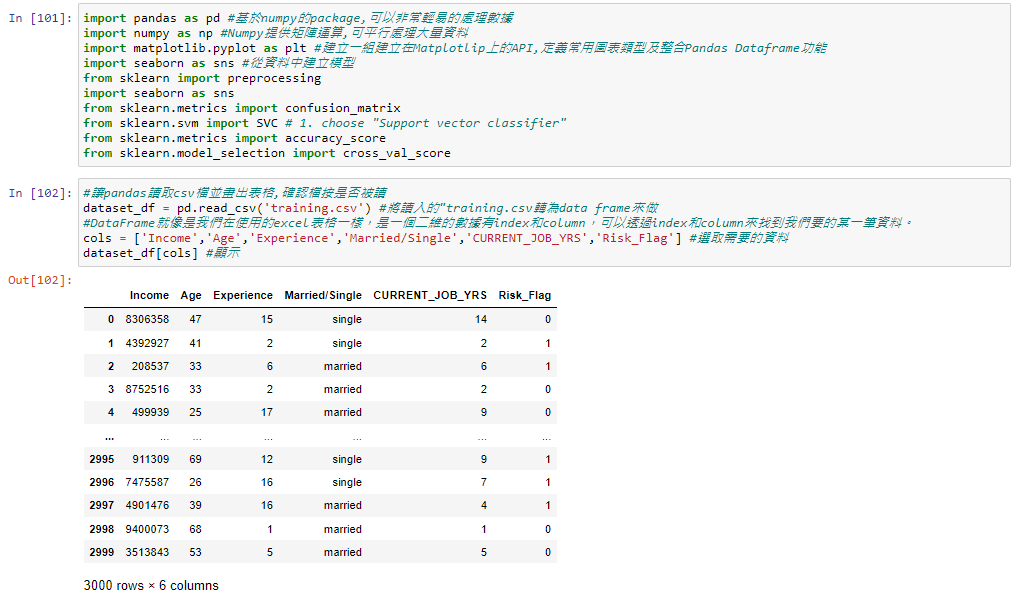
首先利用python內建的matplotlib模組將CSV檔中的”Income”,”Risk Flag”, ”Age”, ”Experience”, ”Married/Single”, ”Current\_job\_yrs” 欄位全部進行視覺化處理，可以協助更輕易的判讀我們所需的資料。在此步驟中，並沒有將id進行視覺化的原因在於此數據僅代表個體，而沒有其他的意義，因此在機器訓練的過程中是無法進行特徵處理得到我們所期待的結果。

* 1. (40%) Use Support Vector Machine to identify possible defaulters, please report training history, confusion matrix and accuracy.

這個題目要求做出”training history”, “confusion matrix” 跟 “accuracy” , 因此我們需要去安裝”numpy”, “matplotlib” 及其下的”seaborn”才有可能完成作圖，在numpy下的sckit-learn可以協助我們去做相關的運算並協助作圖。



讀取我們待會要訓練的csv檔案，並且將其轉成data frame才可以讓python可以讀取的檔案格式，利用col將我們需要輸入的資料選去病輸入，並顯示出來以便我們確認輸入的檔案沒有問題。(在這個地方我們發現id還是被讀取出來，所以下一步要再清楚不必要的資料)。



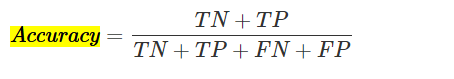
在機器訓練前須要剔除沒有必要的資料，因此我們需要先定義好模組將我們輸入的原始資料先進行處理。首先，我們需要將ID欄下的資料全部刪除，接下來因為空格會使機器訓練發生問題，因此我們需要再將空格全部以零來填空(利用fillna(0));再來，因為機器訓練無法以字串進行訓練，所以我們分別將married跟single改成1和0;第三步我們需要去抓取資料中的特徵值，並且將數值標準化; 第四步我們將資料分為訓練資料及測試資料,再進行處理;第五步，設定好我們的model、並將我們的訓練資料丟入model中進行預測，就可以與我們剩下的非訓練資料進行比對。

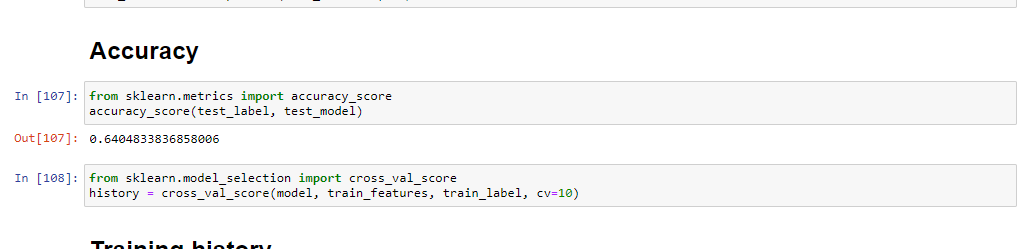
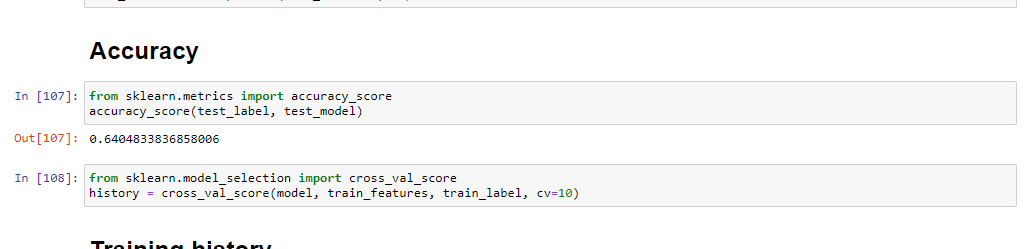


機器訓練完以後，我們就可以利用Scikit learn中的模組來檢測我們的準確度。

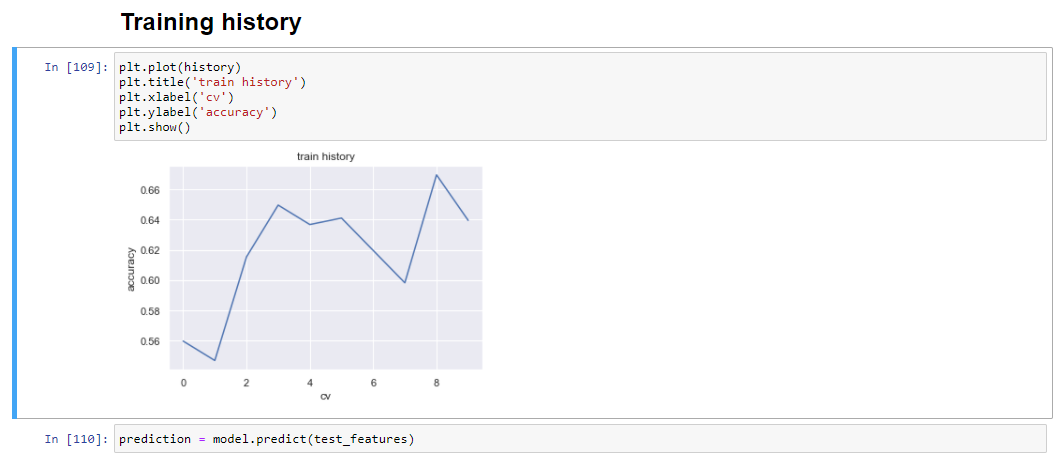
Ref: https://yaojenkuo.io/ml-newbies/07-performance.html

準確率（Accuracy）是類別預測任務最常用的評估指標，分子是正確分類的觀測值個數，即真陰性加真陽性；分母是四個象限的觀測值個數總和，即目標向量的長度，準確率愈高代表分類器的表現愈好、反之則代表分類器的表現愈差。

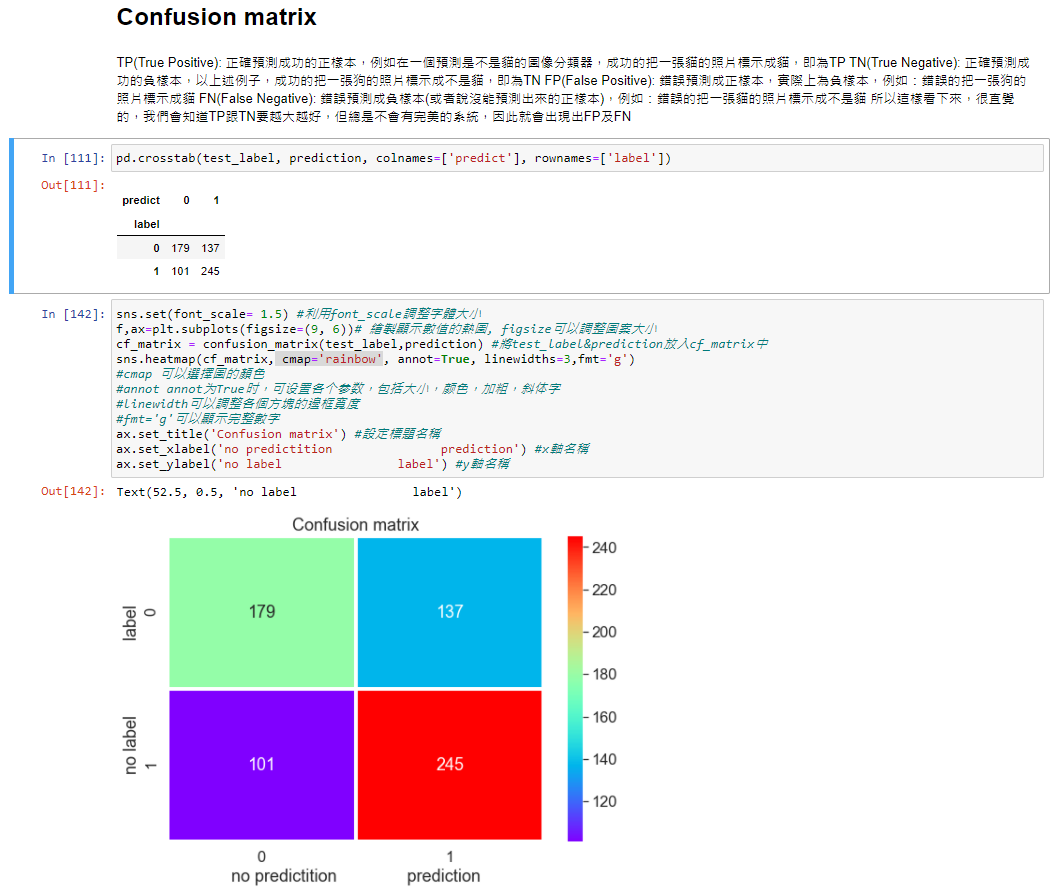




利用Matplotlib畫出Training history,是模型訓練時的歷史數據，可以透過model.fit()獲取，可以了解我們機器訓練準確度各個時間的狀態(網路上都用Keras, 我不確定為什麼我畫得出來)。



我先用crosstab可以看出資料的分布，這樣在畫出渾沌矩陣時就能確定資料是否一致，而我在jupyter中也有將Confusion matrix不同格代表的意義放在上面，以此類推。相關的coding也有在旁邊加上註解。



**結論:**

好險後來改成有整整一天可以寫，我快死掉了(幾乎整晚都沒睡…)。我以為數據視覺化的東西很簡單，但常常重新跑了以後就出現問題，然後就要debug、有時候還會卡很久。這次的功課除了練習了一下python,更在最後進入機器學習的內容，所以pandas,matplotlib,numpy以及特徵工程的部分我需要自己回去多練練手才行。