Network Monitoring & Anomaly Detection using Machine Learning

By: Akshay Shembekar

1

Why is there a need to monitor Networks and have anomaly detection?

The “Why”

Bookish

- Find which users or applications may be causing a network slowdown
- Identify your network’s top talkers and determine the best interventions to minimize their impact on the network as a whole
- Determine if you’re distributing bandwidth effectively across your network
- Get insights into the actual state of your network, including accessing data on bandwidth usage by type of traffic, bandwidth usage by application, usage patterns over time, performance statistics, and end-user experience

More vivid Happenings

IN

Nov 2008

FR

Jan 2015

Podcast

2

What do I wish to accomplish in this study?

The “What”

Wireshark

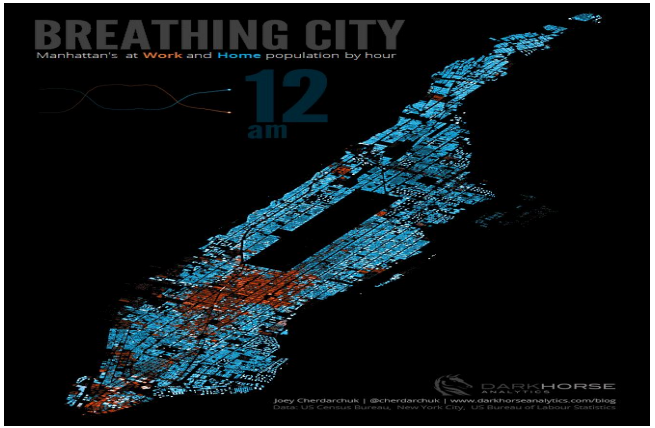
Play around more with wireshark as a tool.

Traffic Monitoring

Get a better sense of traffic flowing over my home network

Pattern ⇨

Traffic over a time series



Datadog - SIEM

Room for expansion to delve into newer Cloud Monitoring Tools

Apply Learnings

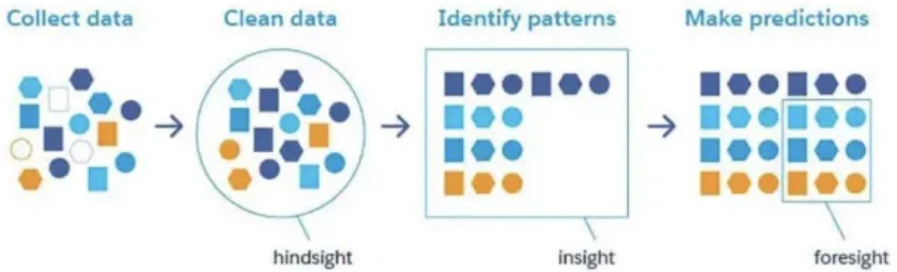
Intersect Cybersecurity + Cloud + AI/ML

Open Source

Make project public and fork-able

ML Ops ⇨

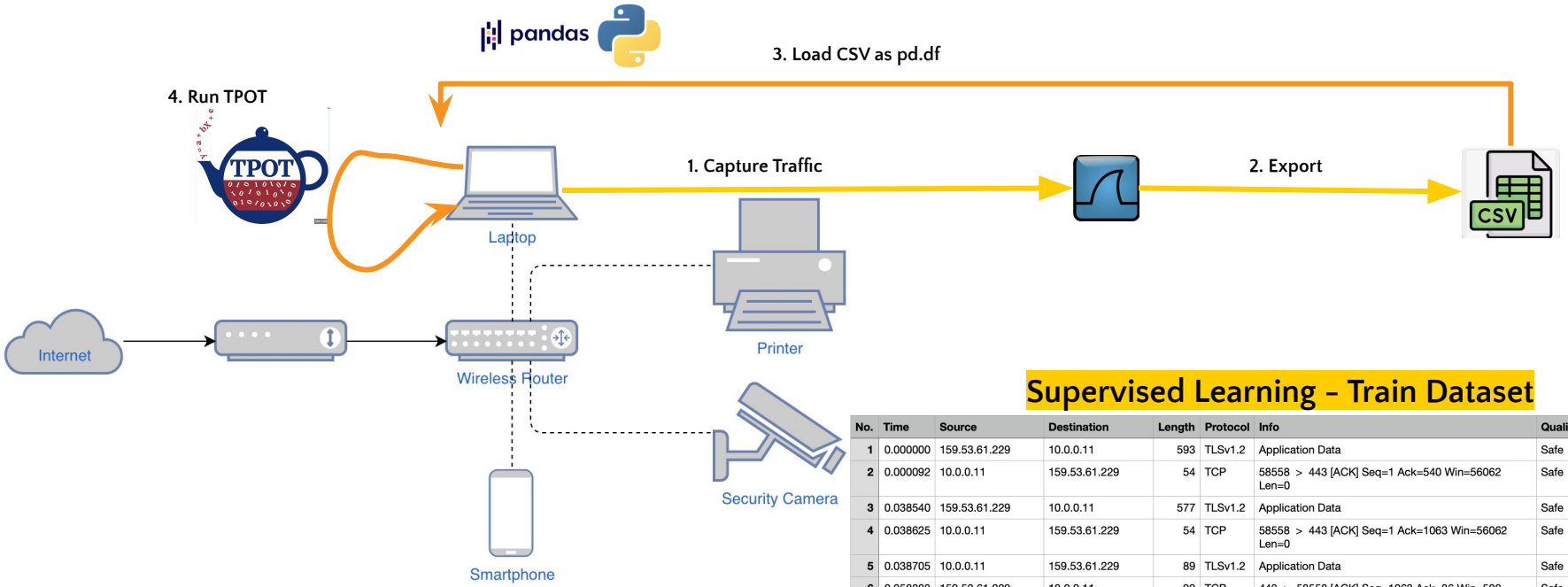
Get hands dirty with ML Ops



3

How do I to accomplish my goal?

The “How”



Supervised Learning - Train Dataset

No.	Time	Source	Destination	Length	Protocol	Info	Qualifier
1	0.000000	159.53.61.229	10.0.0.11	593	TLSv1.2	Application Data	Safe
2	0.000092	10.0.0.11	159.53.61.229	54	TCP	58558 > 443 [ACK] Seq=1 Ack=540 Win=56062 Len=0	Safe
3	0.038540	159.53.61.229	10.0.0.11	577	TLSv1.2	Application Data	Safe
4	0.038625	10.0.0.11	159.53.61.229	54	TCP	58558 > 443 [ACK] Seq=1 Ack=1063 Win=56062 Len=0	Safe
5	0.038705	10.0.0.11	159.53.61.229	89	TLSv1.2	Application Data	Safe
6	0.058803	159.53.61.229	10.0.0.11	92	TCP	443 > 58558 [ACK] Seq=1063 Ack=36 Win=500 Len=0	Safe
7	0.065277	159.53.61.229	xx.yy.zz.aaa	593	TLSv1.2	Application Data	Unsafe
8	0.065398	xx.yy.zz.aaa	159.53.61.229	54	TCP	58558 > 443 [ACK] Seq=36 Ack=1602 Win=56062 Len=0	Unsafe

Up Next...

4

Model Training, Tuning, and Selection

Methods Tested

DECISION TREE

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed.

XGBOOST

Gradient Boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

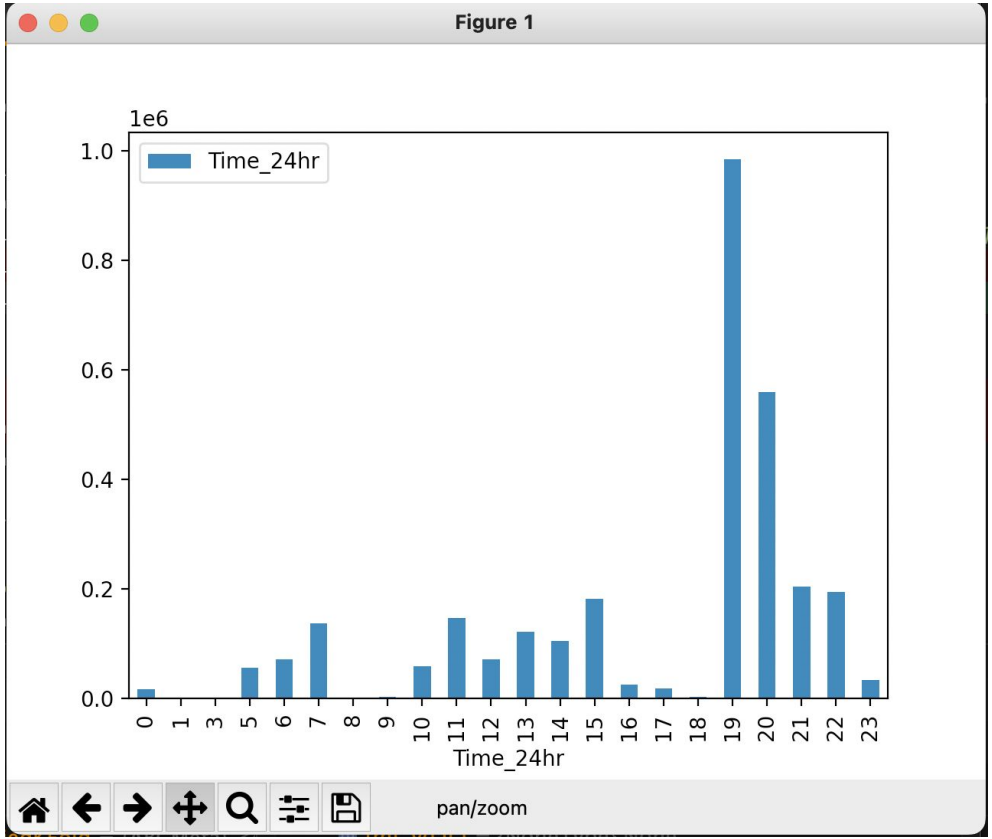
RANDOM FOREST

The PCA does an unsupervised dimensionality reduction, while the logistic regression does the prediction.

The PCA converts data from high dimensional space to low dimensional space by selecting the most important attributes that capture maximum information about the dataset and logistic regression method is used to train the model.

LOGISTIC REGRESSION WITH PCA

24hr - Usage Pattern

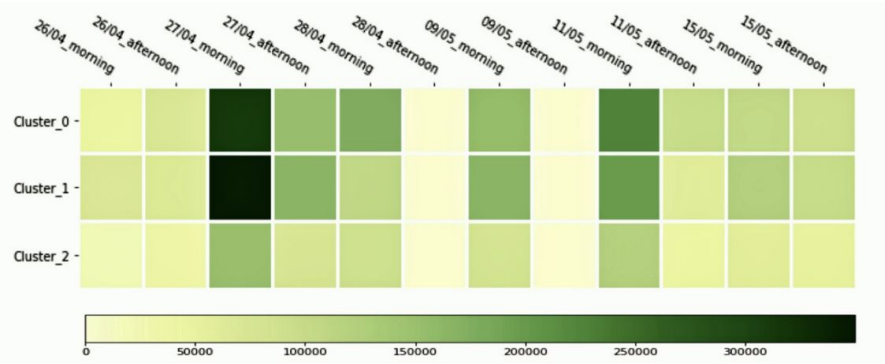



```
"""Plot Histogram"""
df_data['Time_Formatted'] = pd.to_datetime(df_data['Time'])
df_hist = df_data[['Time_Formatted']]   df_hist: Time_24hr [0        2022-1
df_hist.rename(columns={'Time_Formatted': 'Time_24hr'})
df_hist.groupby(df_hist["Time_24hr"].dt.hour).count().plot(kind="bar")
```

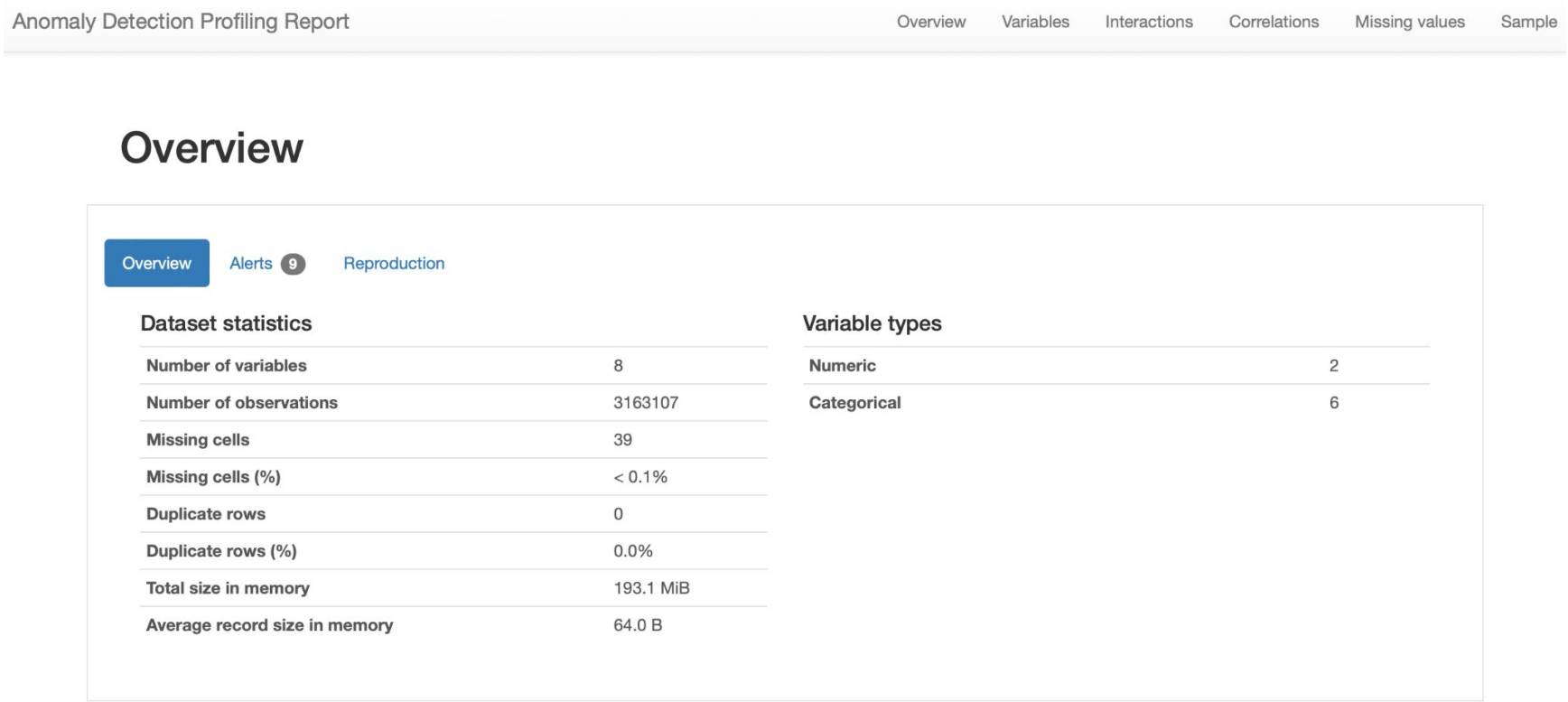
Observations:

- Heavy traffic around 8:30 AM ET, lasts for ~2 hrs.
- Stedy traffic between 10:30 AM ET to 12:30PM ET

Unsupervised Learning (K-Means Clustering)



Exploratory Data Analysis



Exploratory Data Analysis (cont..)

Sample

First rows

	no.	time	source	destination	length	protocol	info
0	1	00:23:49.159985	e673.dsce9.akamaiedge.net	Akshays-MacBook-Pro.local	66	TCP	443 > 61724 [ACK] Seq=1 Ack=1 Win=506 Len=0 TSval=2951765
1	2	00:23:49.159986	e673.dsce9.akamaiedge.net	Akshays-MacBook-Pro.local	353	TLSv1.2	Application Data
2	3	00:23:49.159986	e673.dsce9.akamaiedge.net	Akshays-MacBook-Pro.local	353	TLSv1.2	Application Data
3	4	00:23:49.160179	Akshays-MacBook-Pro.local	e673.dsce9.akamaiedge.net	66	TCP	61724 > 443 [ACK] Seq=2693 Ack=575 Win=2039 Len=0 TSval=3

Last rows

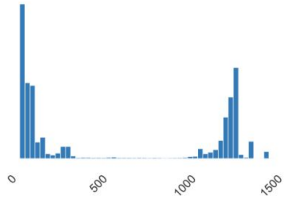
	no.	time	source	destination	length	protocol	info
3163097	15499	00:22:53.838925	Akshays-MacBook-Pro.local	play.google.com	81	UDP	57538 > 443 Len=39
3163098	15500	00:22:53.840538	play.google.com	Akshays-MacBook-Pro.local	220	UDP	443 > 57538 Len=178
3163099	15501	00:22:53.850833	Akshays-MacBook-Pro.local	play.google.com	75	UDP	57538 > 443 Len=33

Exploratory Data Analysis (cont..)

length
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	1419	Minimum	42
Distinct (%)	< 0.1%	Maximum	1514
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	593.4594919	Memory size	24.1 MiB

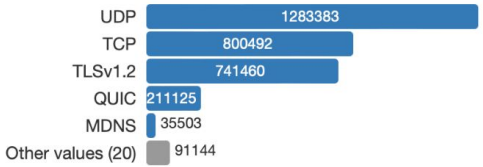


Toggle details

protocol
Categorical

HIGH CORRELATION

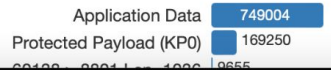
Distinct	25
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	24.1 MiB



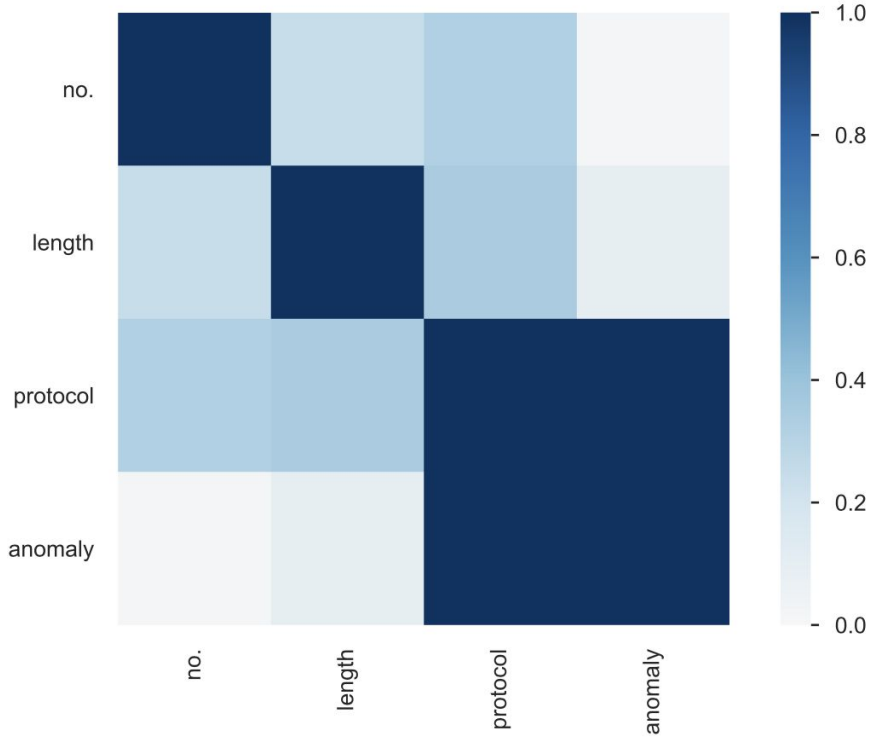
Toggle details

info
Categorical

Distinct	826793
Distinct (%)	26.1%



Exploratory Data Analysis (cont..)



Observations:

- High Correlation between Protocol and anomaly
- Good degree of correlation between length and anomaly

Why?

Supervised Learning

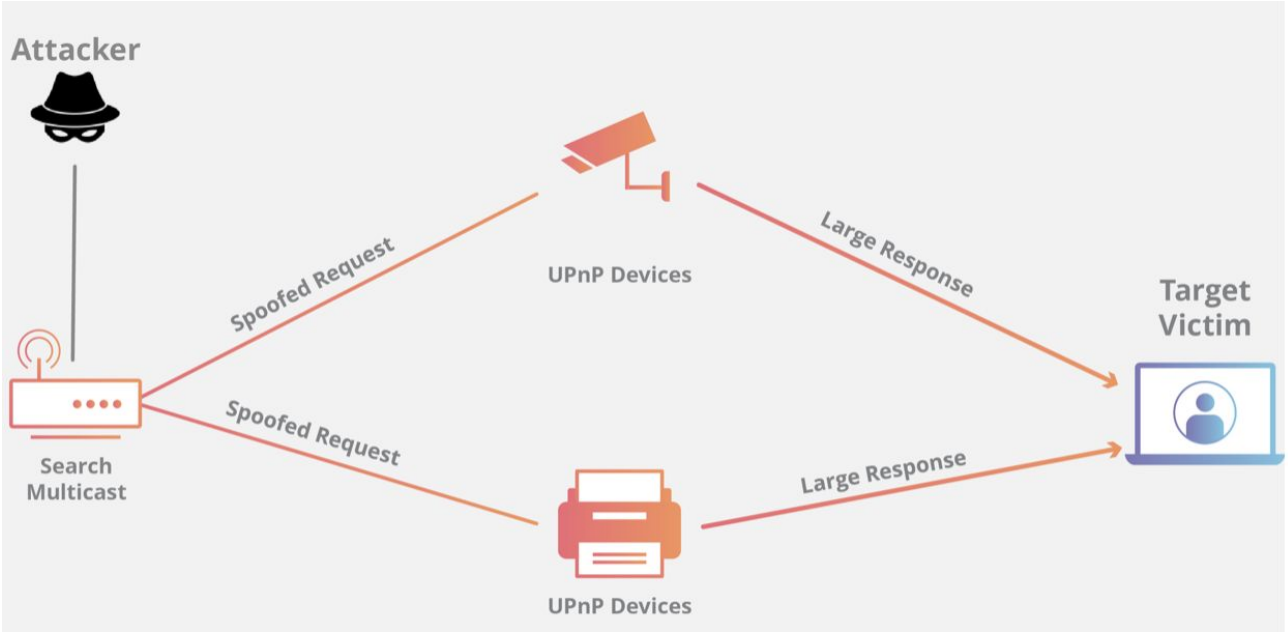
- System is taught/trained by labelled data to classify or predict the outcome

```
"""Apply Classifier Rules"""

df_data['anomaly'] = 0
df_data.loc[df_data['protocol'] == 'NBNS', 'anomaly'] = 1
df_data.loc[df_data['protocol'] == 'SSDP', 'anomaly'] = 1
df_data.loc[df_data['protocol'] == 'TELNET', 'anomaly'] = 1
df_data.loc[(df_data['protocol'] == 'FTP') & (df_data['length'] >= 1000), 'anomaly'] = 1
```

Interesting Protocols Observed

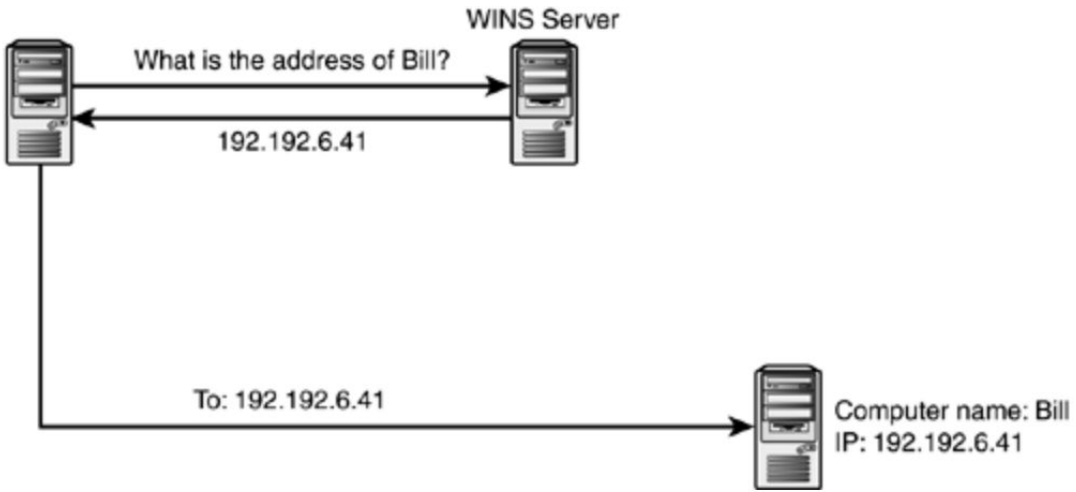
Simple Service Discovery Protocol (SSDP)- Discover available devices (and their capabilities) in a local area network / piconet.



- SSDP attack exploits that final request for services by asking the device to respond to the targeted victim.
- For network administrators, a key mitigation is to block incoming UDP traffic on port 1900 at the firewall.

Interesting Protocols Observed (cont..)

NetBIOS Name Service (NBNS) - NBNS serves the same purpose as DNS does: translate human-readable names to IP addresses (e.g. www.wireshark.org to 65.208.228.223).

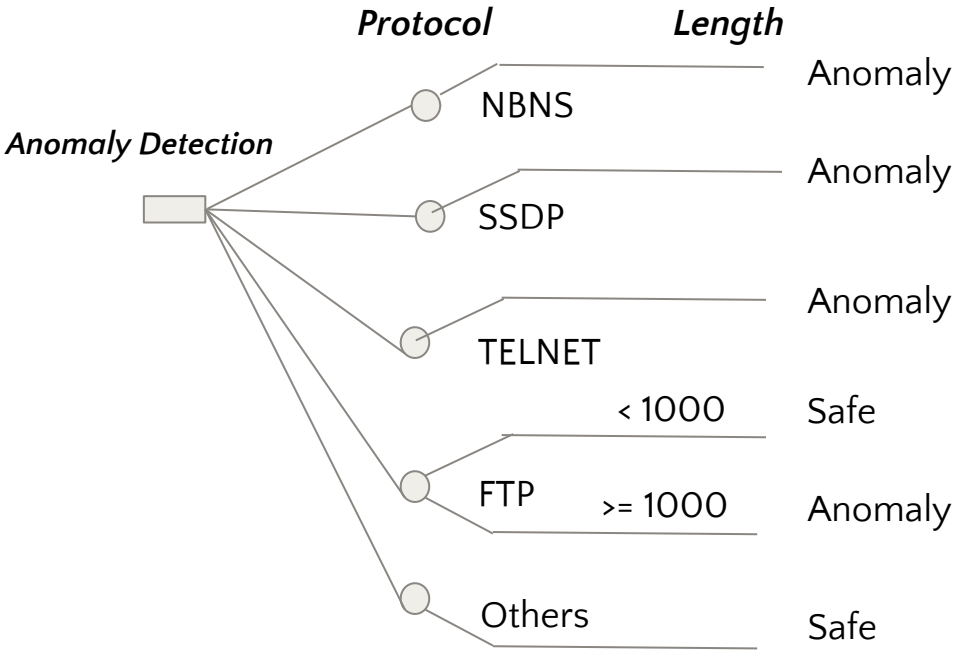


- SMB relies on NetBIOS for communication with devices that do not support direct hosting of SMB over TCP/IP.
- Commonly abused to perform Man-in-the-middle attack.
- NBNS Spoofing can be carried out via metasploit.
- Disabling NBNS support on network & devices is recommended.

Models Explored

Decision Tree

- Probabilistic tree that enables to make a decision



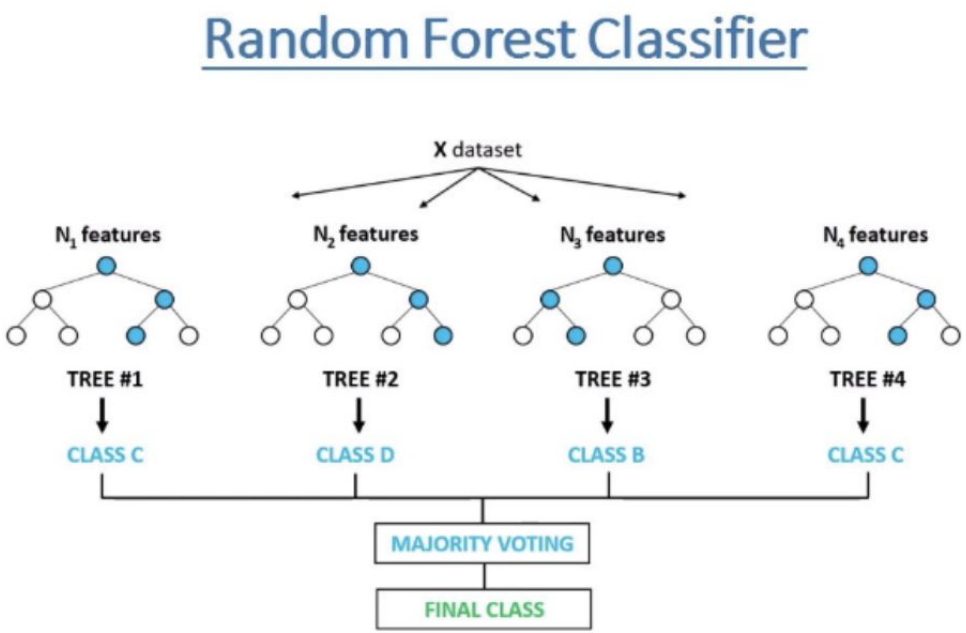

```
"""Apply Classifier Rules"""

df_data['anomaly'] = 0
df_data.loc[df_data['protocol'] == 'NBNS', 'anomaly'] = 1
df_data.loc[df_data['protocol'] == 'SSDP', 'anomaly'] = 1
df_data.loc[df_data['protocol'] == 'TELNET', 'anomaly'] = 1
df_data.loc[(df_data['protocol'] == 'FTP') & (df_data['length'] >= 1000), 'anomaly'] = 1
```

Models Explored (cont..)

Random Forest

- Forest of Decision Trees



- Considering only Protocol as a feature, a Decision Tree (DT1) can be constructed
- Considering only Length as a feature, a Decision Tree (DT2) can be constructed
- Considering Protocol & Length, both, as a feature, a Decision Tree (DT3) can be constructed

Based on consensus voting of DT1, DT2, DT3 – a record will be classified as safe or anomaly. This model is seen to perform better by providing higher precision & accuracy.

References

Git

https://github.com/ax-shay/Network-Monitoring-Anomaly-Detection-using-Machine-Learning

References

1. Network Traffic Analysis using Machine Learning: an unsupervised approach to understand and slice your network
   - Ons Aouedi, Kandaraj Piamrat, Salima Hamma, J K Menuka Perera

   https://hal.archives-ouvertes.fr/hal-03344361/document

2. Decision Tree: Definition and Example

   https://www.statisticshowto.com/decision-tree-definition-and-examples/

3. SSDP DDoS Attack

   https://www.cloudflare.com/learning/ddos/ssdp-ddos-attack/