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

number of estimators (number of classifier = 5)	result
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