A 98% Accurate Model With Three Critical Vulnerabilities

How OrbTech audited a production-grade credit card fraud detection model — and what it found beyond the accuracy score.

Dataset	Model	Scan time	Risk level
Kaggle Credit Card Fraud 284,807 transactions	Random Forest (sklearn) 100 estimators, max_depth 10	Under 60 seconds 8 checks automated	MEDIUM (42/100) 2 HIGH findings

1. Background

This case study documents an OrbTech security audit of a credit card fraud detection model trained on the Kaggle Credit Card Fraud dataset — one of the most widely used public benchmarks in financial ML. The model represents the kind of production-grade classifier that Indian fintech teams routinely deploy: high accuracy, real transaction data, binary classification.

The audit was run to answer a single question: does an excellent accuracy score mean a model is safe to deploy under DPDP Act, GDPR, and EU AI Act requirements?

The short answer: no.

2. Model & Dataset Details

Parameter	Value
Dataset	Kaggle Credit Card Fraud (public)
Transactions	284,807 (492 fraudulent — 0.17% fraud rate)
Features	28 PCA-anonymised features + Amount + Time
Model type	Random Forest (scikit-learn 1.8.0)
Training config	100 estimators, max_depth=10, class_weight=balanced
Baseline ROC-AUC	0.9867
Baseline Accuracy	98.0%

3. Audit Findings — All 8 Checks

OrbTech ran all 8 security checks automatically in a single scan. Here are the findings:

Check	Result	Severity	Regulatory Relevance
Baseline Performance	ROC-AUC 0.9867, Accuracy 98.0%	PASS	EU AI Act Art.9
Feature Perturbation Attack	High accuracy drop under perturbation	HIGH	EU AI Act Art.15
Boundary Search Attack	Minimal boundary exploitability	MINIMAL	EU AI Act Art.15
Membership Inference (MIA)	73.3% inference accuracy	MEDIUM	GDPR Art.35, DPDP Act
Feature Integrity	No single-feature over-reliance	PASS	EU AI Act Art.13
Data Poisoning Detection	No poisoning anomalies detected	PASS	EU AI Act Art.10
Model Inversion Attack	Moderate inversion risk	MEDIUM	GDPR Art.5, Art.25
Model Stealing Attack	100% surrogate agreement	HIGH	IP / Trade Secret

4. Key Findings Explained

Finding 1 — Model Stealing (HIGH)

The surrogate model achieved 100% agreement with the original across test predictions. This means an attacker could reconstruct a functionally identical copy of the model purely by querying its prediction API — no access to training data or source code required. For a fintech team that has invested months of data science work, this represents a direct IP and competitive risk.

Regulatory relevance: while model stealing is not yet explicitly named in DPDP Act or EU AI Act, it constitutes misappropriation of a proprietary AI system and triggers IP protection obligations under Indian IP law.

Finding 2 — Membership Inference (MEDIUM)

The scanner achieved 73.3% membership inference accuracy — meaning it could determine whether a specific transaction record was part of the training dataset with 73.3% accuracy (random chance would be 50%). This is a real privacy vulnerability: if the training data included real customer transactions, an attacker could probe the model to confirm whether a specific individual's data was used.

Regulatory relevance: GDPR Article 35 requires a Data Protection Impact Assessment (DPIA) when processing personal data at scale. The DPDP Act 2023 imposes similar obligations on Indian data fiduciaries. A 73.3% MIA result is direct evidence that a DPIA is required.

Finding 3 — Feature Perturbation (HIGH)

Under deliberate input perturbation — adding small, crafted noise to transaction features — the model's fraud detection accuracy dropped significantly. This simulates the kind of adversarial manipulation a sophisticated fraudster would attempt: carefully crafting transactions to evade detection without triggering obvious anomalies.

Regulatory relevance: EU AI Act Article 15 requires that high-risk AI systems be designed to withstand adversarial manipulation. A fintech fraud model is a strong candidate for high-risk classification under Annex III.

> ***"A model that scores 98% accuracy can still leak private data, be stolen, and be fooled by adversarial inputs. Accuracy measures performance. It says nothing about security."***

5. Regulatory Mapping

Each finding maps to specific regulatory obligations relevant to Indian fintech and healthtech teams:

Regulation	Article / Section	Triggered By	Required Action
GDPR	Article 35 (DPIA)	73.3% MIA result	Conduct Data Protection Impact Assessment
GDPR	Article 25 (Privacy by Design)	Model inversion risk	Implement data minimisation in training pipeline
DPDP Act 2023	Section 8 (Data Fiduciary)	MIA + inversion	Document processing basis; DPIA if sensitive data
EU AI Act	Article 9 (Risk Management)	Overall MEDIUM risk	Maintain risk register; document mitigation measures
EU AI Act	Article 15 (Robustness)	Feature perturbation HIGH	Test adversarial robustness; document results
ISO 42001	Clause 6.1 (Risk Assessment)	All findings	Include AI security in formal risk assessment

6. Recommendations

Model Stealing	Implement prediction rate limiting and output perturbation on the inference API. Do not return raw probability scores — return binned outputs (e.g. LOW/MEDIUM/HIGH) where possible.
Membership Inference	Apply differential privacy techniques during training, or increase regularisation (reduce max_depth, increase min_samples_leaf). Re-run MIA check after changes to verify improvement.
Feature Perturbation	Add adversarial training examples to the training pipeline. Introduce input validation to reject statistically anomalous feature vectors before inference.
Ongoing Compliance	Run a security audit after every major model retraining or dataset update. Document results as part of your AI risk register for EU AI Act and ISO 42001 compliance.

7. About OrbTech

OrbTech is an AI security auditing tool built for Indian fintech and healthtech ML teams. It runs 8 automated security checks — adversarial attacks, privacy analysis, model theft simulation, data integrity — on any sklearn, XGBoost, LightGBM, or Keras model, in under 60 seconds.

Every scan generates a PDF audit report with risk scores, plain-English findings for compliance teams, technical details for engineers, and regulatory mapping to EU AI Act, GDPR, ISO 42001, and DPDP Act

2023.

Run OrbTech on your own model

DM Shubham Kumar on LinkedIn or email security@orbtech.in to request a free invite code. The scan takes about 60 seconds.

orbtech.in