Cybersecurity: Suspicious Web Threat Interactions Project Report

Project Overview

Objective:
To detect and analyse anomalies in web interaction logs using data science and machine learning techniques to identify suspicious behaviour, potential threats, and patterns indicative of cyberattacks.

Tools Used:

- Python (Pandas, NumPy, Seaborn, Matplotlib, scikit-learn, NetworkX)
- Machine Learning: Random Forest Classifier, Isolation Forest
- Dataset: CloudWatch web traffic logs

Dataset Summary

- **Source:** AWS CloudWatch logs
- **Size:** ~X records, Y columns (fill in actuals)
- **Features Included:**
  - Timestamps (creation_time, end_time, time)
  - Source/Destination IP and Ports
  - Bytes transferred (bytes_in, bytes_out)
  - Protocol, response code, detection type
  - Country code (from IP)

Data Preprocessing

- **Duplicates Removed:** All duplicate rows were eliminated.
- **Datetime Parsing:** Converted timestamps to proper datetime format.
- **Missing Values:** Handled missing bytes_in with median and dropped
  records missing critical IPs.
- **Text Normalization:** Country codes were uppercased for consistency.

Feature Engineering

- **Duration Calculation:** duration_seconds = end - start time
- **Average Packet Size:** Calculated as total bytes / duration
- **Scaling:** StandardScaler applied to numerical columns
- **Encoding:** OneHotEncoder used for country code, protocol, response code

Exploratory Data Analysis

- **Traffic Trends:**
  Time-series plot showed spikes in bytes exchanged — potential attack windows.
- **Detection Type vs Country:**
  Stack bar plot revealed countries associated with specific detection types like ‘waf_rule’.
- **Correlation Matrix:**
  Highlighted strong relationship between ‘bytes_in’, ‘bytes_out’, and duration.
- **Top Destination Ports:**
  Port 443 (HTTPS) dominated traffic — reinforcing web-based behavior.
- **Network Graph:**
  Visualized IP interactions using NetworkX to show source-destination communication paths.

Anomaly Detection

Used **Isolation Forest** to detect traffic behavior anomalies.

- **Features Used:** Scaled bytes, duration, and packet size
- **Result:** Tagged ~5% of traffic as anomalous
- **Visualization:** Scatter plot of bytes_in vs bytes_out with color-coded anomaly status

Classification model

Built a **Random Forest Classifier** to predict if traffic is suspicious (waf_rule).

- **Target:** is_suspecious (binary)
- **Features:** Scaled bytes_in, bytes_out, duration_seconds, avg_packet_size
- **Accuracy:** ~XX% (fill with actual)
- **Cross-Validation Accuracy:** ~YY%
- **Feature Importance:**
    - Duration and average packet size had the most predictive power.

Key Insights

1. Sudden spikes in traffic often link to malicious detection labels.
2. ‘waf_rule’ frequently appears in traffic from specific countries and port combinations.
3. Port 443 is the dominant entry point — potential exploitation target.
4. Isolation Forest was effective in flagging atypical traffic patterns not covered by existing rules.

Conclusion

This project demonstrates how machine learning can assist in real-time detection of anomalous web traffic patterns and potential cyber threats. The combination of anomaly detection and supervised classification adds robustness to traditional rule-based systems.