# Doc

## 概要

由於工作繁忙，大概零零散散參賽了2週，所以特徵處理跟EDA只有做了一些基本的，在這次競賽中主要想分享2個部分，同時也是我學習到的部分。

1. 驗證集的實驗
2. 紀錄每一次改動

## 驗證集的實驗

這次大部分都把心力花在驗證集上面，也因為這樣讓我當初public已經掉到100多名了，最終private落在70幾名，也代表我的驗證方式非常的穩定。

我一開始是用train_test_split，發現怎麼訓練，驗證分數都會輕鬆達到0.8以上（tree model），然而測試集上面卻很低分，因此我開始想著怎麼做出好的驗證集，之後嘗試了幾個不同的驗證方法分享在下面。

1. Kfold
2. Stratified Kfold
3. Groupbykfold
4. Adversarial validation

我自己平常最常用的驗證方法即是Kfold 和 Stratified Kfold，但結果也是非常overfit，代表了我們的驗證集和測試集非常不一樣。經過了思考和爬文(Kaggle)，也許這種類型的資料分布，每個月的分佈應該是很類似的，因此我選擇嘗試了groupkfold，將1~30天當成第一個月，31~60當成第二個月，以此類推。果然最終結果驗證分數非常接近public的分數，但在我的這個方法中，會有一個月特別差，我估計那個月是2月，因為當時是年節，因此分佈又更不一樣，如果有人做EDA找出原因，希望可以和我分享EDA的結果。除了以月來做，我也嘗試過如果以一週一週來看呢，最終也是overfitting,

最後也研究了一個在kaggle上很常被使用的驗證方法，非常的有趣，叫做對抗驗證，步驟如下：

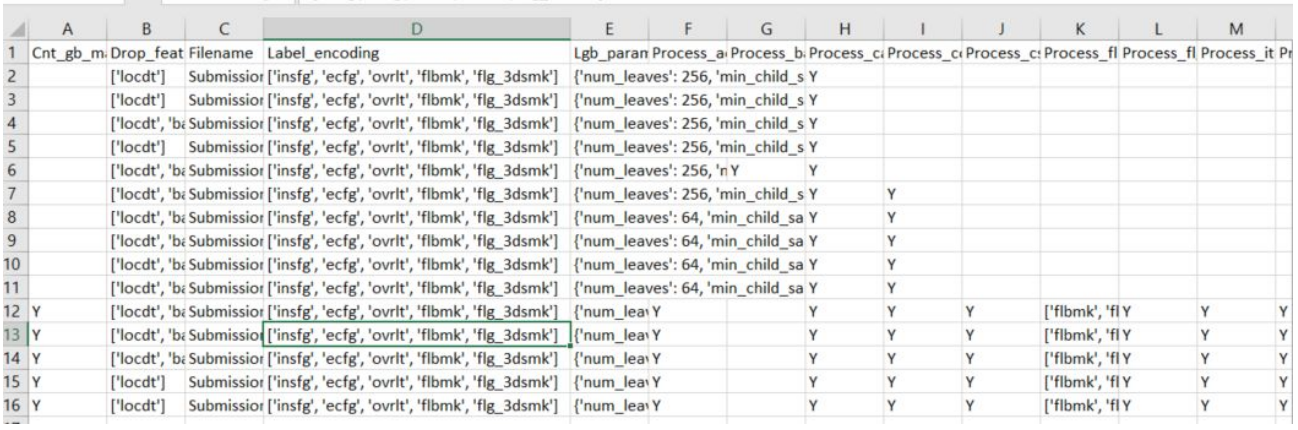
1. train[‘is_test’] = 0, train[‘is_test’] = 1
2. 並做一個分類器去預測 is_test, metric = auc

3. 正常來說，auc 應該是要落在0.5上下，如果不是，代表train 跟 test 有很明顯地分布差異。
4. 最後透過預測出來的機率，去做排序，可以挑出最像test dat的一部分當作驗證集。

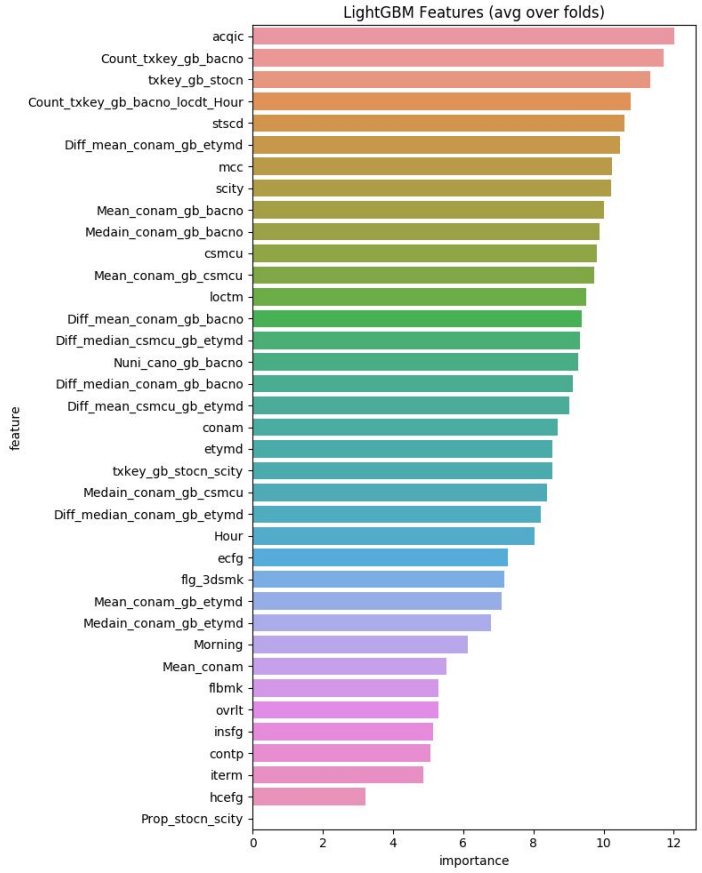
# 紀錄每一次的改動

有別於以往的競賽，我這次想要做點不一樣的，我想要把我每次的改動調整，不論是資料清洗、特徵工程、調整參數...等都記錄下來，目的也是為了知道每次的調整驗證分數和測試分數的變化。

我有一個Rocord.csv，每次會自動存取了做了什麼樣的改動，因此每一次的改動我都能清楚知道，例如：這次的版本我選則對類別特徵作one-hot、下一個版本可能是使用lgb的內建categorical_feature、下下一個版本可能是label encoding...等，並且每個版本的結果，我能更清楚的知道我該怎麼選擇，當然這個方法不是百分之百正確，只能確定說在目前的條件設定下，這個版本比其他版本好。
除了上述的csv，我也記錄了每一次的特徵重要性，當做了一些新的特徵，不只關注分數的波動，也得關心特徵的排名，是不是也跟著調整了，也許能從中找出新的想法。



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Cnt_gb_m | Drop_feat | Filename | Label_encoding | Lgb_paran | Process_a | Process_b | Process_c | Process_c | Process_c | Process_fl | Process_fl | Process_it |
| 2 | | ['locdt'] | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'min_child_s | | | Y | | | | | |
| 3 | | ['locdt'] | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'min_child_s | | | Y | | | | | |
| 4 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'min_child_s | | | Y | | | | | |
| 5 | | ['locdt'] | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'min_child_s | | | Y | | | | | |
| 6 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'n | Y | | Y | | | | | |
| 7 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 256, 'min_child_s | Y | | Y | Y | | | | |
| 8 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 64, 'min_child_sa | Y | | Y | Y | | | | |
| 9 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 64, 'min_child_sa | Y | | Y | Y | | | | |
| 10 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 64, 'min_child_sa | Y | | Y | Y | | | | |
| 11 | | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_leaves': 64, 'min_child_sa | Y | | Y | Y | | | | |
| 12 | Y | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_lea | Y | | Y | Y | Y | ['flbmk', 'fl | Y | Y | Y |
| 13 | Y | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_lea | Y | | Y | Y | Y | ['flbmk', 'fl | Y | Y | Y |
| 14 | Y | ['locdt', 'ba | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_lea | Y | | Y | Y | Y | ['flbmk', 'fl | Y | Y | Y |
| 15 | Y | ['locdt'] | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_lea | Y | | Y | Y | Y | ['flbmk', 'fl | Y | Y | Y |
| 16 | Y | ['locdt'] | Submissio | ['insfg', 'ecfg', 'ovrlt', 'flbmk', 'flg_3dsmk'] | {'num_lea | Y | | Y | Y | Y | ['flbmk', 'fl | Y | Y | Y |


LightGBM Features (avg over folds)
acqic
Count_txkey_gb_bacno
txkey_gb_stocn
Count_txkey_gb_bacno_locdt_Hour
stscd
Diff_mean_conam_gb_etymd
mcc
scity
Mean_conam_gb_bacno
Medain_conam_gb_bacno
csmcu
Mean_conam_gb_csmcu
loctm
Diff_mean_conam_gb_bacno
Diff_median_csmcu_gb_etymd
Nuni_cano_gb_bacno
Diff_median_conam_gb_bacno
Diff_mean_csmcu_gb_etymd
conam
etymd
txkey_gb_stocn_scity
Medain_conam_gb_csmcu
Diff_median_conam_gb_etymd
Hour
ecfg
flg_3dsmk
Mean_conam_gb_etymd
Medain_conam_gb_etymd
Morning
Mean_conam
flbmk
ovrlt
insfg
contp
iterm
hcefg
Prop_stocn_scity
feature
0
2
4
6
8
10
12
importance