Network Traffic Botnet Classification

Kate Stadelman

Agenda

- Introduction
- Data
- Random Forest
- References

science. engineering. analytics. insights.

DATA

Research Question

Introduction

Network Traffic Anomaly Detection

Can we detect Botnet activity among normal network traffic?

Botnet Definition

Introduction

“Botnet” by Tom-b – Own work.
Licensed under CC BY-SA 3.0 via Wikimedia Commons.

What is a Botnet?

- Network of Hijacked Devices
- Most are Home Computers
- Used for Cyber-Crime
  - Distributed Denial-of-Service (DDoS)
  - Stealing Personal Data
  - Sending Spam
  - Bitcoin Mining

Motivation

Introduction

Cyber Security: A Primary Concern for US Businesses

“71% of US CEOs said they are ‘extremely concerned’ about cyber threats

-- ahead of pandemics and other health crises (46%)”

— PwC’s 2021 CEO Survey

The average cost of a Distributed Denial-of-Service (DDoS) attack is

$123K for small business and $2.3M for enterprises.

— Kaspersky Lab’s IT Security Risks Survey 2017

science. engineering. analytics. insights.
DATA

Data Set

Data

"The CTU-13 is a dataset of botnet traffic that was captured in the CTU University, Czech Republic, in 2011."[1]

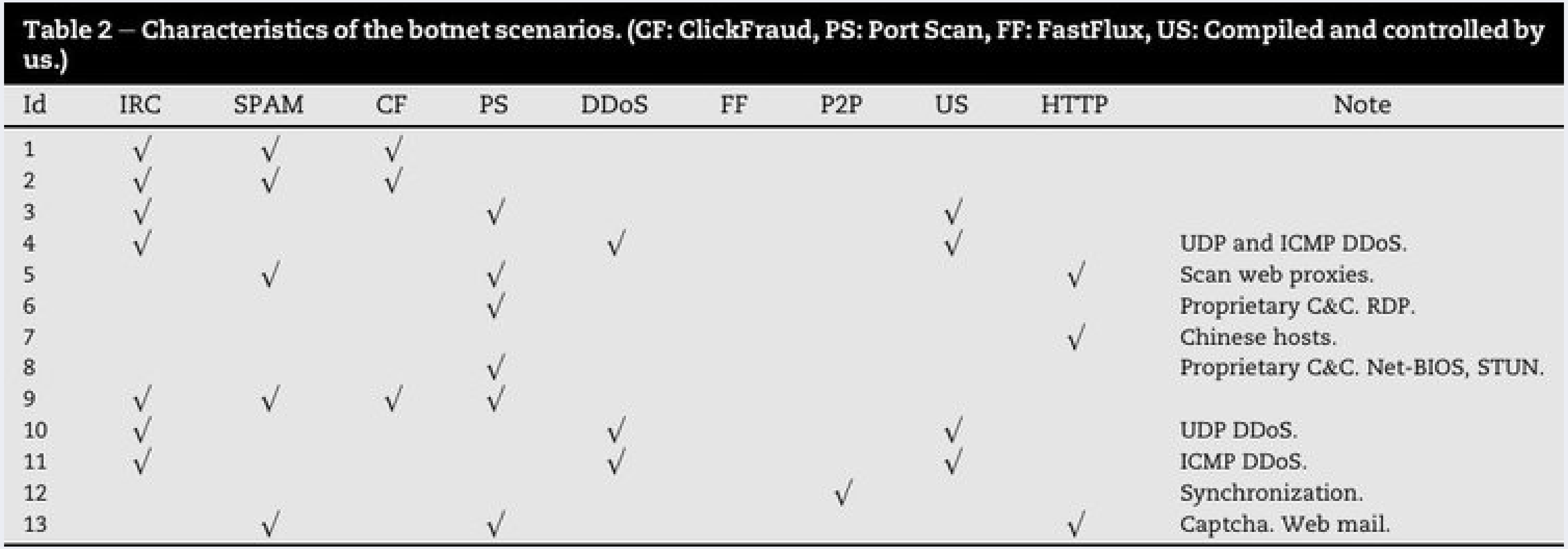


Table 2 – Characteristics of the botnet scenarios. (CF: ClickFraud, PS: Port Scan, FF: FastFlux, US: Compiled and controlled by us.)

Id	IRC	SPAM	CF	PS	DDoS	FF	P2P	US	HTTP	Note
1	√	√	√							
2	√	√	√							
3	√			√				√		
4	√				√			√		UDP and ICMP DDoS.
5		√		√					√	Scan web proxies.
6				√						Proprietary C&C. RDP.
7									√	Chinese hosts.
8				√						Proprietary C&C. Net-BIOS, STUN.
9	√	√	√	√						
10	√				√			√		UDP DDoS.
11	√				√			√		ICMP DDoS.
12							√			Synchronization.
13		√		√					√	Captcha. Web mail.

This research project utilizes CTU-13 #10.

Data Description & Preparation

Data

Data Fields	Mail Delivery Analogy
Traffic Label	Mail Item Name
Start Time	Time First Package Sent
Last Time	Time Last Package Sent
IP Address*	Office Building Address
Protocol*	FedEx/UPS/USPS
Port*	Mailroom Slot
Duration*	Total Transit Time
Packets*	Number of Packages
Bytes*	Combined Size of Packages
Rate*	Package Speed in Transit

*Both Source & Destination

Reduced Sample Period

- 5 Hours to 1 Hours
- 1.3M to 208K Records
- Maintained ~8% Total Botnet Flows

Removed Sparse Fields

- Hops & Time-to-Live

Randomly Split Sample

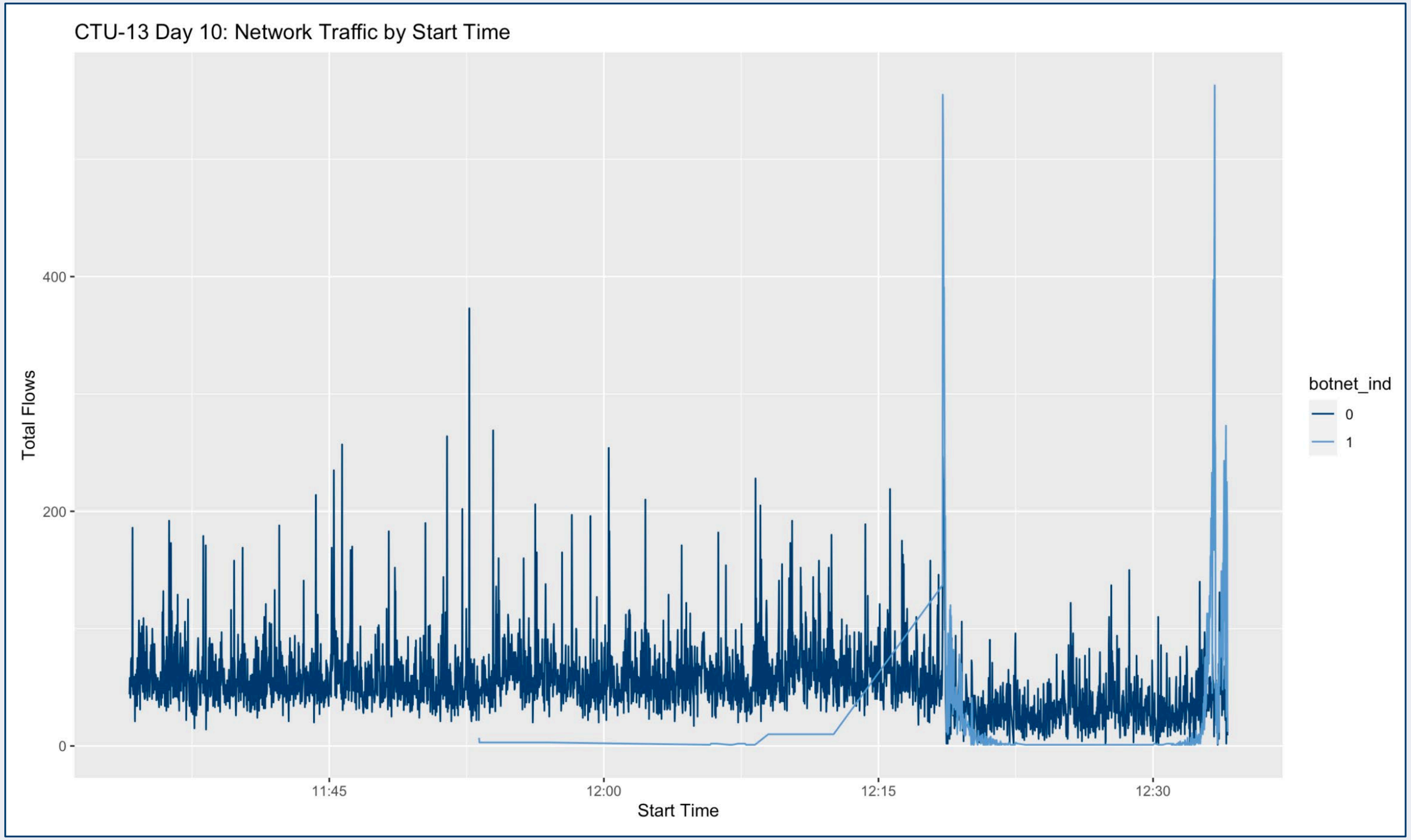
- 75% Training & 25% Validation

Selected Features

- Highlighted in Blue

Data Visualization

Data

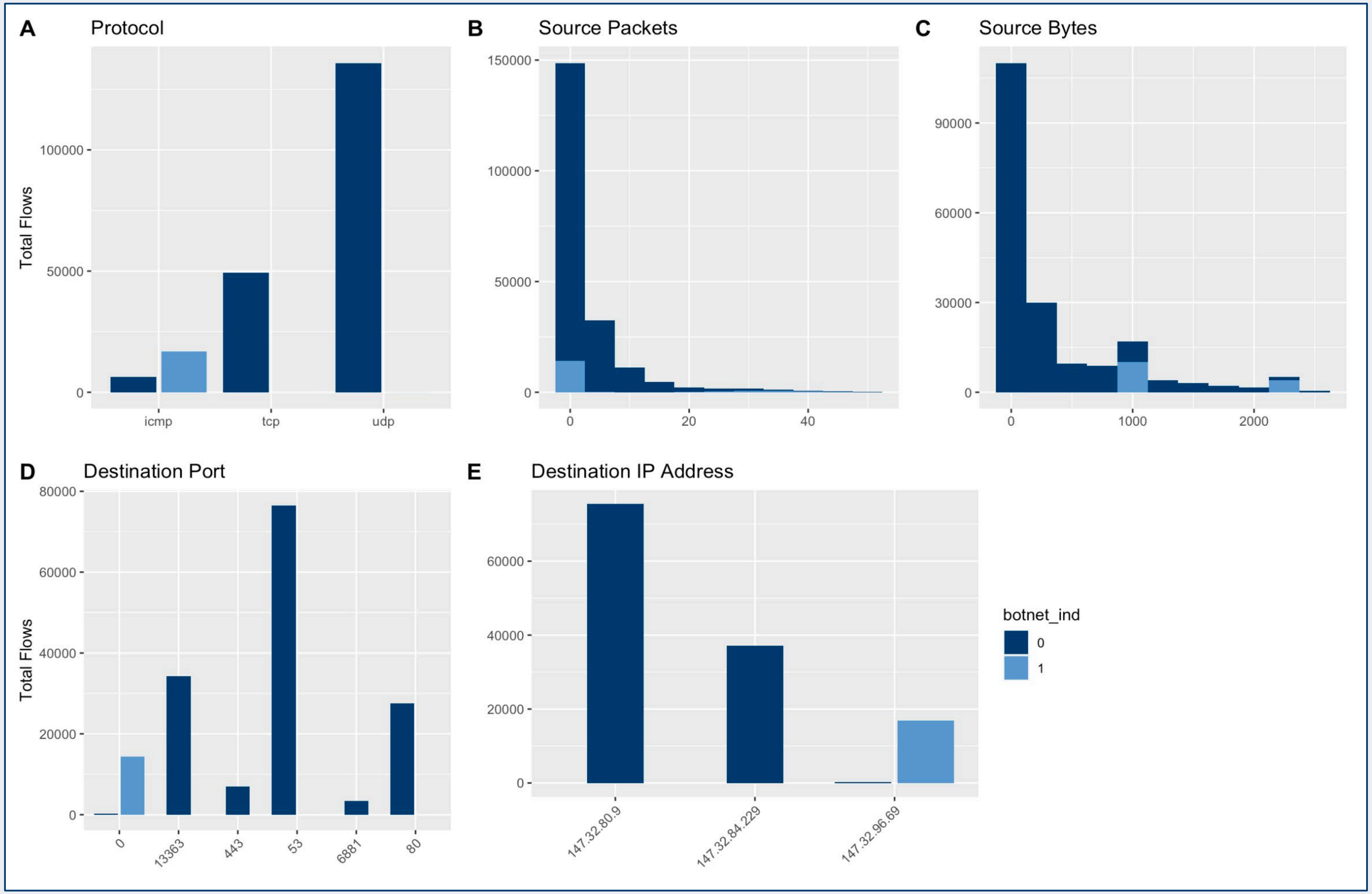


Total Flows

Type	Flows	%
Normal	208,360	92.5%
Botnet	16,810	7.5%

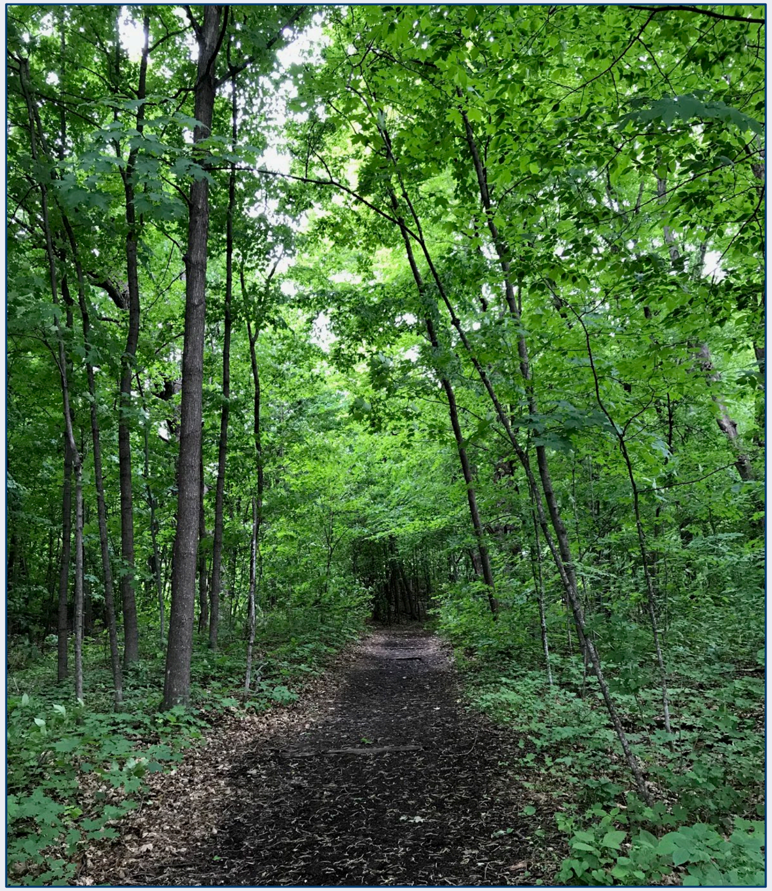
Data Visualization

Data



Model Overview & Description

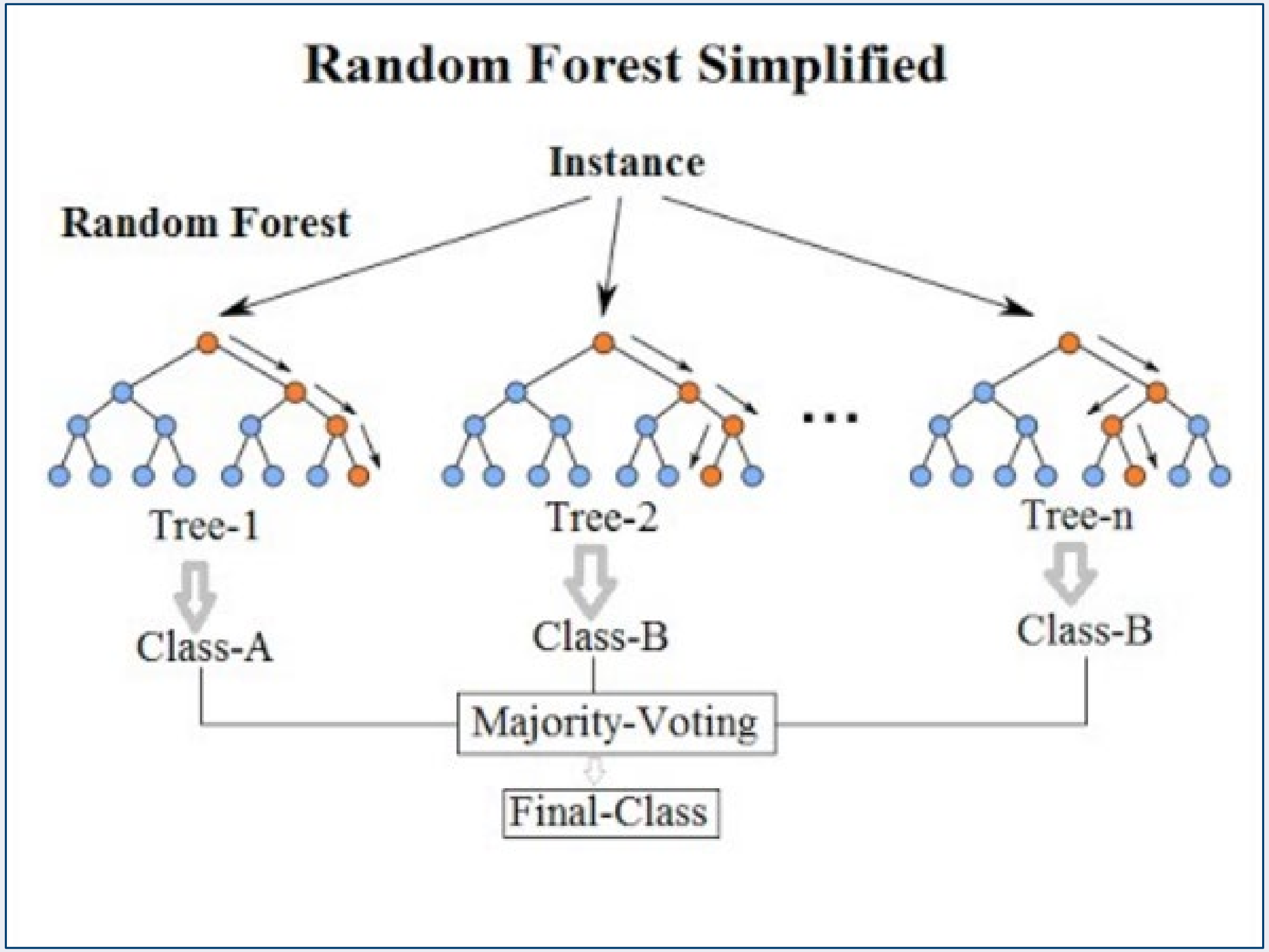
Random Forest



- First Proposed by Tim Kam Ho in 1995
- Ensemble Machine Learning Algorithm
- Performs Classification & Regression Tasks
- Handles Large Data Sets with High Dimensionality
- Bagging (Bootstrapping Aggregation)

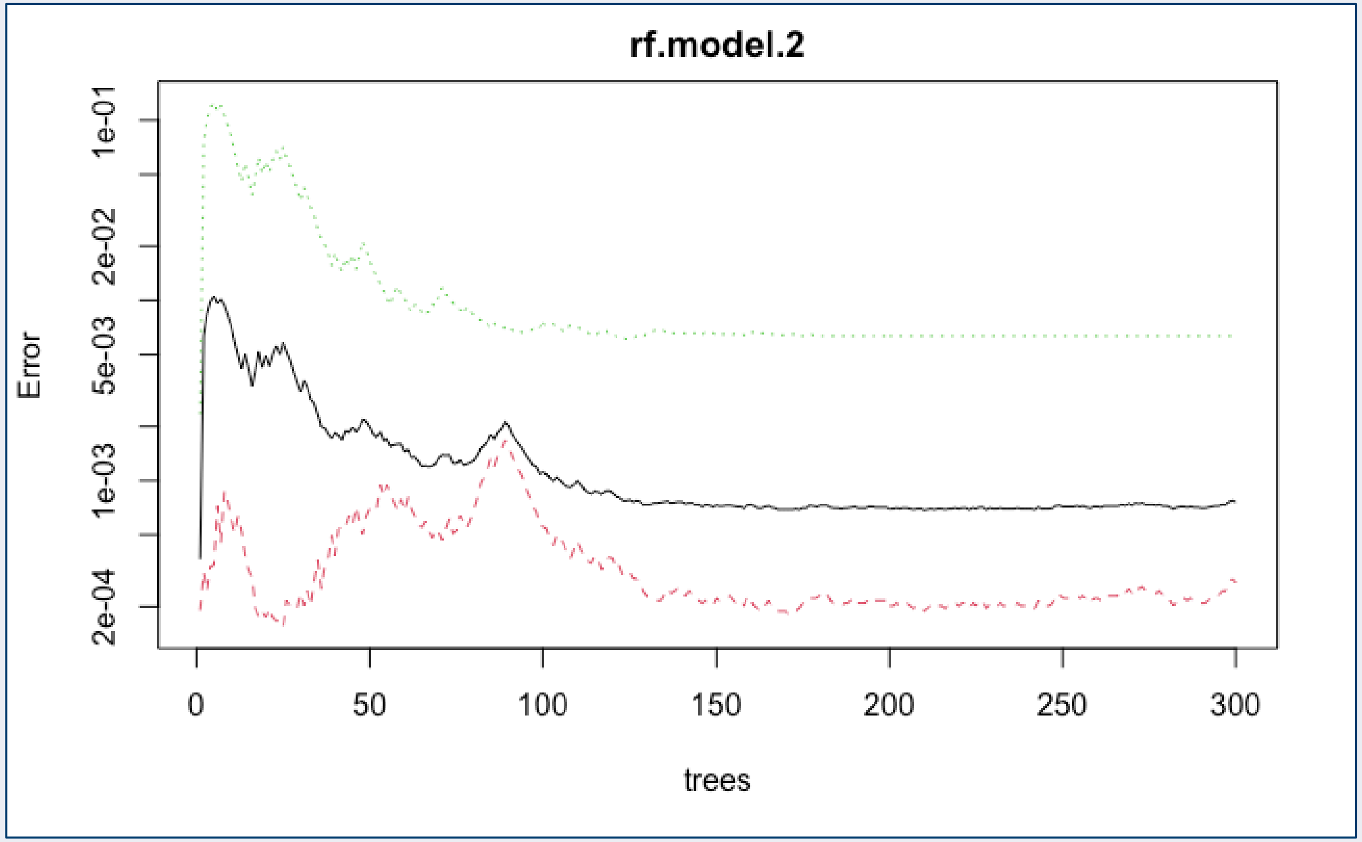
Model Overview & Description

Random Forest



Model Training

Random Forest



Confusion Matrix

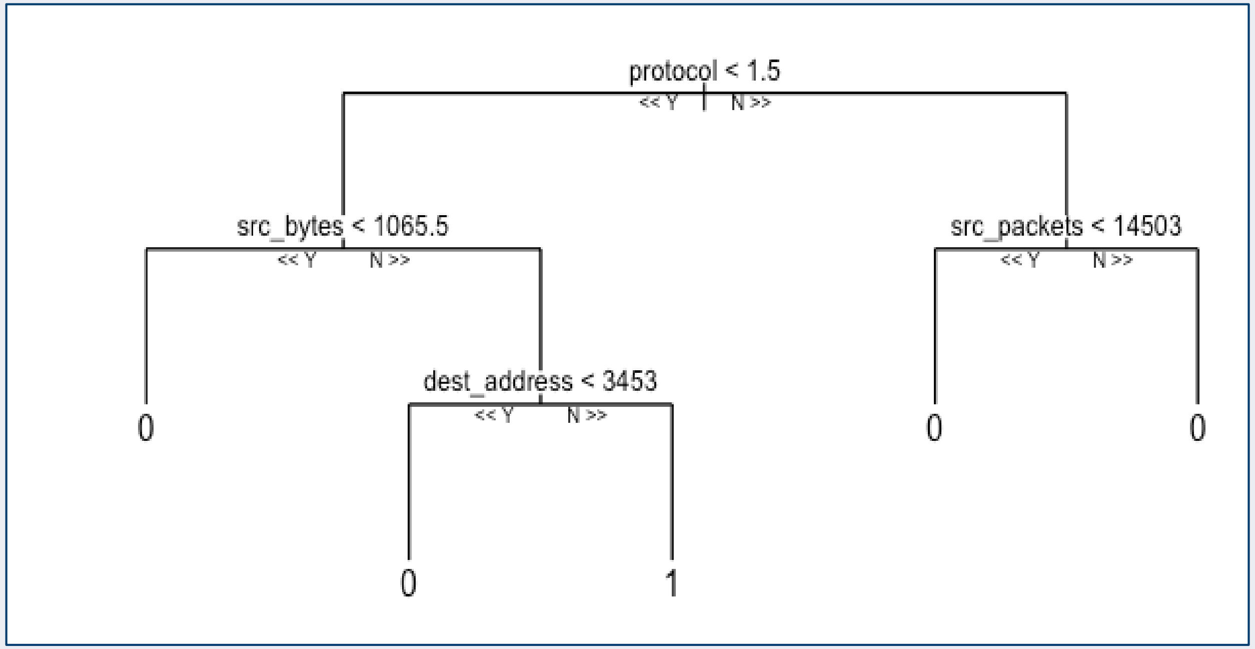
	R. Normal	R. Botnet	Error
P. Normal	143,623	39	2.71e-4
P. Botnet	80	12,528	6.34e-3

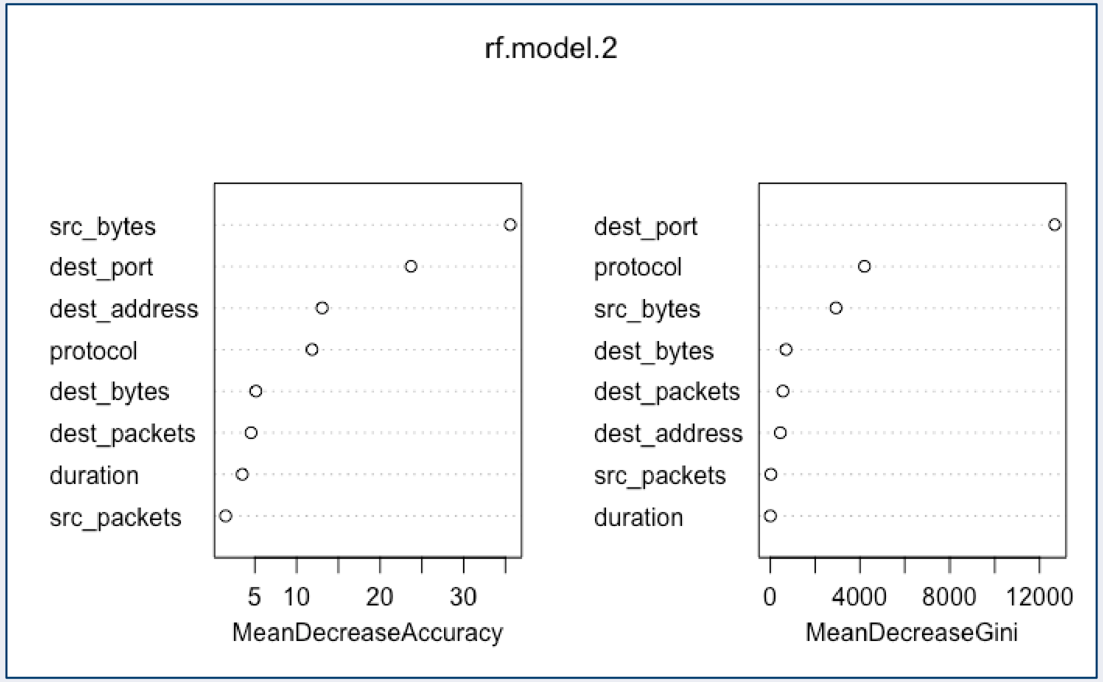
Parameters

- Variables @ Each Split: 5
- Trees: 300
- Max Nodes: 5

Model Diagram & Important Features

Random Forest

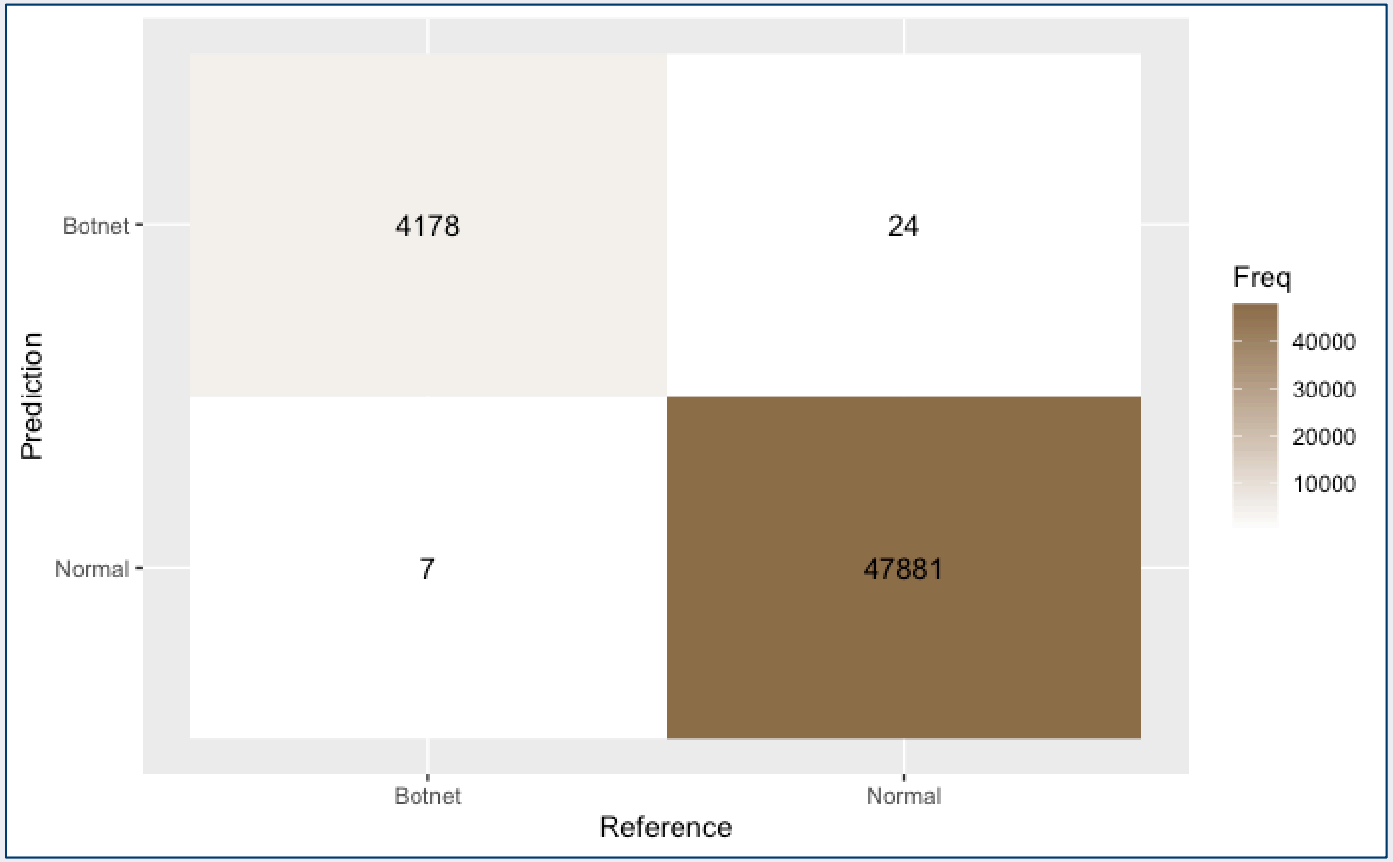


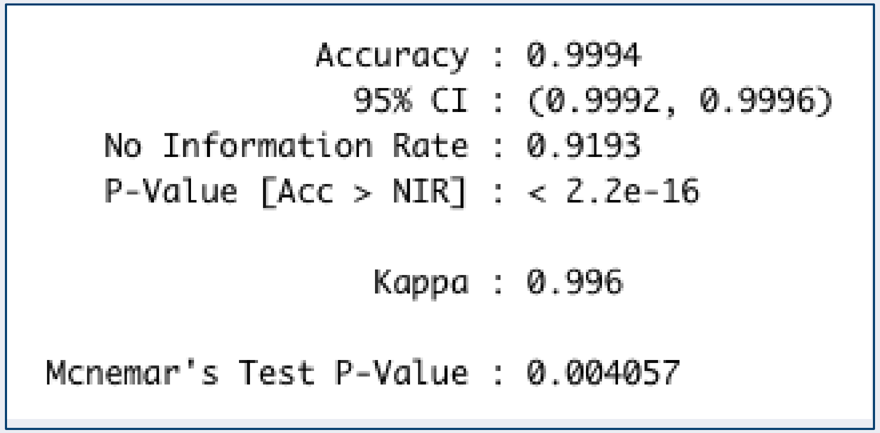


Model Validation

Random Forest

Confusion Matrix





```
                  Accuracy : 0.9994
                    95% CI : (0.9992, 0.9996)
       No Information Rate : 0.9193
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.996

    Mcnemar's Test P-Value : 0.004057
```

Discussion

Q & A

References

Farnaaz, N. & Jabbar, M. A. (2016). Random Forest Modeling for Network Intrusion Detection System. Procedia Computer Science 89, (213–217). https://www.sciencedirect.com/science/article/pii/S1877050916311127

Garcia, S., Grill, M., Stiborek, J., & Zunino, A. (2014). An empirical comparison of botnet detection methods. Computers and Security Journal, Elsevier, 45, (100–123). http://dx.doi.org/10.1016/j.cose.2014.05.011

Marchette, D. (1999, April, 9–12). A Statistical Method for Profiling Network Traffic. Proceedings of the Workshop on Intrusion Detection and Network Monitoring, Santa Clara, California. https://www.usenix.org/conference/id-99/statistical-method-profiling-network-traffic