42

RISK LEVEL: MEDIUM

Model: fraud_credit_model | Scanned: 2026-05-21 08:44:49

What Was Assessed

This report covers an automated security audit of the AI model "fraud_credit_model". The audit tested the model against 8 categories of security threats used by real-world attackers, including adversarial manipulation, privacy extraction, and model theft. The audit was completed in under 60 seconds.

Key Findings

Input Manipulation Risk — HIGH

The model's accuracy dropped under adversarial input manipulation. Attackers who understand your model can craft inputs that systematically reduce its effectiveness.

Privacy Risk - Training Data Exposure — MEDIUM

With 73.33% accuracy, an attacker can determine whether specific individuals were in the training dataset. This is a potential GDPR violation if personal data was used in training.

Model Theft Risk — HIGH

A competitor or attacker achieved 100.0% agreement by querying your model repeatedly. Your proprietary model logic and training investment could be replicated without access to your data or code.

Immediate Actions Required

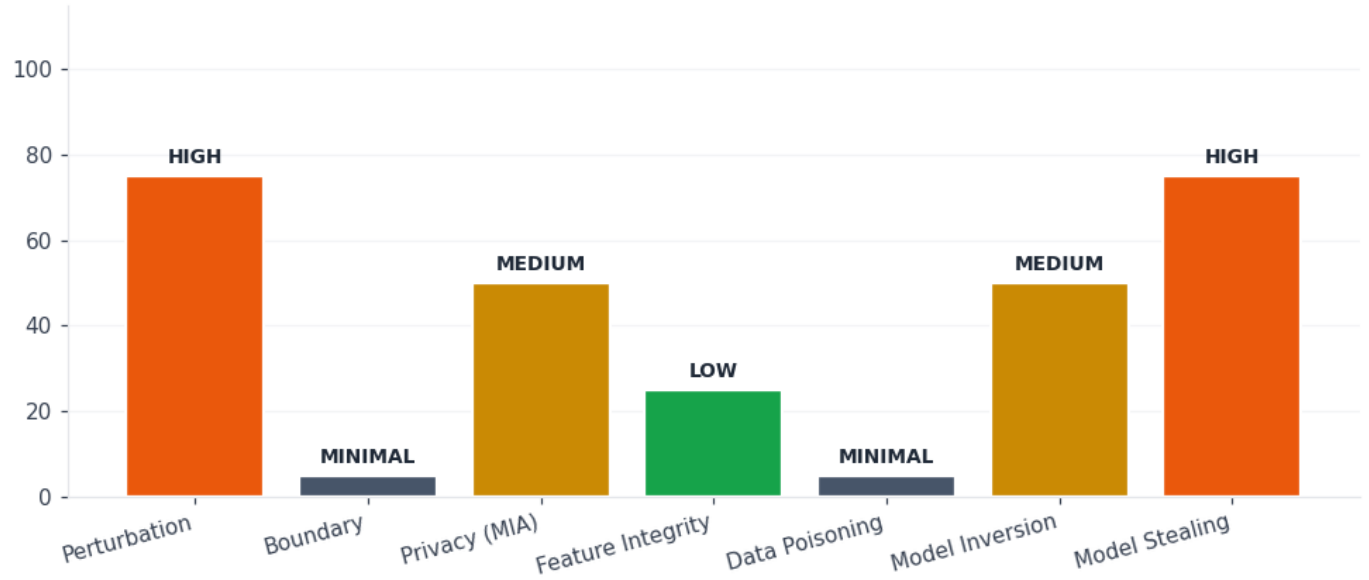
1. Rate limit your prediction API to prevent model theft via systematic querying.
2. Review GDPR obligations - training data privacy risk detected. Consider a Data Protection Impact Assessment (DPIA).

Regulatory Flags

Based on the findings in this report, the following regulatory frameworks may be relevant to your organisation. Full mapping is provided in Section 9 of this report.

EU AI Act 2024	Articles 9, 10, 13, 15, 72 may apply depending on risk classification of your AI system.
GDPR	Articles 5, 25, 32, 35 are relevant if personal data was used in model training.
ISO/IEC 42001	Clauses 6.1, 8.4, 9.1 documentation obligations can be partially satisfied by this report.

This executive summary is intended for non-technical stakeholders. Full technical findings begin on the following page.



This report presents the results of an automated AI security audit conducted on: fraud_credit_model. The scanner performed 8 security checks including adversarial attack simulation, privacy analysis, model stealing detection, and feature integrity assessment. The model received an overall risk score of 42/100 with a risk level of MEDIUM.

1. Baseline Model Performance

ROC-AUC Score	**0.9867**	PASS
Accuracy	**98.0%**	
Status		PASS

2. Adversarial Attack Results

Attack 1 - Feature Perturbation

AUC after attack	**0.9733**	
AUC drop	**1.35%**	
Vulnerability		HIGH

Attack 2 - Boundary Search

AUC after attack	**0.9867**	
AUC drop	**0.0%**	
Vulnerability		MINIMAL

3. Privacy Risk - Membership Inference

Inference accuracy	**73.33%**	
Confidence threshold	**0.9834**	
Privacy risk		MEDIUM

Note: 50% = random guessing. Scores above 70% indicate the model leaks information about its training data, creating potential GDPR compliance risk.

4. Feature Integrity Analysis

Dominant features	**1**
Top feature index	**V14**
Top feature weight	**20.68%**
Anomaly level	**LOW**

5. Data Poisoning Detection

Risk level	**MINIMAL**
Tests run	**3**

Findings:

- No poisoning indicators detected

6. Model Inversion Attack

Risk level	**MEDIUM**
Inverted confidence	**0.8867**
Min distance to real data	**5.2687**
Feature exposure	**0.309**

Note: High inverted confidence means an attacker can craft inputs the model confidently classifies as the target class, revealing information about the decision boundary.

7. Model Stealing Attack

Risk level	**HIGH**
Model agreement	**100.0%**
Stolen model AUC	**1.0**
Fidelity score	**1.0**

Note: High agreement means an attacker can replicate your model logic using only prediction queries. This exposes your training investment and proprietary model architecture.

8. Recommendations

Privacy Protection - Membership Inference

The model shows a 73.33% membership inference accuracy, indicating training data leakage. Recommended actions: (1) Apply differential privacy during training using DP-SGD. (2) Limit prediction API confidence scores - return only class labels, not probabilities. (3) Implement output perturbation to add calibrated noise to prediction outputs. This is critical for GDPR compliance.

Feature Dependency Risk

Feature V14 accounts for 20.68% of model decisions. This creates a single point of failure - if an attacker identifies and manipulates this feature, model performance degrades significantly. Recommended: retrain with max_features='sqrt' and increase min_samples_leaf to reduce single-feature dominance.

Adversarial Robustness

Although evasion attacks showed minimal impact on this model, production models face more sophisticated threats. Recommended: (1) Implement adversarial training by including perturbed samples in retraining. (2) Add input validation to detect out-of-distribution inputs before they reach the model. (3) Monitor prediction confidence distributions in production for anomaly detection.

Ongoing Monitoring

Deploy model monitoring to detect distribution shift in production inputs. Set alerts for: (1) sudden drop in prediction confidence, (2) unusual feature value distributions, (3) spike in low-confidence predictions which may indicate adversarial probing.

Re-audit Schedule

AI models should be re-audited: (1) every 6 months as standard, (2) immediately after any retraining on new data, (3) after any significant change in input data distribution, (4) before deployment to a new environment or use case.

Model Stealing Protection

The stolen model achieved 100.0% agreement with your original model. Recommended actions: (1) Rate limit your prediction API - restrict number of queries per user per day. (2) Return only class labels instead of probability scores where possible. (3) Add prediction perturbation - add small random noise to output probabilities to make cloning harder. (4) Monitor for unusual query patterns that may indicate systematic probing.

Model Inversion Protection

Model inversion confidence reached 0.8867. Recommended actions: (1) Reduce output precision - round probability scores to 2 decimal places maximum. (2) Implement query budgets per user session. (3) Add input validation to reject out-of-distribution inputs before they reach the model. (4) Consider using prediction confidence thresholds - only return high confidence predictions.

9. Regulatory & Compliance Mapping

The following table maps each security finding to relevant regulatory frameworks. This is intended to assist compliance teams in prioritising remediation based on legal obligations.

EU AI Act (2024)

Article 9 - Risk Management System — HIGH

Relevant checks: Perturbation Attack, Boundary Attack

High-risk AI systems must have a risk management system identifying and analysing known and foreseeable risks. Adversarial vulnerability must be documented and mitigated.

Article 10 - Data Governance — MINIMAL

Relevant checks: Data Poisoning Detection

Training data must be examined for possible biases and errors. Data poisoning indicators directly trigger obligations under data governance requirements.

Article 13 - Transparency & Explainability — LOW

Relevant checks: Feature Integrity Analysis

High-risk AI systems must be transparent enough for users to interpret output. Dominant single features reduce explainability and may breach this obligation.

Article 15 - Accuracy & Robustness — HIGH

Relevant checks: All adversarial attack checks

High-risk AI systems must be designed to achieve appropriate levels of accuracy and be resilient to errors, faults, and adversarial inputs.

Article 72 - Post-market Monitoring — HIGH

Relevant checks: Model Stealing Attack

Providers must implement post-market monitoring. High model stealing fidelity indicates the model can be replicated, undermining IP and monitoring integrity.

GDPR (General Data Protection Regulation)

Article 5 - Principles of Data Processing — MEDIUM

Relevant checks: Membership Inference, Model Inversion

Personal data must be processed with integrity and confidentiality. A membership inference accuracy of 73.33% was recorded. This indicates potential training data leakage and may violate data minimisation principles.

Article 25 - Data Protection by Design — MEDIUM

Relevant checks: Model Inversion Attack

Controllers must implement data protection by design. Model inversion vulnerability means the system was not designed with sufficient privacy protections from the outset.

Article 32 - Security of Processing — MEDIUM

Relevant checks: Membership Inference, Data Poisoning

Appropriate technical measures must ensure security of personal data. Membership inference and data poisoning vulnerabilities represent failures of this obligation and may require notification to supervisory authorities.

Article 35 - Data Protection Impact Assessment — MEDIUM

Relevant checks: All privacy-related checks

A DPIA is required for high-risk processing. The privacy vulnerabilities identified in this report are likely to trigger DPIA requirements if personal data was used in model training.

ISO/IEC 42001 - AI Management System Standard

Clause 6.1 - Actions to Address Risks — MEDIUM

Relevant checks: All attack checks

Organisations must identify AI-related risks and plan actions to address them. This report provides the technical evidence required to satisfy Clause 6.1 documentation requirements.

Clause 8.4 - AI System Impact Assessment — MEDIUM

Relevant checks: Privacy checks, Model Stealing

Impact assessments must consider how AI systems affect individuals and organisations. Membership inference and model stealing findings directly inform this assessment.

Clause 9.1 - Monitoring & Measurement — LOW

Relevant checks: Baseline Performance, Feature Integrity

Organisations must monitor AI system performance and security over time. Baseline AUC and feature integrity metrics in this report serve as documented benchmarks for ongoing monitoring obligations.

Important Notice

This regulatory mapping is provided for informational purposes and should be reviewed by a qualified legal or compliance professional before making regulatory decisions. Applicability depends on your organisation's specific use case, data types, and jurisdictions.