洗錢防制

廖峻毅 資工三B 108305006

研究背景與目的

洗錢防制對於金融產業是一項重大的挑戰。犯罪集團會利用各種方式將非法資金洗白，若金融機構不積極審查其所經手的交易，則會變成犯罪集團的洗錢渠道，損害自身商譽。

本次研究將結合玉山銀行在T-Brain(AI cup)所舉辦的「你說可疑不可疑？疑似洗錢交易預測」，利用機器學習的方式預測可疑交易名單候選人可能性，降低可疑活動的誤報率、更精準的篩選出應進行申報的可疑行為。希望藉由機器學習技術的應用，將人力資源留給較為艱難且複雜的案件審核作業中，為健康的金融環境盡一份心力，並獲得比賽的成績。

資料前處理

Alert 時間 alert date	
alert_key	alert主鍵
date	alert主鍵發生日期

Label y(通報sar與否) y	
alert_key	alert主鍵
sar_flag	alert主鍵報SAR與否

顧客資訊custinfo	
cust_id	顧客編號
alert_key	alert主鍵
risk_rank	風險等級
occupation_code	職業
total_asset	行內總資產
AGE	年齡

個人行為

顧客信用借款資訊 ccba	
cust_id	顧客編號
lupay	上月繳款總額
byymm	帳務年月
cycam	信用額度
usgam	已使用額度
clamt	本月分期預借現金金額
csamt	本月預借現金金額
inamt	本月分期消費金額
cucsm	本月消費金額
cucah	本月借現金額

簽帳金融卡消費 cdtx	
cust_id	顧客編號
date	消費日期
country	消費地國別
cur_type	消費地幣別
amt	交易金額-台幣

外匯remit	
cust_id	顧客編號
trans_date	外匯交易日(帳務日)
trans_no	交易編號
trade_amount_usd	交易金額(折合美金)

付款交單D/P	
cust_id	顧客編號
debit_credit	借貸別
tx_date	交易日期
tx_time	交易時間
tx_type	交易類別
tx_amt	交易金額
exchg_rate	匯率
info_asset_code	資訊資產代號
fiscTxId	交易代碼
txbranch	分行代碼
cross_bank	是否為跨行交易
ATM	是否為實體ATM交易

資料前處理-Missing value

D/P交易金額遺失資料共有22015筆遺失。使用顧客為群組將遺失資料設為顧客的中位數。但有17894筆交易內容中該顧客並未有其他有值的交易資訊，使用當天日期的交易中位數作為遺失資料值。

```
[ ]  print(train.tx_amt.isnull().sum())
     print(train["tx_amt"].mean())

     22015
     144236.58446875788
```

D/P最初資料(missing data)個數與平均

註：D/P資料共有1969918筆資料

```
     print(train.tx_amt.isnull().sum())
     print(train["tx_amt"].mean())

     17894
     144198.68389829728
```

D/P補入顧客交易中位數後
(missing data)個數與平均

```
[ ]  print(train.tx_amt.isnull().sum())

     0

[ ]  print(train["tx_amt"].mean())

     142930.52789242458
```

D/P當天交易交易中位數後
(missing data)個數與平均

cust_id	顧客編號
alert_key	alert 主鍵
date	日期
sar_flag	是否通報SAR
risk_rank	風險等級
occupation_code	職業類別
total_asset	行內總資產
AGE	年齡
lupay	上月繳款總額
cycam	信用額度
usgam	已使用額度
clamt	本月分期預借現金金額
csamt	本月預借現金金額
inamt	本月分期消費金額
cucsm	本月消費金額
cucah	本月借現金額

remit_trade_amount_usd	近五日外匯總金額
remit_Count	近五日外匯交易次數
TW_amt	近五日簽帳金融卡境內消費總金額
TW_count	近五日簽帳金融卡境內消費次數
Foreign_amt	近五日簽帳金融卡境外消費總金額
Foreign_count	近五日簽帳金融卡境外消費次數
dta	DB五日內境內總金額
dtc	DB五日內境內次數
dfa	DB五日內境外總金額
dfc	DB五日內外內次數
cta	CR五日內境內總金額
ctc	CR五日內境內次數
cfa	CR五日內境外總金額
cfc	CR五日內外內次數

共30欄

模型簡介-隨機森林樹

隨機森林樹(Random Forest Tree)是一種集成學習(ensemble learning)的模型。它是由許多決策樹(decision tree)組成的森林，通常用於分類和回歸。

具有以下優點:

- 可以處理很多種類的資料。
- 能夠處理大量的訓練樣本，並且不需要很高的計算能力。
- 不容易過擬合(overfitting)。
- 可計算各個特徵對於預測結果的貢獻程度，更好地理解哪些特徵對於預測結果有重要的貢獻。
- 它的效果通常很好，在許多場合下都可以得到很高的準確率。

模型簡介-Extreme Gradient Boosting Regressor

XGBoost(Extreme Gradient Boosting)是一種集成學習(ensemble learning)的演算法, 通常用於分類和回歸。XGB Regressor是XGBoost的回歸版本, 用於進行回歸分析。

XGBoost的工作原理是通過梯度提升(gradient boosting)的方法來建立弱學習器(weak learner)的有力集合。這是將多個弱學習器組合成一個強學習器(strong learner), 以提高分類或回歸的準確性。

XGBoost Regressor有許多優點, 包括:

- 訓練速度很快, 因為它使用了平行計算。
- 效果通常很好, 在許多場合下都可以得到很高的準確率。
- 可以處理高維度的資料。
- 可以自動處理類別特徵。

模型簡介-Extreme Gradient Boosting Pairwise Ranker

XGB Ranker是XGBoost的另一種版本, 用於進行排名分析。

XGB Ranker pairwise是XGB Ranker的其中一種, 總共有三種類別, 分別是pointwise、pairwise與listwise。

在排名分析中, pairwise是一種常用的方法。它的原理是對於每兩個資料之間建立一個關係, 並將這些關係看為一對(pair)。

XGB Ranker pairwise是針對pairwise排名分析而設計的XGB Ranker版本。它使用特殊的損失函數和評分函數來評估模型的效果, 以便能準確的預測每對之間的相對重要性。

遭遇問題

模型實作:在本學期才開始深入機器學習相關的領域, 不熟悉一些既有的算法, 如:Random Forest抑或是XGBoosting的方式來進行預測, 但在課堂中老師有一一提及, 才慢慢熟悉。

對於題目領域不熟悉:在一開始對於金融界是一無所知, 不知道該如何將資料做整理, 甚至是不知道要如何將資料丟進模型中。

評分方式

評分方式將採用 Recall@N-1的Precision，

$$\text{Recall@N}-1\text{的Precision} = \frac{N-1}{\text{抓到}N-1\text{個真正報}SAR\text{案件數所的名單量}}$$

N=該月所有真正報SAR的案件數

例如：

N=11(該月有11筆真正有通報SAR的案件)

該月總名單量=3000

預測機率值由高排至低後，在第1000筆時剛好抓到第10筆真正有通報SAR案件

$$\text{Recall@N}-1\text{的Precision} = \frac{11-1}{1000} = 0.01$$

成果

在公開測資中有1845筆名單，有11筆真正通報的案件。

使用模型	Random Forest	XGB Regressor	XGB Ranker Pairwise
獲得成績	0.006514	0.010373	0.008064

Public成績

final(v1_xgbRegressor).csv XGBRegressor 上傳成員 Liao Jun-Yi	2022-12-25 02:27:47	1.0	0.0053163	Scoring success.

Private成績

心得

第一次接觸到關於機器學習的領域, 從完全不懂Random Forest到後續使用延伸的模組函式庫完成資料的預測。過程中遇到許多問題, 像是不知道如何進行資料前處理、模型的選取或是模型的輸入該如何調整, 像是XGB Ranker需要先將資料進行分組, 但透過GOOGLE或是在上課老師的講述下, 大部分問題都有找到解決方式, 也因為本次比賽時程後續有所改動, 變為最後兩天才公布private data使得進度被打亂, 但在之前有先寫好資料前處理的方式所以還是能夠將資料合併後進行預測。最後, 在本學期學習到很多關於機器學習的相關知識, 在未來進行不同的數據分析上將增加一項不同的分析選擇, 也了解不同機器學習的優缺點以及使用的方式, 並學到要如何進行資料前處理。

未來規劃

在2023年01月28日玉山銀行將進行頒獎典禮，頒獎典禮將會有前六名得獎者分享比賽心得與使用作法。希望能夠去聽其他人的做法，比較不同手法的優缺點。

在本學期並未能夠完成對於Transformer的學習與應用，對於這種具有時間序列的預測分析，我認為老師在課堂上所說的RNN、LSTM都是一種方式，但老師上課有講述到上述兩種方式的缺點，因此希望未來能夠使用Transformer將客戶每一筆交易資訊作為一個輸入，完成有時間序列的輸入預測。