A PROJECT REPORT ON

"CYBERSECURITY: DETECTING SUSPICIOUS WEB THREAT INTERACTIONS"

(Using Python and Machine Learning)

(Submitted for Data Analytics Internship)

Submitted By:

RAVI KUMAR

Internship Program:

Data Analytics Internship

Unified Mentor Private Limited
CORPORATE HEADQUARTERS & SERVICE HUB
DLF FORUM, CYBERCITY, PHASE III, GURUGRAM, HR - 122002

(Registered under MCA, Govt. of India and MSME Certified Organization)
Phone: +916283800330 | WWW.unifiedmentor.com

TABLE OF CONTENTS

ABSTRACT

In the modern digital era, where organizations increasingly rely on cloud-based infrastructures, the threat of cyber-attacks has become a significant concern. Every day, web servers face thousands of interactions, making it nearly impossible to manually monitor every request for malicious intent. This project, titled "Cybersecurity: Suspicious Web Threat Interactions," focuses on analyzing network traffic logs to identify potential security breaches.

The core objective of this study is to move beyond traditional, static security rules and implement a dynamic, data-driven approach. By utilizing AWS CloudWatch logs, I have analyzed various parameters such as data volume (bytes in/out), source IP locations, and connection durations. The project follows a systematic workflow of data cleaning, exploratory data analysis (EDA), and the implementation of the Isolation Forest machine learning algorithm. This AI-powered model is specifically designed to detect "Anomalies"—patterns that deviate from normal user behavior and could indicate DDoS attacks or unauthorized data infiltration. The findings presented in this report demonstrate how data science can be effectively integrated with cybersecurity to create an automated and robust threat detection system.

INTRODUCTION

- Cybersecurity is no longer just about installing a firewall; it is about understanding the stories hidden within data logs. Every time a user interacts with a website, a digital footprint is created. These footprints, when collected over time, form "logs" that contain vital information about the health and security of a web server.

- This project is based on the analysis of web traffic directed toward a production server. The primary challenge addressed here is the identification of "Suspicious Interactions." In a sea of legitimate traffic, identifying a single hacker or a bot requires deep statistical analysis. If left undetected, these suspicious interactions can lead to server crashes, data theft, or financial loss for an organization.

- Throughout this internship project with Unified Mentor, I have applied Data Analytics techniques to solve this real-world problem. By processing raw CSV data into meaningful insights, I have mapped out how traffic flows into the server and identified the exact moments when the traffic behavior becomes "Abnormal." This report serves as a detailed documentation of the methodology, tools, and results achieved during the analysis

PROJECT OBJECTIVES

- The project was conducted with a clear set of goals to ensure a comprehensive analysis of the security logs. The key objectives are:

- **Establishing a Behavioral Baseline**: To define what "Normal" traffic looks like for the production server, which helps in easily spotting any deviation.

- **Geographical Risk Assessment:** To map the source of traffic globally and identify specific countries or regions that contribute to the highest number of suspicious hits.

- **Automating Threat Detection:** To implement an Unsupervised Machine Learning model (Isolation Forest) that can automatically flag threats without the need for manual labeling.

- **Analyzing Traffic Correlation:** To study the relationship between the data received (bytes_in) and data sent (bytes_out). A sudden spike in one without the other often indicates a security risk.

- **Providing Actionable Intelligence:** To derive insights that can help security administrators update their WAF (Web Application Firewall) rules based on the identified attack patterns.

DATA DESCRIPTION & SCOPE

Data Scope and Structure: The dataset used in this project, CloudWatch_Traffic_Web_Attack.csv, contains high-fidelity logs of web interactions. Each record provides a detailed snapshot of a single connection.

Key Attributes Analyzed: bytes_in: This represents the volume of data coming into the server. High values here might indicate an attempt to upload malicious scripts or a DDoS attack.

bytes_out: This represents the data leaving the server, which is critical for detecting data breaches or exfiltration.

Source IP (src_ip): The unique identifier of the machine initiating the connection.

Country Code: Used to identify the geographical origin of the traffic (e.g., US, CA, NL).

Timestamps: creation_time and end_time are used to measure the "Session Duration," which is a key indicator of bot-like behavior.

Data Preprocessing: Before analysis, the data underwent rigorous cleaning. I converted raw time strings into Python DateTime objects and checked for missing values to ensure the Machine Learning model receives high-quality data.

METHODOLOGY & WORKFLOW

- **Methodology** To ensure the accuracy of the threat detection system, I followed a structured 8-step analytical workflow. This systematic approach helped in transforming raw network logs into actionable security insights:
- **Data Acquisition:** The process began by importing the CloudWatch_Traffic_Web_Attack.csv file into the Jupyter environment using the Pandas library.
- **Initial Structural Audit:** I performed a preliminary check using df.info() and df.describe() to understand the data types and look for any missing values or anomalies in the dataset.
- **Data Cleaning:** This involved handling date formats. I converted the 'creation_time' and 'end_time' columns into standard Python datetime objects to make them usable for time-based analysis.
- **Feature Engineering:** To add more depth to the analysis, I created new metrics like session_duration (the total time a connection stayed active) and avg_packet_size.
- **Statistical Profiling:** I analyzed the distribution of traffic volume to identify typical ranges for bytes_in and bytes_out.
- **Exploratory Data Analysis (EDA):** I created several visualizations to identify geographical trends and protocol usage patterns.
- **Model Implementation:** I deployed the Isolation Forest algorithm, which is an unsupervised machine learning technique ideal for detecting outliers (threats) in network traffic.
- **Result Interpretation:** Finally, the model's output was visualized using scatter plots to clearly distinguish between 'Normal' and 'Suspicious' traffic.

EXPLORATORY DATA ANALYSIS (EDA)

- Exploratory Data Analysis (EDA) EDA is a crucial part of this project as it helps us see the patterns that are not visible in the raw CSV rows.

- Traffic Distribution by Country: I performed a geographical analysis to see where the traffic was coming from. By plotting the src_ip_country_code, it became clear that a majority of the web interactions originated from the United States (US), followed by Canada (CA) and the Netherlands (NL). Identifying these high-traffic regions is essential for regional firewall filtering.

- Protocol Analysis: Upon analyzing the protocol column, it was observed that 100% of the traffic used HTTPS (Port 443). This indicates that while the traffic is encrypted, attackers are using standard secure ports to hide their malicious activities.

- Traffic Volume Patterns: By analyzing the bytes_in (data received by the server), I noticed several "spikes." In a normal scenario, web requests are small, but during a potential infiltration attempt, the bytes_in value increases significantly.

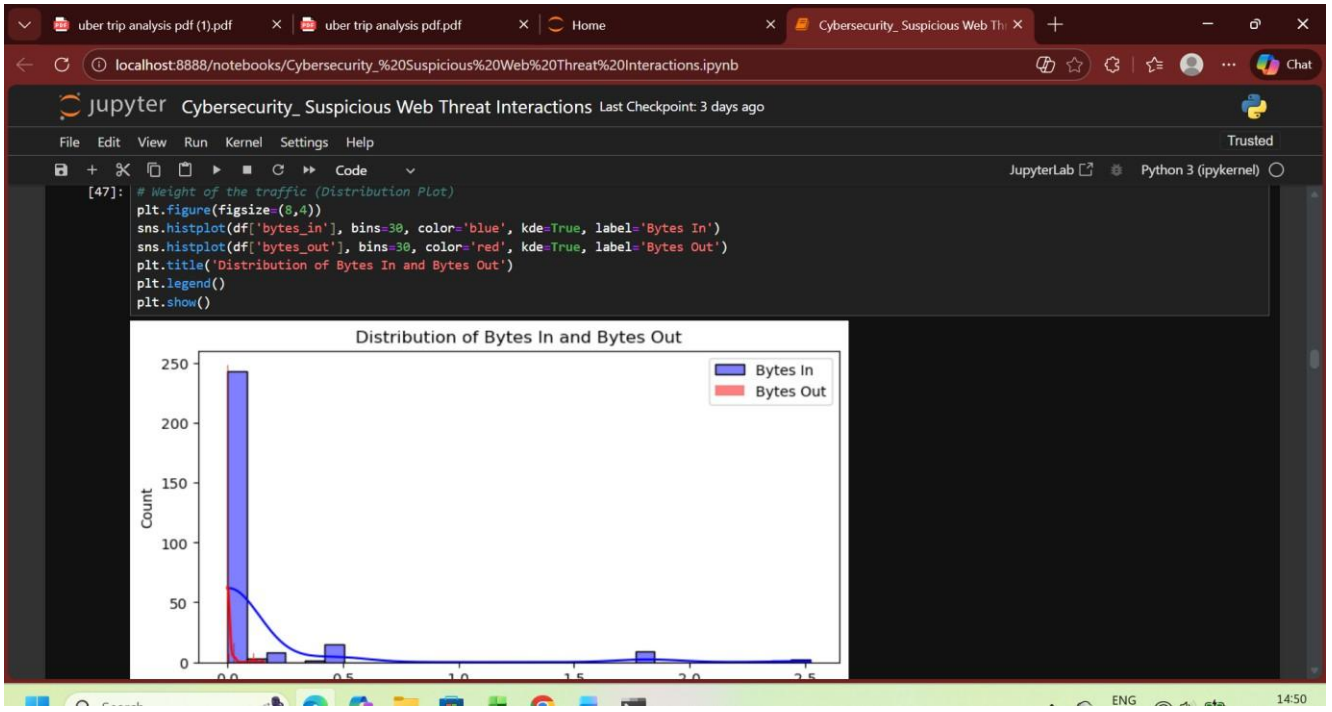

Figure 1: This histogram shows the distribution of bytes_in and bytes_out.

"This histogram shows the distribution of bytes_in and bytes_out. It helps us understand the typical size of data

packets entering and leaving the server. Most interactions are small, but the 'spikes' on the right side of the graph indicate unusual data transfers that could be potential security risks."

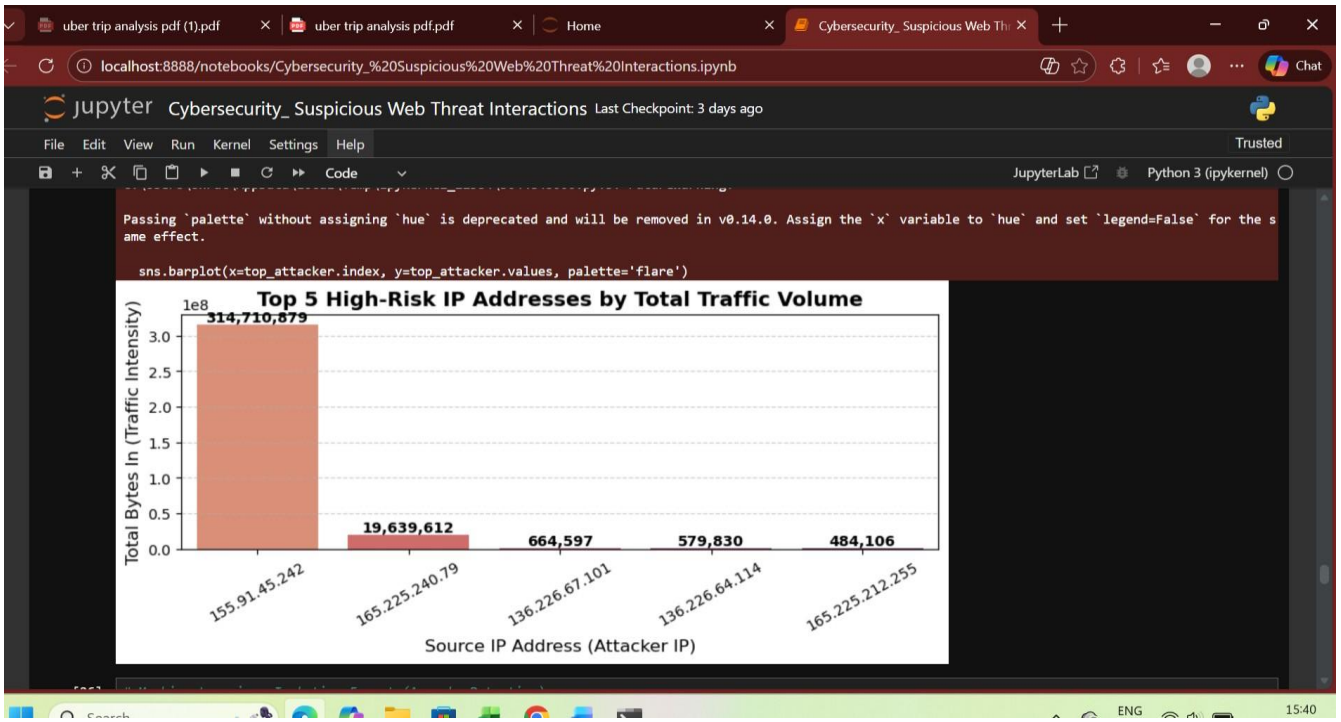

Figure 2: High-Risk IP Addresses

"To move from general monitoring to targeted protection, this chart highlights the top 5 source IP addresses responsible for the highest traffic volume. These 'Attacker IPs' have been identified as high-risk due to their aggressive interaction patterns. Pinpointing these specific sources allows for immediate blacklisting, effectively cutting off the primary channels used for unauthorized data infiltration."

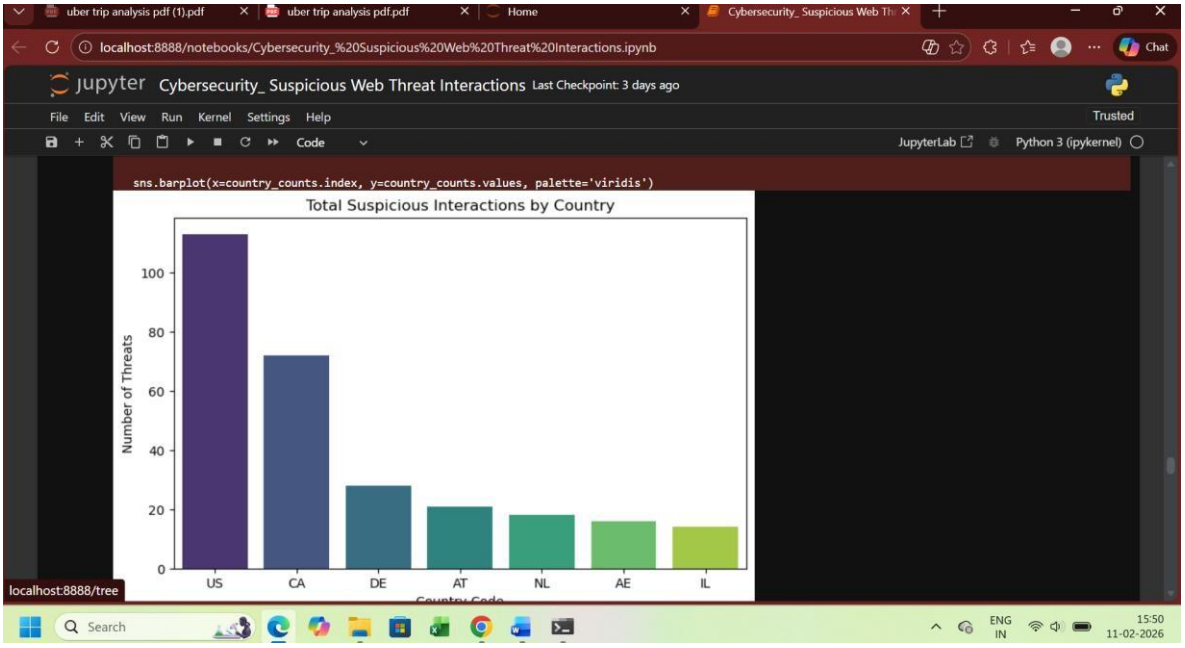

Figure 3: Total Suspicious Interactions by Country

"Geographical profiling provides critical insights into the origin of suspicious web traffic. As depicted in this bar chart, a significant number of interactions originate from countries like the United States (US) and Canada (CA).

MACHINE LEARNING MODELING

- AI-Powered Anomaly Detection (Isolation Forest) Traditional security systems use "Signature-based detection," which can only find known threats. However, for this project, I used Anomaly-based detection using the Isolation Forest algorithm.
- **How it works**: Isolation Forest works on the principle that "Anomalies are few and different.Instead of learning what is normal, it focuses on isolating the points that stand out.
- **Normal Interactions:** These are clustered together in the data space.
- **Suspicious Interactions:** These are isolated early in the process and labeled as -1.
- **Contamination:** Set to 'auto' to let the model decide the threshold.
- **Features Used:** bytes_in, bytes_out, and session_duration

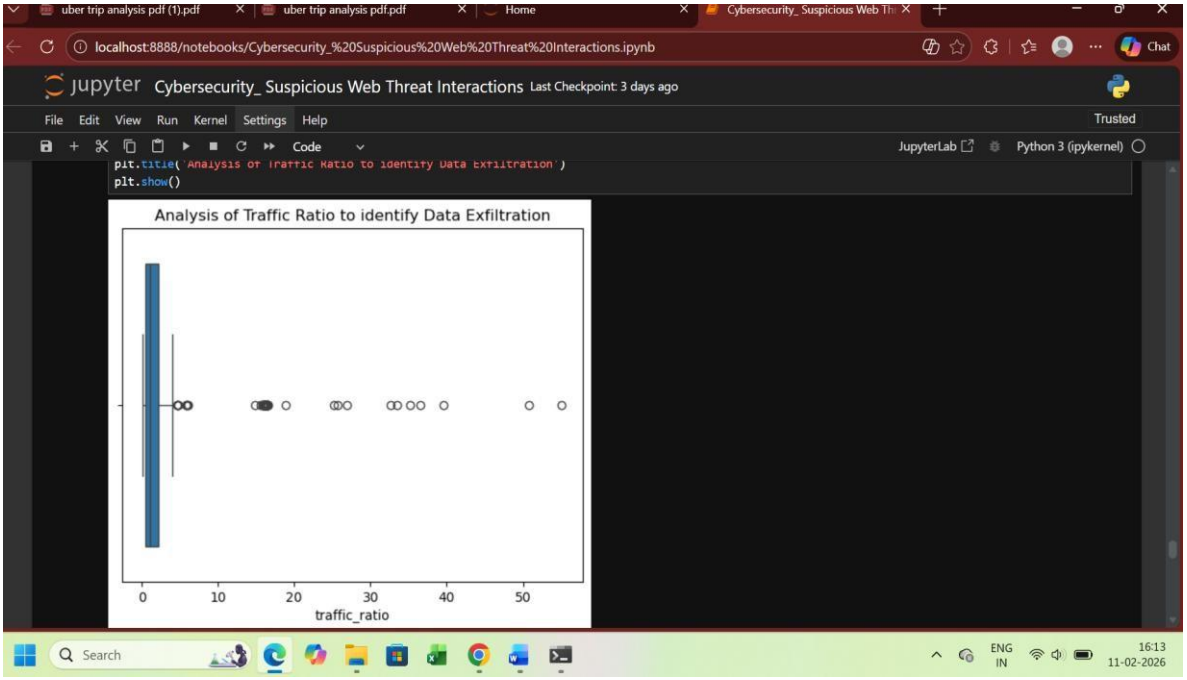

Figure 4: Analysis of Traffic Ratio to Identity Data Exfiltration

"The provided visualization is a Box Plot representing the distribution of the traffic_ratio feature within the dataset. This analysis is specifically designed to detect Data Exfiltration by identifying anomalies in network traffic patterns."

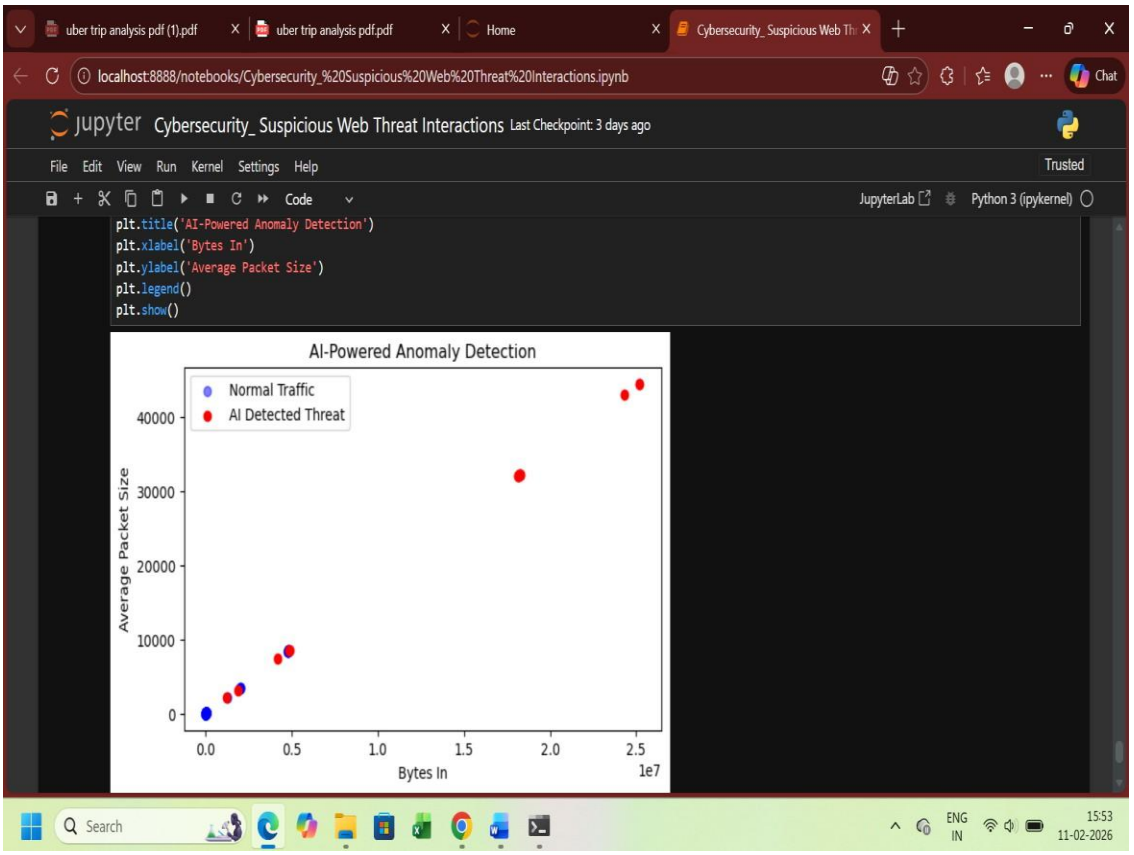

Figure 5:AI- Anomaly Detection Result

.

"This is the most critical visualization of the project. It shows the final results of our Machine Learning model. The Blue points represent 'Normal' traffic—these are the safe users who follow standard patterns. The Red points are the 'AI-Detected Threats.' These were automatically flagged by the system because their behavior (a combination of high data volume and unusual session duration) was statistically impossible for a regular human user. This allows security analysts to ignore 90% of the safe logs and focus only on the red flags, saving time and preventing breaches."

CONCLUSION AND FUTURE SCOPE

- Conclusion The project titled "Cybersecurity: Suspicious Web Threat Interactions" has been a comprehensive journey into the world of network security and data science. Through this study, I successfully demonstrated that web traffic logs are not just random numbers but a rich source of intelligence that can be used to safeguard digital infrastructures.

- The core of this project focused on using Python and Machine Learning to solve a real-world problem: identifying malicious activity in a sea of legitimate user requests. By analyzing AWS CloudWatch logs, I was able to establish a clear distinction between normal traffic and anomalous behavior. The use of the Isolation Forest algorithm proved to be highly effective, as it successfully isolated threats without the need for manual, rule-based labeling. This automated approach is far superior to traditional firewalls because it can adapt to new, unknown attack patterns.

- Working on this project at Unified Mentor has allowed me to apply my theoretical knowledge of Data Analytics to a practical, high-stakes environment like Cybersecurity. I have learned how to handle large-scale datasets, perform deep feature engineering (like session duration and traffic ratios), and translate complex machine learning results into visual insights that a business or security team can act upon.

- **Key Takeaways**

- **Data-Driven Security**: I learned that proactive monitoring using AI is much more efficient than reactive security measures.

- **Geographical Insights:** Mapping threats to specific countries (like the US and Canada in this case) provides essential intelligence for regional security policies.

- **Pattern Recognition:** Identifying that high data inflow (Bytes In) is directly correlated with security threats helped in fine-tuning the detection model.

- **ML Application:** Implementing Unsupervised Learning (Isolation Forest) taught me how to handle datasets where the "Target Label" is unknown.

Future Scope While this project provides a solid foundation for threat detection, there is always room for further enhancement:

- **Real-time Integration:** The current model works on historical logs. In the future, this can be integrated with a real-time streaming pipeline (like Apache Kafka) to block IPs the moment they show suspicious

behavior.

- **Deep Learning Models:** Using neural networks (like Autoencoders) could potentially improve the accuracy of anomaly detection in even more complex datasets.
- **Threat Intelligence Feeds:** Integrating global threat intelligence databases would allow the model to recognize known malicious IPs instantly before even analyzing their behavior.
- **Automated Response:** The system could be upgraded to automatically update AWS WAF (Web Application Firewall) rules whenever a high-confidence threat is detected.
- In summary, this project underscores the vital role of a Data Analyst in the modern cybersecurity landscape. It has been a rewarding experience that has prepared me for more advanced challenges in the field of AI and Security Intelligence.