Challenge Theme: Defence and Security

Network Traffic Anomaly Detection Challenge

Problem statement

Cyber Security is an important field to ensure the safe handling of data by governments and organisations. In today's day and age, there are many sophisticated ways to attack cyber security systems to either shut down a network or to take confidential information. Even a minor mistake can lead to significant repercussions. That is why being able to spot unwanted network traffic automatically is an important problem. One such example is DDoS attacks. DDoS Attack means "Distributed Denial-of-Service (DDoS) Attack" and it is a cybercrime in which the attacker floods a server with internet traffic to prevent users from accessing connected online services and sites. Being able to identify when a DDoS attack is happen can help keep critical services and sites functioning.

Data scientists can help create these anomaly detection systems, but there are several challenges one must overcome when implementing such a system. The goal is to be able to predict when network traffic is either regular traffic or suspicious/unusual traffic. This can be for purposes such as detecting denial of service (DDoS) attacks to shut down a network or detecting unusual/unauthorised data transfers on a network, for example.

One such issue is that many datasets will have a class imbalance. Many datasets (see below) will have many more examples of regular network traffic rather than suspicious network traffic. How would your team deal with the imbalanced dataset? Another such issue is the high rate of false positives. With the data being imbalanced, it may seem tempting to train a model such that it is able to detect all the anomalies but at a cost of a high false positive rate. Typically, human intervention is needed when the system detects an anomaly, so a high false positive rate would lead to a lot of wasted time. How do you tune a model so that it minimizes its false positives? Lastly, because of the sheer quantity of network traffic and the urgency of stopping a potential threat, the system should be able to work in real-time or near real-time. How can you develop such a system to be able to do this?

Data description

Data for this challenge would be drawn from research institutes that develop such datasets. A prime example is the DDoS Evauation Dataset. In this dataset, there contains information such as source and destination IP, packet size, and more. The goal is to determine whether the network traffic sent belongs to a DDoS attack or not. Below is a description of the features.

1. Timestamp: Each data record includes a timestamp indicating when the network activity occurred. The timestamp helps in understanding temporal patterns in the network traffic data.

2. Source IP Address: The IP address of the sender or source of the network traffic.

3. Destination IP Address: The IP address of the recipient or destination of the network traffic.

4. Source Port: The port number associated with the source of the traffic.

5. Destination Port: The port number associated with the destination of the traffic.

6. Protocol: The network protocol used for the communication (e.g., TCP, UDP).

7. Packet Length: The length (in bytes) of the network packet.

8. Data Payload: The actual content of the network packet, which may include information such as HTTP requests, DNS queries, or other application-specific data.

9. Flag: Flags associated with the packet, indicating characteristics like SYN, ACK, FIN, etc.

10. User Agent: User agent information for web traffic, indicating the type of client or browser used.

11. Network Flow ID: A unique identifier for each network flow, which groups related packets together.

12. Packet Time-to-Live (TTL): The Time-to-Live value for each packet, indicating the number of hops a packet can make before it is discarded

13. Label: Each record is labelled as either "normal" (indicating legitimate network traffic) or "anomalous" (indicating suspicious or potentially malicious activity).

Audience roles

- Private companies: Rely heavily on online services, websites, and data infrastructure. DDoS attacks can disrupt these services, resulting in financial losses, damage to their reputation, and a decrease in customer trust. Successfully identifying and mitigating DDoS attacks is crucial to safeguard their online assets.
- Public Sector – Government websites and online services are often targeted by DDoS attacks. Ensuring the availability and security of these services is essential for the public's trust and convenience. As well, many aspects of critical infrastructure, such as energy grids, water supply systems, and transportation, rely on digital technology. Successful DDoS detection and mitigation are crucial to protect these essential services from disruption.

Suggested outputs

Some potential routes for expressing findings that you may wish to consider:

- Ideas on how to deal with imbalanced data to train a model
- Lowering the False Positive Rate means less human intervention. How can we achieve this?
- Research into how to deploy real-time systems to show a proof of concept
- Explainability: How can organisations know why certain network traffic is anomalous? How can one visualise this to stakeholders?