Capstone Project Proposal

03.09.2024

Bruce Walker, UCSD MLE/AI Bootcamp

Network Traffic Analysis for Intrusion/Malware Detection

The Problem

Internet access has become an essential part of life for many households and businesses. With so many computers, phones, tablets, appliances, control systems, smart home devices, game consoles, televisions, and more accessing and being accessed by other systems over the internet, it makes LANs (personal, corporate, government, etc.) vulnerable to access by bad actors. How can someone differentiate between expected/authorized network traffic and unauthorized/malicious network traffic?

If machine learning can be applied to detect potentially malicious network traffic, it would be possible to put safeguards in place that could stop cyber attacks before they can do significant harm. This could prevent stress, save time, and save money for anyone that may be affected by a network breach.

The Data

For initial training and proof-of-concept, I will use the IoT-23: A labeled dataset with malicious and benign IoT network traffic[1] dataset available from Kaggle[2]. This is a labeled dataset of network traffic from Internet of Things (IoT) devices.

Criteria for Success

At model initiation, a baseline model will be established. An acceptable baseline model will predict with an accuracy of 60% or better. As the final model evolves I will look for an accuracy approaching or exceeding 75%. The goal will be to minimize false negatives while maintaining an acceptably low level of false positives.

[1] Sebastian Garcia, Agustin Parmisano, & Maria Jose Erquiaga. (2020). IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4743746

[2] https://www.kaggle.com/datasets/agungpambudi/network-malware-detection-connection-analysis/

The Approach

The project will use a supervised approach training its model on network traffic data that has previously been labeled as either benign or malignant. The model will attempt to predict if network traffic records are normal network traffic or malignant network traffic. The model will classify each record as either normal or malignant.

End Product

Once a viable model with sufficient accuracy is developed, it will be deployed as a web service. Users will be able to access a web page where they can upload a properly formatted file of network traffic records. Along with their network traffic file, they will provide an email address to send the results to. A properly formatted file will include specific fields (to be determined during model development) as well as a unique identifier for each record. If the service determines the input file was not properly formatted, the user will receive an email stating such. If the file is properly formatted, the user will receive an email with an attached file that includes each record's unique identifier and a label as to whether the record was determined to be normal or malignant traffic.

SMART Goals

- Establish a baseline model with 60% accuracy or better by October 1, 2024.
- Evolve and fine-tune the model to have better than 70% accuracy by November 15, 2024.
- Deploy model as web service by January 15, 2024.

Resource Requirements

Data wrangling and model development will be done on a combination of a high-end laptop, Google CoLab and/or AWS SageMaker. Deployed web service will be implemented using AWS web services including SageMaker.