JUPITER PRIME

Consulting • Training • Advisory

جوبيتر برايم للاستشارات والتدريب

COURSE BROCHURE

Adversarial Security for AI Fraud Detection Training Course

jupiterprime.uk

info@jupiterprime.uk

Registered in the United Kingdom and Saudi Arabia

Adversarial Security for AI Fraud Detection Training Course

Adversarial Security for AI Fraud Detection is a 1-week training course for fraud analysts, security officers and risk managers who must keep detection models trustworthy when criminals probe them, and it produces a fraud model threat register and hardening plan. Fraudsters now test scoring systems deliberately, feeding crafted transactions, poisoned labels and stolen credentials to slip past machine learning controls. Jupiter Prime Training runs Adversarial Security for AI Fraud Detection as a security-first course.

Why Adversarial Security for AI Fraud Detection Matters Now

Three groups shape the subject. Organised fraud groups share evasion techniques and adapt within days when a model blocks them. Security teams, used to protecting networks and endpoints, now own models and training pipelines as new attack surfaces. Risk committees and auditors expect evidence that detection still works after attacks and data drift. A hardened model limits losses and false declines; an exposed one silently waves through fraud while customers absorb friction. The course draws on ISO/IEC 42001, the AI management system standard, and the NIST AI Risk Management Framework AI RMF to structure threat assessment and controls for detection models.

Who Gains Most from Adversarial Security for AI Fraud Detection

- Fraud analysts who investigate cases and feed labels back to models
- Security officers who protect data pipelines, credentials and model endpoints
- Risk managers who set loss appetite and review control effectiveness
- Data scientists who retrain and validate anomaly and scoring models
- Operations leads who run investigation queues and service levels

A combined fraud and security group attending together gains most, since attack scenarios and model controls then sit in one shared register.

Adversarial Security for AI Fraud Detection Capabilities

Participants practise viewing a fraud model as an asset under attack and putting proportionate controls around it.

- Assess attack paths against detection models with a structured threat model — security architects
- Test model robustness with evasion and poisoning simulations — data science leads
- Specify data protection controls for training sets and feature stores — privacy officers
- Design case triage rules that route suspicious bypass patterns to investigators — fraud operations
- Build monitoring for drift, attack signals and score manipulation — risk committees
- Communicate residual threats and hardening priorities — senior management

Key Adversarial Security for AI Fraud Detection Terms

- Evasion attack — Crafting inputs, such as split or disguised transactions, so a model scores fraud as legitimate.
- Data poisoning — Inserting false or mislabelled records into training data to weaken future detection.
- Model extraction — Querying a model repeatedly to copy its behaviour or learn its thresholds.
- Anomaly detection — Methods that flag behaviour unusually far from normal patterns without needing labelled fraud cases.
- Concept drift — A change in real-world behaviour that makes a trained model less accurate over time.
- Red teaming — Planned, authorised attempts to defeat controls in order to find weaknesses before criminals do.

Adversarial Security for AI Fraud Detection Day by Day

Day 1: The Fraud Model as an Attack Surface

- Detection Pipeline Mapping — Locating Data, Model and Decision Points
- Threat Modelling With STRIDE — Listing Attacks Against Each Component
- Insider and Credential Risks — Protecting Access to Rules and Scores
- ISO/IEC 42001 and NIST AI RMF — Structuring Model Risk Controls
- Fraud Model Threat Register — Recording Assets, Threats and Owners

Day 2: Detection Methods and Their Weak Points

- Supervised Scoring Versus Unsupervised Anomaly Detection — Choosing Coverage
- Isolation Forests and Autoencoders — Spotting Unseen Patterns
- Feature Engineering Exposure — Judging Which Signals Attackers Can Fake
- Threshold Setting and Score Bands — Balancing Losses and False Declines
- Label Quality and Feedback Loops — Preventing Corrupted Ground Truth

Day 3: Adversarial Attacks and Defences

- Evasion Simulations — Testing Split, Disguised and Velocity-Shaped Transactions
- Data Poisoning Scenarios — Detecting Tampered Training Records
- Model Extraction and Probing — Rate Limiting and Response Shaping
- Adversarial Training and Ensembles — Hardening Models Against Manipulation
- Red Team Exercises — Planning Authorised Attacks on Controls

Day 4: Data Protection, Triage and Monitoring

- Training Data Encryption and Access Control — Securing Feature Stores
- Pseudonymisation of Customer Data — Limiting Exposure in Analysis
- Bypass Pattern Triage — Routing Suspected Evasion to Investigators
- Drift and Attack Signal Dashboards — Watching Score Distributions
- Incident Response for Model Compromise — Rolling Back and Retraining

Day 5: Practising the Fraud Model Threat Register and Hardening Plan

- Case Pipeline Threat Model — Applying STRIDE to a Detection Flow
- Case Attack Simulation — Applying Evasion and Poisoning Tests
- Case Control Selection — Applying Data Protection and Monitoring
- Risk Committee Briefing — Presenting Residual Threats and Priorities
- Fraud Model Threat Register and Hardening Plan — Finalising Controls and Owners

Applying Adversarial Security for AI Fraud Detection After the Course

- Run a STRIDE threat modelling session on the live fraud detection pipeline with security and data science colleagues.
- Schedule a small red team evasion test on one scoring model and log the findings in the threat register.

Adversarial Security for AI Fraud Detection Questions FAQ

What is Adversarial Security for AI Fraud Detection and why does it matter now?

It is the practice of protecting fraud detection models from evasion, poisoning and theft. It matters now because organised fraud groups test and adapt to models quickly, turning machine learning controls into targets.

Which roles gain most from Adversarial Security for AI Fraud Detection?

Fraud analysts, security officers and risk managers gain most, along with data scientists who retrain models and operations leads who run investigation queues.

How does Adversarial Security for AI Fraud Detection differ from a banking fraud analytics course?

It treats the model itself as the asset at risk. A fraud analytics course builds detection and tunes alerts; this course covers threat modelling, adversarial testing, data protection and monitoring for compromise.

What do you bring back from Adversarial Security for AI Fraud Detection?

You bring back a threat register and hardening plan built on a case, with attack scenarios, test results, chosen controls and named owners.

Adversarial Security for AI Fraud Detection: The Way Forward

The fraud model threat register and hardening plan first informs the risk committee's decision on which controls to fund. Security, fraud and data science staff then implement and test them. After the first red team exercise, revisit attack scenarios and drift alerts, since tactics in sectors such as e-commerce payments or insurance claims shift quickly.