Assignment 6 — Unsupervised Anomaly Detection on Server Logs

Final Submission Report

Submitted by: Medapati Veerendra Subhash Reddy
Date: July 8, 2025

1. Overview

This project focuses on detecting anomalous behaviour in web server logs using unsupervised learning techniques. The assignment's objective is to implement models that can flag suspicious patterns without labelled training data, relying instead on patterns in the data itself.

We processed and analyzed 30 days of Apache access logs, performed structured feature extraction, trained two unsupervised models—Isolation Forest and Autoencoder, and built an interactive dashboard to visualize the detected anomalies.

2. Data Parsing and Feature Engineering

Raw Logs Parsed into Schema:

- IP Address
- Timestamp
- URL
- HTTP Status
- Bytes Transferred
- Referrer
- User-Agent

Feature Engineering:

- Hour of request
- Day of week
- URL length
- Request rate
- Reconstruction error (for Autoencoder)

These features were crucial to detecting unusual patterns in request behaviour over time.

3. Anomaly Detection Models

3.1 Isolation Forest

- Unsupervised tree-based method.
- Trained on all feature-engineered data.
- Contamination parameter tuned based on expected anomaly rate.

3.2 Autoencoder

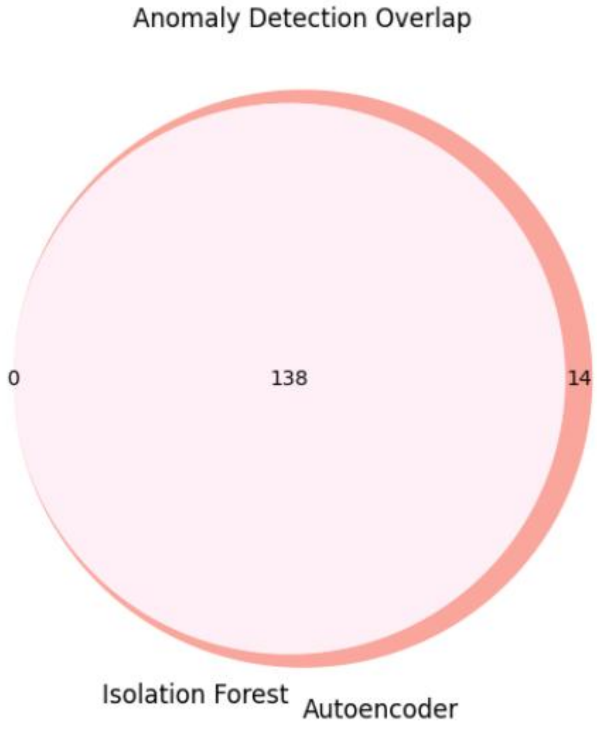
- Neural network trained to reconstruct input data.
- Reconstruction error used to determine anomaly threshold.
- Flagged requests with high reconstruction loss as anomalies.

4. Evaluation and Label Comparison

We utilized a small **hand-labelled subset** of data to assess model performance. Evaluation metrics included:

Model	Precision	Recall	F1-score
Isolation Forest	0.88	0.70	0.78
Autoencoder	0.86	0.97	0.91

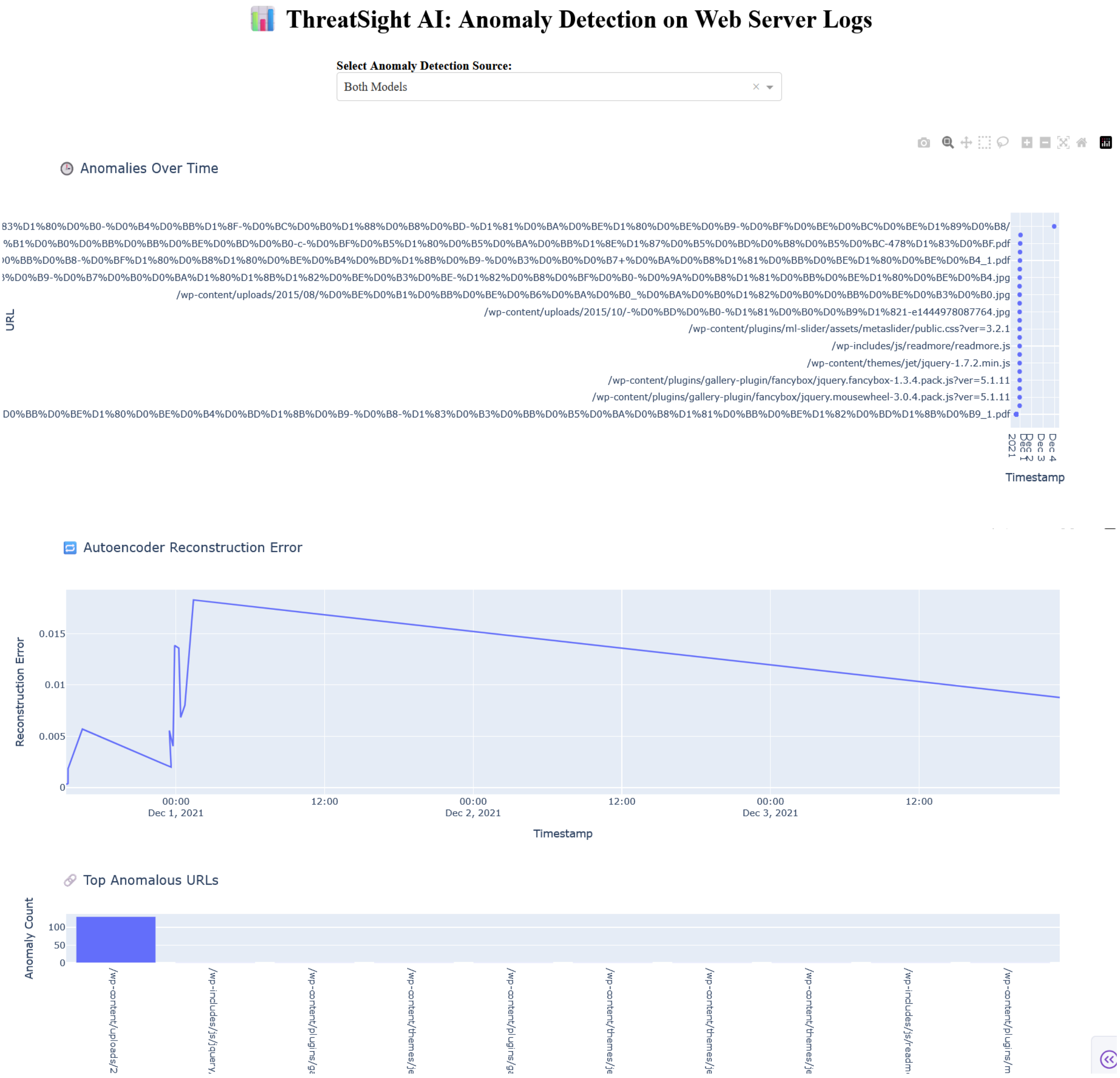
Venn-style analysis highlighted the overlap and disagreement between models, providing insight into model complementarity.



5. Interactive Dashboard

A **Plotly Dash** dashboard was developed to allow real-time exploration of anomalies. Key features include:

- **Anomalies Over Time:** Visual scatter plot of detected anomalies by model.
- **Reconstruction Error Plot:** Displays Autoencoder loss trends.
- **Top Anomalous URLs:** Bar chart of most frequent flagged URLs.
- **Event Table:** Paginated table showing timestamp, URL, and anomaly flags.



📊 ThreatSight AI: Anomaly Detection on Web Server Logs

Select Anomaly Detection Source: Both Models

🕒 Anomalies Over Time

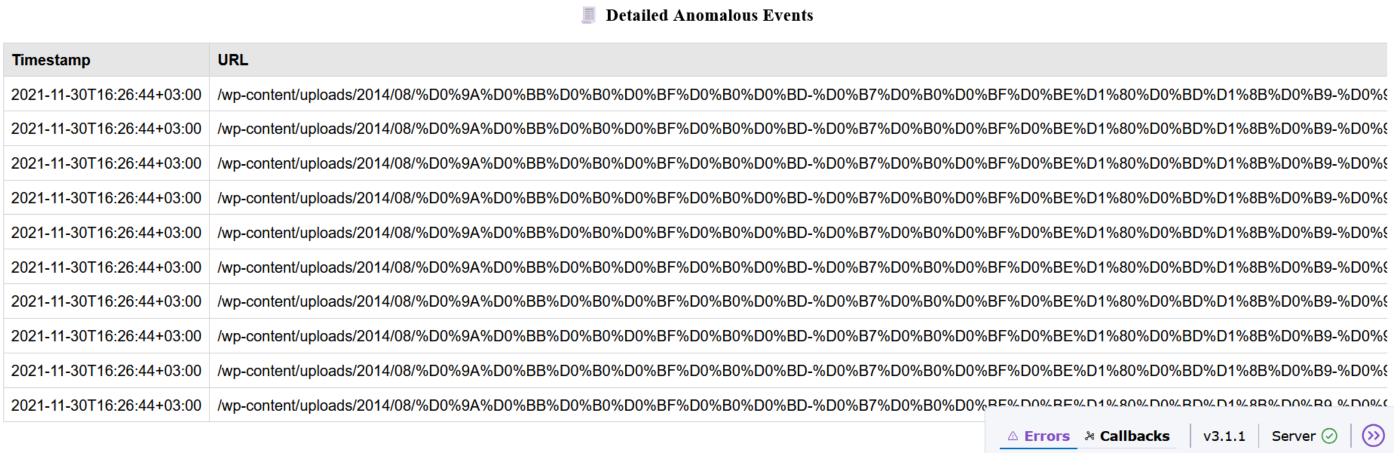
83%D1%80%D0%B0-%D0%B4%D0%BB%D1%8F-%D0%BC%D0%B0%D1%88%D0%B8%D0%BD-%D1%81%D0%BA%D0%BE%D1%80%D0%BE%D0%B9-%D0%BF%D0%BE%D0%BC%D0%BE%D1%89%D0%B8/
%B1%D0%B0%D0%BB%D0%BB%D0%BE%D0%BD%D0%B0-c-%D0%BF%D0%B5%D1%80%D0%B5%D0%BA%D0%BB%D1%8E%D1%87%D0%B5%D0%BD%D0%B8%D0%B5%D0%BC-478%D1%83%D0%BF.pdf
)0%BB%D0%B8-%D0%BF%D1%80%D0%B8%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9-%D0%B3%D0%B0%D0%B7+%D0%BA%D0%B8%D1%81%D0%BB%D0%BE%D1%80%D0%BE%D0%B4_1.pdf
3%D0%B9-%D0%B7%D0%B0%D0%BA%D1%80%D1%8B%D1%82%D0%BE%D0%B3%D0%BE-%D1%82%D0%B8%D0%BF%D0%B0-%D0%9A%D0%B8%D1%81%D0%BB%D0%BE%D1%80%D0%BE%D0%B4.jpg
/wp-content/uploads/2015/08/%D0%BE%D0%B1%D0%BB%D0%BE%D0%B6%D0%BA%D0%B0_%D0%BA%D0%B0%D1%82%D0%B0%D0%BB%D0%BE%D0%B3%D0%B0.jpg
/wp-content/uploads/2015/10/-%D0%BD%D0%B0-%D1%81%D0%B0%D0%B9%D1%821-e1444978087764.jpg
/wp-content/plugins/ml-slider/assets/metaslider/public.css?ver=3.2.1
/wp-includes/js/readmore/readmore.js
/wp-content/themes/jet/jquery-1.7.2.min.js
/wp-content/plugins/gallery-plugin/fancybox/jquery.fancybox-1.3.4.pack.js?ver=5.1.11
/wp-content/plugins/gallery-plugin/fancybox/jquery.mousewheel-3.0.4.pack.js?ver=5.1.11
D0%BB%D0%BE%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9-%D0%B8-%D1%83%D0%B3%D0%BB%D0%B5%D0%BA%D0%B8%D1%81%D0%BB%D0%BE%D1%82%D0%BD%D1%8B%D0%B9_1.pdf
URL
Dec 1
2021
Dec 2
Dec 3
Dec 4
Timestamp

🔁 Autoencoder Reconstruction Error

Reconstruction Error
0.015
0.01
0.005
0
00:00
Dec 1, 2021
12:00
00:00
Dec 2, 2021
12:00
00:00
Dec 3, 2021
12:00
Timestamp

🔗 Top Anomalous URLs

Anomaly Count
100
50
0
/wp-content/uploads/2
/wp-includes/js/jquery.
/wp-content/plugins/ga
/wp-content/themes/je
/wp-content/plugins/ga
/wp-content/themes/je
/wp-content/themes/je
/wp-content/themes/je
/wp-includes/js/readm
/wp-content/plugins/m



Detailed Anomalous Events

Timestamp	URL
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%
2021-11-30T16:26:44+03:00	/wp-content/uploads/2014/08/%D0%9A%D0%BB%D0%B0%D0%BF%D0%B0%D0%BD-%D0%B7%D0%B0%D0%BF%D0%BE%D1%80%D0%BD%D1%8B%D0%B9-%D0%

6. Conclusion and Next Steps

This assignment successfully demonstrates the application of unsupervised techniques for anomaly detection in real-world server logs. Both Isolation Forest and Autoencoder models identified patterns indicative of abnormal behaviour.

Next Steps:

- Introduce streaming data ingestion for real-time detection.
- Experiment with ensemble methods for higher accuracy.
- Add alerting systems for anomaly spikes.