Detect Fraud from Customer Transactions

Fraud Detection

- help the banks to save money
- help the government to crack down economic crimes
- protect customers benefits

It's a binary classification problem.

Use machine learning algorithms to solve this problem.

Data acquisition and cleaning

- The data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features.

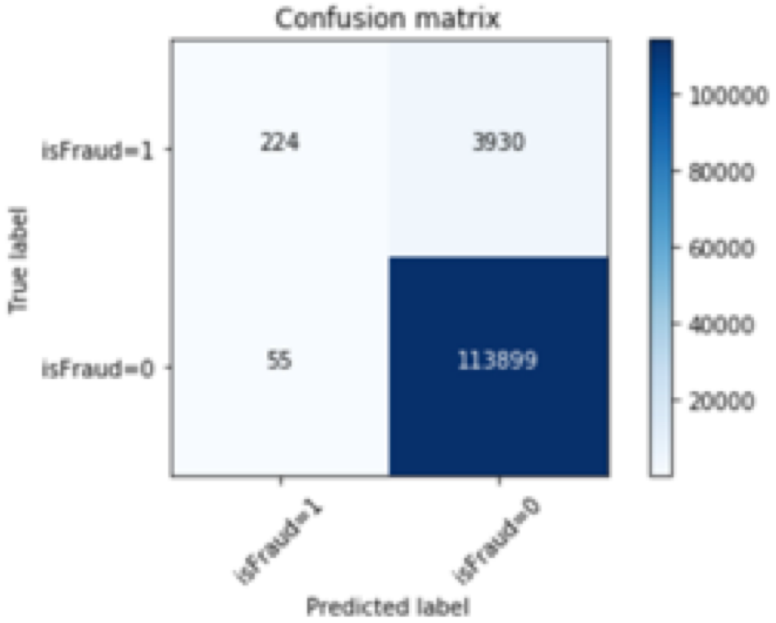
Data Sources

sample_submission.c...	507k x 2
test_identity.csv	142k x 41
test_transaction.csv	507k x 393
train_identity.csv	144k x 41
train_transaction.csv	591k x 394

Data acquisition and cleaning

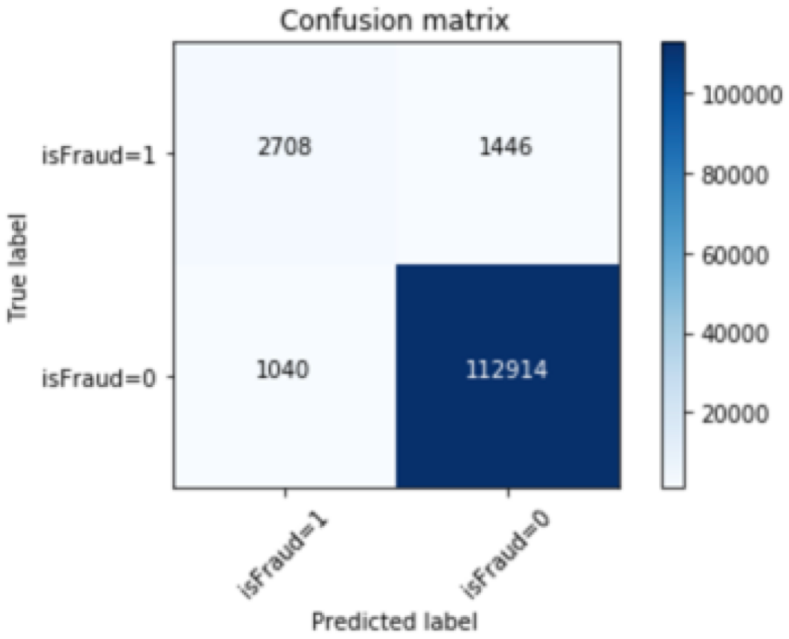
- Merge the data to get train
- Drop V1~V339 and id_01~id_38 columns
- Handle missing values
- Handle categorical values
- Select Features

```
X = train[['TransactionDT', 'TransactionAmt', 'card1', 'card2', 'card3', 'card5', 'C1', 'C2', 'C3', 'C4', 'C5', 'C6', 'C7', 'C8',
       'C12', 'C13', 'C14', 'card4_american express', 'card4_discover', 'card4_mastercard',
       'card4_visa', 'card4_nan', 'card6_charge card', 'card6_credit',
       'card6_debit', 'card6_debit or credit', 'card6_nan', 'ProductCD_C',
       'ProductCD_H', 'ProductCD_R', 'ProductCD_S', 'ProductCD_W']]
y = train['isFraud']
```



Logistic Regression

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- Like all regression analyses, the logistic regression is a predictive analysis.



LightGBM

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.

Conclusion and future directions

LightGBM achieved better performance

Select better parameters for LightGBM

Feature engineering