AI Model Documentation & Regulatory Compliance Summary

1. Overview

This document outlines the performance characteristics and compliance alignment of the deployed AI model used to detect anomalies in cybersecurity threat data and generate cyber risk scores.

The model is deployed behind a FastAPI interface and accepts structured inputs related to cyber incidents. It returns both an anomaly prediction and a numeric risk score.

2. Model Description

- Type: Isolation Forest (unsupervised anomaly detection)
- Framework: scikit-learn
- Purpose: Detect abnormal patterns in cyber incidents to flag suspicious cases
- Training Dataset: Historical cybersecurity threat dataset (2015-2024)
- Total records used: ~3,000 entries

3. Input Features

Feature	Type	Description
Financial Loss ($M)	Numerical	Estimated financial loss from incident
Number of Affected Users	Numerical	Impacted users
Resolution Time (hours)	Numerical	Time to resolve incident
Country	Categorical	Origin of attack
Attack Type	Categorical	Type of attack (e.g., ransomware, phishing)
Attack Source	Categorical	Internal or external source
Vulnerability Type	Categorical	e.g., Zero-Day, SQL Injection

4. Preprocessing Summary

- All categorical features encoded using LabelEncoder
- Numerical features scaled using StandardScaler
- Features used for both anomaly detection and risk scoring

5. Model Performance Summary

Metric	Value
Contamination Rate	0.05 (5% assumed anomalies)
Anomalies Detected	~150 of 3,000
Prediction Time	~5-10 ms per record
Evaluation Method	Visual + Manual Inspection

No labeled ground truth was available for anomaly classification. Output interpretation is based on isolation scores and downstream risk impact.

Additional performance considerations: - Model performance was verified using internal testing scripts and visual analysis of anomaly clusters. - Risk scoring was validated through rule-based scoring criteria and reviewed for consistency.

---

6. Risk Scoring

A supplemental rule-based scoring system was developed to quantify incident severity:

- Score range: 0 to 100
- Key contributing factors:
  - Financial Loss magnitude
  - Number of users affected
  - Resolution time
  - High-risk countries (e.g., Russia, Iran)
  - Critical attack types (e.g., ransomware, APTs)

---

7. Regulatory Compliance

Regulation / Principle	Status	Notes
Data Minimization	✅	Only operational metadata is processed; no PII included
GDPR / Data Privacy	✅	No personal identifiers collected, stored, or returned
Explainability (Right to Explanation)	✅	Risk scoring is transparent, rules are published in API logic
Model Versioning & Reproducibility	✅	Model and feature files version-controlled and stored separately
Logging & Auditing	🟡	Logs can be enabled in API server, currently not persisted

Documenting compliance includes: - Identifying applicable regulatory frameworks (e.g., GDPR, IRDAI for cyber insurance) - Demonstrating non-PII data usage - Verifying that models produce interpretable output (e.g., risk levels) - Ensuring model artifacts (joblib files) are versioned for auditability

---

8. Model Management

- Model is saved as: isolation_forest_model.joblib
- Label encoders and scaler saved in separate joblib files
- Feature columns stored and versioned: feature_columns.joblib
- Retraining planned quarterly based on new threat data
