Q1. Supervised Outlier Detection

Just run the jupyter notebook of `Q1.ipynb`. It uses the given dataset. The program has one output file: * **Q1_output.csv.**

a. preprocessing
- remove the outlier by IsolationForest

b. choose model
- **Random Forest**

    class_weights = {0: 121277 , 1: 6573 }

```
Random Forest evaluate on validation set
              precision    recall  f1-score   support

           0       0.95      0.99      0.97      6280
           1       0.47      0.12      0.19       343

    accuracy                           0.95      6623
   macro avg       0.71      0.56      0.58      6623
weighted avg       0.93      0.95      0.93      6623
```

- **SMOTEENN**

    Before SMOTEENN Counter({0: 127656, 1: 6573})
    After SMOTEENN Counter({1: 123078, 0: 118077})

```
Random Forest evalute on validation set
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      6280
           1       0.46      0.24      0.31       343

    accuracy                           0.95      6623
   macro avg       0.71      0.61      0.64      6623
weighted avg       0.93      0.95      0.94      6623
```

- **Random Undersampling**

    After RU Counter({0: 6573, 1: 6573})

```
Random Forest evalute on validation set
              precision    recall  f1-score   support

           0       0.97      0.80      0.88      6280
           1       0.14      0.61      0.23       343

    accuracy                           0.79      6623
   macro avg       0.56      0.71      0.56      6623
weighted avg       0.93      0.79      0.85      6623
```

- **SMOTE**

    After SMOTE Counter({0: 127656, 1: 127656})
    class_weight = {0: 24 , 1: 1 }

```
evalute on validation set
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      6280
           1       0.47      0.27      0.34       343

    accuracy                           0.95      6623
   macro avg       0.71      0.62      0.66      6623
weighted avg       0.94      0.95      0.94      6623
```

c. Train the model

● Introduce Focal loss
FocalLoss(alpha=0.949, gamma=2.1)

	precision	recall	f1-score	support
0	1.00	0.97	0.98	6280
1	0.64	0.96	0.77	343
accuracy			0.97	6623
macro avg	0.82	0.97	0.88	6623
weighted avg	0.98	0.97	0.97	6623

sReference

[1]https://zhuanlan.zhihu.com/p/258506276 focal_loss
[2]https://blog.csdn.net/qq_41917697/article/details/117471604
[3] https://zhuanlan.zhihu.com/p/404798546 PR curve