FRAUD DETECTION ON FAKE VIRTUAL MOBILE PHONE NUMBER

A. Data:

There are 2 types of labels for 2 models:

+ Type 1 label: Subscribers who are defrauded according to current rules. Label is Positive if charged according to existing rules, Negative otherwise.

+ Type 2 label: Observing transaction arising / or not arising in the next month after receiving the fee. The label is Positive (virtual) if there is no transaction (financial) in the next month after receiving the fee.

Data for machine learning is aggregated from July 2020 to the end of December 2020. With label type 2, December is omitted because no transaction has been observed until the end of January. Label and model data Will be updated to get richer over the months.

We will use 2 models, respectively model 1 and model 2 for label types 1 and 2 respectively.

B. Model

Extreme gradient boosting (XGBoost)

Extreme gradient boosting was developed on the framework of gradient boosting [9]. Boosting, also known as the 'sequential ensemble method', creates a sequence of models in which the models attempt to correct the mistakes that were made in the previous model in the sequence. The first model is built based on the training data, then the second model attempts to improve on the first model, after which the third model attempts to improve on the second model, and so on.

Figure 1 shows that the original data is passed to the first classifier. The yellow area represents the predicted blue hyphen, and the blue area represents the predicted red cross. Thus, for this first attempt, the classifier misclassified the three circled instances. After this, the weights of these incorrectly classified instances are adjusted and sent to the second classifier. The second classifier then correctly predicts the three instances incorrectly predicted by the first classifier, but incorrectly predicts three different instances. This process is then repeated until the specified number of iterations is reached, or a certain threshold is reached by the classifier.

Gradient boosting uses an approach where a new model is created that predicts residuals (errors) of the prior models that, when added together, make the final prediction. Gradient boosting uses gradient descent to minimize the loss function (the difference between the predicted value and the actual value).

The objective function of the XGBoost model can be calculated by adding the loss function with a regularization component. This means that the loss function has predictive power, and the regularization term controls the simplicity and the overfitting of the model.

XGBoost can be used for both regression and classification problems, with a loss function such as root means squared error for regression versus a loss function such as log loss for binary classification. One major advantage of XGBoost is that it parallelizes the tree-building components of the boosting algorithms; it is thus very fast to train and test.

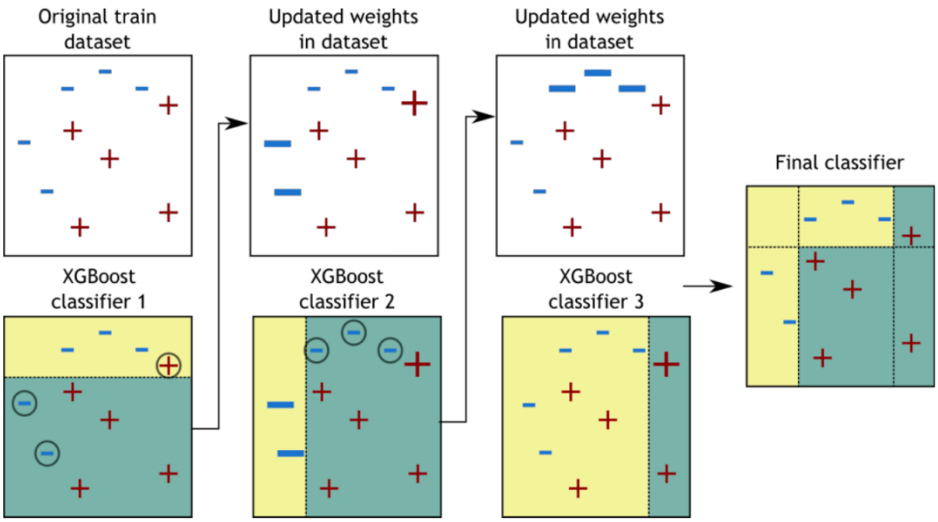


Figure 1. Boosting

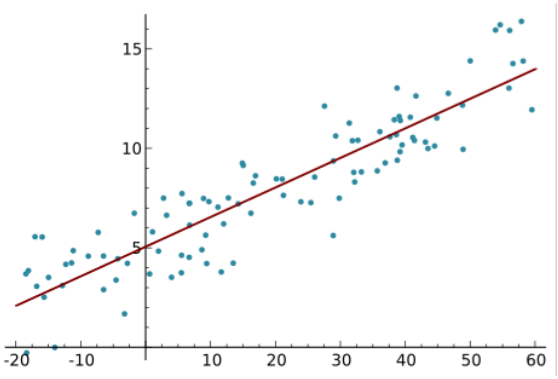
C. Results:

+ With label type 1, gini on backtest set is currently 0.83, with best cutoff criteria, giving recall=62.4% and precision=65%

+ With type 2 labels, the gini on the backtest set is currently 0.58, with the best cutoff criteria, giving recall 72.5% and precision 55.6%

D. Detailed parameters

1. **Model Accuracy**
   a. **Model 1**



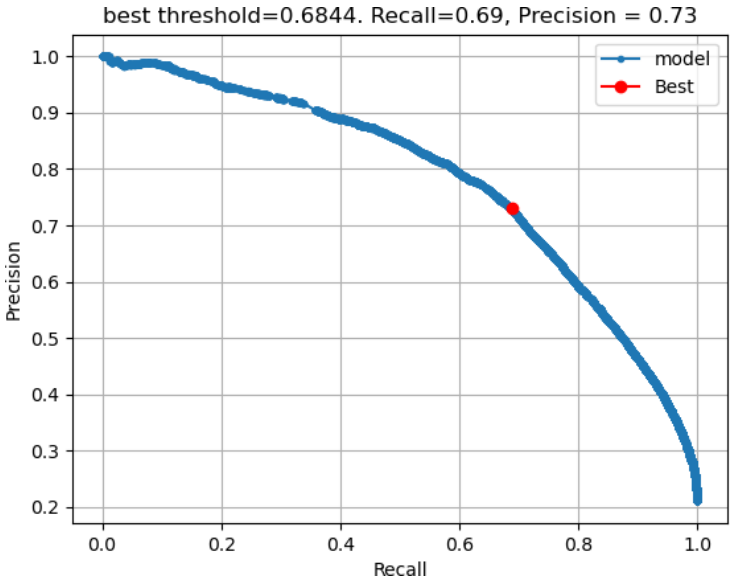
```
Best threshold finding on this test set = 0.6851
on TRAIN
AUC = 0.920 | Gini = 0.840  | Recall score=0.699 | Precision score=0.732
confusion matrix:
                Predict Negative  Predict Positive     SUM
Actual Negative            369462             26401  395863
Actual Positive             31144             72181  103325
SUM                        400606             98582  499188
********************************************************************************
Best threshold finding on this test set = 0.6844
on TEST
AUC = 0.913 | Gini = 0.827  | Recall score=0.687 | Precision score=0.731
confusion matrix:
                Predict Negative  Predict Positive    SUM
Actual Negative             41078              2907  43985
Actual Positive              3588              7893  11481
SUM                         44666             10800  55466
********************************************************************************
Best threshold finding on this test set = 0.7087
on BACK TEST
AUC = 0.914 | Gini = 0.829  | Recall score=0.630 | Precision score=0.649
confusion matrix:
                Predict Negative  Predict Positive     SUM
Actual Negative            614593             26684  641277
Actual Positive             29038             49366   78404
SUM                        643631             76050  719681
```

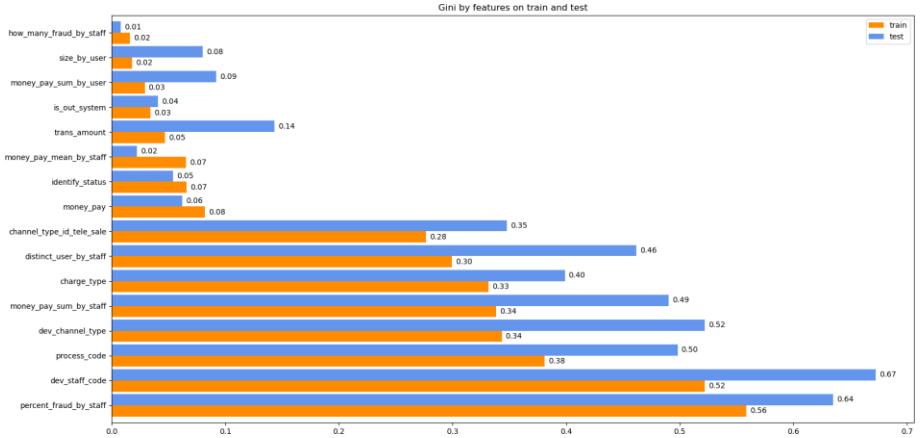
b. Model 2

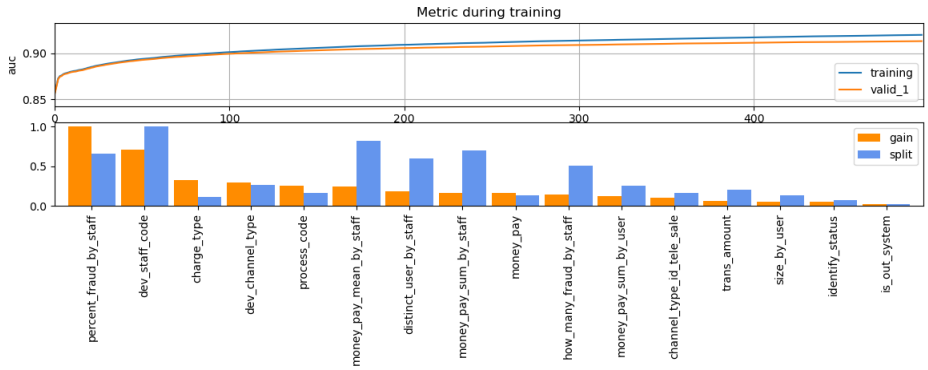
```
********************************************************************************
Best threshold finding on this test set = 0.4814
on TRAIN
AUC = 0.828 | Gini = 0.655  | Recall score=0.773 | Precision score=0.603
confusion matrix:
                Predict Negative  Predict Positive     SUM
Actual Negative            281729            110925  392654
Actual Positive             49463            168313  217776
SUM                        331192            279238  610430
********************************************************************************
Best threshold finding on this test set = 0.4814
on TEST
AUC = 0.821 | Gini = 0.641  | Recall score=0.763 | Precision score=0.597
confusion matrix:
                Predict Negative  Predict Positive    SUM
Actual Negative             31170             12458  43628
Actual Positive              5723             18475  24198
SUM                         36893             30933  67826
********************************************************************************
Best threshold finding on this test set = 0.4516
on BACK TEST
AUC = 0.792 | Gini = 0.584  | Recall score=0.725 | Precision score=0.556
confusion matrix:
                Predict Negative  Predict Positive     SUM
Actual Negative            160025             65964  225989
Actual Positive             31260             82507  113767
SUM                        191285            148471  339756
```

2. Other parameters

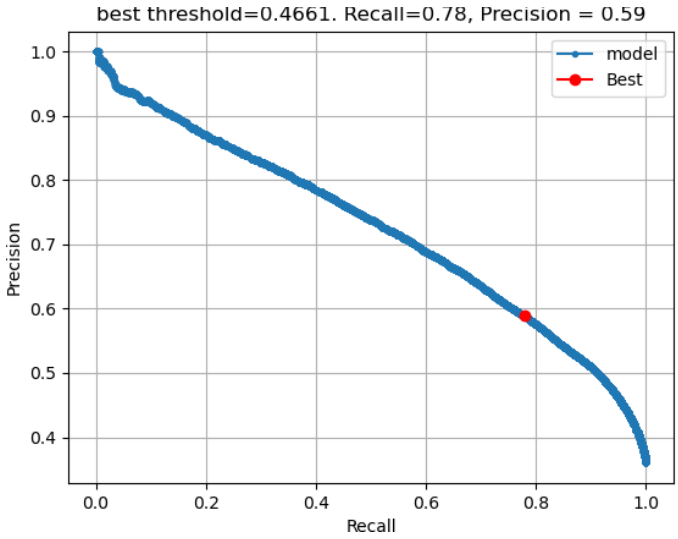
a. Model 1

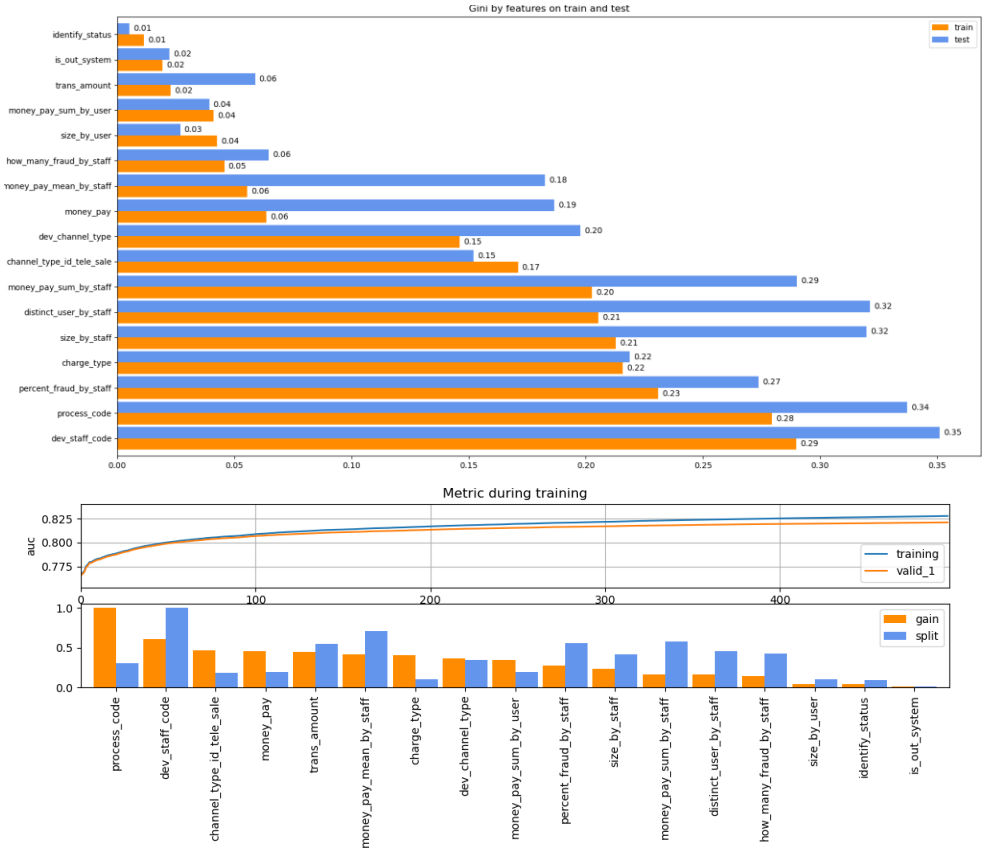

best threshold=0.6844. Recall=0.69, Precision = 0.73
model
Best
Precision
Recall


Gini by features on train and test
train
test
how_many_fraud_by_staff
size_by_user
money_pay_sum_by_user
is_out_system
trans_amount
money_pay_mean_by_staff
identify_status
money_pay
channel_type_id_tele_sale
distinct_user_by_staff
charge_type
money_pay_sum_by_staff
dev_channel_type
process_code
dev_staff_code
percent_fraud_by_staff


Metric during training
auc
training
valid_1
gain
split
percent_fraud_by_staff
dev_staff_code
charge_type
dev_channel_type
process_code
money_pay_mean_by_staff
distinct_user_by_staff
money_pay_sum_by_staff
money_pay
how_many_fraud_by_staff
money_pay_sum_by_user
channel_type_id_tele_sale
trans_amount
size_by_user
identify_status
is_out_system

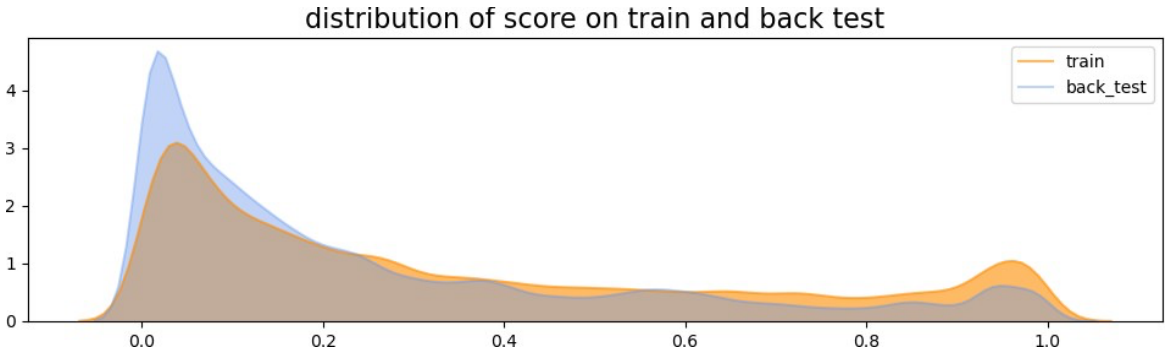
b. Model 2

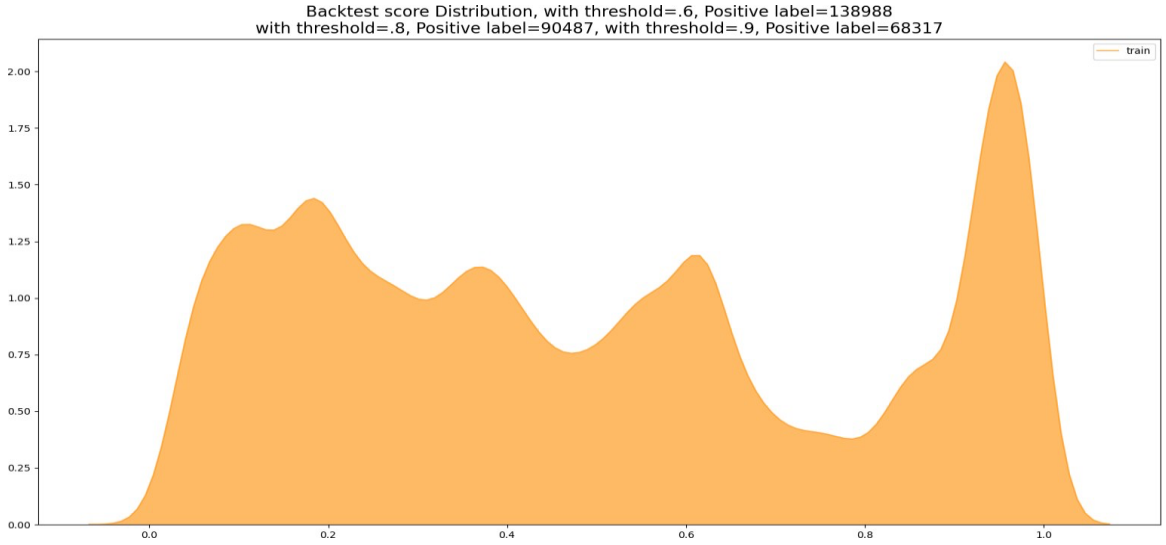




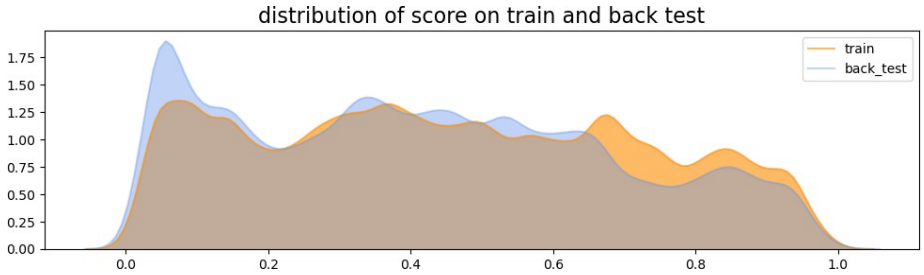
3. **Compare the distribution of the data fields on the train set and the backtest set.**
4. **Distribution of the point spectrum**

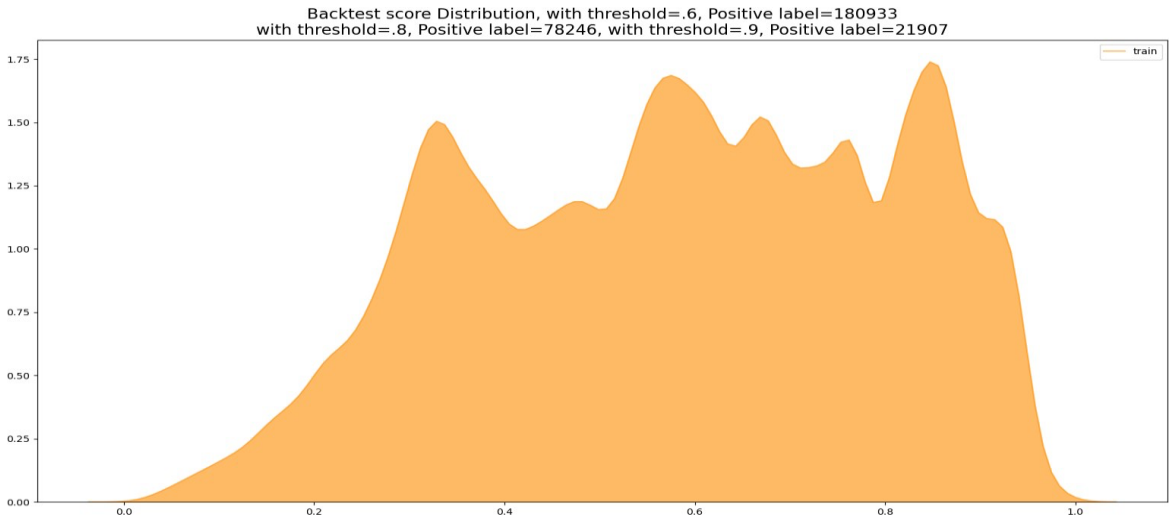
a. **Model 1**


distribution of score on train and back test
train
back_test
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

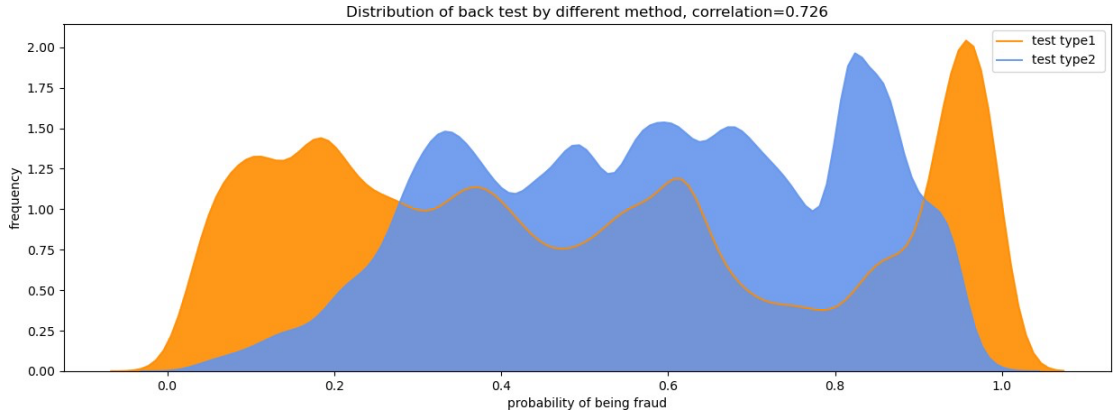

Backtest score Distribution, with threshold=.6, Positive label=138988
with threshold=.8, Positive label=90487, with threshold=.9, Positive label=68317
train
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0.0
0.2
0.4
0.6
0.8
1.0

b. Mô hình 2

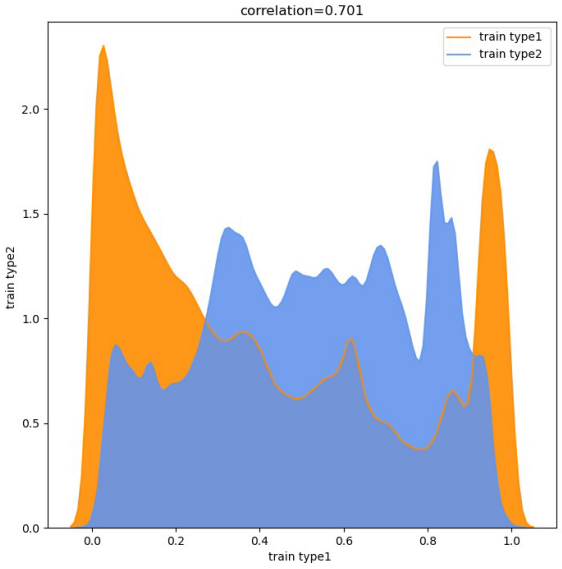

distribution of score on train and back test
train
back_test
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.0
0.2
0.4
0.6
0.8
1.0

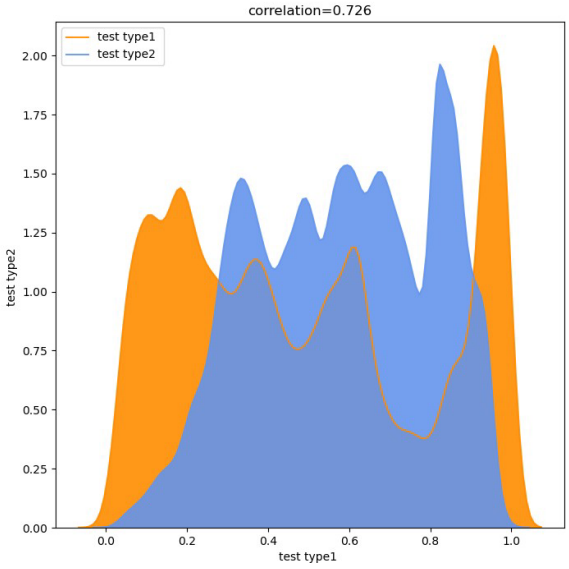


c. Score correlation between the two models:



comparison score distribution by method





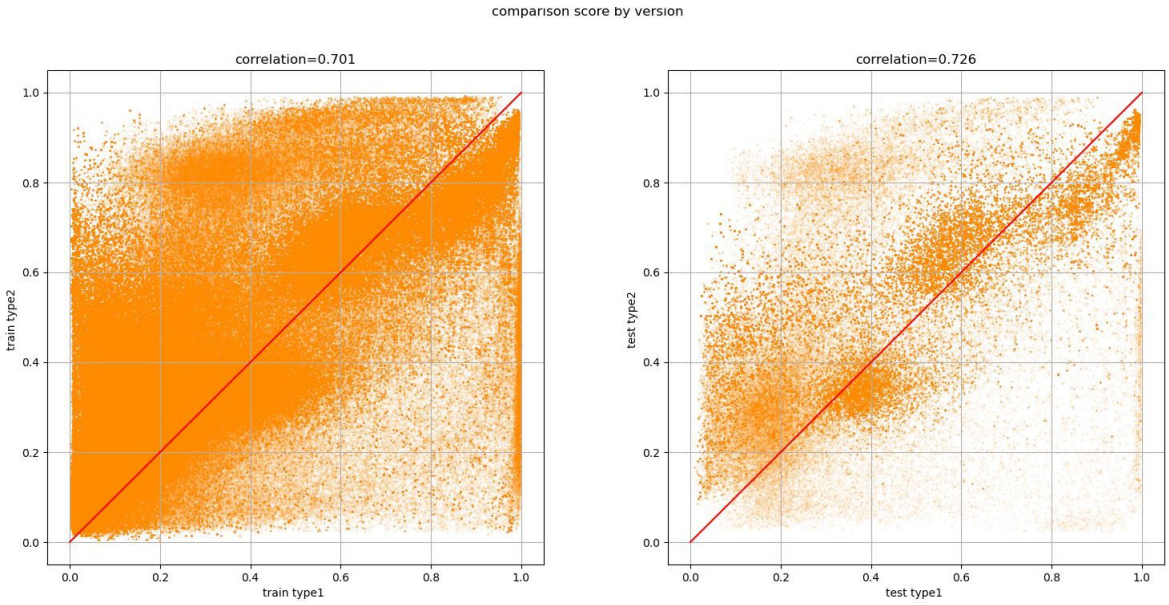


Figure 1 So sánh giá trị của score trên cùng tập train/back test theo các phương pháp gán nhãn khác nhau

E. Conclusion

- Data quality is relatively good, resulting in models that work stably on different sets on different parameters and algorithms.
- The score attributes of both models are very correlated with each other, which proves that the two labels currently have a close relationship. These 2 results can be used to complement each other in two directions:

+ Consolidation: Retrieve the set of positives if both models vote positive.
+ Expanding objects: Subscribers are identified virtual if at least one of the two identification models is virtual.

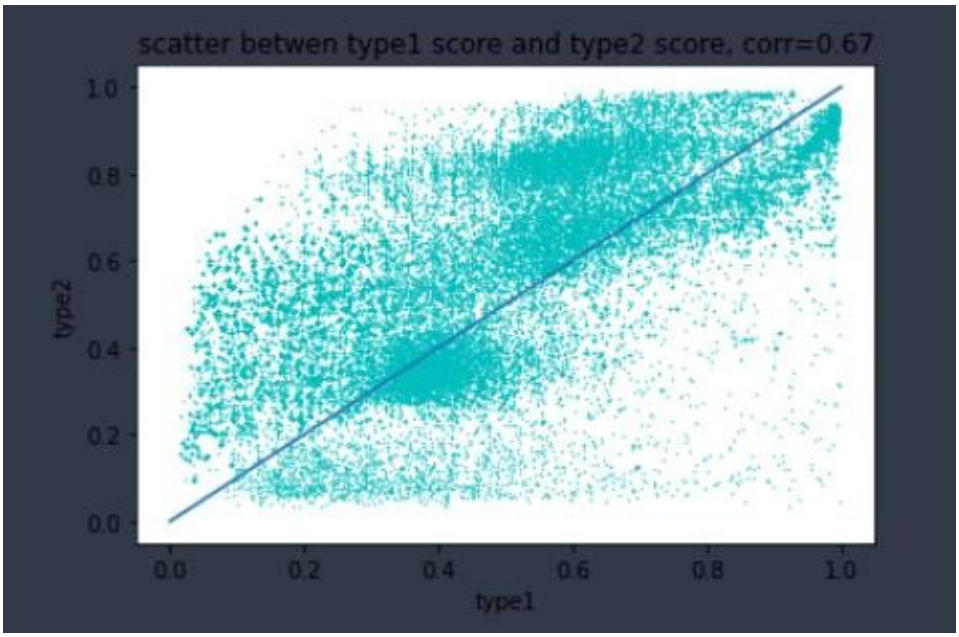
- Model 1 has high accuracy, and model 2 has acceptable accuracy. It is recommended to use both models to run alarms and tolls.

F. Model Post-test

In February 2020, a list of 18293 subscribers suspected of virtual development for profiteering, up to 87% of them had check_ok as null (potentially not generating any event to get paid for development). 10.67% have already defrauded the fee, and 1.98% have still received the fee. The fact that the check_ok field occupies a very large proportion is because events of type check_ok are nan were not included in the model for training (because only training to distinguish check_ok=1 and check_ok=-1 statuses).

Overall, the gini index is doing quite well between the model's score and the premium/premium list.

	Model 1	Model 2	Max Score	Min Score
Gini vs true target	0.69	0.55	0.66	0.64


scatter betwen type1 score and type2 score, corr=0.67
type2
type1
0.0
0.2
0.4
0.6
0.8
1.0

G. Phụ lục