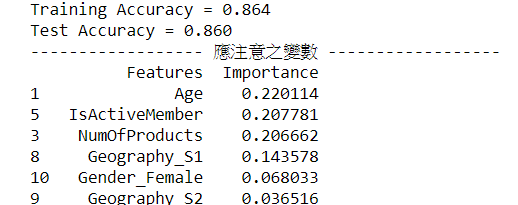
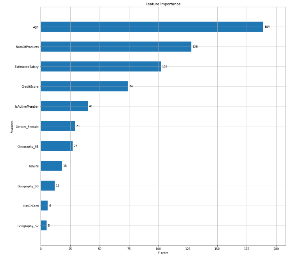
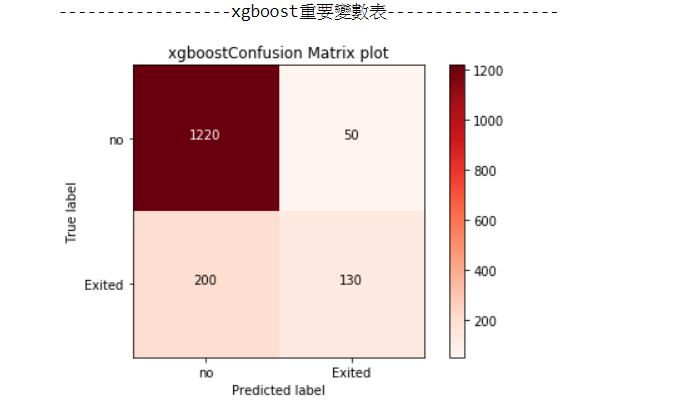
1. 統計111 H24076029劉米婷 H24071029常思為 H24071011曹舒瑜
2. 給了部分銀行客戶的訓練資料，包括客戶的信用指數、地區、年齡、不動產數量、銀行存款、金融商品、是否有信用卡、銀行活躍度、預估薪水等，以及最後會不會繼續留在這家銀行，並且給予新的測試資料，要來評估測試資料中的客戶是否會繼續留在該銀行。
3. 先使用 將非數值的變數訂新的數字分類，並判斷該數字是有否為啞變數，避免代表分類的數字列入計算影響預測結果，並分析每個變數對是否留下的關聯性，將關聯性較小的變數不列入考量
4. 使用xgboost中的XGBClassifier進行分析各個變數的重要性，畫出plot\_importance的比較圖，可以從下圖得知，因此我刪除’Balance’ 這項參數，並且，因為顧客資料有多種代號表示，所以我刪除了'RowNumber', 'CustomerId', 'Surname', 'Unnamed: 0'這幾項參數。



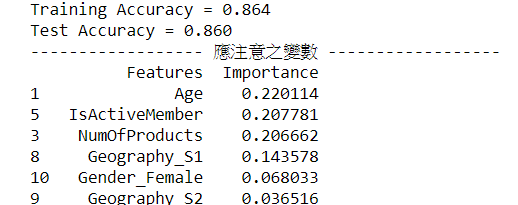
1. 將train.csv中的資料拆成訓練及測試資料，避免過度擬合的問題，使用xgboost中的XGBClassifier進行資料的訓練，可以觀察計算出我們訓練資料與測試資料的準確率

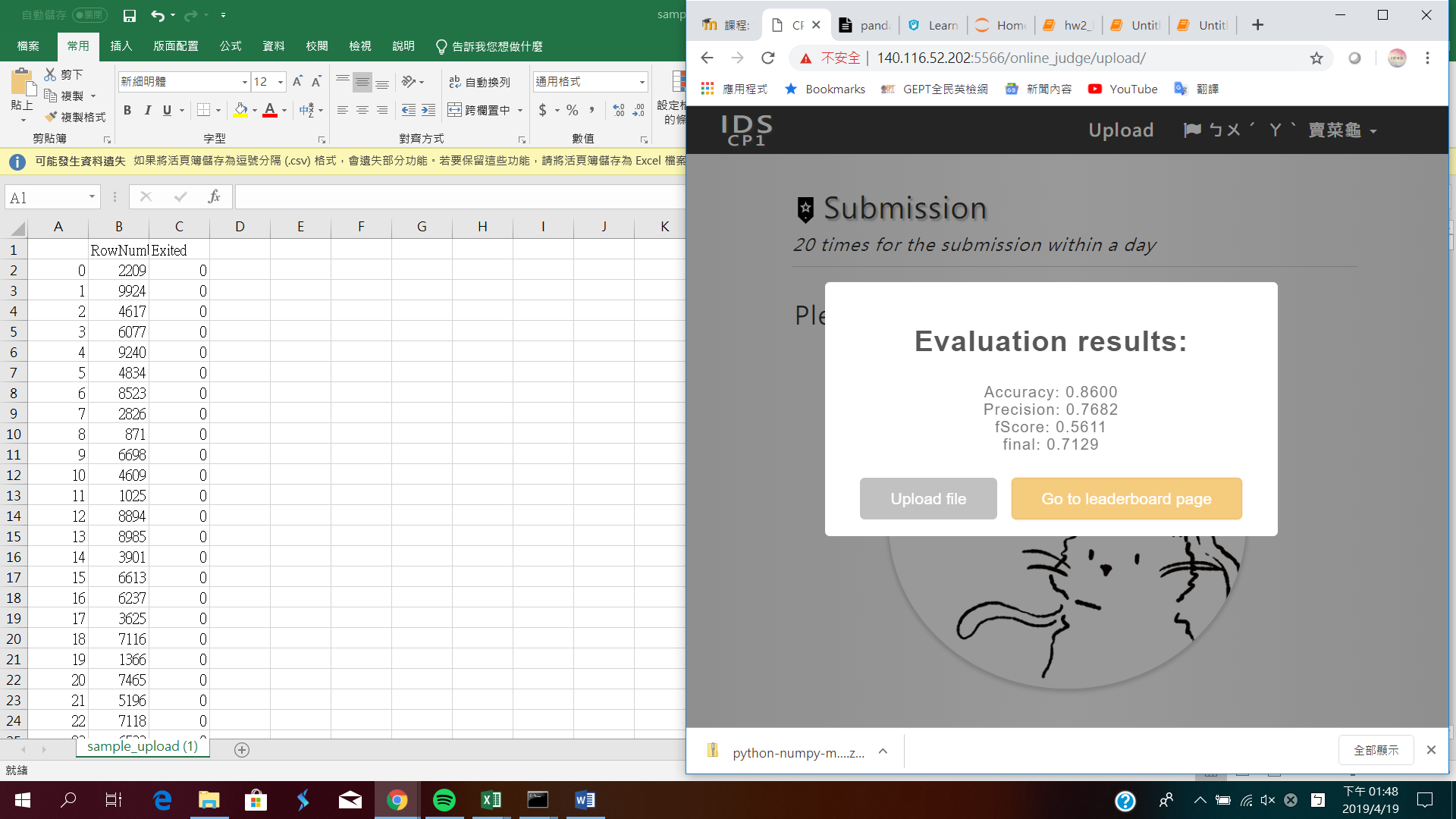
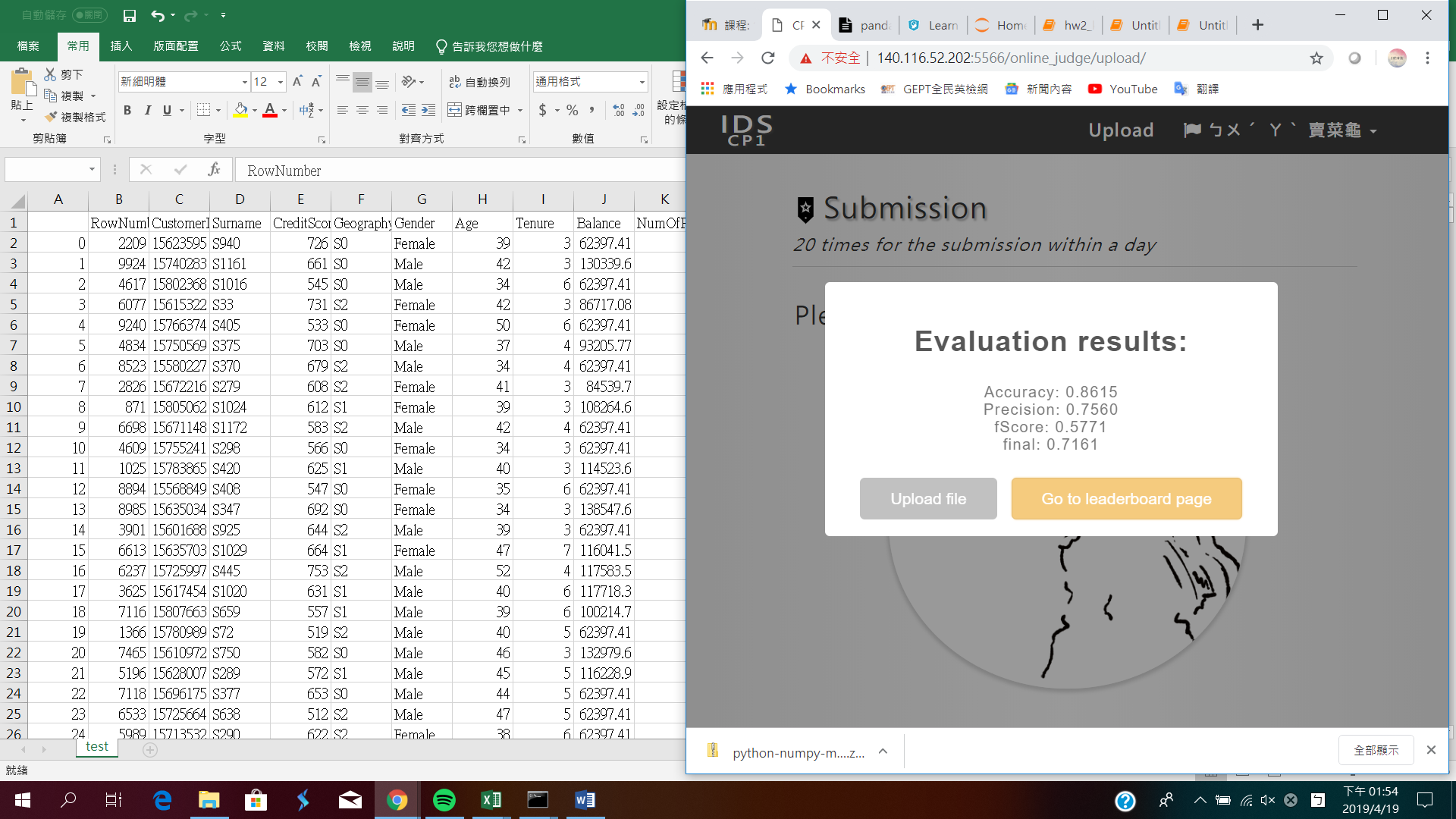
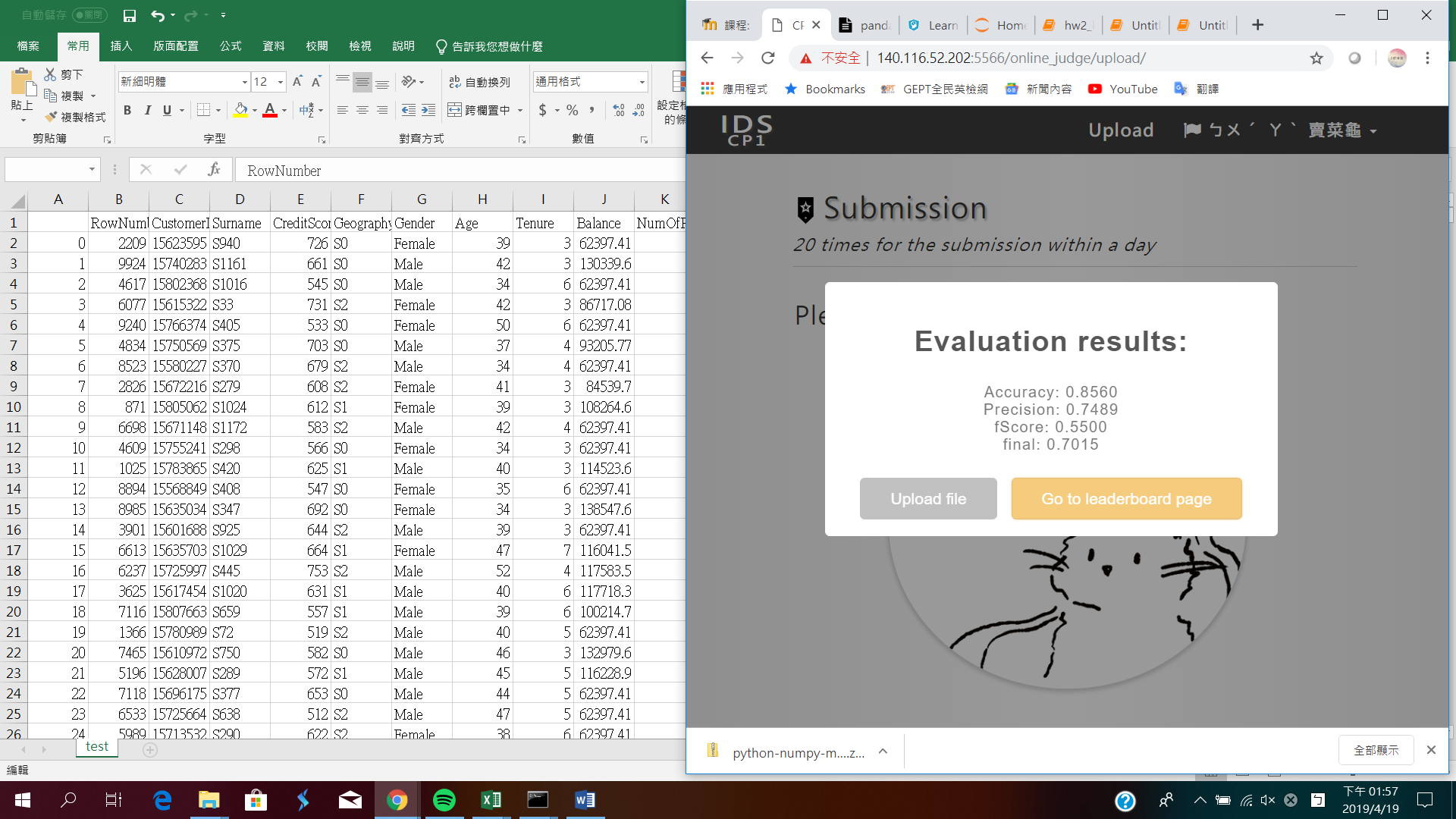
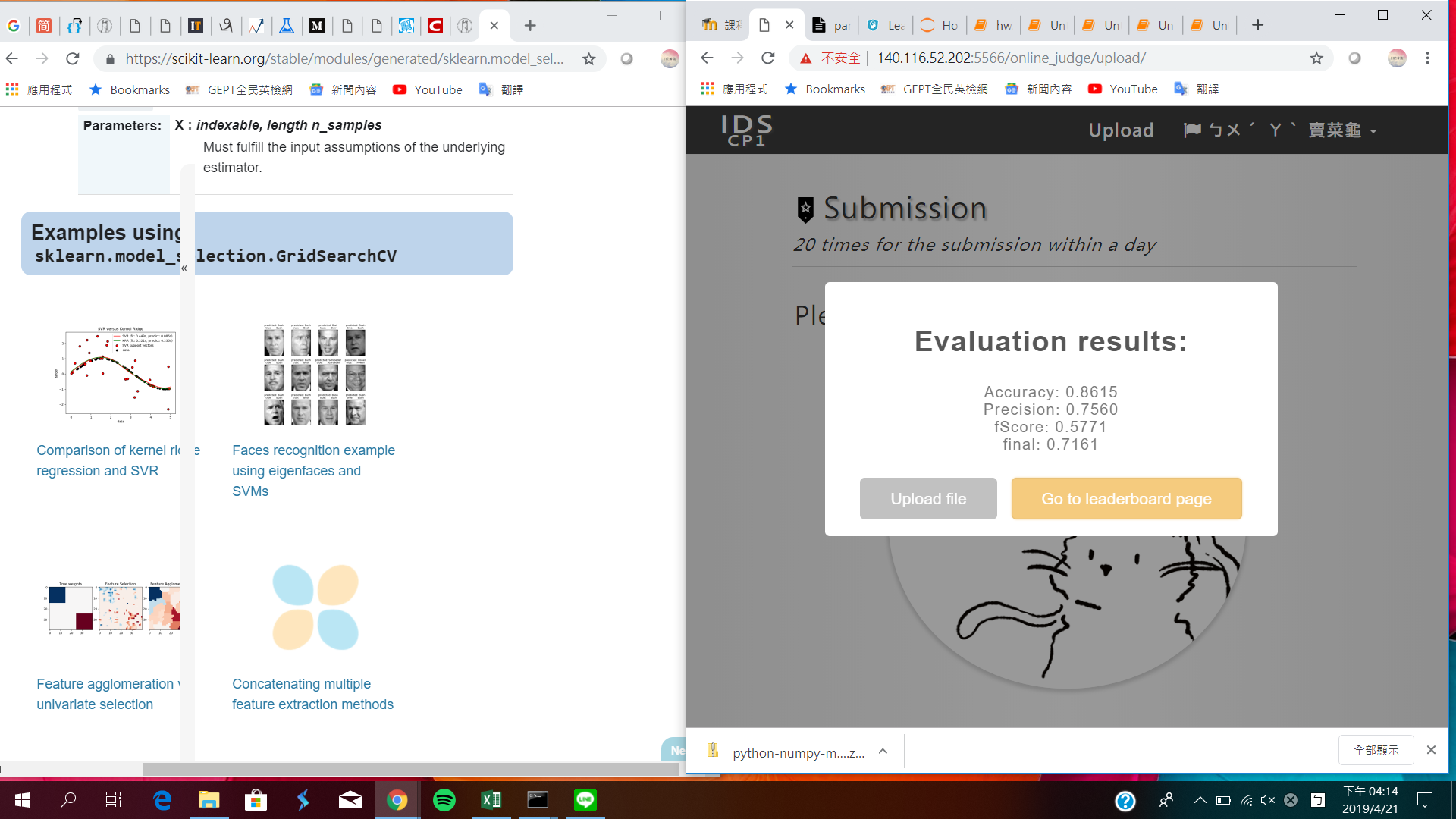
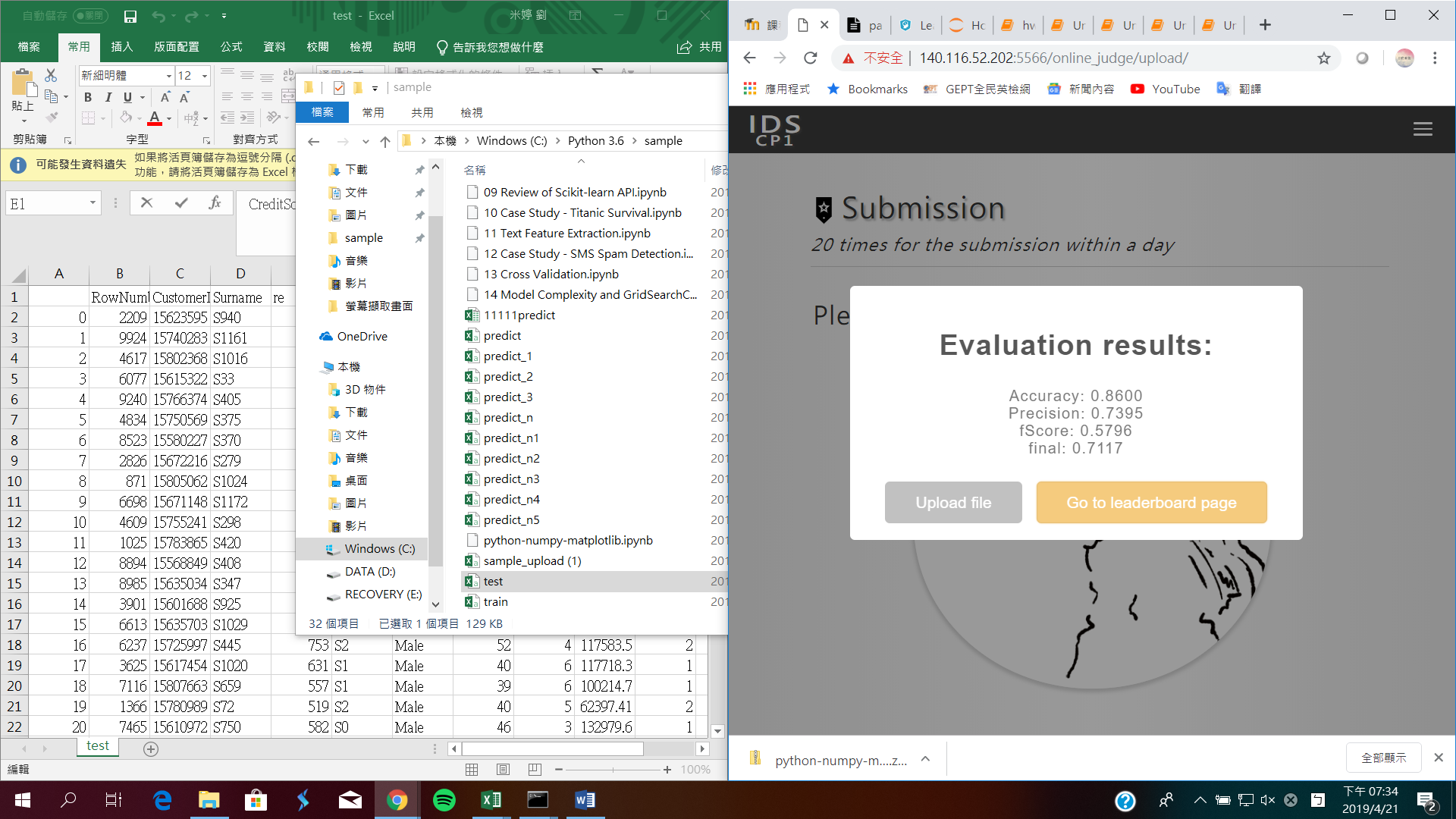


1. [方法二]

使用RandomForestClassifier 進行資料的訓練與測試，但因為沒有好好的做資料變數的篩選與處理，可能與’Exited’ 不太相關的變數我也一起考量進去了，所以沒有很好的結果，有試著調過參數，但不夠大膽的調整使的表現差異不大，如果調到100、1000等可能會好一點，但相對的也可能就出現過度擬合的問題。

[方法二]

這是我先將train.csv中的資料拆成訓練與測試資料進行模型訓練後算出的準確率，並依找此訓練出來的麼行進行 test.csv 的資料測試。

以下結果是將不同變數排出考量的結果，最後歸納出只刪除 ‘Balance’ 這項變數測出來的準確率較高。  


從先前統整出來的feature importance來看，’Tenure’, ‘Geography’, ‘Gender’ 等的變數與客戶是否流失的關係不大，但一旦將這些變數的資料排除考量，測試出來的結果卻沒有變好，我覺得可能是模型的參數沒有調到最好。

1. 統計111劉米婷 H24076029

剛開始寫的時候會想要用老師上課有提到的模型處理資料，那時講的都是迴歸的方法，但會發現不知道怎麼定義x軸和y軸，所以花了很多時間在想要怎麼用迴歸來分析，直到之後的幾堂課才發現有除了迴歸以外的方法，才剛開始真正有結果出現。剛開始會想要先計算每個變數與結果的關係，然後依照相關程度進行比重的加乘，接著用randomforest試了以後改了些參數，但由於不夠大膽的改，所以沒有什麼明顯的差異，就沒有繼續在這個方法努力了。接著我就想要嘗試用不同的模型處理，所以就用了xgboost，並且學會分析相關程度，跑出個結果後就想要嘗試提高準確率，所以就試圖要用交叉檢驗的方法，但因為對很多參數格式都不夠清楚，所以一直試不出來，有點可惜。我覺得這次最大的問題是我一直想要找新的模板來套來提高分數，卻忽略原本的模型其實還可以做很多調整。下次我希望可以進行多一點的資料分析工作再套模板，試著自己努力研究模型運作的方式，或者自己令參數等跑跑看結果，然後要使用交叉檢驗法多一點準確率，並且在自己尋找新的模板之餘還是要再把老師的ppt看熟一點才不會本末倒置。