IoT Network Traffic Malware Detection with Spark Pipeline

University of Minnesota
Big Data Analytics

Will Lewis, Rithvik Bhonagiri, Claire Liu, Rajesh Kumar Routray, Yufei Shen, Hyemin Yu

Presentation Outline

01 Cyber Security

Importance
Scope of the Problem

02 Key Considerations

Threat Facts
Machine Learning Benefits

03 Approach

Dataset
Framework
Model Capabilities

04 Business Case

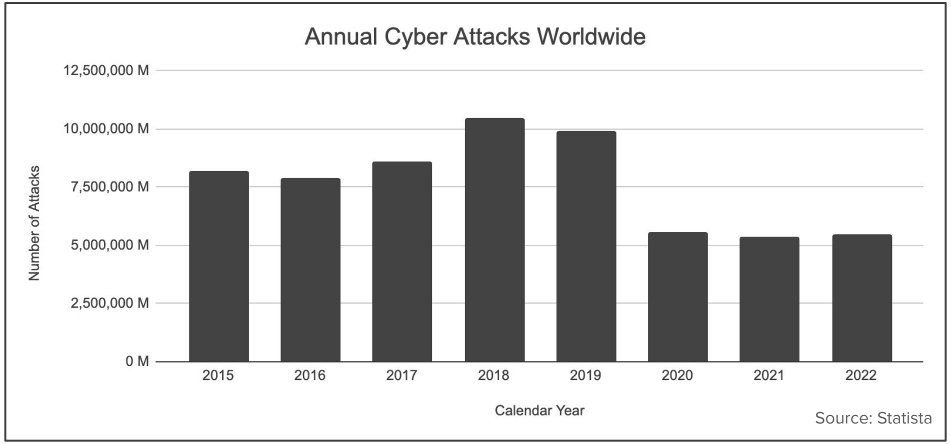
Summary
Sophistication

Cyber Security

Cybersecurity for any business or entity is critical in safeguarding sensitive information and digital assets from unauthorized access, attacks, and data breaches.

It plays a pivotal role in protecting individuals, businesses, and governments from the evolving and sophisticated threats that can compromise privacy, financial stability, and intellectual property.

- During 2022, the worldwide number of malware attacks reached 5.5 billion.



Source: Statista

Key Considerations

Threat Facts

- In 2022, **76% of organizations were targeted by a ransomware attack**, out of which 64% were actually infected.
- Every day, 560,000 new pieces of malware are detected, adding to the over 1 **billion malware programs already in circulation**.
- In its 2022 State of Cybersecurity Report, ISACA found that **69 percent of cybersecurity professionals believe their organization’s cybersecurity team is understaffed**

Machine Learning Benefits

- **With limited resources and funds**, machine learning can be used to automatically analyze real time traffic data and determine a response and recovery plan.
- This autonomous defense, using an in-house malware detection system can be **tailored specifically to an organization's security requirements**.
- Machine learnings ability to run on **a distributed computing platform eliminates a single point of failure** and further underscores the benefit for cyber security.
- Conventional systems fail to handle big data efficiently as they lack the inherent **scalability and distributed processing capabilities** essential for effectively managing large datasets.

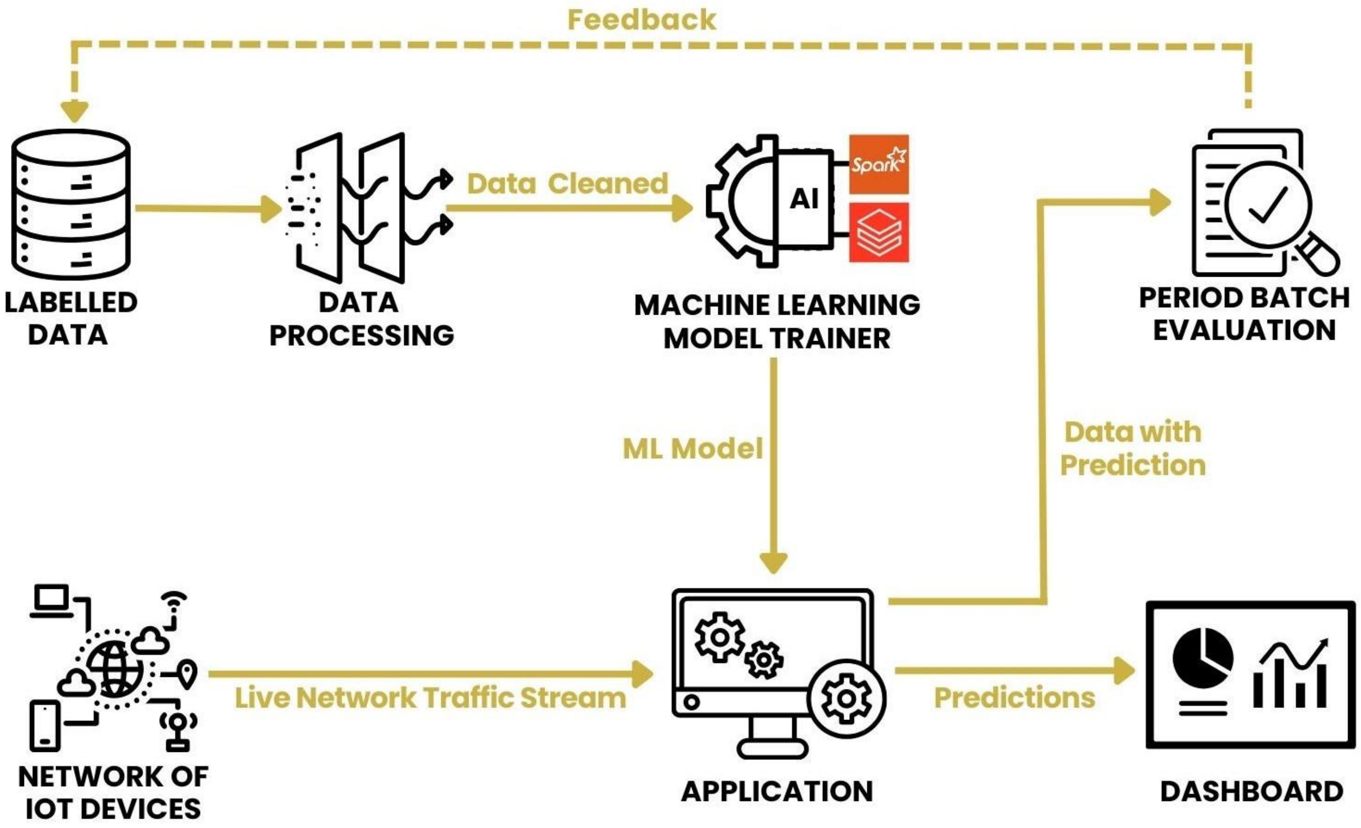
Approach

Utilizing ~5GB of public data from Kaggle, the team was able to develop a machine learning tool that can detect malware in real-time.

The dataset originates from the **Avast AIC laboratory** and consists of 23 different IoT networks with each record labelled as malicious or benign.

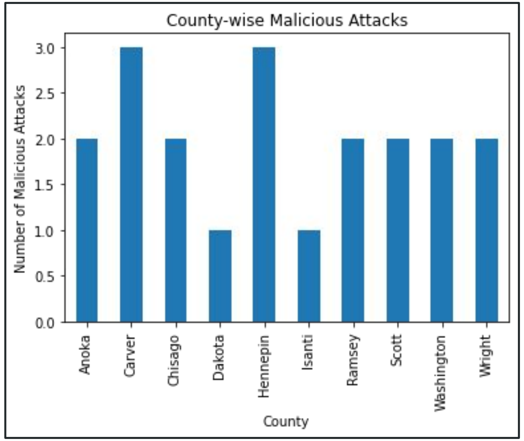
Both malicious and benign samples were **created in a controlled network environment with unrestricted internet access** to behave like any other real world device, a key consideration for future applications of our model.

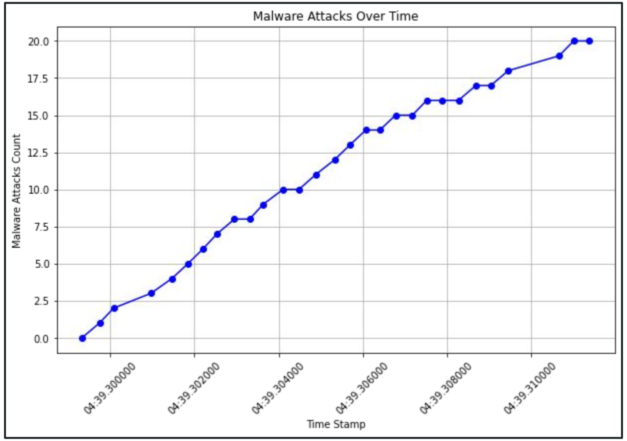
Approach

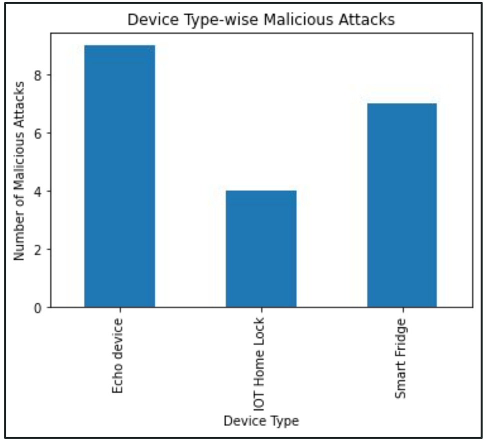

Feedback
LABELLED DATA
DATA PROCESSING
Data Cleaned
AI
Spark
MACHINE LEARNING MODEL TRAINER
PERIOD BATCH EVALUATION
ML Model
Data with Prediction
NETWORK OF IOT DEVICES
Live Network Traffic Stream
APPLICATION
Predictions
DASHBOARD

Approach

- The model developed showcases remarkable performance with an impressive Area Under the Curve (AUC) value of 99.73%, a high accuracy rate of 99.26%, and a robust F1 score of 99.18%.
- In addition to the model developed the team also developed dashboarding capabilities seen in the figures below, to identify to origin of an attack, the number of attacks over time and the types of devices creating the attacks.







Business Case

- In summary, cybersecurity stands as a pivotal element in protecting the information of any business or entity, particularly in the face of the ongoing technological advancements.
- Given that “69 percent of cybersecurity professionals believe their organization’s cybersecurity team is understaffed” the team has successfully crafted a model capable of addressing this challenge in real-time.
- Through the creation of a machine learning model able to handle streaming data sources leveraging a distributed computing environment the model can protect entities of any size from cyber attacks 24/7.

Thank you