MALWARE DOMAIN DETECTION USING MACHINE LEARNING

Tomáš Ebert

Supervisor
Ing. Radek Hranický, Ph.D.

1. GET EXTERNAL DATA

mongoDB

TLS **RDAP** **DNS**

Geolocation **Reputation**

BENIGN DOMAINS

Cisco Umbrella

MALWARE DOMAINS

ThreatFox, MISP, URLHaus, Rescure, GitHubs, blacklists

RESULTS (F1)

Logistic Regression **0.917571**
SVM **0.958890**
Decision Tree **0.962631**
Random Forest **0.952157**
AdaBoost **0.948923**
LightGBM **0.980454**
K-Nearest Neighbors **0.915303**
CatBoost **0.973000**

2. TRAINING MODELS

Data → **XGBoost**

Precision **0.991440**
Recall **0.985669**
F1 0.980511
False Positive Rate **0.00385**

3. FEATURE IMPORTANCE

score for most-abused Top-level domain

A and AAAA records ratio to all

number of IPv4 addresses in RDAP

entropy of autonomous systems IP prefixes

hash of the Top-level domain

LET'S DETECT MALWARE

DOMAIN DETECTION EXAMPLE

yytoutube.com → **XGBoost model** → **MALWARE**

External resources

Probabilities

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024