SUBMISSION NOTES

Ngay	gio	luot submit	score	thay doi	Tăng
17/11/2023	12h	5	0.48031	- Baseline application train - Đang fill na = SimpleImputer - Đang scale = MinMaxScaler() - Code nằm trong file big_join, folder PREPROCESSING - tham số bỏ vào Logistic regression `(penalty='l2', solver='saga', C=100, max_iter=5000)`	
18/11/2023	17h	7	0.48591	- Data file big join. Chưa thay đổi gì so với lần nộp thứ 5. - Tune param  - Best grid: `{'C': 1, 'penalty': 'l2', 'solver': 'saga'}`	0.00563
	10h	8	0.48591	- Data file big join - Param tune như submission 7 nhưng với 30 fold - Best grid như submission 7 => Tăng thêm fold không ảnh hưởng gì nhiều	0
19/11/2023	9h30	12	0.5198	- Chạy file test trong ggcolab của mai, data tên my_csv - Đã preprocess lại data, bổ sung thêm các biến từa	0.03389

				các bảng, scale bằng MinMaxScaler() - Fill na bằng SimpleImputer, strategy = mean - param tune như lần 7 - `Best param: {'C': 10, 'penalty': 'l2', 'solver': 'saga'}` - `AUC: 0.775476654207589` - `File data submit tên test_tune`	
20/11	5h21 chiều	13	0.53396	Nộp file updated_ver_tune1.csv - Aggregate thêm ở bảng previous	0.01416
21/11	23h13	15	0.54838	Nộp file tune5.csv được tune trên kaggle notebook Thay MinMaxScaler bằng StandardScaler	0.01442
23/11	16h21	18	0.55093	Nộp file updated_ver_tune3.csv được tune trên google colab Thay MinMaxScaler bằng StandardScaler	0.00255
25/11	3h03	19	0.55148	Nộp updated_ver_tune4.csv Thêm feature selection với lgbm	0.00055
	9h17	20	0.55196	Nộp updated_ver_tune5.csv Thay solver ‘saga’ bằng solver ‘liblinear’	0.00048
	13h06	21	0.55227	Nộp updated_ver_tune6.csv Đổi fill na bằng ‘mean’ sang ‘median’	0.00031
	16h02	22	0.55724	Nộp after_fs1.csv Giống file nộp lần 24 nhưng được tune ở kaggle notebook, có một số feature bị thiếu so với lần nộp 24	0.00497
	17h52	24	0.55947	Nộp file fs_tune.csv Thêm bước xử lí outlier bằng 3 sigma	0.00223

26/11	12h 16	26	0.55992	Nộp file fs_tune1.csv Bổ sung các feature liên quan đến EXT_SOURCE	0.00045
28/11	0h0 6	29	0.56131	Nộp file fs_tune5.csv Bổ sung thêm feature vào các bảng	0.00277
1/12/2023	11h 49	33	0.56408	Nộp file fs_tune10.csv Thêm các feature trên bảng application về tương tác giữa EXT_SOURCE và DAYS_BIRTH, trên bảng bureau về chênh lệch thời gian, aggregate thêm các feature được thêm từ các bảng	0.00277
	16h 46	34	**0.56655**	Nộp file fs_tune11.csv Thay LabelEncoder trên bảng application train bằng TargetEncoder	0.00247
	18h 16	35	0.56662	Nộp file fs_tune12.csv Thay TargetEncoder bằng WOEEncoder	7x10^-5
4/12/2023	16h 09	43	0.56704	Nộp file fs_tune18.csv Bỏ các feature liên quan đến OBS_DAYS, SOCIALS, HOUSING,... trên bảng application	0.00042
	19h 35	44	0.56715	Nộp file test12.csv là blending của fs_tune18 và fs_tune16 với hệ số 0.9, 0.1	0.00011

Tune GG colab

- **File tune2.csv:**
  Trong bộ param_grid thay C = [10, 15, 20, 25]
  `Best param: {'C': 20, 'penalty': 'l2', 'solver': 'saga'}`
  `AUC: 0.7755256414465773 -> Không lệch nhiều so với AUC ở submission 12 nên t nghĩ nếu submit thì khả năng sẽ bằng hoặc thấp điểm hơn submission 12.` ***File này chưa được submit***
- **File updated_ver_tune1.csv**
  Preprocess thêm một số feature và drop một số cột. Data giảm từ 798 cột xuống 662 cột
  Bộ param grid giống submission 12
  `Best param: {'C': 10, 'penalty': 'l2', 'solver': 'saga'}`
  `AUC: 0.7817378769762595`

- **File tune3.csv:**

```
 Chỉ thay bộ param grid, các thứ khác giữ nguyên như submission 12
param_grid = {
              'penalty' : ['l2',
                           'l1',
                           'elasticnet'],
              'C' :[
                  # 0.001,0.01,
                    0.1, 1,
                    10, 15, 20, 25],
              'solver' : [
#                     'liblinear',
                    'saga'
#                     'lbfgs',
#                     'newton-cg',
#                     'sag'
              ],
              # 'max_iter' : [100, 500]
         }
Best param: {'C': 10, 'penalty': 'l1', 'solver': 'saga'}
AUC: 0.7754737465087335

penalty l1 tốt hơn nhưng AUC thấp hơn submission 12 =((
```

- **File updated_ver_tune2.csv:**
  Như updated_ver_tune1, nhưng có scale weight của logistic regression với tỉ lệ 0:1, 1:11

```
Best param: {'C': 0.1, 'penalty': 'l2', 'solver': 'saga'}
AUC: 0.7672582130393761
-> AUC giảm cực mạnh
```

- **File updated_ver_tune3.csv:**
  Bộ param grid mới, thay MinMaxScaler bằng StandardScaler

```
param_grid = {
              'penalty' : ['l2'
                           #  'l1',
                           #  'elasticnet'
                            ],
              'C' :[
                  0.001,0.01,
                    0.1, 1,
                    10, 100
                  # 15, 20, 25
                  ],
```

```
            'solver' : [
                # 'liblinear',
                 'saga'
                # 'lbfgs',
                #'newton-cg',
#                 'sag'
            ],
            'max_iter' : [500]
        }
```

```
Best param: {'C': 0.01, 'max_iter': 500, 'penalty': 'l2',
'solver': 'saga'}
AUC: 0.7817134844776721
```

- **File updated_ver_tune4.csv**

```
 param_grid = {
            'penalty' : ['l2'
                         #  'l1',
                         #  'elasticnet'
                          ],
            'C' :[
                0.001,
                0.01,
                  0.1, 1,
                  10
                # 15, 20, 25
                ],
            'solver' : [
                'liblinear',
                # 'saga'
                # 'lbfgs',
                # 'newton-cg',
#                 'sag'
            ],
            # 'max_iter' : [1000]
        }
```

Có sử dụng thêm feature selection with lightgbm. Số feature giảm từ 662 xuống 649

Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
AUC: 0.7837421907972548

- **File updated_ver_tune5.csv:**
  Param grid như trên, thay saga bằng liblinear
  Có add thêm feature của phanh. Tổng sau khi add là 680 feature. Feature selection nên giảm xuống 670 feature
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}

AUC: 0.7844822040495625

- **File updated_ver_tune6.csv:**
  + Thay C trong bộ param_grid thành 0.01, 0.03, 0.06, 0.09
  + Đổi fill nan từ mean sang median
  + Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  + AUC: 0.7842890573330374

- **File fs_tune.csv:**
  + Fill nan bằng median
  + Có thêm handling outlier

```
+   param_grid = {
+                'penalty' : [
+                                'l2'
+                               #  'l1',
+  #                               'elasticnet'
+                                ],
+                'C' :[
+                    0.001,
+                    0.01,
+                      0.1, 1,
+                      10
+                    # 15, 20, 25
+                    ],
+                'solver' : [
+                    'liblinear',
+                    # 'saga'
+                    # 'lbfgs',
+                    # 'newton-cg',
+  #                   'sag'
+                ],
+                # 'max_iter' : [1000]
+           }
```

```
Best param: {'C': 0.01, 'penalty': 'l2', 'solver':
'liblinear'}
AUC: 0.7858824665164515
```

- File fs_tune1.csv:
  Mọi thứ như fs_tune.csv.
  Thêm một số polynomial feature đến bậc 2 của EXT_SOURCE_1,2,3
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7860554279870758

- File fs_tune2.csv:
  Như fs_tune1.csv nhưng đã code lại phần feature selection
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7860554279870758

- File fs_tune5.csv:

```
param_grid = {
            'penalty' : [
                        'l2'
                       #  'l1',
#                         'elasticnet'
                        ],
            'C' :[
                0.001,
                0.01,
                  0.1, 1,
                  10
                # 15, 20, 25
                ],
            'solver' : [
                'liblinear',
                # 'saga'
                # 'lbfgs',
                # 'newton-cg',
#                   'sag'
            ],
            # 'max_iter' : [700]
        }
```

Code ở file feature_selection_tune1: data là my_csv_updated3
Đã add thêm 100 feature, tổng là 780 feature. Sau khi giảm thì còn 770 feature.
Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
AUC: 0.7869874606079782

- File fs_tune10.csv: my_csv_updated4
  Add thêm feature, số lượng feature ở bảng data là 805, sau khi giảm còn 792
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7883160217886012
- File fs_tune11.csv: my_csv_updated5
  Add thêm feature, có thêm TargetEncoder trên bảng application
  Số lượng feature ở bảng data là 823 feature, sau giảm còn 810
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7902278267778124

- File fs_tune12.csv: my_csv_updated6
  Thay TargetEncoder bằng WOEEncoder
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7901988588084952

- **File fs_tune18.csv: my_csv_updated9**
  Bỏ bớt các feature thừa ở bảng application.
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
  AUC: 0.7901924028671304

```
Application dataframe shape:  (307511, 124)
application_train and application_test - done in 22s
Bureau dataframe shape:  (263491, 173)
Bureau and bureau_balance data - done in 32s
<ipython-input-18-990c300c53d1>:504: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  approved['DAYS_LAST_DUE_DIFF'] = approved['DAYS_LAST_DUE_1ST_VERSION'] - approved['DAYS_LAST_DUE']
Previous dataframe shape:  (291057, 244)
previous_application - done in 42s
Pos-cash dataframe shape:  (289444, 30)
Installments dataframe shape:  (180733, 97)
Credit card dataframe shape:  (86905, 123)
previous applications balances - done in 680s
Initial df memory usage is 1881.88 MB for 802 columns
Final memory usage is: 746.07 MB - decreased by 60.4%
(307511, 802)
Return df - done in 211s
Pipeline total time - done in 1031s
```

Final shape of data of final submission

Tune Kaggle Notebook

- **File tune3.csv: Data my_csv**
  Thay SimpleImputer bằng KNNImputer -> Không work. KNN quá lâu
  Thay SimpleImputer with strategy mean -> median
  Best param: {'C': 20, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.7759668899207741 -> ***Cũng hơi lên lên nma chưa submit thử***

- **File tune4.csv:**
  Update thêm 2 cột house_score và doc_score vào my_csv
  param thêm C = 15, 20
  Best param: {'C': 20, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.776247055794161

  -> Nộp xong score là 0.5185 -> bị giảm so với lần chưa thêm 2 cột (là lần nộp thứ 12)

- **File tune5.csv:**
  Thay MinMaxScaler bằng StandardScaler
  param như tune4
  Best param: {'C': 1, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.7781549529998615
  AUC thấp hơn submission 12 NHƯNG submit score cao hơn =))). T đánh giá score này sẽ ít bị overfit vì khi CV t thấy score giữa các fold khá đều và có chung distribution.

- **File tune6.csv:**
  Scale data MinMaxScaler sau đó StandardScaler, param như tune4
  Best param: {'C': 20, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.7781575414280874

-> Đã submit thử và score = 0.54836 (giảm so với submission 15). Tức là thêm MinMaxScaler không có nhiều tác động

- **File tune7.csv:**
  Scale data StandardScaler then RobustScaler, param như tune 4
  Best param: {'C': 0.1, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.49450494820394925
- **File tune8.csv:**
  Scale data with MaxAbsScaler
  Best param: {'C': 20, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.7817013953564677
  => Ko work. Nộp AUC giảm

- **File tune9.csv:**

```
 param_grid = {
          'penalty' : ['l2'],
          'C' :[
                0.001,
               0.01, 0.1, 1, 10,
               15, 20
              ],
          'solver' : [
             'liblinear',
#               'saga'
#               'lbfgs',
#               'newton-cg',
#               'sag'
          ],
          'max_iter' : [2500]
       }
```

  Cái này fail r
- **File after_fs1:**
  Giống cái fs_tune ở gg colab. Nhưng fill na bằng mean
  Best param: {'C': 0.01, 'penalty': 'l2', 'solver': 'saga'}
  AUC: 0.78332927887074

- after_fs2.csv:
  Giống fs_tune1 ở gg colab. Fill na bằng mean