Doc

由於工作繁忙，大概只有零零散散參賽了2週，在過程中為了做出一個好的驗證集，所以特徵處理跟EDA只有做了一些基本的，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，然發現怎麼訓練，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，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。

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。