Web Traffic Anomaly Detection Using AWS CloudWatch Dataset

==========

Project Overview

In today's cloud-based environments, real-time detection of cybersecurity threats is crucial. This project aims to analyze and detect suspicious web traffic using a dataset derived from AWS CloudWatch, which logs and monitors traffic on cloud-based servers.

Objective:

- To build a machine learning model to detect and classify anomalous (suspicious) traffic from normal traffic based on server logs.
- To analyze and visualize patterns that can indicate potential threats or attack attempts.

Dataset Description

This dataset contains records of web traffic monitored through AWS CloudWatch. Each record represents a web interaction session with associated metadata and indicators.

Key Columns in the Dataset:

- bytes_in: Total incoming data to the server (in bytes).
- bytes_out: Total outgoing data from the server (in bytes).
- creation_time: Timestamp of when the record was created.
- end_time: Timestamp of when the connection ended.
- src_ip: IP address of the source (client).
- dst_ip: IP address of the server (destination).

- src_ip_country_code: Country code of the source IP.

- protocol: The protocol used (e.g., TCP, HTTP).

- response.code: HTTP status code returned.

- dst_port: Destination port used for the connection.

- rule_names: Detection rules that flagged the traffic as suspicious.

- observation_name: Observation linked to suspicious activity.

- source.meta: Metadata about the traffic source.

- source.name: Name of the data source.

- time: Timestamp when the suspicious activity was detected.

- detection_types: Type of detection (e.g., anomaly, signature-based).

Approach

---

1. Data Preprocessing:

   - Load and clean the dataset.

   - Handle missing or null values.

   - Encode categorical variables for machine learning processing.

2. Anomaly Detection using Isolation Forest:

   - Apply unsupervised learning technique (Isolation Forest) to detect anomalies.

   - Assign labels: -1 (suspicious), 1 (normal).

3. Data Visualization:

   - Count plot of normal vs. suspicious traffic.

   - Visualization of bytes in/out patterns.

   - Source country-wise suspicious IP distribution.

   - Time series plot of detected anomalies over time.

4. Result Analysis:

- Insights into patterns of attacks.
- Identification of common suspicious IPs or countries.
- Understanding the impact of different protocols and ports.

Libraries and Tools Used

- Python
- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn (IsolationForest)
- fpdf (for PDF generation)

How to Use

---

1. Install required libraries using:

   pip install -r requirements.txt

2. Run the Python script or Jupyter notebook to load the data, train the model, and visualize results.

3. Review the generated plots and PDF report for insights.

Use Cases

- Security Information and Event Management (SIEM) enhancement.
- Training intrusion detection systems (IDS).
- Cloud-based real-time security monitoring.
- Academic and research studies on anomaly detection in networks.

Conclusion

This project demonstrates the effective use of AWS CloudWatch traffic logs in detecting suspicious activities using machine learning. By visualizing and modeling the data, we gain deeper insights into cybersecurity threats and can develop stronger defenses.