# Features

We identify the following features and apply them to our classifier

**selected features:**

payment\_account\_prefix\_same\_as\_address\_prefix

**ip related features:**

'ips\_per\_bidder\_per\_auction\_median',

'ips\_per\_bidder\_per\_auction\_mean', 'ip\_only\_one\_user\_counts',

'on\_ip\_that\_has\_a\_bot', 'on\_ip\_that\_has\_a\_bot\_mean',

'ip\_entropy', 'dt\_change\_ip\_median', 'dt\_same\_ip\_median', ‘num\_first\_bid’,

**bids related features:**

'bids\_per\_auction\_per\_ip\_entropy\_median', 'bids\_per\_auction\_per\_ip\_entropy\_mean',

'ips\_per\_bidder\_per\_auction\_median', 'ips\_per\_bidder\_per\_auction\_mean',

'bids\_per\_auction\_median', 'bids\_per\_auction\_mean',

'countries\_per\_bidder\_per\_auction\_median', 'countries\_per\_bidder\_per\_auction\_mean',

**url related features:**

'n\_bids', 'n\_bids\_url',

'n\_urls', 'f\_urls', ‘url\_entropy',

**countries related features:**

'countries\_per\_bidder\_per\_auction\_median', 'countries\_per\_bidder\_per\_auction\_mean', 'countries\_per\_bidder\_per\_auction\_max',

**address related features:**

'address\_rare\_address', ‘address\_infrequent\_address',

**payment related features:**

'payment\_account\_rare\_account', 'payment\_account\_infrequent\_account', 'only\_one\_user'

# Data cleaning and processing:

We first read the features.csv file and extract our selected features.

Then we fill up the data with NULL value and convert boolean string “True” and “False” to 0 and 1.

# Classifier

We tried to apply our classifier with different parameters.

| number of estimators (number of classifier = 5) | result |
| --- | --- |
| 10 | 0.89054 |
| 100 | 0.91206 |
| 200 | 0.90655 |
| 500 | 0.90830 |

| number of classifier (number of estimators = 100) | result |
| --- | --- |
| 1 | 0.90591 |
| 2 | 0.90511 |
| 5 | 0.91099 |
| 10 | 0.90473 |

There are two criteria supported by decision tree

| criteria method | result |
| --- | --- |
| gini | 0.89453 |
| entropy | 0.91206 |