Applying Machine Learning to Network Security Monitoring

Alex Pinto
Chief Data Scientist | MLSec Project
@alexcpsec
@MLSecProject
• Almost 15 years in Information Security, done a little bit of everything.
• Most of them leading security consultancy and monitoring teams in Brazil, London and the US.
  – If there is any way a SIEM can hurt you, it did to me.

• Researching machine learning and data science in general for the 2 years or so and presenting about its intersection with Infosec for more then an year now.

• Created MLSec Project in July 2013
Agenda

- Definitions
- Network Security Monitoring
- PoC || GTFO
- Feature Intuition
- MLSec Project
Big Data + Machine Learning + Data Science
Big Data + Machine Learning + Data Science

big data is
- like teenage
- the new oil
- watching you
- dead
- on the rise
- opening doors but maybe too many
- a big deal

machine learning is
- the future
- a branch of which scientific discipline
- hard
- not as cool as it sounds
- just statistics

data science is
- statistics on a mac
- the new black
Big Data

Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Oozie Workflow
Pig Scripting
Mahout Machine Learning
R Connectors Statistics
Hive SQL Query
HBase Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System
(Security) Data Scientist

“Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.”

-- Josh Willis, Cloudera

Data Science Venn Diagram by Drew Conway
Kinds of Machine Learning

- "Machine learning systems automatically learn programs from data" – CACM 55(10) Domingos 2012

- Supervised Learning:
  - Classification (NN, SVM, Naïve Bayes)
  - Regression (linear, logistic)

- Unsupervised Learning:
  - Clustering (k-means)
  - Decomposition (PCA, SVD)

Source – scikit-learn.github.io/scikit-learn-tutorial/general_concepts.html
Classification Example

VS

THQUIRREL!
Considerations on Data Gathering

- Models will (generally) get better with more data
  - Always have to consider bias and variance as we select our data points
  - Am I selecting the correct features to describe the entities?
  - Have I got a representable sample of labeled data I can use?

- “I’ve got 99 problems, but data ain’t one”

---

Figure 1: Bias and variance in dart-throwing.

Domingos, 2012

Abu-Mostafa, Caltech, 2012
Security Applications of ML

- Fraud detection systems (not security):
  - Is what he just did consistent with past behavior?

- Network anomaly detection:
  - Good luck finding baselines
  - ML is a bit more than rolling averages

- User behavior anomaly detection:
  - My personal favorite, 2 new companies/day
  - Does fraud detection follow the CLT?

- SPAM filters
Considerations on Data Gathering (2)

- Adversaries - Exploiting the learning process
- Understand the model, understand the machine, and you can circumvent it
- Any predictive model on InfoSec will be pushed to the limit
- Again, think back on the way SPAM engines evolved

Posit: “Intrinsic features of malicious actors cannot be masked as easily as behavioral features"
Network Security Monitoring
Kinds of Network Security Monitoring

• Alert-based:
  – “Traditional” log management
  – SIEM
  – Using “Threat Intelligence” (i.e. blacklists) for about a year or so
  – Lack of context
  – Low effectiveness
  – You get the results handed over to you

• Exploration-based:
  – Network Forensics tools (2/3 years ago)
  – ELK stacks
  – High effectiveness
  – Lots of people necessary
  – Lots of HIGHLY trained people
  – Much more promising

• Big Data Security Analytics (BDSA):
  – Basically exploration-based monitoring on Hadoop and friends
  – Sounds kind of painful for the analysts involved
Alert-based + Exploration-based
Using robots to catch bad guys
• We developed a set of algorithms to detect malicious behavior from log entries of firewall blocks

• Over 6 months of data from SANS DShield (thanks, guys!)

• After a lot of statistical-based math (true positive ratio, true negative ratio, odds likelihood), it could pinpoint actors that would be 13x-18x more likely to attack you.

• Today reducing amount of clutter in log files to less then 0.5% of actors worth investigating, and having less than 20% false positives in participant deployments.
Feature Intuition: IP Proximity

- Assumptions to aggregate the data
- Correlation / proximity / similarity BY BEHAVIOR
- “Bad Neighborhoods” concept:
  - Spamhaus x CyberBunker
  - Google Report (June 2013)
  - Moura 2013

- Group by Geolocation
- Group by Netblock (/16, /24)
- Group by BGP prefix
- Group by ASN information
Map of the Internet

- (Hilbert Curve)
- Block port 22
- 2013-07-20
Feature Intuition: Temporal Decay

- Even bad neighborhoods renovate:
  - Attackers may change ISPs/proxies
  - Botnets may be shut down / relocate
  - A little paranoia is Ok, but not EVERYONE is out to get you (at least not all at once)

- As days pass, let’s forget, bit by bit, who attacked
- Last time I saw this actor, and how often did I see them

GO HOME CLOCK
YOU'RE DRUNK
Feature Intuition: DNS features

- Who resolves to this IP address – pDNS data + WHOIS
- Number of domains that resolve to the IP address
- Distribution of their lifetime
- Entropy, size, ccTLDs
- Registrar information
- Reverse DNS information
- History of DNS registration
- (Thanks, Farsight Security!)
Training the Model

- YAY! We have a bunch of numbers per IP address/domain!
- How do you define what is malicious or not?
  - Curated indicator feeds
  - OSINT indicator feeds – with some help from statistical-based curating
  - Top X lists of visited sites.
  - Feedback from security tools (if you trust them)
MLSec Project

- Working with several companies on tuning these models on their environment with their data
- Looking for participants and data sharing agreements

- Visit https://www.mlsecproject.org, message @MLSecProject or just e-mail me.
MLSec Project - Current Research

• Inbound attacks on exposed services (BlackHat 2013):
  – Information from inbound connections on firewalls, IPS, WAFs
  – Feature extraction and supervised learning

• Malware Distribution and Botnets (hopefully BlackHat 2014):
  – Information from outbound connections on firewalls, DNS and Web Proxy
  – Initial labeling provided by intelligence feeds and AV/anti-malware
  – Some semi-supervised learning involved

• User Impersonation in Web Applications (early days):
  – Inputs: logs describing authentication attempts (both failed and successful), click stream data
  – Segmentation of users by risk level
Thanks!

- Q&A at the end of the webinar

Alex Pinto
@alexcpssec
@MLSecProject
https://www.mlsecproject.org/

"Essentially, all models are wrong, but some are useful."
- George E. P. Box