Arghyajit Debnath
AI/ML Intern of Team 74

TOPIC : Support post-launch AI tuning and troubleshooting
Status: Complete

Report Summary:

This project simulates a complete post-deployment AI monitoring and tuning system within the cybersecurity domain.
A synthetic dataset representing potential security threats (based on features like login attempts, unusual locations,
and data transfer volumes) is used to build an XGBoost classification model. The system follows a realistic MLOps flow,
which includes model training, live prediction simulation, monitoring for performance degradation and data drift,
retraining logic, and explainability using SHAP.

Key Features Implemented:

1. Initial Model Training:
  - Synthetic data is generated to simulate normal behavior.
  - A baseline XGBoost model is trained and evaluated using accuracy and classification reports.

2. Live Simulation:
  - The model performs predictions on a simulated stream of events (both normal and drifted).
  - It mimics a real-world scenario where feedback and real-time predictions are collected.

3. Monitoring and Retraining:
  - Accuracy and data drift are continuously monitored using statistical techniques like KS-test.
  - If performance drops or drift is detected, alerts are triggered.
  - A retraining mechanism replaces the current model only if the new one performs better.

4. Explainability and Troubleshooting:
  - SHAP (SHapley Additive exPlanations) is used to explain predictions locally.

- This enables transparent and auditable model decisions, which are critical in cybersecurity applications.

Outcome:
The model successfully adapts to incoming feedback and drifted data, demonstrating how MLOps practices such as
feedback loops, retraining strategies, and explainability tools can ensure the AI system remains effective post-launch.

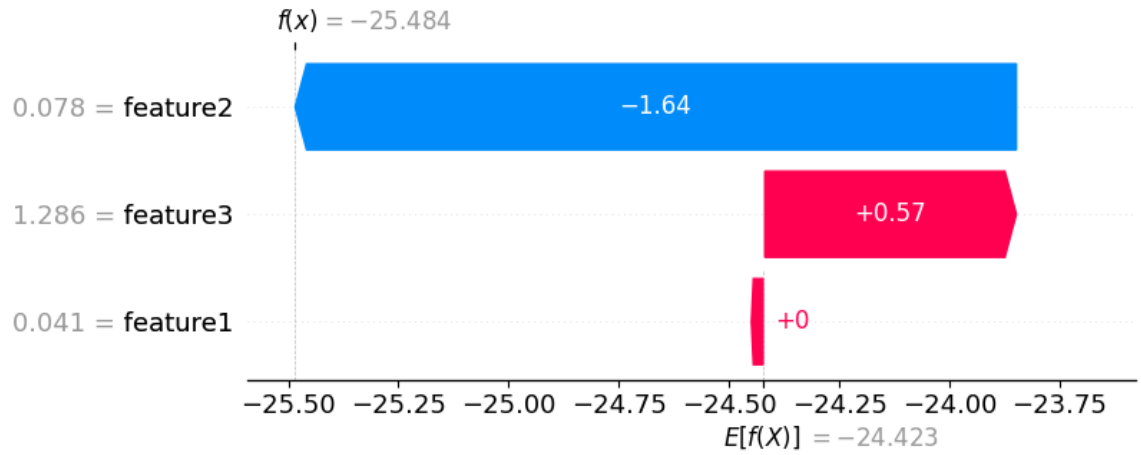
This report concludes that the simulation of post-launch AI tuning and troubleshooting is functionally complete and successful.

- Live Data Simulator

```
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 453.1/453.1 kB 15.7 MB/s eta 0:00:00
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.9/1.9 MB 45.0 MB/s eta 0:00:00
  Building wheel for lime (setup.py) ... done
2025-07-13 16:40:59.498 | INFO     | __main__:generate_new_live_data_batch:41 - Simulating a batch of 5 new live data points.
2025-07-13 16:40:59.524 | INFO     | __main__:generate_new_live_data_batch:51 - Generated 5 new data points. 5 with immediate actuals.
2025-07-13 16:40:59.535 | INFO     | __main__:<cell line: 0>:68 - Generated new data batch:
   feature_1  feature_2  feature_3
0   0.496714         74   0.650888
1  -0.138264         74   0.056412
2   0.647689         87   0.721999
3   1.523030         99   0.938553
4  -0.234153         23   0.000779
2025-07-13 16:40:59.538 | INFO     | __main__:<cell line: 0>:69 - Actuals available in this batch: [(0, np.int64(1)), (1, np.int64(0)),
```

- Monitoring and Troubleshooting Pipeline

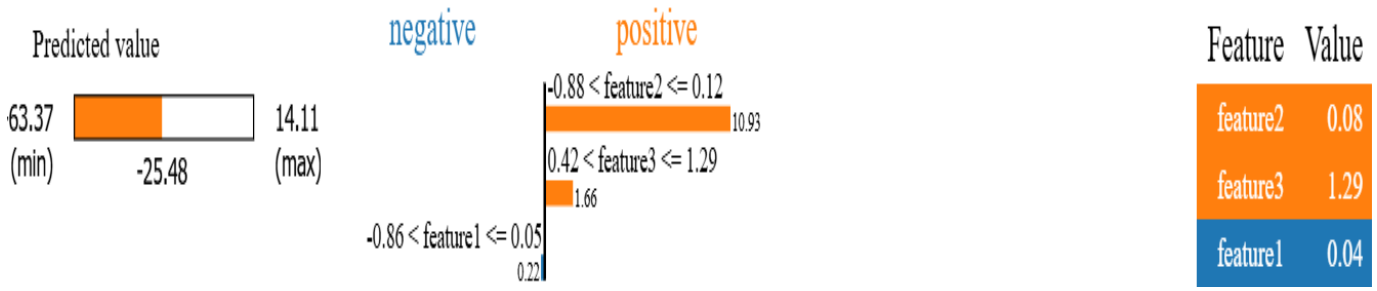
MSE: 2.56, R2: 0.99



Intercept -26.654547751167225

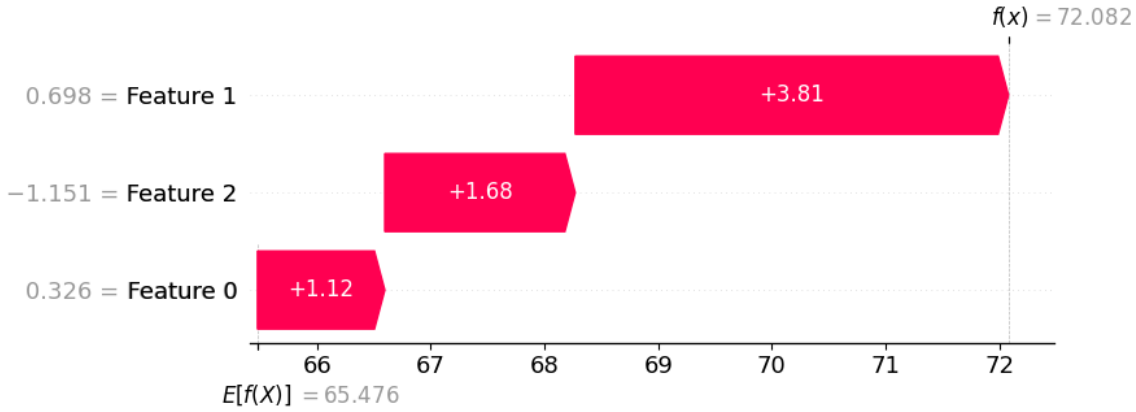
Prediction_local [-14.28463568]

Right: -25.483995



- Main Orchestrator for Continuous AI Loop

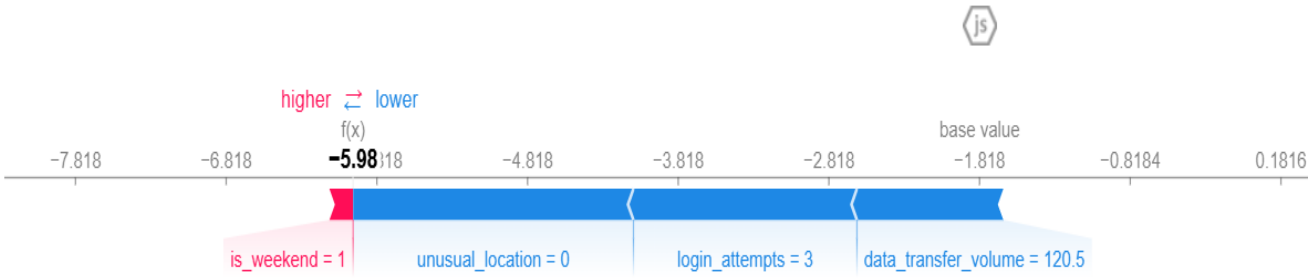
```
✅ Packages loaded
[INFO] Initial Model MSE: 30.54
[INFO] ---- Running Live Predictions ----
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Pred: 67.79, Actual: 69.38
[INFO] Collected feedback points: 20
[INFO] New model did not improve enough. Keeping old model.
[INFO] Generating SHAP explainability on 1 test instance
ExactExplainer explainer: 2it [00:11, 11.50s/it]
```



- AI Tunin And Monitoring System
- FOR SIMULATION - 100 EVENT

```
✅ Packages loaded
[INFO] AI Tuning and Monitoring System initialized.
[INFO] Starting initial model training...
[SUCCESS] Initial model trained. Accuracy: 1.00

========================================
    Initial Model Performance Report
========================================
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       160
           1       1.00      1.00      1.00        40

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

========================================


--- Simulating 100 predictions under NORMAL conditions ---
[INFO] Starting live simulation for 100 events...
[INFO] Live simulation finished. Starting monitoring and tuning cycle.
[INFO] Live monitoring: Accuracy over last 100 events = 1.00
[INFO] Retraining threshold reached. Attempting to retrain model with new data.
[INFO] Retraining complete. Old Model Accuracy: 1.00, New Model Accuracy: 1.00
[INFO] New model did not outperform the old one. Keeping the current model.

--- Simulating 100 predictions under DRIFTED conditions ---
[INFO] Starting live simulation for 100 events...
```

```
--- Simulating 100 predictions under DRIFTED conditions ---
[INFO] Starting live simulation for 100 events...
[INFO] Live simulation finished. Starting monitoring and tuning cycle.
[INFO] Live monitoring: Accuracy over last 100 events = 0.99
[WARNING] ALERT: Data drift detected in 'login_attempts' feature (p-value: 0.0000). Model may be inaccurate.
[INFO] Retraining threshold reached. Attempting to retrain model with new data.
[INFO] Retraining complete. Old Model Accuracy: 1.00, New Model Accuracy: 1.00
[INFO] New model did not outperform the old one. Keeping the current model.

--- Troubleshooting a sample prediction ---
[INFO] Generating SHAP explanation for a specific instance.
```




```
Explanation:
- The plot above shows which features pushed the prediction higher (in red) or lower (in blue).
- The base value (-1.82) is the average prediction over the entire dataset.
- Features in red increased the likelihood of a malicious prediction.
- Features in blue decreased the likelihood.
```

- AI Tunin And Monitoring System
- FOR SIMULATION – 50 EVENTS

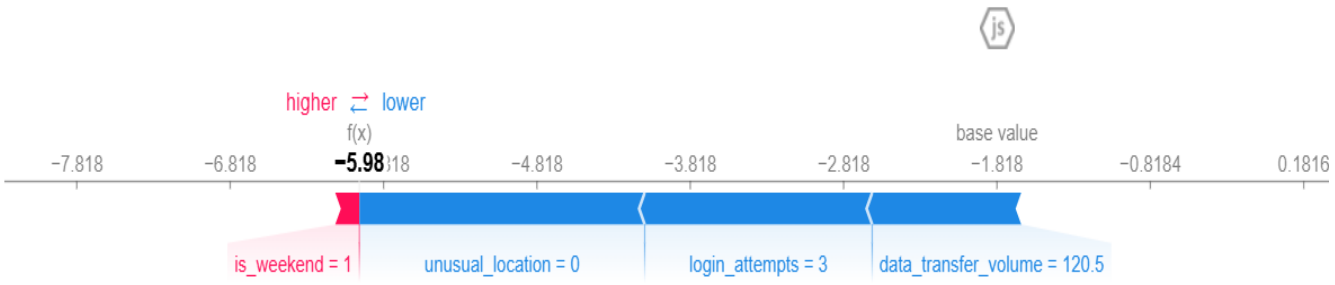
✅ Packages loaded
[INFO] AI Tuning and Monitoring System initialized.
[INFO] Starting initial model training...
[SUCCESS] Initial model trained. Accuracy: 1.00

```
========================================
    Initial Model Performance Report
========================================
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       160
           1       1.00      1.00      1.00        40

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

========================================
```

--- Simulating 50 predictions under NORMAL conditions ---
[INFO] Starting live simulation for 50 events...
[INFO] Live simulation finished. Starting monitoring and tuning cycle.
[INFO] Live monitoring: Accuracy over last 50 events = 1.00
[INFO] Retraining threshold reached. Attempting to retrain model with new data.
[INFO] Retraining complete. Old Model Accuracy: 1.00, New Model Accuracy: 1.00
[INFO] New model did not outperform. Keeping current model.

--- Simulating 50 predictions under DRIFTED conditions to trigger tuning ---
[INFO] Starting live simulation for 50 events...

```
--- Simulating 50 predictions under DRIFTED conditions to trigger tuning ---
[INFO] Starting live simulation for 50 events...
[INFO] Live simulation finished. Starting monitoring and tuning cycle.
[INFO] Live monitoring: Accuracy over last 50 events = 1.00
[WARNING] ALERT: Data drift detected in 'login_attempts' feature (p-value: 0.0000). Model may be inaccurate.
[INFO] Retraining threshold reached. Attempting to retrain model with new data.
[INFO] Retraining complete. Old Model Accuracy: 1.00, New Model Accuracy: 1.00
[INFO] New model did not outperform. Keeping current model.

--- Troubleshooting a sample prediction ---
[INFO] Generating SHAP explanation for a specific instance.
```




```
Explanation:
- The base value (-1.82) is the model's average prediction.
- Features in red (like high login_attempts) push the prediction towards 'malicious'.
- Features in blue pull the prediction away from 'malicious'.
```