Detecting malicious DNS packets

Team Hatomugi:
Yu Miyauchi

About

Classify types of attack within malicious dns packets

Targets : 5 different attacks

EDA
and
Feature Engineering

Workflow:

Domain Knowledge

↓

Assumption

↓

Test (Plot)

↓

Result + Feature Engineering

1. Count of different attacks for each ip address

Domain knowledge:

Same ip address for same attack

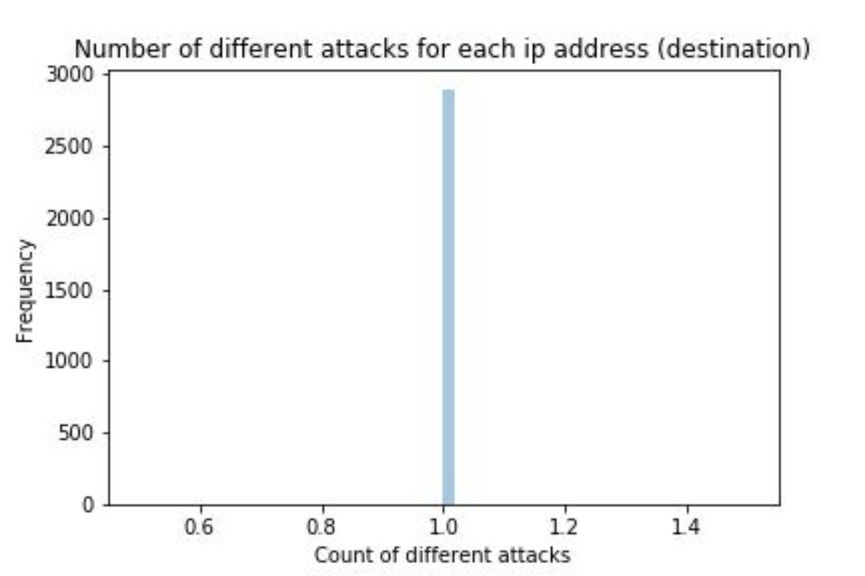
Assumption:

Same ip address = same attacks

Result: True.

If we know either destination or source ip address, we know the type of attack id 100%.

Ip address is strong predictor.



How much ip address do we know in test data?

About 52%.

This means that it is highly likely that we know 52% of answers in the test data.

Question

How do we predict rest 48% of data without knowing ip address?

2. Size of packets

Domain knowledge:

Some attacks such as DoS attack sends many packets to shut down the website. Therefore, they just use small amount of data while other types of attack use relatively larger size of data.

Assumption: Types of attack depends on size of packets transferred.

2. Size of packets

Result:

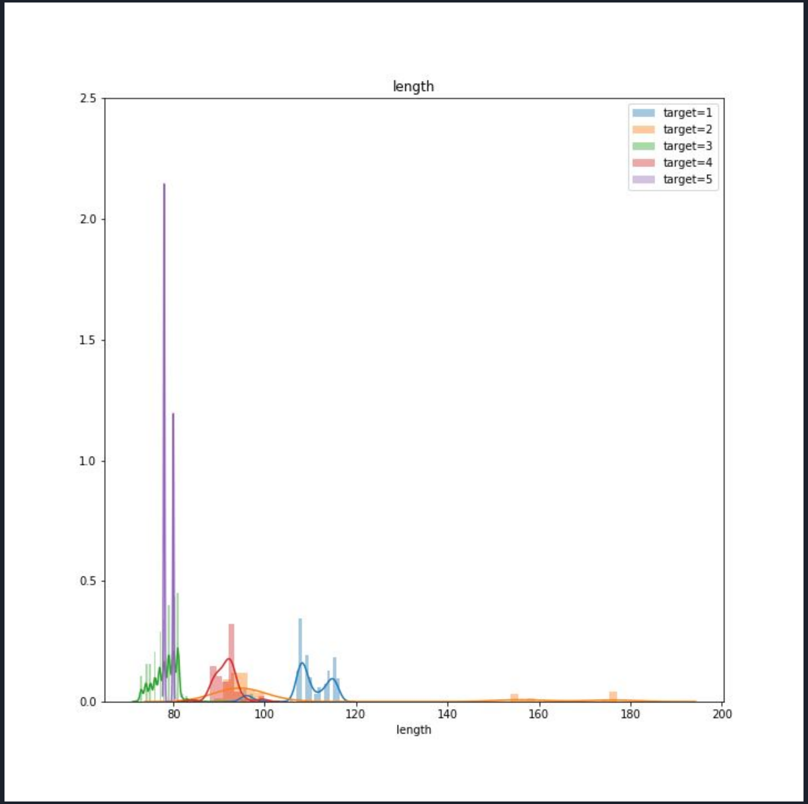
Attack id = 5:

Small and same size + many packets

Attack id = 1:

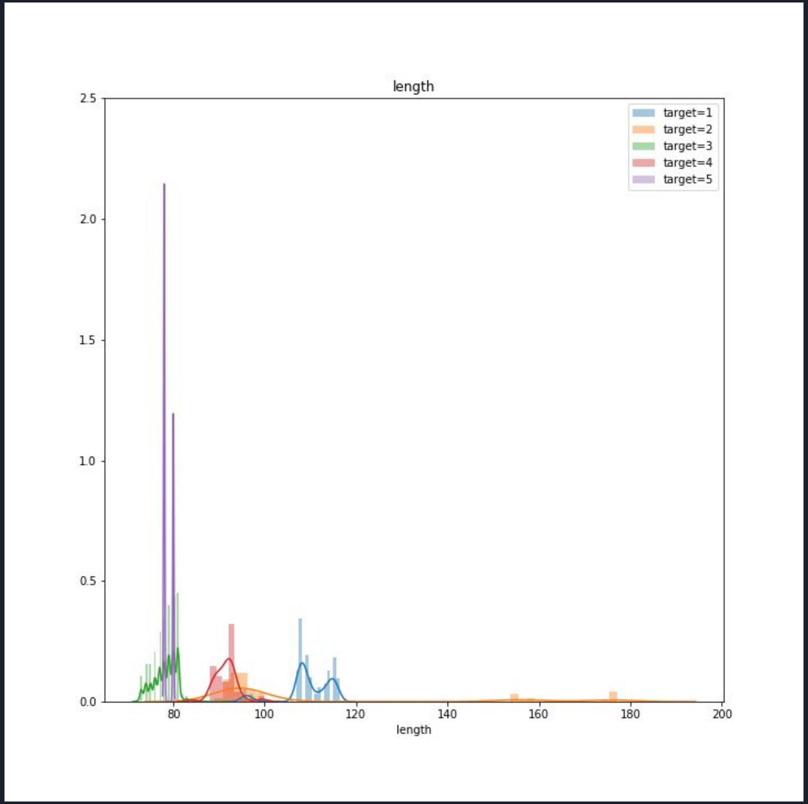
Large and different size + not many packets

Size of packets is strong predictor.



Question

Some packets like attack id = 5 have very high frequency.



2. Frequency of packets transferred by each destination ip address

Assumption:

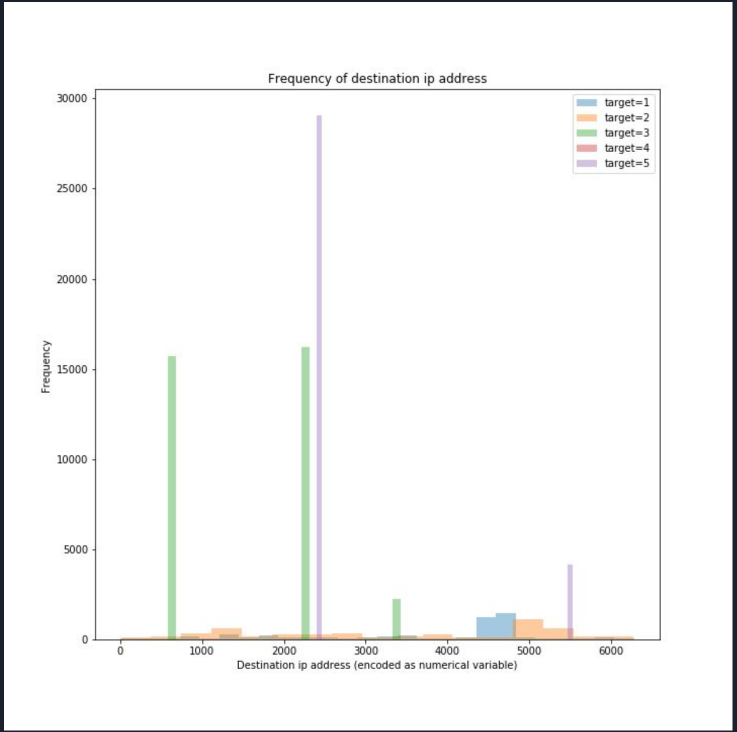
Count of packets are related to

types of attacks (Ex: DoS Attack)

Result: True.

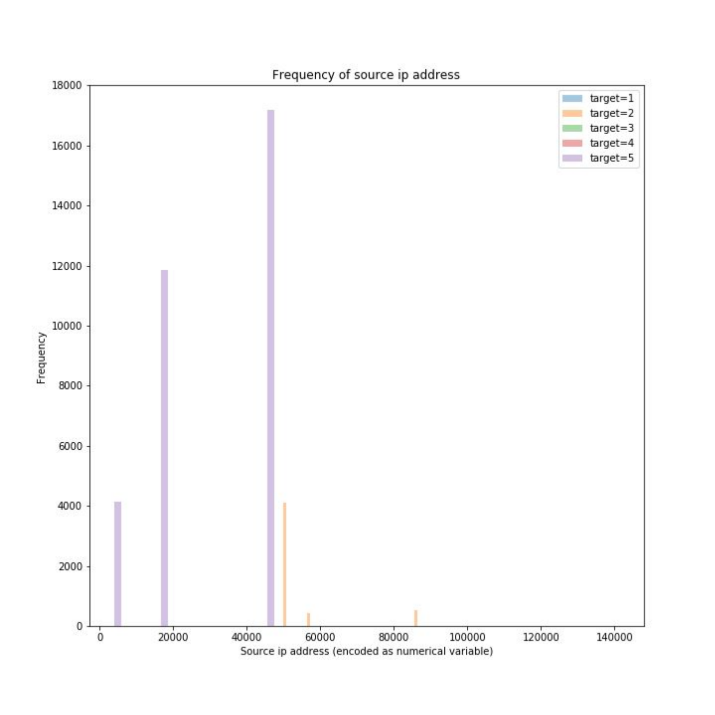
Plot - Frequency of packets for each destination ip address

Count encoding is strong predictor.



2. Frequency of packets transferred by each destination ip address

Frequency of packets for each source ip address



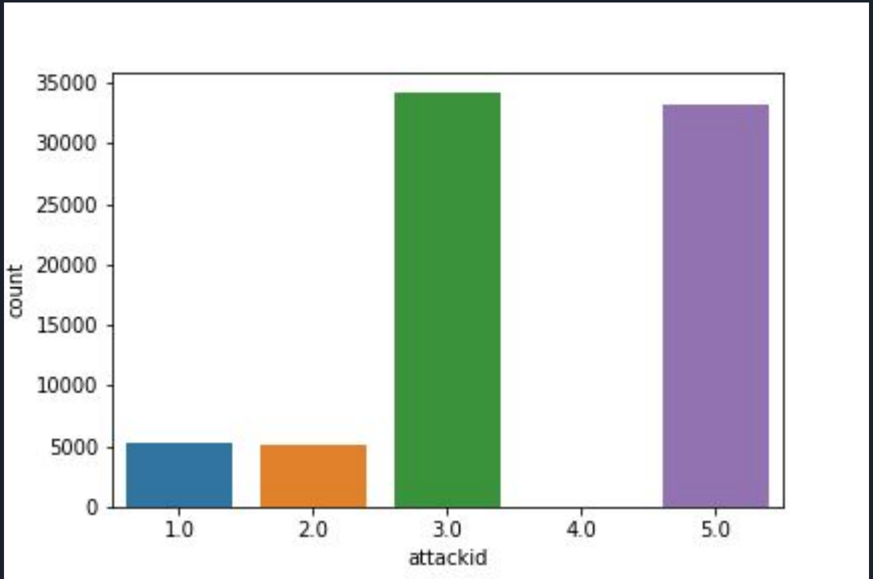
Model

Target Distribution

Target is not uniformly distributed.

↓

Needs to stratify targets



Model

LightGBM with Stratified K-Fold (K=5)

Three reasons

1. Lack of samples for attacid = 4
2. Large amount of data → Needs to process faster
3. GBM models almost always surpass others models in tabular data except deep learning

LightGBM with Stratified K-Fold (K=5)

Validation Accuracy: 0.9999 ≈ 1.0

Validation Log Loss: 0.0061762

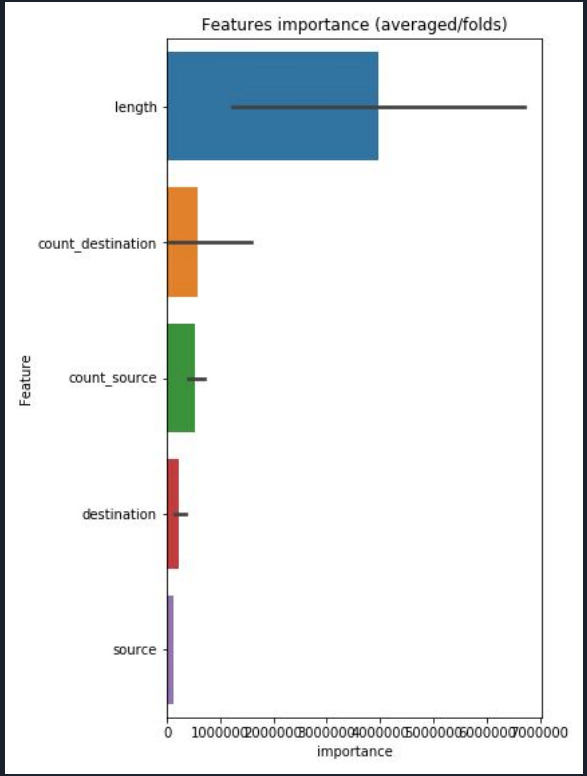
5 simple features

Feature Importance

Points:

Count of destination ip address and source address win over pure ip address.

*ip address is not one hot encoded. If one hot encoded, there will be many features so ip addresses should come down naturally, but this is not the case in my model.



Results

Almost 1.0 accuracy with simple model and features

After Hackathon,

1. Learn more domain knowledge and try to apply to the machine learning.