FINAL PROJECT REPORT

ENGR - E 516 Engineering Cloud Computing

Topic:

AnomalyFinder

(Real-Time Anomaly Detection in Network Systems)

Team Members:
Shreenidhi Vishwanath Shetty
Hiren Rupchandani
Adityaraj Jain

1. Introduction

In today's hyper-connected digital landscape, the reliance on networking systems for seamless operations across industries has become indispensable. However, the complexity and scale of these networks bring forth a formidable challenge: the constant threat of system anomalies, failures, and security breaches. The repercussions of these disruptions can range from compromised data integrity to substantial financial losses and even pose severe risks to the continuity of critical services.

In response to this pressing need, this project endeavors to introduce an innovative approach: a robust, cloud-based system fortified by machine learning algorithms. The primary aim is to proactively identify, analyze, and mitigate anomalies within the networking system's metrics and logs in real time. By harnessing the power of machine learning, this system seeks to transcend traditional reactive measures, offering preemptive solutions to impeding system failures and security breaches.

Objectives

The fundamental objective of this project is to conceptualize, design, and implement an advanced anomaly detection system that operates within the cloud infrastructure. This system will continuously monitor and scrutinize an array of metrics and logs generated by the networking environment, employing sophisticated machine learning algorithms. By analyzing historical data patterns and discerning deviations in real time, this system aims to predict potential anomalies and deviations from normal operations. Moreover, it endeavors to provide timely alerts and actionable insights to enable swift intervention, thus preventing system failures or security incidents.

Significance and Scope

The significance of this project lies in its potential to revolutionize the paradigm of system monitoring and security enhancement. By adopting a proactive stance through predictive anomaly detection, organizations can substantially reduce the likelihood of critical system disruptions. This system is envisioned to cater to a diverse spectrum of industries, including but not limited to finance, healthcare, telecommunications, and manufacturing, safeguarding their networks against unforeseen adversities.

The scope of this project encompasses the design, development, and deployment of the cloud-based anomaly detection system. It will involve the selection and integration of appropriate machine learning models, the establishment of a scalable and efficient cloud infrastructure, and the implementation of robust monitoring and alerting mechanisms.

2. Background and Related Work:

The evolution of networking systems has been marked by a constant quest for enhancing performance, reliability, and security. As technology advances, so do the complexities and challenges associated with maintaining these networks at peak efficiency while safeguarding them against potential threats.

- **Networking System Challenges**
  The burgeoning scale and intricacies of modern networking systems have intensified the need for comprehensive monitoring and preemptive measures. Traditionally, reactive approaches to system monitoring have proven insufficient, often leading to significant downtimes, compromised data integrity, and vulnerabilities to security breaches. The dynamic nature of network behavior and the sheer volume of data generated pose substantial hurdles in detecting anomalies and potential threats in real time.

- **Previous Approaches and Research**
  A plethora of research and initiatives have been dedicated to addressing these challenges. Previous studies have explored various techniques, including rule-based systems, statistical analysis, and clustering algorithms, to detect anomalies within networking systems. However, these approaches often falter in adapting to the evolving nature of anomalies, resulting in high false-positive rates or delayed identification of critical issues.

- **Machine Learning and Anomaly Detection**
  The advent of machine learning has introduced a paradigm shift in anomaly detection. Its ability to discern intricate patterns within vast datasets and adapt to changing behaviors aligns well with the dynamic nature of network anomalies. Researchers have explored diverse machine learning models, such as neural networks, support vector machines, and ensemble methods, showcasing promising results in anomaly detection across different domains.

- **Cloud-Based Solutions and Anomaly Detection**
  Simultaneously, the migration of systems to cloud-based architectures has opened new avenues for anomaly detection. Leveraging the scalability, flexibility, and computational power of cloud infrastructure offers a promising environment for deploying real-time anomaly detection systems. Cloud-based solutions enable the processing of extensive data streams, facilitating quicker analysis and response times to potential threats.

- **Limitations and Areas of Improvement**
  Despite these advancements, challenges persist, including model interpretability, scalability concerns, and the need for continuous model adaptation. Achieving a balance between accuracy and computational efficiency remains a crucial area of improvement.

- **Gap and Rationale for Current Project**
  Amidst these advancements and challenges, this project aims to bridge the gap by devising an innovative cloud-based system. This system harnesses the prowess of machine learning algorithms, capitalizing on their adaptive capabilities to perform real-time anomaly detection in networking systems' metrics and logs. By amalgamating the strengths of cloud infrastructure and cutting-edge machine learning, this project endeavors to overcome existing limitations and provide a robust, proactive solution to impending system failures and security incidents.

3. Architecture

The architecture of the proposed anomaly detection system is designed as a cohesive amalgamation of cloud-based services and frameworks aimed at providing robust, real-time anomaly detection capabilities. Leveraging the scalability and security of Amazon Web Services (AWS), the system orchestrates a seamless flow of data processing, model training, predictions, and user interaction.

Amazon S3 Bucket
At the core of the system lies the Amazon S3 bucket, serving as the secure repository for storing both the initial and processed data. This scalable cloud storage solution ensures the reliability and accessibility of the data required for model training and real-time predictions.

Amazon EC2 - Flask and React Integration
The architecture consists of Amazon EC2 instances responsible for managing the back-end operations using Flask, a Python-based microframework, and the front-end user interface developed using React. The EC2 instances facilitate efficient data processing, API handling, and seamless integration between the back-end and front-end components.

Methodology

- **Data Preparation**
  The system initiates with a meticulous data preparation phase. This involves rigorous data cleaning procedures to ensure the accuracy and relevance of the data used for anomaly detection. Once cleaned, the data is fetched securely from the Amazon S3 storage service.

- **Model Training and Storage**
  Trained machine learning models play a pivotal role in anomaly detection. The prepared data undergoes model training to identify potential anomalies effectively. Post-training, these models are securely stored back into the Amazon S3 bucket for efficient retrieval and reuse.

- **Prediction and Analysis**
  Upon user request, the system retrieves the stored model from Amazon S3. Simultaneously, the relevant testing data is fetched for analysis. The system employs the retrieved model to analyze the testing data, executing predictions regarding potential anomalies.

- **Data Visualization and Reporting**
  Real-time visualizations and reports are generated using APIs, offering users an interactive and informative interface. The system utilizes API calls to create graphs that visually represent the analysis and predictions. Additionally, the responses from the model are formatted in JSON for seamless interoperability and ease of data handling.

System Architecture

- **Front-End (React)**
  The user interface, developed using React, ensures a responsive and engaging user experience. React's capabilities enable the creation of an intuitive interface, facilitating user interaction and data visualization.

- **Back-End (Amazon EC2)**
  Amazon EC2 instances serve as the backbone of the back-end infrastructure. These instances handle data processing, model execution, and API interactions, ensuring efficient management of the system's functionalities.

Collaboration and Integration

- **Seamless Integration**
  The architecture is meticulously designed for seamless integration between the front-end and back-end components. This integration ensures smooth data flow, allowing users to interact with the handling of data processing tasks.

4. Result Analysis

The culmination of the anomaly detection system's functionality is reflected in the comprehensive visualization dashboard, AWS monitoring, and webpage screenshots. These elements collectively showcase the system's efficacy in analyzing network metrics and logs, providing valuable insights into potential anomalies and security threats.

Amazon S3 Dashboard and AWS EC2 Monitoring
The Amazon S3 Dashboard serves as a centralized hub, housing crucial data and models for the anomaly detection system. Simultaneously, AWS EC2 monitoring ensures the efficient operation of the back-end infrastructure, ensuring optimal performance and reliability.

Graphs and Prediction Results

- **Distribution of Intrusions**
  The initial bar graph delineates the distribution of various intrusions within the network. Notably, it reveals the prevalence of the "Benign" category as the highest, with "Botnet-Attempted" registering as the lowest intrusion type.

- **Detailed Categorization**
  A ring chart supplements the first graph, offering a detailed breakdown of intrusion categories. This visualization provides a more nuanced view, highlighting the proportions of different intrusion types within the network.

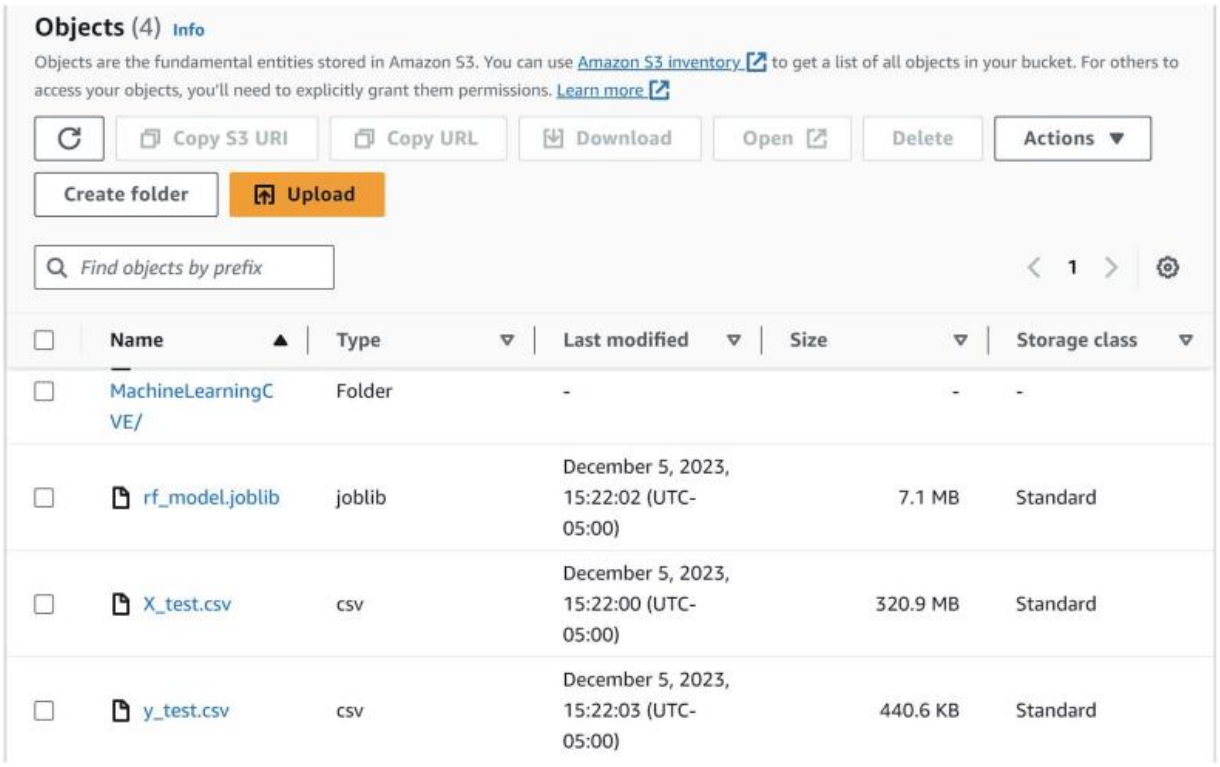
- **Distribution of Safe vs. Harmful Devices**
  The subsequent bar chart depicts the distribution between safe and harmful devices operating within the network. This visualization sheds light on the ratio of benign to malicious devices present.

- **Malicious and Benign Device Percentage**
  A ring chart complements the distribution chart, illustrating the percentages of malicious and benign devices. This visualization showcases a clear differentiation, with malicious devices accounting for 24.6% and benign devices comprising the majority at 75.4%.

- **Prediction Table**
  Below these graphs, the prediction table unfolds, displaying critical information. It outlines predicted attacks, the probability of occurrence, and the possible attack names. At the conclusion of the table, the accuracy metric is presented, signifying the system's efficacy in predicting potential attacks.

These visual representations and prediction insights corroborate the system's ability to not only identify anomalies but also categorize and predict potential threats within the network environment. The accuracy metrics validate the system's reliability in discerning potential attacks, contributing to enhanced preemptive measures and system fortification against security breaches and failures.

S3 Dashboard



Objects (4) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Copy S3 URI | Copy URL | Download | Open | Delete | Actions

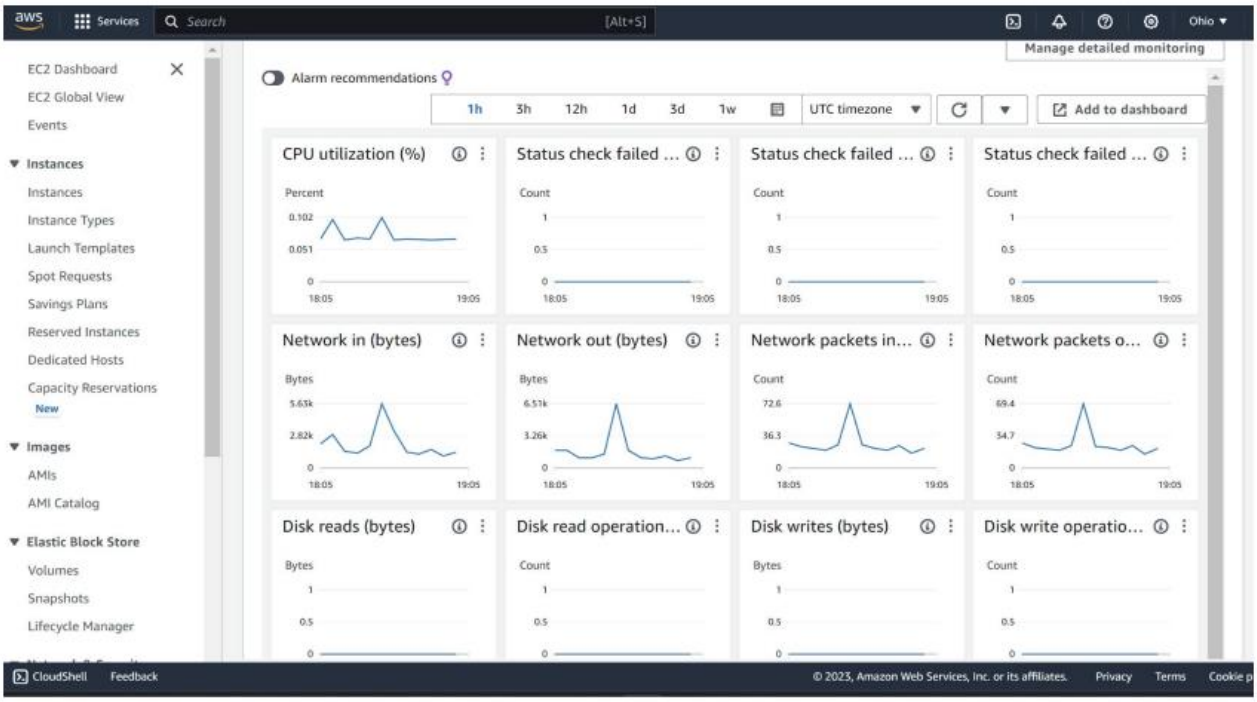
Create folder | Upload

Find objects by prefix

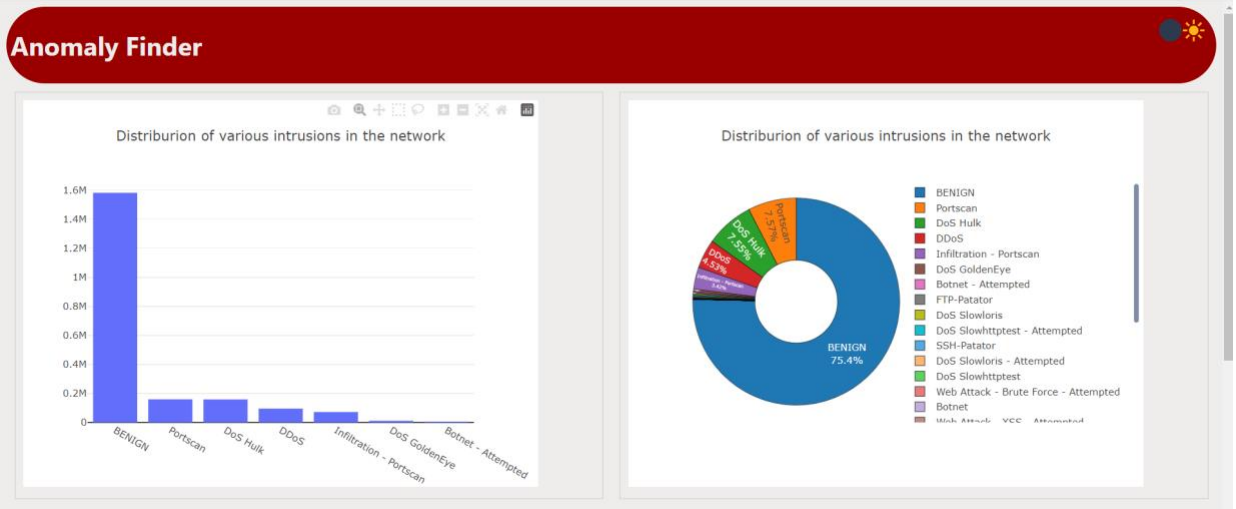
1

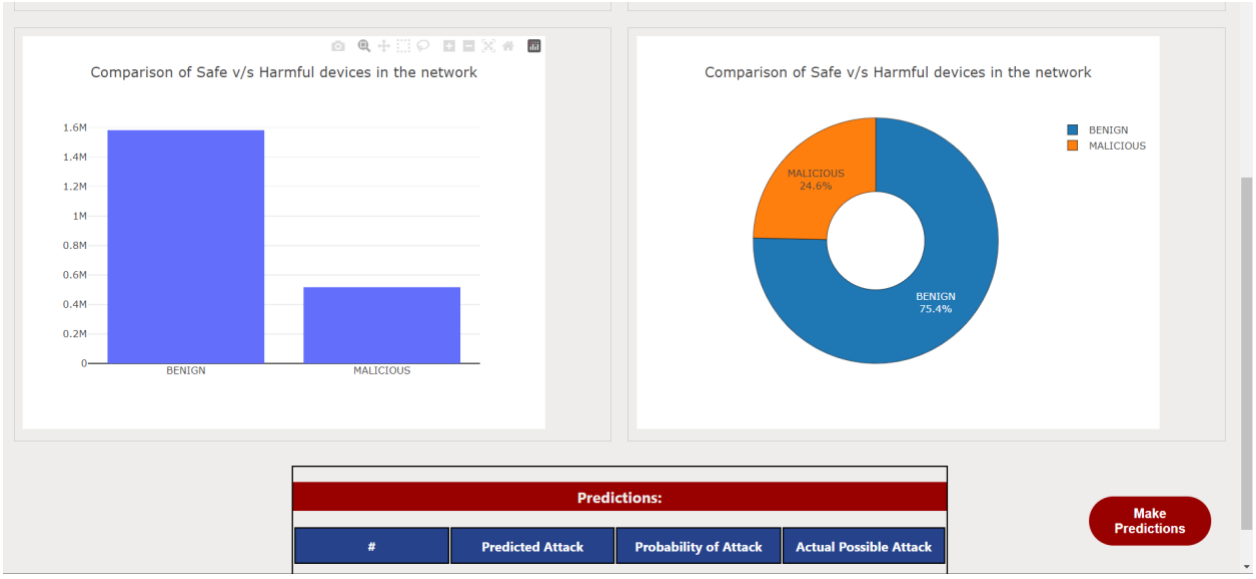
Name	Type	Last modified	Size	Storage class
MachineLearningCVE/	Folder	-	-	-
rf_model.joblib	joblib	December 5, 2023, 15:22:02 (UTC-05:00)	7.1 MB	Standard
X_test.csv	csv	December 5, 2023, 15:22:00 (UTC-05:00)	320.9 MB	Standard
y_test.csv	csv	December 5, 2023, 15:22:03 (UTC-05:00)	440.6 KB	Standard

EC2 Monitoring



Application Snapshots





Predictions:

#	Predicted Attack	Probability of Attack	Actual Possible Attack
1	DoS Slowloris - Attempted	0.9912	DoS Slowloris - Attempted
2	DoS Slowloris - Attempted	0.9995	DoS Slowloris - Attempted
3	BENIGN	0.9973	BENIGN
4	DoS Slowloris - Attempted	0.9688	DoS Slowloris - Attempted
5	BENIGN	0.997	BENIGN
6	BENIGN	0.9971	BENIGN
7	DoS Slowloris	0.9987	DoS Slowloris
8	DoS GoldenEye	1	DoS GoldenEye
9	DoS GoldenEye	1	DoS GoldenEye
10	Botnet - Attempted	1	Botnet - Attempted
11	DoS Slowloris	1	DoS Slowloris
12	DoS Slowloris - Attempted	1	DoS Slowloris - Attempted

Make Predictions

8	DoS GoldenEye	1	DoS GoldenEye
9	DoS GoldenEye	1	DoS GoldenEye
10	Botnet - Attempted	1	Botnet - Attempted
11	DoS Slowloris	1	DoS Slowloris
12	DoS Slowloris - Attempted	1	DoS Slowloris - Attempted

13	DoS Slowloris - Attempted	1	DoS Slowloris - Attempted
14	Botnet - Attempted	1	Botnet - Attempted
15	DoS Slowloris - Attempted	0.9695	DoS Slowloris - Attempted
16	BENIGN	0.9989	BENIGN
17	DoS GoldenEye	1	DoS GoldenEye
18	Botnet - Attempted	1	Botnet - Attempted
19	Botnet - Attempted	1	Botnet - Attempted
20	DoS Slowloris - Attempted	0.9945	DoS Slowloris - Attempted

Accuracy: 98.8%

Make Predictions

5. Future Work

The current anomaly detection system lays a robust foundation, yet several avenues for future enhancements and advancements exist. The following focal points delineate potential directions for augmenting and refining the system's capabilities:

- **Exploration of Diverse Data Sources**
  Future endeavors could involve an extensive exploration of diverse data sources to bolster the anomaly models. Incorporating additional data streams from various sources, including IoT devices, cloud logs, and user behavior patterns, can enrich the models.

- **Proactive Anomaly Prediction and Prevention**
  Emphasizing proactive measures for anomaly prediction and prevention stands as a pivotal aspect for future development. The system could evolve to not only detect anomalies but also forecast potential threats with greater precision. Implementing advanced predictive analytics and leveraging historical patterns can enable the system to anticipate anomalies before they escalate, thus fortifying preemptive measures.

- **Utilization of Ensemble Learning Advancements**
  Harnessing the continual advancements in ensemble learning techniques holds promise for improving the system's accuracy. Future iterations could delve deeper into ensemble methodologies, exploring ensemble models that combine diverse algorithms.

- **Development of Resilience Strategies Across Domains**
  Expanding the system's resilience strategies across diverse domains remains a crucial avenue for future work. Tailoring anomaly detection models to accommodate the specific nuances and challenges within various domains, such as finance, healthcare, or critical infrastructure, can elevate the system's adaptability and effectiveness in safeguarding against domain-specific threats.

- **Integration of Real-Time Response Mechanisms**
  Incorporating real-time response mechanisms aligned with anomaly detection could be an area of focus. Future iterations might explore automated responses or adaptive actions triggered upon anomaly detection, allowing for immediate mitigation of identified threats before they escalate.

- **Enhancement of Scalability and Efficiency**
  Efforts to optimize scalability and efficiency should persist in future developments. Streamlining the system's architecture, employing distributed computing, and leveraging cloud-native technologies can further enhance the system's ability to handle vast data streams while maintaining real-time anomaly detection capabilities.

These future directions stand as potential pathways to elevate the anomaly detection system's capabilities, fortify its resilience across diverse domains, and foster a proactive stance in averting potential threats within cloud-based systems.