Machine Learning Final Report

**Format:**

規定最少6頁A4長度(5頁報告 + 1頁的專案開發個人心得)，以pdf格式繳交，文字字型12，單行間距。

A minimum of 6 pages of A4 length (5 pages of report + 1 page of personal experience in project development) is required, submitted in pdf format, with a font size of 12 and single-line spacing.

**Content include：**

1. Project Title
2. Motivation and Goal
3. Related Work/Market Survey
4. Solution Introduction
   1. Solution Architecture (Including data preprocessing and model architecture)
   2. Detail of data prepocessing and the model.
5. Experiment Result
6. Introduce dataset
7. Introduce experiment environment
8. Evaluation of the proposed solution and comparison with other models.
9. The difficulties of this project and why it is the most difficult part?
10. The project meeting picture
11. Personal experience in project development (Each team member has to write)
12. Team member contribution ratio (Each team member has to write)

Machine Learning Final Report

Anti-Money Laundering

Group 3  
111522071 張凱東

111527002 陳丕中

111525025 劉洮語

1. Motivation and Goal

對於金融業而言，洗錢防制是一個必然面對的難題與挑戰。犯罪者利用各種洗錢管道將非法資金漂白，以逃避司法機關的查緝與追訴，甚至將該犯罪所得再次利用於其他不法行為當中。金融機構若不積極審查各種由其經手的交易行為，則將淪為犯罪集團的漂白管道，除了損及自身商譽之外，又因為金融機構具有集合廣大人民資金的特性，而擾亂了金融市場秩序。除此之外，金融犯罪者不斷以新興科技或渠道從事不法所得的掩飾或隱匿行為，洗錢態樣推陳出新，致使金融業單靠人力顯然不足以辨識出可疑的犯罪活動。

在現代金融系統的飛速發展和複雜的金融交易，已經導致各類詐騙和洗錢活動更加嚴重。為了有效防止洗錢活動，政府和銀行們正在尋求更有效的技術，以便檢測和監控可疑交易。機器學習技術可以提供一種更有效的方式，通過採用大數據技術和複雜的分析算法，可以更加有效地檢測和檢查可疑的洗錢活動。另外，機器學習技術可以提供更多的精準度和穩定性，更好地支持銀行對可疑交易的檢查和監控。因此，我們希望能透過機器學習的方式實作出疑似洗錢交易預測模型，減少人力成本同時增加預測效率，替金融業解決這個棘手的問題。

1. Related Work/Market Survey

我們總結了洗錢防制預測領域的現有工作，洗錢防制開始於1970年，當時金融機構開始向政府報告大額交易。一些統計方法像Bayesian models跟時間序列的匹配載1990年後期被用於檢測洗錢活動，而在2004年之後，機器學習方法開始應用時洗錢預測的模式裡。

我們這次的研究提出了使用機器學習技術，來偵測洗錢活動的可行性。在根據上述對市場上的商業解決方案進行了分析後，我們發現它們集中在金融業務處理，並且在偵測洗錢活動方面存在許多局限性，例如低準確度和無法檢測新的洗錢模式。為了解決上述問題，我們提出使用機器學習技術來偵測洗錢行為，深度學習技術可以比其他技術更有效地檢測洗錢活動，並且可以更快速地偵測新的洗錢模式。

隨著新一代投資產品的普及，以及COVID-19所造成的遠距、無接觸金融服務持續增加，使得金融機構的洗錢預測能力日漸增強。為此我們研發的這一個模型將可以用來提供給金融機構，使其可以更有效率的檢測洗錢的行為，並提高檢測的精準度。此外，我們的模型也可以用於檢測加密貨幣及虛擬資產等新一代的投資新寵，從而為金融機構帶來更多的安全保證。

|  |  |  |
| --- | --- | --- |
|  | 傳統人工方法 | 機器學習方法 |
| 準確率 | 人為檢查一定會出現失誤，不準也不穩定 | 準確率可達人類水準之上且穩定 |
| 耗時 | 一次一筆，每筆交易至少需要檢查5分鐘 | 同時處理大量交易，每筆檢查可壓縮到毫秒 |
| 投入資金 | 每個檢查員都每天需支薪，最低時薪160 | 訓練過程需要資金，真正部屬後只需維護費用 |

1. Solution Introduction
2. Raw Data

原始資料裡六個csv當作為訓練特徵，一個csv為Ground Truth，圖一說明了全部特徵所代表的意義。y會給定某個alert\_key是否有報sar(sar\_falg=1為有報)，alert\_date則是提供了alert\_key發生的日期，剩下的就是依序從各個csv提取有用的資訊。

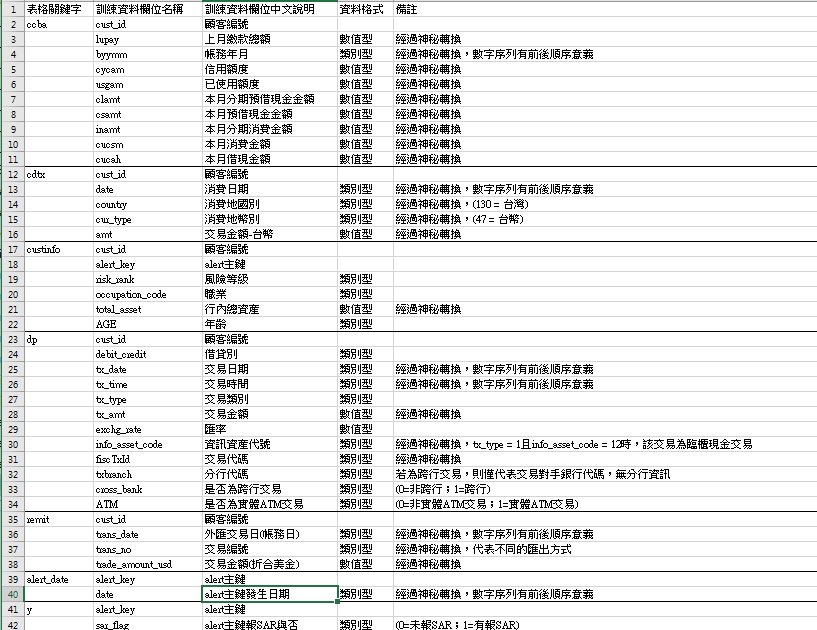


圖 一、csv欄位說明

1. Date Preprocessing

拿到原始資料後我們發現資料非常不平衡，我們使用的y.csv 有高達23672個資料類別為0，僅有234個資料類別為1，為解決這個問題我們嘗試使用ibmlearn函式庫內的上採樣器以及下採樣器，但上採樣會把1的個數提升到跟0一樣多而下採樣會把0縮減到跟1一樣少，最後我們決定裁類別0隨機採樣10000筆搭配上234筆類別1的資料(圖二)。再來將y.csv以及alert\_key.csv做合併，即可得到alert\_key發生時的日期並且知道是否有報sar。(圖三)

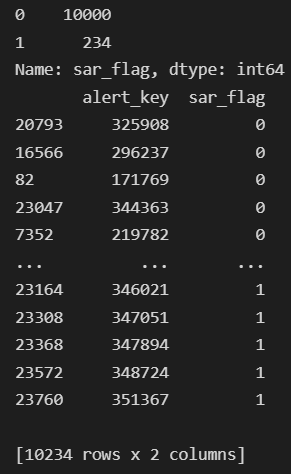


圖 二、resample後的資料

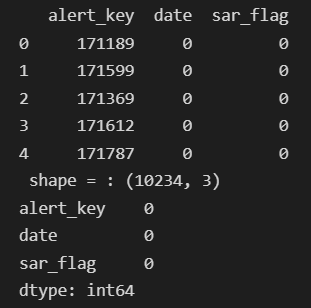


圖 三、y以及alert\_key合併後結果

接下來要處理的5個csv分別為custinfo、ccba、cdtx、dp、remit，custinfo提供了alert\_key對應到的cust\_id以及該顧客更多的個人資訊，這邊直接合併。然後ccba、cdtx、dp、remit我們有先查看內容，並確保把太多null的column drop掉，再來ccba、cdtx、dp、remit都有類似日期的欄位，我們就依序把這些欄位當作key去合併(圖四)。稍微使用 corr() 去查看各個特徵的相關性(圖五)。最後要輸入進模型訓練前有做資料的正規化，並且使用0.05去切割training data 以及 validation data(圖六、圖七)

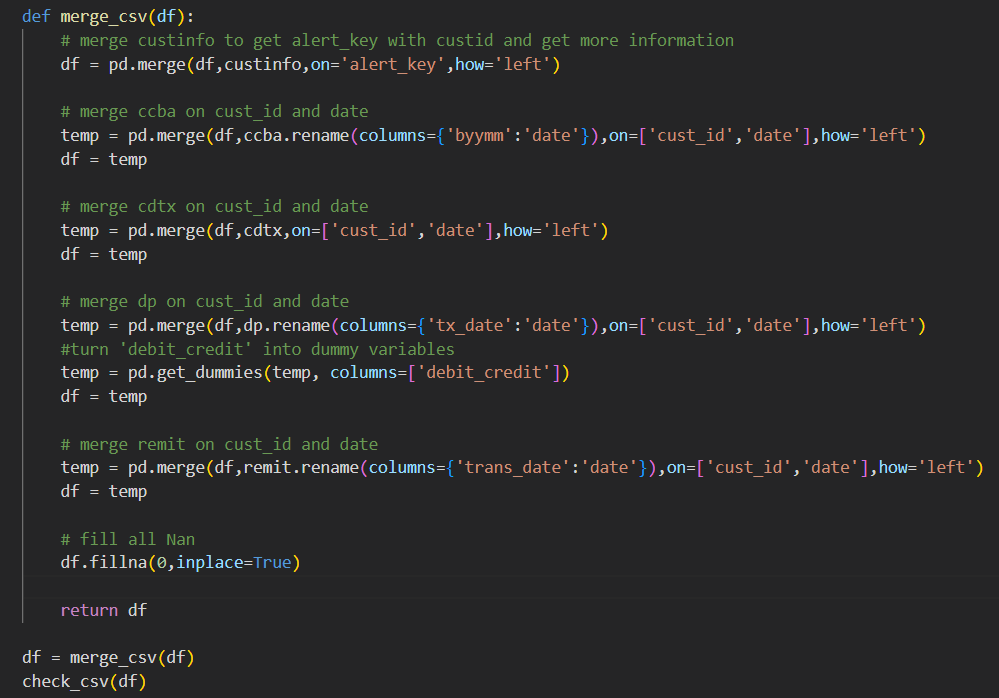


圖 四、csv合併

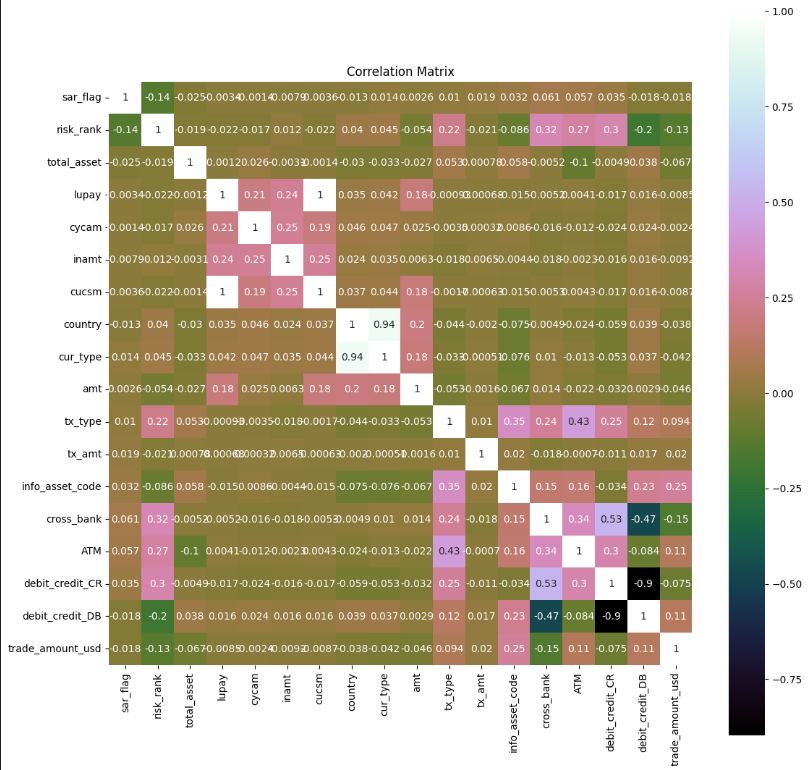


圖 五、相關性熱圖

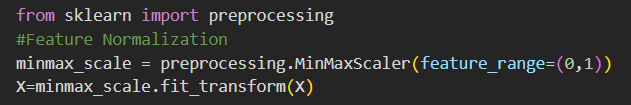


圖 六、正規化



圖 七、切割訓練資料以及驗證資料

1. Model

模型方面選擇使用DNN，輸入為上面處理好的資料並且把’cust\_id’、’alert\_key’ drop掉，留下19個Feature的交易資訊，架構方面以三層Dense Layer下去做預測，每層節點數量依序為100、10、1，激活函數選擇LeakeyRelu，輸出層選擇sigmoid，優化器選擇 Adam，損失函數選擇 Binary Cross-Entropy。

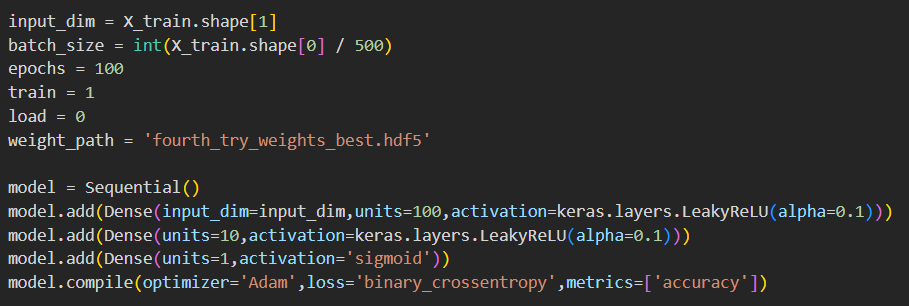


圖 八、Model架構以及訓練參數

1. Experiment Result

如圖八，我們的訓練參數使用500個Batch以及100個Epochs，使用GTX 3060去跑模型訓練，最後在訓練集得到0.977的精準度以及0.0582的loss(圖九)。利用Validation data跑評分得到還行的成績(圖十)。繳交至競賽網頁後取得0.00956的預測分數(圖十一)，分數的評分標準如圖十二。

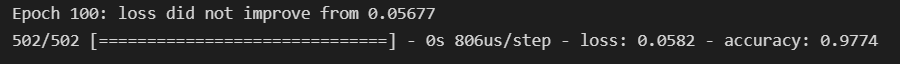


圖 九、訓練損失以及精準度

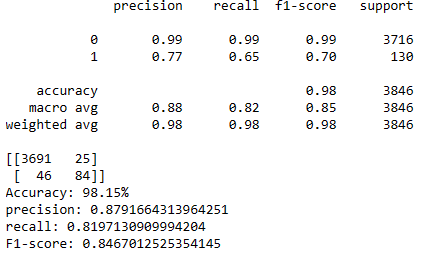


圖 十、使用Validation data得到的評分

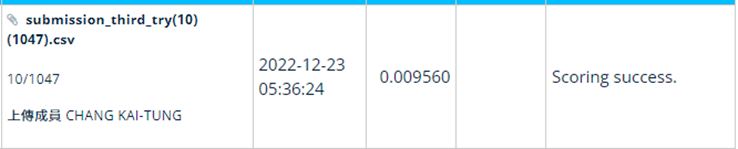


圖 十一、競賽成績

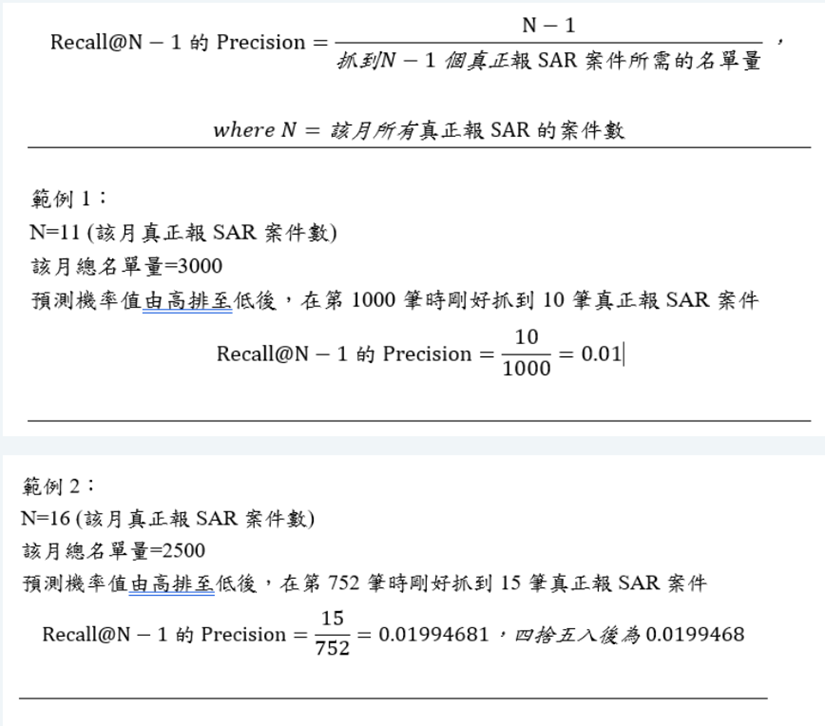


圖 十二、競賽評分標準

1. The difficulties of this project

第一個難題就是資料不平衡，資料不平衡會導致模型趨向預測資料比較多的類別，資料少的類別則很難學習到如何預測，後來使用隨機採樣有稍微解決這個問題，讓模型的表現稍微上升一些。

第二個難題是尋找特徵之間的關聯性，使用correlation查看各個特徵與答案的相關性後發現沒有特別具有影響力的特徵，那就只能盡量把每個特徵都留下來。

第三個就是模型的選擇，我們有嘗試過套用Randomforest、Logistic Regression、SVM等，但成效都比不過我們用Sequential()自己兜一個模型，至於導致這個現象的原因，我們覺得資料前處理還是比較重要的，模型的選擇反而不是特別重要，也導致了後續花費大量時間在資料前處理上。

1. The project meeting picture
2. Personal experience in project development (Each team member has to write)
3. Team member contribution ratio (Each team member has to write)

- Each team member writes down their own and team member's contribution ratio so the result can be different.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Student ID | Work | Contribution Ratio |
| 張凱東 | 111522071 |  | 100/3% |
| 陳丕中 | 111527002 |  | 100/3% |
| 劉洮語 | 111525025 |  | 100/3% |