Malicious Domain Detection from External Data Sources

Bc. Adam Horák

supervisor

Ing. Radek Hranický, Ph.D.

Implementing a Domain Classifier

1. Gathering Data

Name lists from reputable sources for both benign and malign domains are gathered and as much relevant data as possible is gathered for each of them to offer maximum flexibility when extracting distinct features.

UMBRELLA TOP 1M

Regularly top-ranking **benign domains** from Cisco Umbrella.

MISP FEEDS

Malicious domains from MISP threat intelligence.

fit.vut.cz
google.com
ftl.netflix.com

eth-wallet.appyour.uk
leadingedgecash.com
first.nova.cz

MongoDB

A AAAA CNAME MX NS SOA T… **DNS RECORDS**

Registrar Exp. Date Owner DNSSE… **RDAP DATA**

Certificates Validity Authorities E… **TLS CHAIN**

Coordinates ASN Country Code **GEOLOCATION**

NERD Score Events Blacklists Com… **REPUTATION**

2. Feature Extraction & Engineering

From Mongo documents to a table of features using a custom automated data transformation pipeline. To find out what combination of things gives away each of the malicious domain category, various features are extracted from the rich data we collected.

PHISHING
MALWARE
DGA
CRYPTO
TRACKERS
…

FEATURE IMPORTANCE

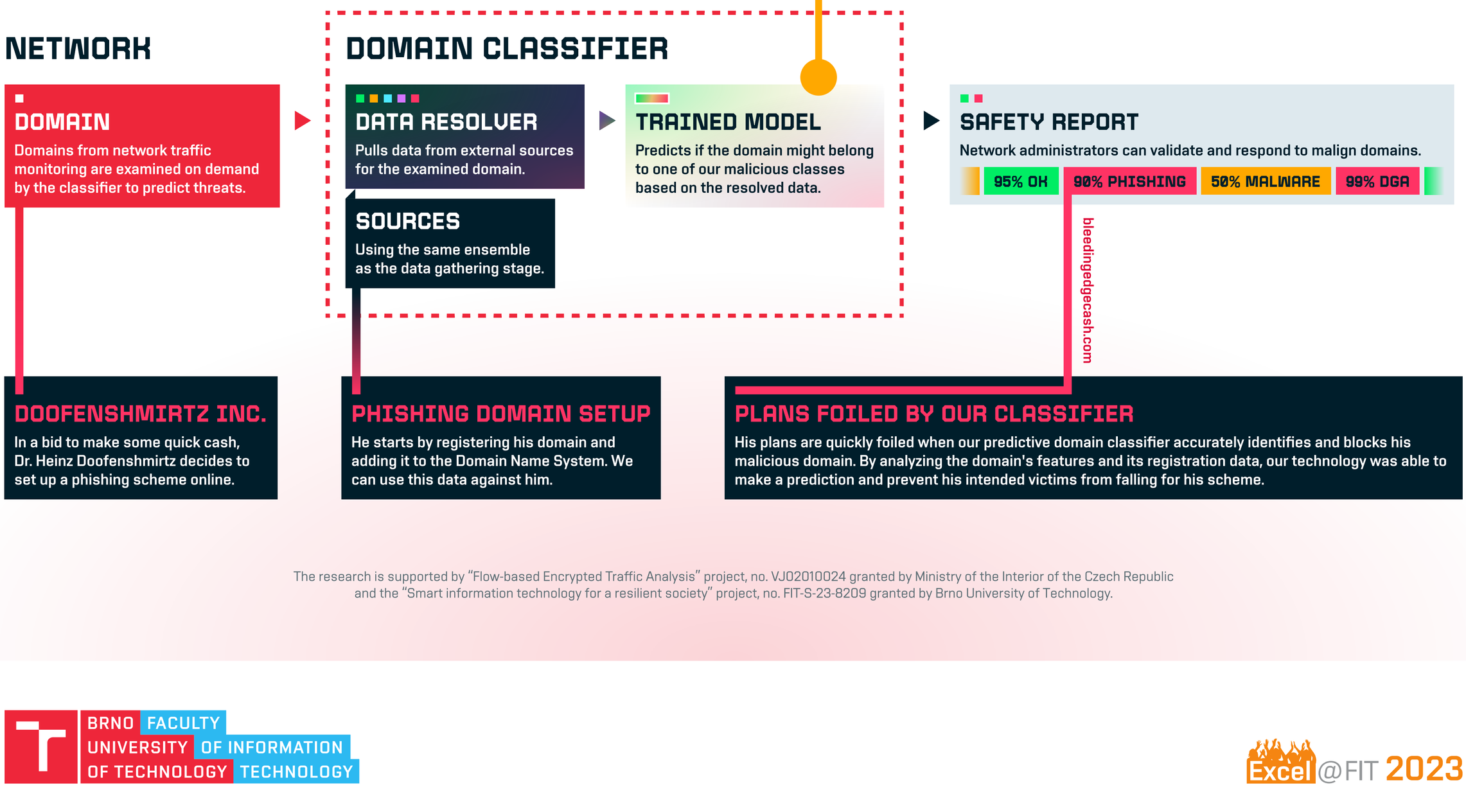
- time before leaf certificate expires
- time before root certificate expires
- domain registration date
- subject alternative name extension name count
- DNS TXT record count
- time before domain expires

3. Model Training & Evaluation

The extracted features are then evaluated for importance by training a gradient boosting model to get their gain scores. Iterating this process helps to identify the most informative features for accurately categorizing malicious domains.

DATASET ▶ XGBOOST ▶ IMPORTANCE

Deploying to Protect a Network



The research is supported by “Flow-based Encrypted Traffic Analysis” project, no. VJ02010024 granted by Ministry of the Interior of the Czech Republic and the “Smart information technology for a resilient society” project, no. FIT-S-23-8209 granted by Brno University of Technology.

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023