Project Report

Cybersecurity: Suspicious Web Threat Interactions

1. Introduction

In contemporary cloud environments, the threat of cyberattacks and evil web traffic is constantly on the increase. Real-time detection and classification of abnormal web activities play an important role in the elimination of risks as well as safeguarding sensitive resources. This project utilizes **machine learning (ML), feature analysis (FA), and data analysis (DA)** methods for suspicious web threat interactions detection based on AWS CloudWatch traffic data.

2. Project Overview

Title: Cybersecurity – Anomalous Web Threat Interactions

Domain: Data Analytics / Cybersecurity

Language & Tools: Python, SQL, Excel, Jupiter Notebook, VS Code

Difficulty Level: Advanced

Dataset: AWS CloudWatch web traffic logs with labelled suspicious activities.

Objective:

To identify and analyse anomalous patterns in web interactions for recognizing suspicious or possibly malicious activities.

3. Dataset Details

Each row in the dataset is a web traffic session with the following key features:

`bytes_in`, `bytes_out` – Volume of network traffic

`creation_time`, `end_time`, `time` – Session dates

`src_ip`, `dst_ip` – Source & destination IP addresses

`src_ip_country_code` – Source IP geolocation

`protocol`, `dst_port` – Server port & protocol

* `response.code` – Status of HTTP response

* `rule_names`, `observation_name` – Detection reason

* `detection_types` – Classification of attacks

4. Methodology

4.1 Data Preparation

Deleted duplicate records (282 unique entries kept).

Time fields converted to datetime format.

Standardized categorical columns like country codes.

Confirmed lack of null values.

4.2 Feature Engineering

Session length: `end_time – creation_time`

Mean packet size: `(bytes_in + bytes_out) / session_length`

Scaled numeric features (`bytes_in`, `bytes_out`, `duration`) using StandardScaler.

One-hot encoded `src_ip_country_code`.

4.3 Exploratory Data Analysis (EDA)

Traffic Analysis: High volatility in `bytes_in` and `bytes_out`.

Protocol/Port Usage: All entries use HTTPS (200 response on port 443).

Country Insights: Traffic primarily from US, NL, CA, AE. Some areas exhibit repeated suspicious traffic.

Correlation Analysis: Strong correlation between `bytes_in` and `bytes_out`.

4.4 Modelling Approaches

1.Anomaly Detection (Isolation Forest): Detected outlier traffic sessions with abnormal byte ratios and session durations. Marked them as **Suspicious** or **Normal.**

2. Classification Models:

Random Forest Classifier: Obtained **100% accuracy** in classifying suspicious sessions **(`is_suspicious`).**

Neural Networks (Dense layers): Attained **100% test accuracy.**

CNN Model: Utilized Conv1D for sequential pattern detection, also attaining **100% test accuracy.**

5. Results & Findings

Model Performance:

Near-perfect classification was offered by Random Forest and Deep Learning models.

Isolation Forest successfully emphasized anomaly patterns.

Suspicious Traffic Insights:

High `bytes_in` with low `bytes_out` → suspected infiltration attempts.

Frequent traffic from certain country codes (i.e., US, NL) → potential botnets or targeted attacks.

Consistent duration (600s) indicates fixed monitoring periods over session-based logging.

Visualization Insights:

Scatterplots of anomalies had clean separation of normal vs. suspicious traffic.

Network graphs depicted interactions btween source and dest IPs.

Detection type distribution differed widely by country.

6. Conclusion

This project effectively showed **how data preprocessing, EDA, feature engineering, anomaly detection, and ML modeling** can be integrated to identify suspicious web threats in real-time traffic data.

The models attained **outstanding accuracy (100%),** but deployment in the real world would involve testing on larger, more heterogeneous datasets to prevent overfitting.

Future Scope:

Integrate live streaming detection pipelines.

Use reinforcement learning for dynamic threat responses.

Increase dataset with unlabelled real-world traffic for semi-supervised learning.

7. References

Dataset: [AWS CloudWatch Suspicious Traffic] (https://drive.google.com/file/d/1-OpnR9FK8EqGuLFB1k45ctPbl-vuZnC-/view?usp=sharing)

GitHub Repository: [Web Threat Analysis – Cybersecurity] (https://github.com/Tharunr0/Web-Threat-Analysis-Cyber-Security)