AI Model Documentation & Regulatory Compliance Summary

1. Overview

This documentation outlines the functionality, performance characteristics, and regulatory compliance of the deployed AI-based cybersecurity threat detection system. The AI model is specifically designed to identify anomalies in cybersecurity incident data and assign a corresponding cyber risk score to each event. The model is deployed behind a FastAPI interface, which serves as the API endpoint for structured incident data input and returns both an anomaly classification and a numeric risk score (0–100 scale).

The system plays a critical role in helping organizations detect potentially suspicious behavior and prioritize incident response efforts in real-time. It is intended to augment security analysts by providing early warning signals for high-risk events without processing personally identifiable information (PII).

2. Model Description

- Type: Isolation Forest (Unsupervised Anomaly Detection)
- Framework: scikit-learn
- Purpose: Detect abnormal patterns in cybersecurity incidents and flag potential risks
- Training Dataset: Historical cybersecurity incident logs (collected from 2015 to 2024)
- Total Records Used: ~3,000 curated entries
- Model Retraining Cycle: Quarterly (based on updated incident logs and new threat vectors)

The model was selected due to its effectiveness in unsupervised environments where labeled anomalies are unavailable. Isolation Forest leverages the principle of isolating rare points to detect outliers efficiently.

3. Input Features

The following structured features are used for both anomaly detection and cyber risk scoring:

- Financial Loss ($M) - Numerical: Estimated monetary loss associated with the incident
- Number of Affected Users - Numerical: Number of impacted users or customers
- Resolution Time (hours) - Numerical: Duration taken to fully resolve the incident
- Country - Categorical: Country of origin for the cyberattack
- Attack Type - Categorical: Category of attack (e.g., ransomware, phishing)
- Attack Source - Categorical: Internal or external origin
- Vulnerability Type - Categorical: Known vulnerability exploited (e.g., Zero-Day)

4. Data Preprocessing Summary

To ensure robust performance and consistency during inference, the following preprocessing pipeline is applied:

- Categorical Features: Encoded using LabelEncoder from scikit-learn
- Numerical Features: Standardized using StandardScaler
- Scalability: Preprocessing pipeline is modular and easily adaptable for new features or transformations

Preprocessing components are version-controlled independently to support reproducibility and streamlined updates.

5. Model Performance Summary

- Contamination Rate: 0.05 (5% expected anomalies)
- Anomalies Detected: ~150 out of 3,000 records
- Prediction Latency: ~5–10 ms per record
- Evaluation Methodology: Visual analysis + expert review

Due to the lack of labeled ground truth, model validation was conducted through manual inspection of anomaly clusters and logical interpretation of the isolation scores. Internal testing scripts visualized the distribution of anomaly scores and confirmed separation of normal vs. abnormal cases. Risk scoring was verified using domain knowledge-based rules aligned with threat intelligence.

6. Risk Scoring Methodology

To provide actionable insights, a rule-based scoring engine complements the unsupervised model. This system quantifies cyber incidents on a 0–100 severity scale.

Scoring Logic Includes:

- Magnitude of Financial Loss
- Extent of Impact (user count)
- Incident Resolution Time
- Geopolitical Risk (e.g., high-risk countries like Russia, Iran, North Korea)
- Criticality of Attack Type (e.g., ransomware, APTs, supply-chain attacks)

Scoring thresholds and rules are embedded in the API layer for full transparency and auditability. The scoring system allows organizations to prioritize incidents without requiring additional manual input.

7. Regulatory Compliance Alignment

The deployed AI system was built with data protection and transparency principles in mind. Its architecture and data handling practices align with global regulatory frameworks:

- Data Minimization: Compliant – Only non-PII, operational metadata is processed
- GDPR / Data Privacy: Compliant – No personal identifiers collected, stored, or returned
- Explainability (Right to Explanation): Compliant – Risk scores derived from published rule sets
- Model Versioning & Reproducibility: In Place – All model and preprocessing components are version-controlled
- Logging & Auditing: Partially Enabled – API logging configurable; persistent logging to be added

The system ensures full traceability through version control of all components. Logs can be enabled for incident-level audits or debugging without exposing sensitive content.

8. Model Management & Deployment

- Retraining Schedule: Every 3 months or upon emergence of major cyber threats
- Deployment Interface: FastAPI with JSON-based input/output
- Scalability: Can be containerized using Docker and deployed on cloud infrastructure (e.g., AWS, Azure, GCP)

Version control ensures reproducibility, rollback, and governance. Updates to the model are reviewed and tested internally before deployment.

9. Future Enhancements

To improve accuracy and regulatory robustness, the following upgrades are planned:

- Integration with real-time threat intelligence feeds
- Automated retraining and model drift detection
- Persistent logging and role-based access controls (RBAC)
- Enhanced visualization dashboard for analysts
- CI/CD pipeline for seamless model deployment and monitoring

10. Conclusion

This anomaly detection and risk scoring system offers an interpretable, fast, and regulation-aligned AI solution for cybersecurity incident analysis. By combining unsupervised learning with rule-based scoring, the platform provides reliable insights while maintaining full transparency and data compliance. Future iterations will focus on automation, scalability, and integration with broader security information ecosystems.