Impact of Adversarial Attacks Against Random Forest Classification

Katherine Carlile

Machine Learning for Cybersecurity Analytics

Problem Statement

Intrusion Detection Systems that implement machine learning algorithms to identify malicious activity, while effective, may themselves vulnerable to adversarial attacks. The goal of this work is to present a statistical analysis of the impact an adversarial attack has on the ability of an IDS to accurately classify malicious network traffic.

- Adversarial Attack Vector: Data Poisoning
  - Altering the training data
  - Real-world example: GMail Spam Misclassification

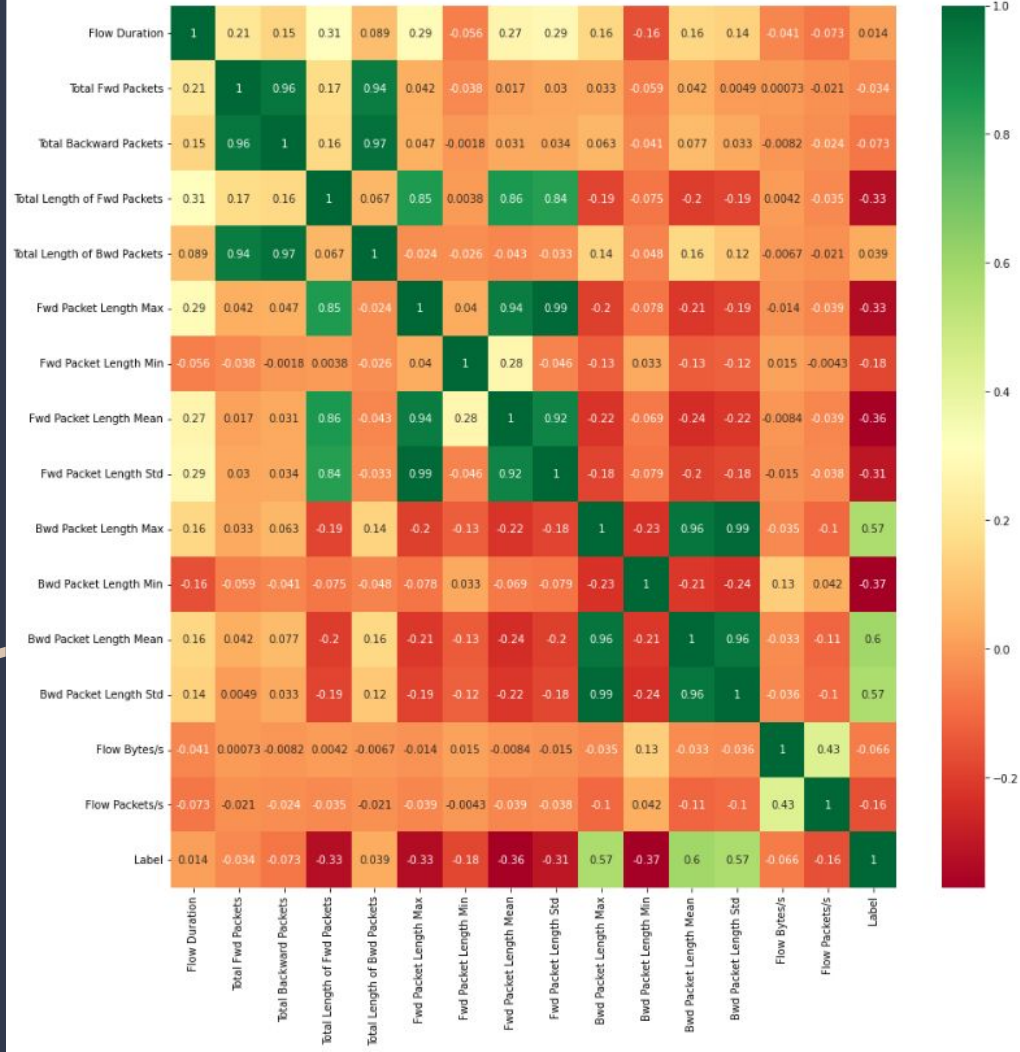
Dataset & Preprocessing

- CIC-IDS-2017
    - Friday-WorkingHours-Afternoon-DDos
- Feature Selection: Removing categorical features and selecting features of interest

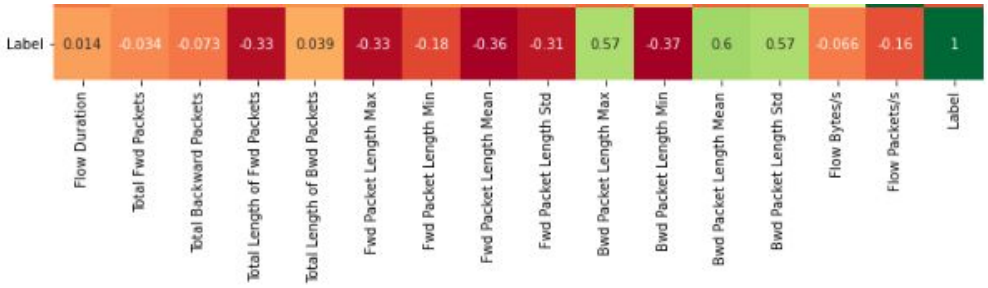
```
"Flow Duration", "Total Fwd Packets", "Total Backward Packets",
"Total Length of Fwd Packets", "Total Length of Bwd Packets",
"Fwd Packet Length Max", "Fwd Packet Length Min",'Fwd Packet Length Mean',
'Fwd Packet Length Std','Bwd Packet Length Max', 'Bwd Packet Length Min',
'Bwd Packet Length Mean', 'Bwd Packet Length Std', 'Flow Bytes/s',
'Flow Packets/s', "Label"
```

- Applied Ordinal Encoding to Label, 0.0 as Benign and 1.0 as DDoS
- Outlier removal beyond the 1st and 99th percentiles for each feature
- Scaled each feature using MinMaxScalar

Feature Selection



Feature Selection, Continued



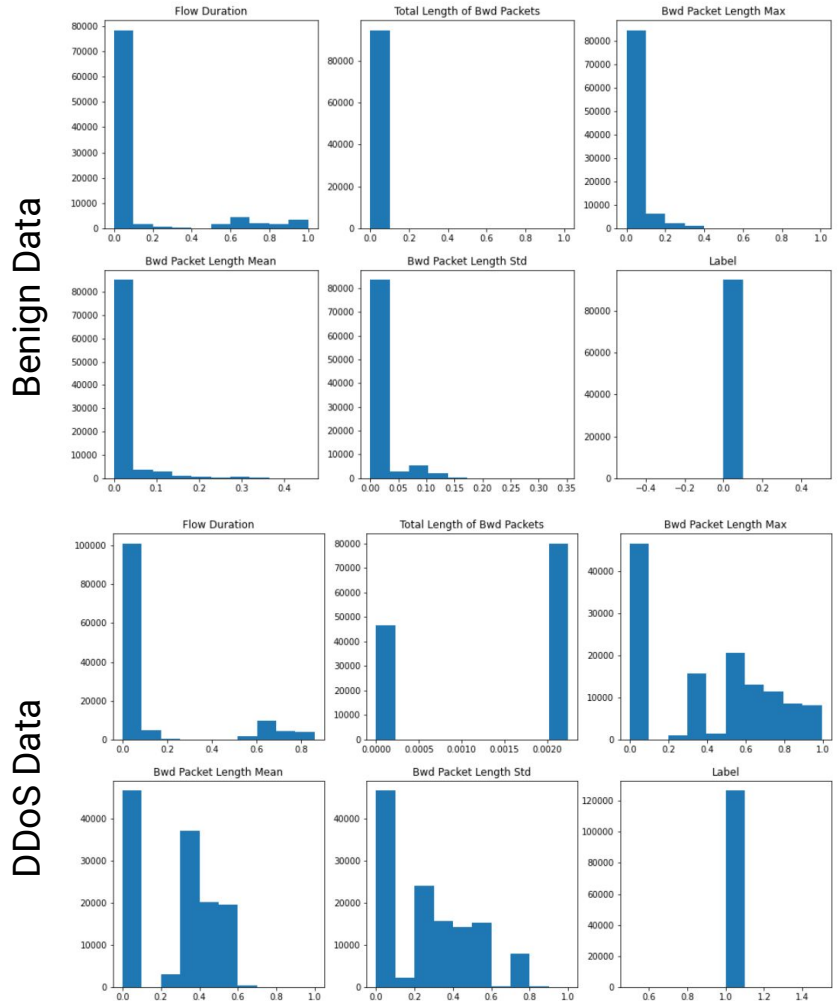
- Only 6 features had a positive correlation with Label
    - Flow Duration, Total Length of Fwd Packets, Bwd Packet Length Max, Pwd Packed Length Mean, and Bwd Packet Length Std

Random Forest Baseline Performance

	precision	recall	f1-score	support
0.0	0.98	0.98	0.98	28476
1.0	0.99	0.99	0.99	37878
accuracy			0.99	66354
macro avg	0.99	0.99	0.99	66354
weighted avg	0.99	0.99	0.99	66354

Visualization

- Flow Duration
  - Benign Range 0 - 80000
  - DDoS range 0 - 100000
- Total Length of Bwd Packets
  - Benign & DDoS Range 0 - 80000
- Bwd Packet Length Max, Mean, and Std
  - Benign mostly 80000
  - DDoS clump ≠ 80000
- According to the Seaborn graph, Bwd Packet Length Max, Bwd Packet Length Mean, and Bwd Packet Length Std

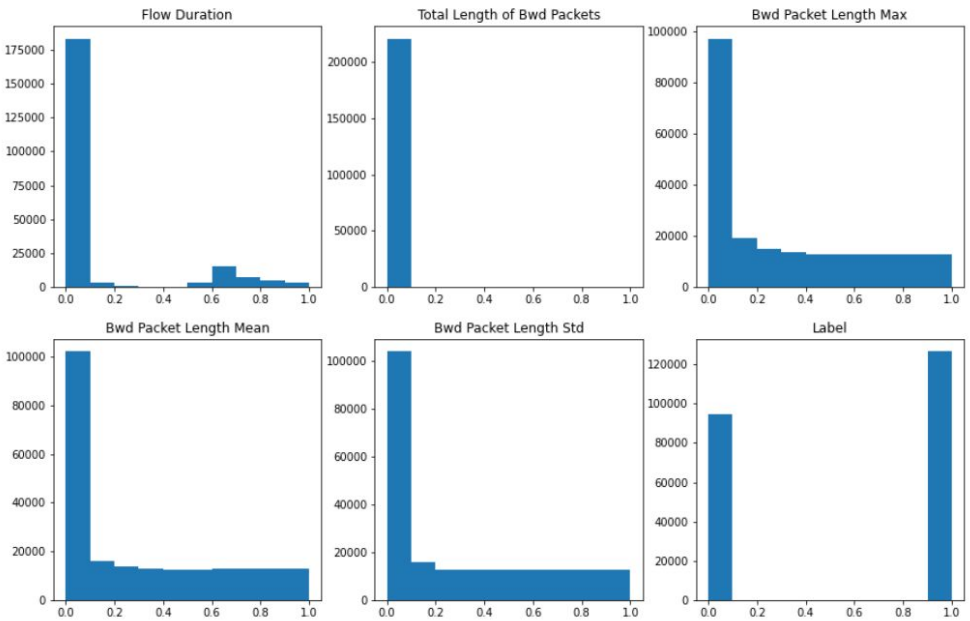


Data Perturbation Strategies

- Randomize the entirety, or a percentage, of the feature values within feature range
- Target feature value randomization to critical points (ex. Data “clumps” seen in DDoS data vs. Benign data)
- Shift feature-range of malicious data into a range associated with benign data

Randomizing Bwd Packet Max, Mean, & Std

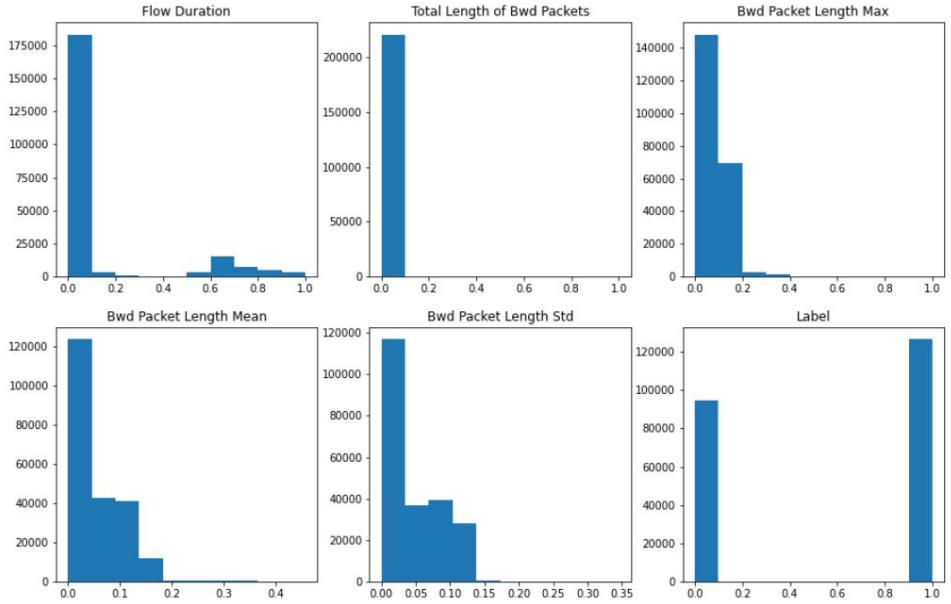
- Indiscriminate randomization within per-feature min-max bounds



	precision	recall	f1-score	support
0.0	0.67	1.00	0.80	28476
1.0	1.00	0.63	0.77	37878
accuracy			0.79	66354
macro avg	0.84	0.82	0.79	66354
weighted avg	0.86	0.79	0.79	66354

Targeted Randomizing of Bwd Packet Max, Mean, & Std

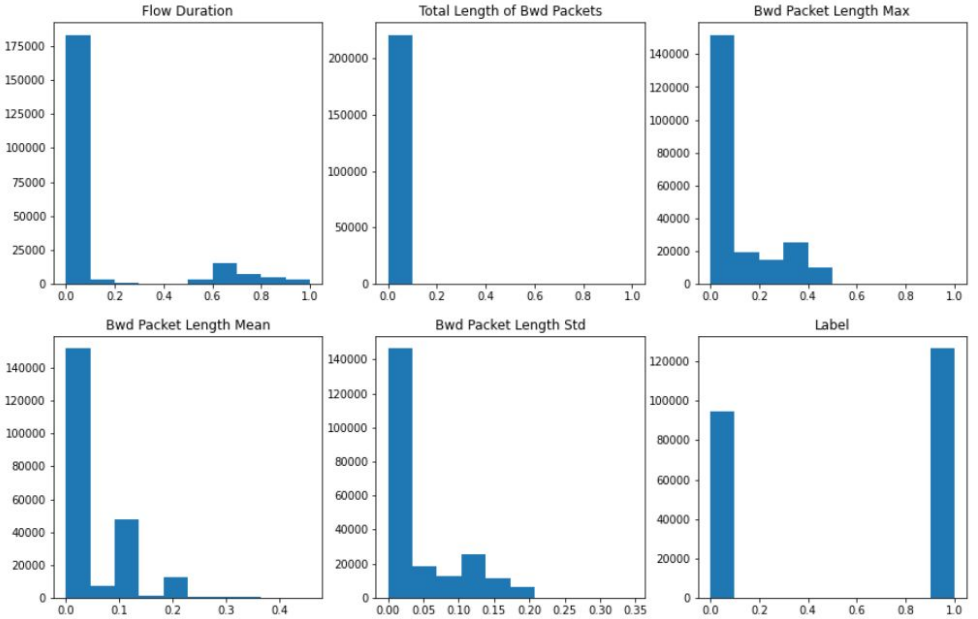
- Focus randomization of malicious data into most-common ranges for benign data



	precision	recall	f1-score	support
0.0	0.43	1.00	0.60	28476
1.0	0.00	0.00	0.00	37878
accuracy			0.43	66354
macro avg	0.21	0.50	0.30	66354
weighted avg	0.18	0.43	0.26	66354

Shift Malicious Features into Benign Range

- Shift existing DoS values into range that may be associated with Benign features



	precision	recall	f1-score	support
0.0	0.53	0.98	0.69	28476
1.0	0.96	0.36	0.52	37878
accuracy			0.62	66354
macro avg	0.75	0.67	0.61	66354
weighted avg	0.78	0.62	0.59	66354

Results

- All of the data poisoning attacks affected the accuracy of the Random Forest classifier negatively
  - Method 1: Indiscriminate Randomization within Bounds, F1-Score Accuracy = 0.79
  - Method 2: Targeted Randomization within Benign “Zones”, F1-Score Accuracy = 0.43
  - Method 3: DoS Entry Shift into Benign “Zones”, F1-Score Accuracy = 0.62
- Experiment is running on assumptions:
  - Attacker knows basic packet characteristics of devices on-network
    - Not incredibly unlikely!
- Proves the vulnerability of ML models to a poisoned training data pool

Questions?

Consider This: **How vulnerable is your model to a poisoning attack?**