My algorithm:

1. Data exploratory:

A. Understand the project. Based on given data set, this project is a binary classification, fraud and non-fraud, with fraud ratio around 0.30.

B. Understand the data set. The number of the train data set, test data set and feature are 14000,6000 and 23 respectively.

C. Classify the features by format. Divided features into continuous features, timestamp features, and also the discrete features.

2. Feature processing:

A. Duplicated and single dimension features.

We drop these columns since they are useless for solving the problem.

Note:

```
duplicated_cols =['ACTVY_DT','TRAN_DT','CUST_ZIP'];
single_dim_cols =['ACTN_CD','ACTN_INTNL_TXT','TRAN_TYPE_CD']
```

B. missing value and its influence on fraud.

Normally, people use mode or mean value to fill the missing. But here is not the case,because by analyzing data, we found missing value is very useful for classification. With special concern, we fill the missing value with a special value.

Note: special value equals to “missing”

C. Discrete features

using 10 as threshold, counting the unique value of the feature, we divide the features into two parts, naming “discrete_cols_over_10” and “discrete_cols_below_10”.

For the “discrete_cols_over_10”, by calculating the fraud ratio of each value, we classify the value into 10 risk level, ranging from no-risk, middle-risk to high-risk. E.g. for feature “CARR_NAME”， one value is “charter communications”, the fraud number is 226, the non-fraud number is 1459, the fraud ratio equals to 226/1459, we label “charter communications” with label “risk-0.3”.

Note:

```
discrete_cols =['CARR_NAME','RGN_NAME','STATE_PRVNC_TXT','CUST_STATE',
    'ALERT_TRGR_CD','DVC_TYPE_TXT','AUTHC_PRIM_TYPE_CD','AUTHC_SCNDRY_STAT_TXT']
discrete_cols_over_10 =['CARR_NAME','RGN_NAME','STATE_PRVNC_TXT','CUST_STATE']
discrete_cols_below_10=['ALERT_TRGR_CD','DVC_TYPE_TXT','AUTHC_PRIM_TYPE_CD',
                        'AUTHC_SCNDRY_STAT_TXT']
discrete_cols_over_10_cls =['CARR_NAME_bin_10_feature', 'RGN_NAME_bin_10_feature',
                            'STATE_PRVNC_TXT_bin_10_feature','CUST_STATE_bin_10_feature']
```

Item in “discrete_cols_over_10_cls” represents the treated feature name.

D. Timestamp feature

Non-existed date. By exploring data, we found several non-exited date,like “1/0/2021” in features ”PWD_UPDT_TS” and “PH_NUM_UPDT_TS”, will cause fraud definitely.

Extract operating hour from existed one. Common sense tells us, different operating hour may contribute to different fraud risk.

Cross features to get high-level information. Manually calculate the updating day gap between transaction and operation, updating phone number or password.

note:

ts_cols =['PWD_UPDT_TS','PH_NUM_UPDT_TS', 'TRAN_TS']

ts_hour_cols =['PWD_UPDT_TS_hour','PH_NUM_UPDT_TS_hour', 'TRAN_TS_hour']

op_day_cols =['PWD_UPDT_TS_day','PH_NUM_UPDT_TS_day']

the item in " ts_hour_cols" represents the new feature' name,so as to "op_day_cols"

E. Continuous features

There is no missing value in these features, so we just keep them originally.

Note:

continuous_features=['TRAN_AMT','ACCT_PRE_TRAN_AVAIL_BAL','CUST_AGE',
'OPEN_ACCT_CT','WF_DVC_AGE']

3. Feature dimension reduction:

Employing random forest to calculate the features' importance, we drop some useless features to strength robust. the feature candidates are the union set of "continuous_features","discrete_cols_below_10","discrete_cols_over_10_cls","ts_hour_cols","op_day_cols". the output show that the feature in "bottom_cols",listed on the note below,is useless. finally, we exclude these features in the following step.

Note:

continuous_features=['TRAN_AMT','ACCT_PRE_TRAN_AVAIL_BAL','CUST_AGE',
'OPEN_ACCT_CT','WF_DVC_AGE']

discrete_cols_below_10=['ALERT_TRGR_CD','DVC_TYPE_TXT','AUTHC_PRIM_TYPE_CD',
'AUTHC_SCNDRY_STAT_TXT']

discrete_cols_over_10_cls =['CARR_NAME_bin_10_feature', 'RGN_NAME_bin_10_feature',
'STATE_PRVNC_TXT_bin_10_feature', 'CUST_STATE_bin_10_feature']

ts_hour_cols =['PWD_UPDT_TS_hour', 'PH_NUM_UPDT_TS_hour', 'TRAN_TS_hour']

op_day_cols =['PWD_UPDT_TS_day', 'PH_NUM_UPDT_TS_day']

bottom_cols =['ALERT_TRGR_CD','CUST_STATE_bin_10_feature','TRAN_TS_hour',
'PWD_UPDT_TS_day','AUTHC_PRIM_TYPE_CD','AUTHC_PRIM_TYPE_CD',
'AUTHC_SCNDRY_STAT_TXT','CUST_AGE']

4. Model selection and training:

compare the reports of xgboost, random forest,DeepFM and also the merged result among them, we choose xgboost as our model.

5. Model predict:

processing the test date set like the training one, and put the features' array into the model, finally we get the predict result.