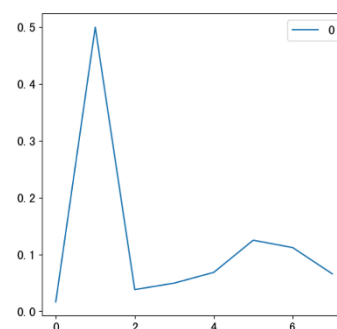


原本依序驗證 2022-03、2022-01、2021-12、2021-11、2021-10...，使用 Sequential

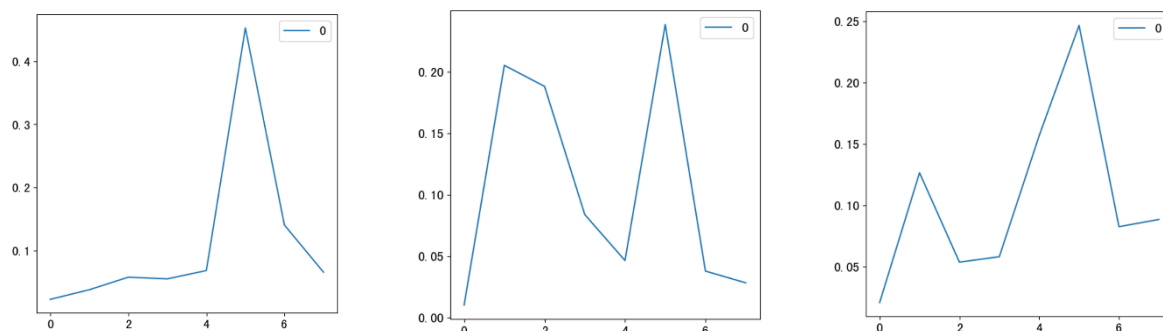
Forward Selection 挑變數，由於這次比賽很困難，常常努力挑的變數只把其中一個月分數

衝高讓平均變高，交叉驗證分數如下圖。

右圖只有 2022-01 分數很高其他月份分數低



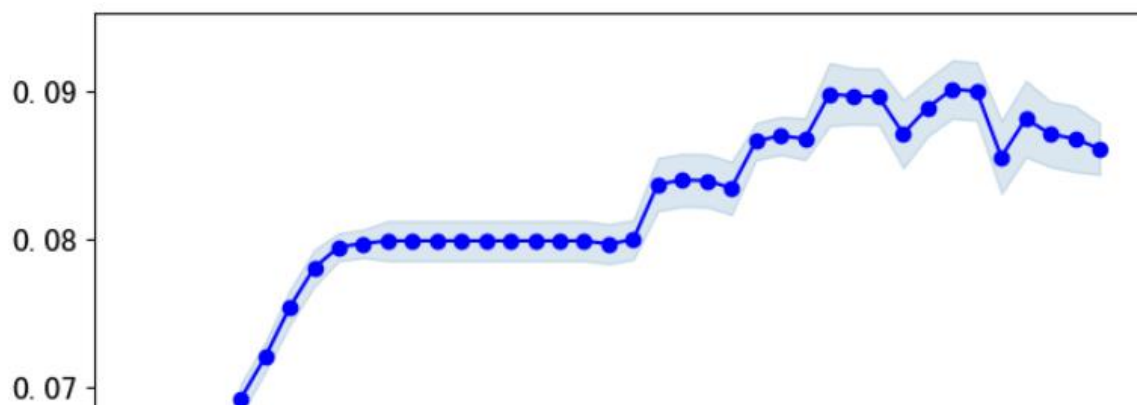
不同模型交叉驗證都有這個現象，只有某幾的月特別高分



甚至有時候可以把某個月準度衝上 1 或 0.75，正確的特徵應該要在多個月份都有幫助，所以

這樣挑選到的特徵很可能只是巧合。

另外每個月 alert 為 1 的數量很少，可能挑變數往好的方向前進，分數也不會前進。



就算 score function 增加 log_loss 項也只稍微改進，但會使和比賽的評分不同。

```
def n1recall(y_test, y_pred, num=1):
    pred1=y_pred[np.where(y_test == 1)]
    pred1.sort()
    return (y_test.sum()-num)/((y_pred>=pred1[num]).sum()) - log_loss(y_test,y_pred)*0.5
```

還是希望以本次比賽評分做優化。

Recall@N - 1 的 Precision = $\frac{N - 1}{\text{抓到}N - 1 \text{ 個真正報 SAR 案件所需的名單量}}$ ，

where N = 該月所有真正報 SAR 的案件數

為了解決少次驗證不穩定的問題，使用 30 次交叉驗證，交叉驗證使用 StratifiedShuffleSplit

以 1 測試 9 訓練比例隨機抽取資料，且確保每折測試資料都同等有 alert 0 有 2,521 筆 alert

1 有 25 筆，每次挑變數跟調整模型都改用不同的 30 次 StratifiedShuffleSplit 切法，把挑選

變數改為最差 10 筆，另外也嘗試了 5 種版本的改法，但都沒有這個方法穩定。

挑選變數

```
for new_subset, cv_scores in work:
    all_avg_scores.append(np.nanmean(cv_scores))
    all_cv_scores.append(cv_scores)
    all_subsets.append(new_subset)

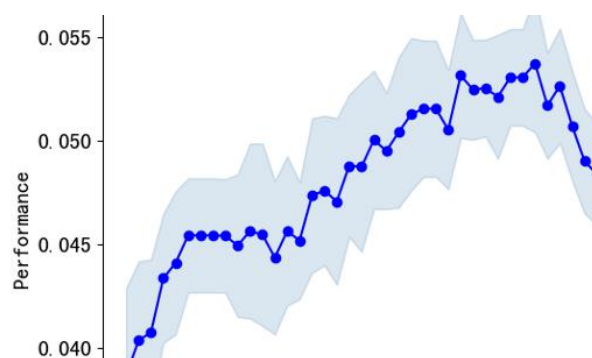
if len(all_avg_scores) > 0:
    best = np.argmax(all_avg_scores)
    out = (all_subsets[best], all_avg_scores[best], all_cv_scores[best])
```

挑選變數由平均修改為最差10筆平均(使用30次交叉驗證)

```
for new_subset, cv_scores in work:

    all_avg_scores.append(np.nanmean(sorted(cv_scores)[:10]) )
```

有了多次交叉驗證 score function 用本次比賽的評分方式也比較不容易卡住不往前進步。



特徵方面

ccba 的帳務日期，可以對到 2021 跟 2022 日期，且對應 2021、2022 國定假日不會有 alert 發生。

日期	ccba 的帳務日期
2021-04-01	0
2021-05-01	30
2021-06-01	61
2021-07-01	91
2021-08-01	122
2021-09-01	153
2021-10-01	183
2021-11-01	214
2021-12-01	244
2022-01-01	275
2022-02-01	306
2022-03-01	334
2022-04-01	365

alert 當天之前如果放假，也要拿連續假期的資料同時做特徵，禮拜一的 alert 有可能是六、

日發生洗錢交易所以 alert 筆數特別多。

星期幾	幾筆 alert
一	5839
二	4733
三	5326
四	4915
五	4532
六	110

幾乎所有 alert 等於 1 是那個顧客在那個月的最後一次出現，除了 2022 年 4 月有同月某個顧客出現兩次，因為有這筆意外，所以可以當作特徵但不能只預測那個月每個顧客的最後一筆。

是否是顧客當月最後一筆 alert	alert 主鍵報 SAR 與否	筆數
0	0	14615
	1	1
1	0	10595
	1	244

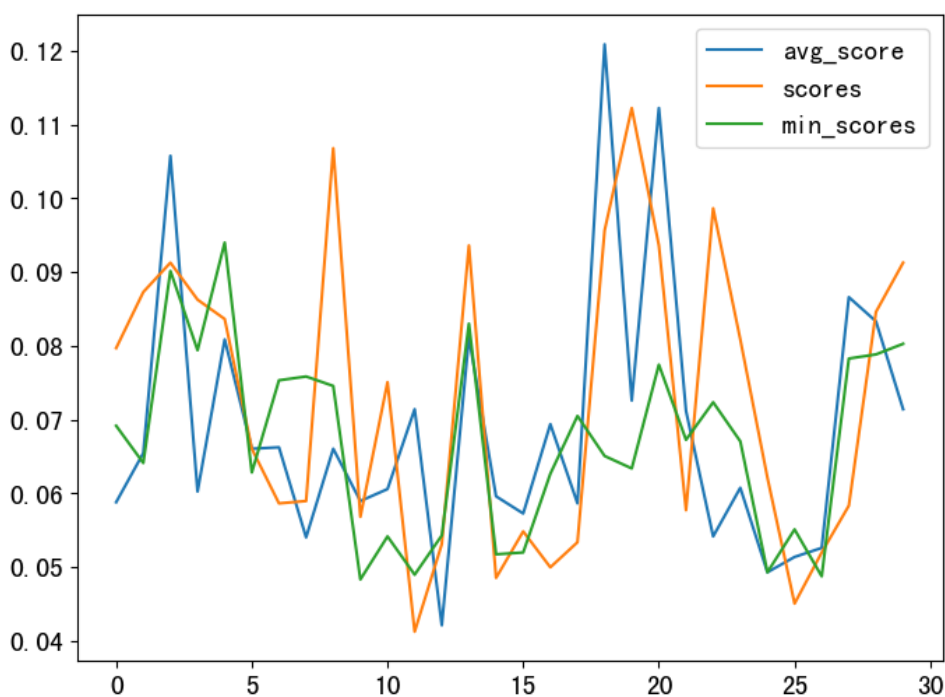
當月才出現第一次 alert 的顧客，alert 等於 1 機會稍微高一點，所以放入第一次 alert 的天數離當筆資料多少天來判斷是不是新 alert 顧客。

alert 日期	新帳號	當月帳號數	新帳號是 1	所有 1
2021-04	2031	2031	19	19
2021-05	551	737	45	51
2021-06	569	775	28	34
2021-07	502	721	25	29
2021-08	454	659	18	20
2021-09	390	631	12	17
2021-10	462	682	16	20
2021-11	434	675	10	12

2021-12	461	731	9	11
2022-01	556	939	9	13
2022-02	383	691	1	1
2022-03	471	832	4	7
2022-04	444	735	7	11

最終模型挑選最佳 30 折平均，訓練兩個模型，每次挑變數跟調整模型都改用不同的 30 次 StratifiedShuffleSplit，最後驗證準度有提升下融合兩個模型結果。

avg_scores:最佳最差 10 折平均、scores: 最佳 30 折平均、min_scores:最佳最差 1 折



使用 30 次隨機交叉驗證，或以月份驗證分數平均有 0.07 以上，如上圖最差一折至少都有 0.04 以上，最後如果運氣不好預期也有 0.05 以上。最後成績 0.054 雖然比預期差，但很好運的獲得第一名。