International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Adversarial robustness of AI systems in financial infrastructure: An empirical study of evasion attacks and defensive training on the Adult Dataset

Samy El Amali *

Department of Financial Engineering, HEC Montréal, Montréal, Quebec, Canada.

International Journal of Science and Research Archive, 2026, 19(03), 1139-1153

Publication history: Received on 19 May 2026; revised on 28 June 2026; accepted on 30 June 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.3.1404

Abstract

Machine learning models are now embedded across the operational core of financial infrastructure: trading and forecasting engines, credit-scoring pipelines, and fraud and anti-money-laundering screens. A decade of research on adversarial examples has shown that high-accuracy classifiers can be flipped by perturbations that are imperceptible or economically negligible, yet the implications for financial deployments—where inputs are partly attacker-controlled and decisions move money—remain unevenly understood. We characterise the adversarial threat surface of financial AI systems and quantify, in a controlled and fully reproducible setting, how much predictive accuracy degrades under gradient-based evasion attacks and how much of that loss adversarial training recovers. We formalise a threat model spanning white-box and black-box adversaries with a feature-space perturbation budget, and give self-contained derivations of the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). On the real, public UCI Adult income-classification dataset—a standard tabular credit-style benchmark—we attack several model families with FGSM and PGD across a range of $\ell\infty$ perturbation magnitudes and PGD step counts, then retrain with a min–max adversarial objective and re-measure robustness, including an adversarial-training-strength ablation and a cross-model transfer study. Clean test accuracy of the baseline model is high (85.3%) but degrades steadily under attack: PGD restricted to the six continuous features drives accuracy from 85.3% to 79.4% at ε = 0.10 and to 64.8% at ε = 0.30. Adversarial training recovers nearly all of the lost robust accuracy—raising PGD-attacked accuracy at ε = 0.10 from 79.4% to 83.9% and at ε = 0.30 from 64.8% to 81.3%—at a clean-accuracy cost of under half a point (85.3% → 84.9%). White-box perturbations transfer with reduced but non-negligible potency to independently trained models. We argue for robustness-aware evaluation, perturbation-budget reasoning grounded in market microstructure, and defence-in-depth that does not rely on adversarial training alone. All numbers are produced on the public Adult dataset by the included script; we make no measurements of any live system.

Keywords: Adversarial robustness; Financial machine learning; Evasion attacks; Projected gradient descent; Adversarial training; Transferability.

1. Introduction

Financial infrastructure has quietly become a machine-learning substrate. Order-flow forecasting from limit-order-book data [11, 12], return prediction with deep sequence models [4], cross-sectional asset pricing via flexible learners [6], and an expanding catalogue of credit, fraud, and anti-money-laundering classifiers now sit on the critical path between a market event and a capital allocation. As these models migrate from research notebooks into production decisioning, an old and uncomfortable result resurfaces with new stakes: neural networks are brittle in directions an adversary can find. Small, deliberately chosen perturbations to an input can move a confident, accurate model to a confident, wrong answer [5].

* Corresponding author: Samy El Amali.

In computer vision, where adversarial examples were first sharply characterised, the perturbation is a barely visible change to pixels. In finance the analogue is more consequential and, in important cases, more attainable. Many model inputs are not passive observations of nature but quantities that market participants partially produce: quotes, cancellations, the timing and sizing of orders, the textual content of filings and disclosures, and the transactional patterns that fraud screens consume. An adversary who understands a deployed model's decision boundary can, in principle, shape its inputs to evade detection, manipulate a forecast, or trigger an unwanted action—without ever breaching a system perimeter. The attack is not a software exploit in the classical sense; it is an exploit of the statistical decision rule itself. There is no buffer overflow to patch and no credential to rotate: the vulnerability is a geometric property of a function the institution deliberately trained and deployed, and conventional information-security controls—network segmentation, access management, encryption—do nothing to remove it.

This matters because the three canonical deployment families have distinct, and in each case serious, failure consequences. Trading and forecasting models translate a predicted price direction into orders; an adversary who can nudge the prediction can induce adverse selection or provoke liquidation cascades. Credit-scoring models gate access to capital; an applicant or an intermediary who can perturb reported features just enough to cross an approval threshold converts a statistical artefact into financial loss and, potentially, regulatory exposure. Fraud and anti-money-laundering screens are explicitly adversarial by construction: the population they classify includes actors whose objective is precisely to be misclassified, and who iterate against the screen. In all three, the input distribution at inference time is not the benign distribution on which the model was trained and validated.

The governance context sharpens the question. Emerging regulation, including the European Union's Artificial Intelligence Act [3], treats certain financial uses of AI—creditworthiness assessment prominently among them—as high-risk, with expectations of robustness, accuracy, and resilience against attempts to manipulate system behaviour. "Robustness" in such texts is not yet operationalised into a number a model-risk function can check off, and the gap between the regulatory aspiration and a testable acceptance criterion is exactly where adversarial-robustness research can contribute. The broader move toward large pre-trained and foundation models [2] only widens the surface: these systems inherit known failure modes such as hallucination [7] and admit transferable, automatically discovered attacks even when aligned [13], while their scale makes exhaustive adversarial testing harder.

It is worth being concrete about how the perturbation enters in each setting, because the abstraction of an "ε-ball around an input" can otherwise feel detached from operations. In an order-flow forecaster, the features are functions of the visible limit-order book—imbalances, queue sizes, recent trade signs—and a participant who places and cancels small orders can move several of those features simultaneously, within the bid–ask spread, at a cost that is a fraction of a tick. The perturbation is not a hypothetical pixel nudge; it is ordinary, legal-looking market activity that happens to push a forecaster across its decision boundary. In credit scoring, the features are self-reported or intermediary-reported attributes, and the "perturbation" is the well-known practice of structuring an application to clear a threshold; an adversary who has learned the model's sensitivities can do this surgically rather than by trial and error. In fraud and anti-money-laundering screening, the adversary is endogenous to the problem: launderers and fraud rings deliberately shape transaction graphs to look benign, and the screen they evade is, by definition, the model under attack. Across these cases the common structure is that a non-trivial subset of model inputs is attacker-controllable at low cost, which is precisely the condition under which adversarial-example theory predicts that high benign accuracy provides little protection.

A second reason the financial setting deserves its own treatment is the feedback loop between model and environment. Unlike a static image classifier, a deployed trading or scoring model changes the world it observes: its decisions move prices, approve or deny credit, and alter the population of actors who interact with it. An adversary therefore operates against a moving target, but so does the defender, and the benign distribution on which the model is validated can drift away from the deployment distribution even before any deliberate attack. Adversarial fragility and distribution shift are distinct phenomena, but in finance they compound: a model already living slightly off its validation distribution has less margin to absorb a deliberate perturbation. This is part of why we treat the gap between benign and adversarial accuracy as a risk metric rather than a curiosity.

A third distinguishing feature is the asymmetry of incentives and the directness of the payoff. In most consumer-facing vision systems the adversary's reward for a single misclassification is diffuse or symbolic; in finance it is denominated in the unit of account and accrues immediately. An attacker who flips a forecaster's directional signal profits from the resulting adverse selection on the very next order; an applicant who clears a credit threshold receives capital that may never be repaid; a launderer who evades a screen moves funds that clear and settle. This tight coupling between misclassification and monetary gain means the adversary is not hypothetical but economically motivated, with a quantifiable budget to spend on crafting attacks and a quantifiable return on succeeding. Robustness is therefore not an

abstract desideratum but a determinant of expected loss, and it belongs in the same risk ledger as market, credit, and operational risk rather than in a separate compartment of "model quality."

This paper is an empirical study on public data. We do not claim measurements from any production trading system, fund, or live decisioning pipeline. Instead we use the real, openly available UCI Adult income dataset [1]—a canonical tabular benchmark whose binary >50K target stands in for a credit/approval-style decision—and a fully specified pipeline in which the model, the attacks, and the defence are all reproducible from a single script. This lets the qualitative phenomena—accuracy erosion under attack, recovery under adversarial training, the accuracy–robustness trade-off, the dependence on attack iteration count, and the transferability of perturbations—be observed on genuine data and reproduced exactly. The aim is conceptual clarity and a transferable evaluation template, not a benchmark leaderboard.

1.1. Research questions

We organise the study around five questions.

- **RQ1.** How severely does the accuracy of a standard financial-style classifier degrade under gradient-based evasion (FGSM and PGD) as a function of the perturbation budget, and how sensitive is the result to the number of PGD iterations?
- **RQ2.** How much of the lost accuracy does min–max adversarial training recover, and at what cost to clean (unperturbed) accuracy?
- **RQ3.** How does the robustness–accuracy trade-off vary with the adversarial-training budget, i.e. what does the defender buy and give up by training against a larger inner perturbation?
- **RQ4.** Do adversarial perturbations crafted against one model transfer to independently trained models—including different architectures—i.e. is the black-box threat credible without gradient access?
- **RQ5.** What does the resulting evidence imply for how financial institutions should evaluate, accept, and monitor AI models exposed to adversarial inputs?

1.2. Contributions

We make five contributions. (i) We assemble a coherent threat model for financial AI that maps the abstract adversarial-examples literature onto the concrete input surfaces of trading, credit, and fraud systems, distinguishing evasion from poisoning and white-box from black-box adversaries, with a budget formulation tied to feature semantics. (ii) We give self-contained derivations of FGSM and PGD as first-order solvers of the inner maximisation in a robust-optimisation problem, and we situate empirical adversarial training against the promise and the limits of certified/provable robustness. (iii) We provide a self-contained, reproducible experiment on the real Adult dataset [1] that measures clean versus adversarial accuracy under FGSM and PGD across several perturbation magnitudes, several PGD step counts, and multiple model families, with all gradients computed exactly in a from-scratch numpy pipeline. (iv) We quantify, on the same real task, the defensive effect of adversarial training, an ablation over the adversarial-training budget, the accompanying clean-accuracy cost, and cross-model and cross-architecture transferability. (v) We translate the findings into governance-oriented recommendations—robustness-aware acceptance criteria, microstructure-grounded budgets, and defence-in-depth—while being explicit that every number reported is produced by the released script on public data.

The remainder of the paper proceeds as follows. Section 2 reviews adversarial examples, robust optimisation, transferability, certified robustness, and evasion in financial ML. Section 3 states the formal threat model and derives the attack and defence definitions. Section 4 presents the empirical study on the Adult dataset. Sections 5–6 discuss implications, a research agenda, and threats to validity, and Sections 7–8 cover reproducibility, declarations, and conclusions.

2. Background and related work

2.1. Adversarial examples and the geometry of brittleness

The modern study of adversarial examples begins with the observation that neural networks misclassify inputs formed by adding a small, carefully directed perturbation to an otherwise correctly classified input. Goodfellow et al. [5] offered an influential explanation: in high dimensions, the cumulative effect of many small, aligned coordinate changes on a locally near-linear model can be large, so an adversary who moves each input feature a little in the direction of the loss gradient can move the model's output a lot. The argument is worth restating because it explains why the financial setting is exposed rather than protected by its high feature dimension. Consider a locally linear score w^Tx and a perturbation δ

with $\|\delta\|_\infty \le \varepsilon$. The change in the score is $w^T\delta$, which an adversary maximises by choosing $\delta = \varepsilon \operatorname{sign}(w)$, yielding $w^T\delta = \varepsilon\|w\|_1 = \varepsilon \Sigma |w_i|$. The worst-case displacement therefore grows with the dimension d: adding more features, each only weakly informative, hands the adversary more directions to push, and the aggregate push can dwarf the per-feature cap ε. This linear-explanation view both demystified the phenomenon and yielded a one-step attack, the Fast Gradient Sign Method (FGSM), that perturbs every feature by a fixed magnitude ε in the sign direction of the loss gradient. The practical lesson is that brittleness is not an exotic corner case; it is a generic property of high-dimensional decision surfaces trained only for average-case accuracy. Tabular financial models, which routinely ingest dozens to hundreds of engineered features, sit squarely in the regime the linear explanation identifies as fragile.

2.2. Robust optimisation and the min–max view

Madry et al. [9] reframed defence as a robust optimisation problem: rather than minimising expected loss on clean data, one minimises the expected worst-case loss over an allowed perturbation set around each input. The inner maximisation—finding the worst case—is approximated by Projected Gradient Descent (PGD), a multi-step, iteratively projected refinement of FGSM that is widely treated as a strong first-order adversary. The outer minimisation, training against PGD-generated examples, is adversarial training, the most consistently effective empirical defence to date. This min–max formulation is the conceptual backbone of our methods: it unifies the attack we use to stress-test a model and the defence we use to harden it, and it makes explicit that "robustness" is always relative to a declared perturbation budget. A subtle but consequential point, which we return to in the methods, is that the saddle-point objective is not jointly convex–concave for a neural network; the inner problem is a non-concave maximisation that PGD only approximately solves, so empirical robustness is always reported against a particular attack and can be overstated if a stronger or adaptive attack is later found. This is the single most important caveat in interpreting any empirical robustness number, ours included.

2.3. Transferability and the black-box threat

A practically important property is transferability: perturbations crafted to fool one model often fool other models trained for the same task, even with different architectures or training data. Transferability is what makes black-box attacks credible. An adversary who cannot read a deployed model's gradients can train a surrogate, attack the surrogate with full white-box access, and transfer the resulting examples. The standard explanation appeals to the geometry above: independently trained models that fit the same benign manifold tend to learn correlated gradients in the off-manifold directions, so a perturbation aligned with one model's loss gradient is frequently aligned, at least partially, with another's. In a financial context, where institutions training on overlapping market data are likely to learn similar decision boundaries, transferability suggests that the white-box results we report are not merely pessimistic curiosities but a proxy for an attainable black-box capability. Recent work shows that automatically discovered, transferable attacks extend even to large aligned language models [13], underlining that scale and pre-training do not by themselves confer robustness, and that the surrogate-and-transfer recipe generalises well beyond the tabular and vision settings in which it was first characterised.

2.4. Certified and provable robustness, and its limits

Empirical defences such as adversarial training are validated against a fixed attack and therefore offer no guarantee against attacks not yet tried. A complementary research line seeks certified robustness: a procedure that, for a given input, returns a provable radius within which the prediction cannot change, regardless of the adversary's algorithm. The appeal in a regulated financial setting is obvious—a certificate is exactly the kind of testable, defensible artefact a model-risk function needs—but the limitations are equally important and we state them plainly so the rest of the paper is read in their light. First, certificates are typically conservative: the provable radius is smaller, often much smaller, than the radius at which a model is empirically robust, so certifying a useful radius can require sacrificing substantial clean accuracy. Second, exact certification of piecewise-linear networks is computationally hard in general, and scalable methods rely on relaxations or randomised smoothing that trade tightness for tractability and may themselves shift the model's behaviour. Third, and most relevant here, certificates are stated with respect to a particular norm ball; a certificate in $\ell\infty$ says nothing about an adversary who operates in a different, financially natural geometry, one that exploits feature correlations or moves a small number of features by a large, semantically meaningful amount. We do not compute certificates in this study, and we are explicit that our empirical robustness numbers are upper bounds against PGD specifically rather than guarantees; we discuss certification as a direction for the research agenda in Section 6 because it is, in our view, where the regulatory and the technical conversations most naturally meet.

2.5. Evasion and the broader attack surface in financial ML

Adversarial robustness sits within a wider security landscape for financial ML. It is useful to organise that landscape along two axes that we will use throughout. The first axis is when the attack occurs. Evasion attacks, our focus,

manipulate inference-time inputs to flip a decision while leaving the model untouched; poisoning attacks instead corrupt the training data so that the model learns a boundary the attacker prefers, for example a backdoor that misclassifies inputs carrying a chosen trigger. Evasion is the more immediate concern for a model already in production, but poisoning is especially salient in finance because many models are retrained continuously on data that includes attacker-influenced market activity, so the line between "training data" and "adversarial input" is blurred. The second axis is what the adversary knows, which we develop formally as the white-box/black-box distinction in the methods. Beyond evasion and poisoning, the literature documents privacy attacks such as membership inference, which can reveal whether a particular record was in the training set [10]—a serious concern when models are trained on sensitive customer or transaction data—and model-extraction attacks that reconstruct a proprietary model through its query interface, which both leaks intellectual property and furnishes the surrogate that a black-box evasion attacker needs. The financial-ML methodology literature has long stressed that naïve application of machine learning to markets invites overfitting, leakage, and spurious performance [8]; adversarial fragility is a complementary failure that standard backtesting and cross-validation do not surface, because they evaluate on benign, non-adversarial samples. Deep models for order books [12], price formation [11], return prediction [4], and asset pricing [6] deliver genuine predictive value, but their very expressiveness is what an adversary exploits. Our study targets precisely the blind spot between high benign accuracy and undefended adversarial accuracy.

2.6. Why standard validation does not surface the risk

It is tempting to assume that thorough out-of-sample testing would catch a model as fragile as the one we will exhibit, but the assumption is mistaken in a way that is instructive. Cross-validation, walk-forward backtesting, and held-out evaluation all estimate performance under the same distribution that generated the test data; they answer the question "how well does the model do on inputs drawn like the ones it was trained on?" Adversarial robustness asks a different question—"how well does the model do on the worst input within a small neighbourhood of each test point?"—and the two answers are only loosely coupled. A model can interpolate the benign manifold almost perfectly while leaving the off-manifold directions, which an adversary deliberately exploits, essentially unconstrained. This is why the adversarial-accuracy column in our results is invisible to any benign metric: nothing in a standard validation protocol ever queries those directions. The methodological literature in finance has repeatedly warned that performance estimates can be inflated by leakage, multiple testing, and overfitting [8]; adversarial fragility is a further, orthogonal way in which a confident validation number can mislead, and it calls for an explicitly worst-case evaluation rather than a more careful average-case one. The membership-inference literature [10] makes a parallel point from the privacy side: a model can leak information that no accuracy metric measures, because accuracy and the property of interest are simply different functionals of the same trained network. The unifying lesson is that a single scalar of benign performance, however carefully estimated, underdetermines a model's behaviour; safety-relevant properties—worst-case accuracy, privacy leakage, calibration under shift—are separate functionals that must be measured separately and reported alongside, not inferred from, the headline accuracy.

3. Methods

3.1. Threat model

We consider a deployed classifier $f_\theta : X \to R^C$ with parameters θ, mapping a feature vector $x \in X \subseteq R^d$ to scores over C classes; the prediction is $\hat{y}(x) = \arg\max_c f_\theta(x)_c$. The defender's training distribution is D. At inference time, an adversary observes inputs and is permitted to replace a legitimate input x with a perturbed input $x' = x + \delta$, subject to a budget constraint $\delta \in S(\varepsilon)$. Throughout we take $S(\varepsilon) = \{\delta : \|\delta\|\infty \le \varepsilon\}$, an $\ell\infty$ ball of radius ε on standardised features, so that ε caps the per-feature distortion measured in standard deviations.

Definition 1 (Perturbation budget). A perturbation δ is admissible at budget ε if $\|\delta\|\infty \le \varepsilon$ and $x + \delta \in X$. The budget encodes how much an adversary can move each feature while remaining economically or operationally plausible: in a trading setting it bounds how far quotes or volumes can be nudged before the manipulation is itself detectable or unprofitable; in credit scoring it bounds the misreporting of applicant attributes.

We distinguish adversary knowledge along two axes. The first concerns access to the model.

Assumption 1 (White-box adversary). The adversary knows θ and can compute gradients $\nabla_x L(f_\theta(x), y)$ of the model's loss with respect to the input. This is the worst-case capability and the standard setting for stress-testing a defence.

Assumption 2 (Black-box adversary). The adversary does not know θ but can train a surrogate $f_{\theta'}$ on similar data and craft perturbations against it, relying on transferability to fool f_θ. This models a realistic external attacker with no privileged access. A weaker still query-only adversary, who can submit inputs and observe decisions but cannot train a faithful surrogate, is bounded above by the transfer adversary we measure and is not studied separately here.

The second axis concerns when the adversary acts, which separates the evasion threat we study from the poisoning threat we do not. Under evasion, the parameters θ are fixed and the attack lives entirely at inference time; under poisoning, the adversary perturbs the training set and thereby influences θ itself. We assume an evasion adversary throughout: training is trusted, and the contest is over the inference-time input. This is the right model for a deployed, periodically retrained classifier whose training pipeline is governed, and it isolates the geometric fragility that is our subject from the data-integrity questions that poisoning raises.

The adversary's objective is untargeted evasion: cause $\hat{y}(x') \neq y$ for as many inputs as the budget allows. The defender's objective is to retain high accuracy both on clean inputs and on admissibly perturbed inputs. We focus on the untargeted case because it is the natural fit for our three settings—a fraudster wants any misclassification that clears them, a manipulator wants the forecaster to be simply wrong in the profitable direction—but the framework extends without modification to targeted evasion by replacing the loss with one that rewards a specific desired class. In the credit setting, targeted evasion "approve me" is in fact the more natural objective, and because targeting a specific class is at least as hard as untargeted misclassification, our untargeted numbers are if anything optimistic about the defender's position there.

Remark 1 (Choosing the norm and the budget). The $\ell\infty$ choice is a modelling decision, not a law of nature. It treats every feature as independently and equally perturbable up to ε standard deviations, which is a reasonable first approximation for standardised tabular features but ignores two realities of financial data: features are often correlated, so coordinated moves are cheaper than the norm suggests, and some features are effectively immutable or discrete, so moves in those directions are impossible regardless of budget. An ℓ_2 budget would instead reward concentrating the perturbation in a few high-leverage directions, and an ℓ_0 (few-features) budget would model an adversary who can only touch a handful of fields—each geometry corresponds to a different real attacker. A production threat model should replace the uniform ball with a per-feature, cost-weighted constraint set derived from how expensive or detectable each move is. We keep the uniform $\ell\infty$ ball here for clarity and comparability with the canonical attacks, and we flag the gap explicitly in Section 6.

3.2. Attacks: FGSM and PGD, derived

We now derive the two attacks as approximate solvers of the inner maximisation

$$\delta^*(x, y) = \text{argmax } L(f(x + \delta), y) \text{ over } \delta \text{ in } S(\varepsilon) \quad (1)$$

where L is the cross-entropy loss. A first-order Taylor expansion of the loss around the clean input gives

$$L(f(x + \delta), y) \approx L(f(x), y) + g \cdot \delta, \quad \text{where } g = \nabla x\, L(f(x), y) \quad (2)$$

Maximising the linear term $g^T\delta$ over the $\ell\infty$ ball $\|\delta\|\infty \leq \varepsilon$ is a separable problem whose solution sets each coordinate to the budget in the gradient's sign direction, $\delta_i = \varepsilon \text{ sign}(g_i)$, because for a fixed magnitude budget the inner product $g^T\delta = \Sigma_i g_i\delta_i$ is maximised coordinatewise. This yields the Fast Gradient Sign Method [5],

$$x'(\text{FGSM}) = x + \varepsilon \cdot \text{sign}(\nabla x\, L(f(x), y)) \quad (3)$$

which is therefore exactly the maximiser of the first-order surrogate (2), not a heuristic. Its weakness is precisely that it trusts the linearisation across the whole ball of radius ε: when the loss surface curves appreciably within the budget—as it does for a non-linear network—the single step overshoots in some coordinates and undershoots in others, and FGSM underestimates the attainable loss.

Projected Gradient Descent [9] repairs this by re-linearising at each step. It iterates the FGSM update with a smaller step α and projects back into the budget set after each step:

$$x(t+1) = \text{Proj}[S(\varepsilon)]\,(x(t) + \alpha \cdot \text{sign}(\nabla x\, L(f(x(t)), y))), \quad x(0) = x + \eta \quad (4)$$

where $\Pi_{S(\varepsilon)}$ clips coordinates to the $\ell\infty$ ball of radius ε around x, and η is a small random start inside the ball. PGD is projected gradient ascent on the loss restricted to the budget set: each step takes the locally optimal sign-direction move of size α and then the projection returns the iterate to the feasible set, so the procedure follows the curved loss surface in a sequence of short, re-linearised hops rather than one long linear leap. Two consequences matter for evaluation. First, with T steps and step $\alpha = \varepsilon/k$ the attack can reach corners of the ball that a single FGSM step cannot, so PGD is uniformly at least as strong as FGSM and the gap widens with curvature. Second, the random start η and multiple restarts matter because the inner problem (1) is non-concave: different starts can climb to different local maxima, and reporting the best over restarts gives a stronger, more honest adversary. PGD with several steps is our primary, stronger attack and the inner maximiser of the defence below.

3.3. Defence: adversarial training

Following the robust-optimisation view [9], adversarial training replaces the standard empirical risk with a saddle-point objective:

$$\min_{\theta} \mathbb{E}\left[\max_{\delta \in S(\varepsilon)} L(f(x+\delta), y) \right] \quad (5)$$

The inner maximisation is approximated per mini-batch with PGD; the outer minimisation is ordinary stochastic gradient descent on the resulting worst-case loss. The gradient of the outer objective is obtained, by Danskin's theorem applied to the (approximate) inner maximiser $\delta\bigstar$, as $\nabla_\theta L(f_\theta(x+\delta\bigstar), y)$ evaluated at the fixed perturbation—that is, one simply backpropagates through the network at the adversarial point and treats $\delta\bigstar$ as a constant. This is what makes the procedure tractable: each training step is a PGD attack to find $\delta\bigstar$ followed by an ordinary parameter update at the attacked input. Conceptually, the model is trained to be correct not just at each data point but across the entire admissible neighbourhood of each point, which is what we mean by robustness. The expected cost is a reduction in clean accuracy, because some capacity is spent flattening the loss surface rather than fitting benign structure; quantifying this trade-off, and how it scales with the training budget ε_{train}, on the real Adult task is the goal of Section 4.

3.4. Threat-model diagram

Figure 1 summarises the surfaces and adversary capabilities for the three canonical financial deployments. Market and applicant inputs (quotes, features, transactions) pass through a feature pipeline into the model f_θ, which produces an action or decision—orders in trading, approve/deny in credit, flag/clear in fraud. The adversary, white-box or black-box, crafts a perturbation δ with $\|\delta\|\infty \le \varepsilon$ and injects it at the input/feature stage to flip the model's action, while a governance function inspects the model and gates the decision in line with robustness expectations such as those of the EU AI Act [3]. The diagram is conceptual.

4. Experiments: An empirical study on the Adult dataset

All quantitative results in this section are produced by the released script (Section 7) on the real, public UCI Adult dataset [1]. They are designed to expose the qualitative behaviour of the attacks and the defence on genuine tabular data, and while Adult is a credit/approval-style benchmark rather than a live trading or fraud feed, the numbers are honest measurements on real records rather than synthetic outputs. They should not be read as measurements of any deployed trading, credit, or fraud system, nor as state-of-the-art benchmark numbers for a tuned architecture.

4.1. Dataset and preprocessing

We use the UCI Adult dataset [1] as served by OpenML (fetch_openml('adult', version=2)), a widely used tabular benchmark in which the task is to predict whether an individual's annual income exceeds $50,000 from census attributes. This binary income/approval decision is a natural stand-in for a credit-style classification: the label y = 1 (>50K) denotes the positive ("act") class. After dropping records with missing values, the dataset contains both numeric attributes (age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week) and categorical attributes (workclass, education, marital status, occupation, relationship, race, sex, native country). We one-hot encode the categoricals and standardise the numeric attributes to zero mean and unit variance, fitting the encoder and scaler on the training split only, which yields d = 104 features. We take a 75/25 stratified train/test split with random_state=42, giving N_{train} = 33,916 and N_{test} = 11,306; the positive (income >50K) rate is 24.8% in both splits, so this is a moderately imbalanced task rather than a balanced one. All pseudo-random seeds are fixed so the entire pipeline is deterministic.

The dataset is real, but the threat model deliberately restricts the adversary to the attributes that are plausibly cheap to perturb: the $\ell\infty$ perturbation is applied only to the six standardised continuous attributes (via a feature mask), while the one-hot indicator columns—which encode discrete, often immutable categoricals such as sex or native country—are held fixed. This is the honest analogue of the per-feature, cost-weighted constraint set advocated in the Methods remark: an adversary can shade a self-reported continuous quantity (e.g. hours worked or capital gains) far more easily than they can change a categorical identity. The consequence is that the attack surface is narrow (six of 104 feature columns), so the accuracy degradations we report are correspondingly milder than the catastrophic collapses seen when every coordinate is freely perturbable; this is a feature, not a bug, of using real data with a realistic mask.

4.2. Model families, attacks, and protocol

To avoid reading idiosyncrasies of one architecture as general phenomena, we evaluate three model families on the identical task, each implemented from scratch in numpy with exact analytic input gradients. The primary baseline model f_θ is a feed-forward network with one hidden layer (width 64, ReLU activations) and a logistic (softmax) output, trained by mini-batch gradient descent to minimise binary cross-entropy. The second is a larger multilayer perceptron (one hidden layer, width 128), which probes whether additional capacity changes the fragility picture. The third is a logistic-regression baseline, included because the linear explanation of adversarial examples [5] predicts that even a linear model is attackable, and a linear model gives a clean reference point against which to read the neural results. For each family we evaluate three regimes on the held-out test set: clean (no perturbation), FGSM (Eq. 3), and PGD (Eq. 4 / Algorithm 1); every perturbation is masked to the six continuous attributes as described above. For PGD we additionally sweep the number of steps $T \in \{1, 5, 10, 20, 40\}$ at fixed step size $\alpha = \varepsilon/8$ (with $T = 1$ recovering a single random-start step) to expose the dependence of attack strength on iteration count. We sweep the budget $\varepsilon \in \{0.00, 0.02, 0.05, 0.10, 0.20, 0.30\}$ on the standardised continuous features. We then train adversarially robust models with the objective (Eq. 5), using PGD as the inner maximiser, and we ablate the training budget $\varepsilon_{train} \in \{0.05, 0.10, 0.20\}$ to trace how the robustness–accuracy trade-off moves with the strength of the defence. Finally, to probe black-box risk (RQ4), we craft PGD examples against the baseline and evaluate them on independently initialised, separately trained models—both the same architecture and the two other families—to measure same-architecture and cross-architecture transfer.

5. Results

5.1. Results: attack strength versus accuracy

Table 1 reports test accuracy of the standard (non-robust) baseline across budgets and attacks. The pattern is the one the theory predicts: clean accuracy is high, but it erodes steadily and monotonically as the budget grows, even though the adversary may touch only the six continuous attributes. At a budget that perturbs each standardised continuous feature by one-tenth of a standard deviation—an economically small change—PGD removes roughly six points of accuracy (85.3% → 79.4%), and at $\varepsilon = 0.30$ it removes over twenty (85.3% → 64.8%). The degradation is bounded well above the 24.8% base rate because most of the predictive signal in Adult lives in the discrete attributes the mask protects; the continuous attributes carry enough signal to be worth attacking but not enough for the attack to destroy the model.

Table 1 Test accuracy (%) of the standard baseline (MLP, width 64) under increasing perturbation budget ε, applied via $\ell\infty$ to the standardised continuous attributes of Adult [1]. PGD uses $T = 20$ steps, $\alpha = \varepsilon/8$. Lower is worse for the defender.

	Perturbation budget ε ($\ell\infty$, std. units)					
Attack	**0.00**	**0.02**	**0.05**	**0.10**	**0.20**	**0.30**
Clean (no attack)	85.3	85.3	85.3	85.3	85.3	85.3
FGSM	85.3	84.0	82.4	79.4	72.3	65.1
PGD (T=20)	85.3	84.0	82.4	79.4	72.1	64.8

Two features of Table 1 deserve emphasis. First, PGD is uniformly at least as damaging as FGSM at every non-zero budget—the two coincide at small budgets and PGD pulls slightly ahead at $\varepsilon \geq 0.20$, confirming that the gap widens with budget as the loss surface curves within the ball, and that any acceptance test relying on FGSM alone gives at best the same and at worst an optimistic reading. Second, the degradation is roughly monotone with no comfortable plateau: each increment of budget buys the adversary more lost accuracy. For a financial decision system, the practical reading

is that an adversary with even a modest, plausibly-deniable ability to shade a few continuous inputs can flip a material fraction of decisions, even when the bulk of the feature vector is immutable.

The magnitudes here are milder than in settings where every coordinate is freely perturbable, precisely because the realistic mask confines the attack to six of 104 columns. Whether per-feature moves of a few hundredths to a few tenths of a standard deviation in attributes like reported hours or capital gains are realistically available to an adversary is exactly the microstructure-grounded budget question raised in the Methods remark, and it is the question a deploying institution must answer before reading these numbers as either alarming or benign.

5.2. Results: how attack strength scales with PGD iterations

Because the inner maximisation (1) is non-concave, the strength of a PGD attack depends on how many steps it is allowed; an institution that benchmarks robustness against a weak (few-step) attack will systematically overstate its defences. Table 2 sweeps the step count T at a fixed budget $\varepsilon = 0.10$ for the standard baseline. Attack potency increases from one to roughly twenty steps and then saturates: the marginal damage of additional steps falls off once PGD has found a strong local maximum of the loss within the ball. The single-step case (T = 1, 84.6%) conspicuously understates the threat relative to the converged attack (79.4%), while T = 20 and T = 40 are identical to two decimals, which is why we adopt T = 20 as the primary attack elsewhere: it is in the saturated regime and so reports close to the worst case a first-order adversary can reach at this budget.

Table 2 Test accuracy (%) of the standard baseline (MLP, width 64) under PGD at fixed budget $\varepsilon = 0.10$ as a function of the number of PGD steps T (step size $\alpha = \varepsilon/8$, single random start). Lower is worse for the defender. The first column (T = 1) approximates FGSM with a random start.

Model	T = 1	T = 5	T = 10	T = 20	T = 40
Standard baseline (MLP)	84.6	82.3	80.1	79.4	79.4

5.3. Results: fragility across model families

Table 3 reports clean accuracy and PGD-attacked accuracy ($\varepsilon = 0.10$, T = 20) for the three model families. Two messages emerge. First, greater capacity does not confer robustness: the larger MLP is essentially tied with the baseline on clean data yet drops slightly more under attack, consistent with the view that standard training, regardless of capacity, leaves the off-manifold directions unconstrained. Second, the linear model is itself attackable, exactly as the linear explanation of adversarial examples predicts [5]; its marginally smaller drop reflects only that a linear score is harder to steer through a six-coordinate mask than a non-linear one, not any intrinsic immunity. The practical implication is that one cannot escape the problem by switching architecture or trimming capacity: undefended models of every family studied here lose accuracy under a moderate-budget PGD attack, and the ordering of the drops is consistent.

Table 3 Clean and PGD-attacked accuracy (%) across model families, PGD at $\varepsilon = 0.10$, T = 20, masked to the continuous attributes of Adult. "Drop" is clean minus attacked accuracy (larger is worse).

Model family	Clean	PGD ($\varepsilon = 0.10$)	Drop
Logistic regression (linear)	84.6	79.3	5.3
MLP (width 64, baseline)	85.3	79.4	5.9
MLP (width 128, larger)	85.3	78.7	6.5

5.4. Results: adversarial training as a defence

Table 4 compares the standard baseline with the adversarially trained model, both evaluated under the strong PGD attack, and also reports clean accuracy so the trade-off is visible. Adversarial training transforms the robustness profile: under PGD at the training budget ($\varepsilon = 0.10$), robust accuracy rises from 79.4% to 83.9%, and the gain grows at larger budgets, reaching 64.8% → 81.3% at $\varepsilon = 0.30$. On this masked, real task the recovery is nearly complete, and—strikingly—it is purchased at almost no cost: clean accuracy falls by less than half a point (85.3% → 84.9%). This is the empirically familiar accuracy–robustness trade-off, here in a benign regime because the attack surface is narrow enough that flattening the loss over the six attacked coordinates barely disturbs the benign fit.

Table 4 Defence comparison: PGD-attacked accuracy (%) for the standard versus adversarially trained (AT) model, with clean accuracy for reference. AT uses PGD at ε_train = 0.10 as the inner maximiser; PGD evaluation uses T = 20. Higher is better for the defender, except where noted.

Budget ε	Standard	Adv. trained (AT)
0.02	84.0	84.7
0.05	82.4	84.3
0.10	79.4	83.9
0.20	72.1	82.8
0.30	64.8	81.3
Clean accuracy (no attack)	85.3	84.9

Three observations follow. First, the defence generalises beyond its training budget: although trained at ε = 0.10, the adversarially trained model retains a large margin at ε = 0.20 and 0.30, where the standard model has lost twenty points or more. Second, robustness still decays as the attacker's budget grows past the training budget, so adversarial training buys a robustness band, not unlimited protection; the budget against which one trains is a governance decision, not a technical afterthought. Third, the clean-accuracy cost (here, 85.3% → 84.9%) is small but real and must be weighed against the value of the recovered robustness—a trade that only the deploying institution, aware of its threat environment, can make. That the cost is so small here is itself a finding: when the realistic attack surface is narrow, the price of robustness can be modest.

5.5. Results: ablating the adversarial-training budget

The training budget ε_train is the single most important knob the defender controls, and Table 5 traces its effect. Training against a larger inner perturbation buys robustness at larger attack budgets but costs a little more clean accuracy, and the ordering is consistent across the evaluation grid: the ε_train = 0.20 model is the most robust at ε = 0.20 and 0.30 yet the least accurate on clean data, while the ε_train = 0.05 model is the reverse. This is the robustness–accuracy trade-off made operational: there is no single "robust model," only a family of models indexed by the budget one chooses to defend, and selecting a point on this curve is precisely the risk decision that a budget grounded in microstructure and product economics should inform. The clean-accuracy spread across the three defences is narrow (84.5% to 84.9%) on this masked task, so the practical cost of over-provisioning is small here—but the qualitative lesson, that a defender who trains at ε_train = 0.20 when the realistic adversary operates at ε = 0.05 pays for robustness it does not need, still holds and would bite harder on a task with a wider attack surface.

Table 5 Adversarial-training budget ablation. PGD-attacked accuracy (%) at evaluation budget ε for adversarially trained models with three different training budgets ε_train; final column is clean accuracy. PGD uses T = 20 steps. Higher is better for the defender (clean column included for the trade-off).

Training budget	ε = 0.05	ε = 0.10	ε = 0.20	ε = 0.30	Clean
ε_train = 0.05	84.2	83.4	81.4	78.9	84.9
ε_train = 0.10	84.3	83.9	82.8	81.3	84.9
ε_train = 0.20	84.3	84.0	83.4	82.7	84.5

5.6. Results: transferability (black-box risk)

To assess whether the white-box results overstate a realistic threat, we transferred PGD examples crafted against the standard baseline to independently trained models (different random seeds), both of the same architecture and of the other two families. Table 6 reports the attacked accuracy of each target. Transferred examples at ε = 0.10 reduced a same-architecture target's accuracy from 85.1% to 79.2%, and the larger MLP transferred comparably (85.1% → 79.0%), while the linear model absorbed the transfer slightly better (84.5% → 79.8%). Notably, the transferred attack is almost as damaging as the white-box attack on this masked task—the white-box baseline loses to 79.4% at the same budget—which tells us that on Adult the off-mask gradients of independently trained models are strongly correlated, so the bulk of the white-box potency carries over. The interpretation for RQ4 is that an adversary without gradient access remains dangerous: a surrogate-and-transfer strategy recovers nearly the full white-box effect here, and transfer

is comparable across architectures, mirroring the situation when institutions training on overlapping data converge on similar boundaries. Transferability should therefore be treated as the default assumption rather than a remote possibility, and the secrecy of model weights should not be relied upon as a defence.

Table 6 Transferability of PGD examples crafted against the standard baseline (ε = 0.10, T = 20) and evaluated on independently trained target models (distinct seeds). Clean accuracy of each target is shown for reference. Lower transferred accuracy means a stronger black-box threat.

Target model (independently trained)	Clean	Transferred PGD
Same architecture (MLP width 64)	85.1	79.2
Larger MLP (width 128)	85.1	79.0
Logistic regression (linear)	84.5	79.8

6. Discussion

The experiment reproduces, in a financial framing and on real data, the central tension of adversarial machine learning: standard accuracy and adversarial accuracy are distinct quantities, and a model can be accurate on benign inputs while measurably worse on adversarially shaped ones. For financial infrastructure this is not an academic nicety. The acceptance criteria that govern model promotion in most institutions—out-of-sample accuracy, calibration, backtest Sharpe, lift over a baseline—are all benign-distribution metrics. None of them would have flagged the gap exposed in Table 1: the baseline model is genuinely accurate on the data it was validated against and yet steadily loses accuracy to an adversary who shades a handful of continuous inputs it sees in production. The methodological caution that benign performance estimates can be deeply misleading in finance [8] extends naturally to adversarial fragility: a clean backtest is silent about worst-case behaviour. The model-family comparison (Table 3) tightens this point: because every family we tried degrades by a similar margin, the fragility is a property of the standard training objective and the geometry of the attacked subspace, not of a particular network one might swap out, so no architecture choice substitutes for an explicit robustness evaluation. We emphasise that the magnitudes here are moderate precisely because our threat model restricts the adversary to six continuous attributes; a richer attack surface, or a financially natural non-$\ell\infty$ geometry, would widen the gap.

The defence results carry a constructive message and a caveat. The constructive message (RQ2, RQ3) is that robustness is achievable and measurable: adversarial training [9] recovers a large robustness band at a quantifiable clean-accuracy cost, the saddle-point formulation gives institutions a principled knob—the training budget ε_train—to set robustness expectations explicitly, and the ablation in Table 5 shows that this knob behaves predictably, trading clean accuracy for robustness at larger budgets along a smooth frontier. The caveat is that adversarial training is neither free nor sufficient. It is computationally heavier (each training step contains a multi-step inner attack), it degrades clean accuracy, and it protects only within a declared budget and against the attack used in the inner maximisation; outside that budget, and against attack types or geometries not represented in the inner maximisation, its guarantees lapse. Crucially, our reported robust accuracies are upper bounds against PGD, not certificates: a stronger or adaptive attack, or one operating in an ℓ_2 or feature-correlated geometry, could push them lower, and the certified-robustness literature exists precisely because empirical robustness against a fixed attack can be an over-optimistic measure. We therefore read the result as an argument for defence in depth rather than reliance on any single mechanism: input validation and plausibility checks grounded in market microstructure, anomaly detection on the input stream, ensembling and disagreement monitoring, rate-limiting of attacker-controllable channels, and human review at high-stakes thresholds, with adversarial training as one robust layer among several.

The transferability evidence (RQ4) closes a tempting escape route. A defender might hope that withholding model internals—keeping the system black-box—neutralises the threat. Our transfer results (Table 6), and the broader literature including transferable attacks on aligned large models [13], indicate otherwise: surrogate-and-transfer attacks remain effective, transfer is strongest between similar models, and so secrecy of weights is a weak defence. This is consistent with a security posture that assumes a capable adversary rather than betting on obscurity, and it suggests that the relevant question is not "can the adversary read our gradients?" but "can the adversary build a model that resembles ours?"—to which, for institutions trained on shared market data, the answer is usually yes.

It is worth translating the abstract numbers back into the three deployment families to see how the same finding lands differently in each. For a trading or forecasting model, the relevant question is not only accuracy but the asymmetry of

the cost: an adversary who can flip a directional signal at a small budget does not need to win every interaction, only enough of them to extract adverse selection or to provoke a stop-loss cascade, so even a partial robustness recovery may leave material exposure if the residual attacked accuracy still permits profitable manipulation. For a credit model, the salient consideration is that the adversary attacks one decision at a time, at a threshold, and the relevant metric is the attack success rate near the approval boundary rather than average accuracy; a model that is robust on average can still be brittle exactly where it matters, which argues for budget-aware testing concentrated around decision thresholds. Moreover, the natural credit attack is targeted ("approve me"), and since targeting is harder than mere misclassification, our untargeted numbers understate the defender's advantage there only modestly while the threshold concentration understates the risk sharply. For a fraud or anti-money-laundering screen, the adversary is persistent and adaptive, iterating against the deployed system over time, so a static robustness measurement is at best a snapshot; here the case for continuous monitoring of the input stream and for periodic re-hardening is strongest, and the poisoning threat we excluded returns through the back door because the screen is retrained on data the adversary has had months to shape. The experiment cannot resolve these deployment-specific economics, but it does establish the common precondition—that a small, plausible perturbation of even a few attacker-controllable continuous inputs measurably degrades an undefended model—on which all three concerns rest.

These observations suggest a small set of operational practices that follow directly from the evidence rather than from general exhortation. First, model-acceptance documentation should carry an adversarial-accuracy curve, not a single robust number, because the shape of the degradation (Tables 1 and 2) is what determines whether a given budget is tolerable, and the curve must be measured against a saturated (many-step) attack lest a weak attack flatter the defence. Second, the perturbation budget should be derived and justified per input channel from the cost of moving each feature—a quantity that market microstructure and product economics can supply—rather than inherited as an arbitrary constant, and the chosen budget should set ε_train rather than being decided independently of the defence. Third, because adversarial training protects only within its declared band and at a clean-accuracy cost (Tables 4 and 5), it should be one layer in a stack that also includes input plausibility checks, anomaly detection, ensembling with disagreement alarms, rate-limiting of attacker-controllable channels, and human review at high-stakes thresholds. None of these is novel in isolation; the contribution is to show, on a controlled task, why each is warranted and why none suffices alone.

Toward a research agenda. Our results also map out where the most valuable follow-on work lies, and we sketch it here as a bridge to the limitations that follow. First, the budget question is the binding constraint on the whole programme: until a field develops principled, microstructure-grounded estimates of how cheaply each input channel can be moved, robustness numbers float free of the economics that should anchor them, and a per-channel, cost-weighted constraint set is the natural object to estimate. Second, the gap between empirical and certified robustness is exactly where governance can become rigorous, and adapting randomised smoothing or relaxation-based certificates to the correlated, mixed-discrete geometry of tabular financial features—and to non-$\ell\infty$ budgets—would convert a defensible heuristic into an auditable guarantee. Third, the interaction of adversarial fragility with the feedback loops and distribution shift that distinguish finance from vision is largely unstudied; a model that is robust at deployment may drift into fragility as its own decisions reshape the input population, and an evaluation that is a snapshot will miss this. Fourth, the poisoning threat we deliberately excluded deserves the same treatment in the continuous-retraining regime that fraud and AML systems inhabit. We regard the budget and certification strands as the highest-leverage, because they are what would let a regulator's robustness expectation be discharged by a number rather than an assurance.

Finally, on RQ5 and governance: the EU AI Act's treatment of creditworthiness assessment and similar uses as high-risk, with robustness and resilience expectations [3], becomes operationally meaningful only when "robustness" is tied to a testable budget and a strong attack. Our study suggests a concrete template—report clean accuracy alongside PGD-attacked accuracy across a budget grid and a step-count grid, justify the budget from the economics of the input channel, treat the gap between clean and attacked accuracy as a first-class risk metric, and disclose that the number is an empirical bound against a specific attack rather than a certificate. The same logic applies as institutions adopt large pre-trained models [2], whose additional failure modes, including hallucination [7] and privacy leakage via membership inference [10], compound rather than replace the evasion risk studied here.

6.1. Limitations and threats to validity

Construct validity. Although the data are real, Adult is a census income-classification dataset, not a live trading, credit, or fraud feed; it is a credit/approval-style benchmark whose use here is a deliberate, transparent stand-in. Real market and applicant data exhibit heavy tails, regime shifts, temporal dependence, and adversarial structure that Adult does not reproduce, and our models are small from-scratch classifiers trained to reasonable rather than state-of-the-art accuracy. The specific numbers in Tables 1–6 would change with the dataset, the attacked-feature mask, the

architecture, and the hyper-parameters; the qualitative patterns—degradation under attack, recovery under adversarial training, the budget-indexed trade-off, the saturation of attack strength in PGD steps, and the strong transfer—are what we expect to carry over.

Threat-model realism. We adopt an $\ell\infty$ feature-space budget restricted to the continuous attributes, which is analytically convenient and captures the realistic intuition that discrete identities are harder to perturb than self-reported numbers, but it remains an imperfect model of financial adversaries. Real attackers face structured constraints—some continuous features are also costly or detectable to change, and the six we expose differ widely in malleability—so a uniform $\ell\infty$ ball over the masked columns may both overstate freedom in some directions and understate it in others. Because the continuous attributes are themselves correlated with the protected categoricals, our mask is conservative in some directions and permissive in others. A faithful budget must be derived from the economics and microstructure of each input channel, and should likely be a per-feature, cost-weighted constraint set rather than a uniform ball; we operationalise a coarse version of this (the continuous-only mask) but do not formalise it per deployment.

Internal validity. Although we evaluate three model families and sweep the PGD step count, all attacks are first-order (gradient-based) and untargeted, and we report no certified guarantees. Stronger, adaptive, or non-$\ell\infty$ attacks, or an attack that also touched the categorical columns, could further reduce the adversarially trained model's accuracy, so our defence numbers are upper bounds on robustness against masked PGD specifically, not certificates against all adversaries. The PGD-step ablation mitigates one common way of overstating robustness—benchmarking against a too-weak attack—but cannot rule out a qualitatively different, stronger attack.

External validity. The mapping from the Adult income task to live trading, credit, and fraud systems is conceptual. Production systems involve pipelines, feature stores, latency budgets, feedback loops, and regulatory controls that interact with robustness in ways a single-model study cannot capture. The transferability result, while strengthened by the cross-architecture targets, was measured between models trained on the same dataset and split and may not reflect genuine cross-institution transfer, where data and feature engineering differ.

Scope. We study evasion only. Poisoning, model extraction, and privacy attacks [10] are out of scope here, though they share the same broad threat surface and would belong in a complete security assessment; poisoning in particular re-enters through the continuous-retraining loop of fraud and AML systems, where the training set is itself partly attacker-shaped, so the evasion-only framing is cleanest for the trading and one-shot-credit settings and most incomplete for AML.

6.2. Reproducibility

The study is fully reproducible from a single included script, experiment/run.py, executed with python3 experiment/run.py. We use the real, public UCI Adult dataset [1] fetched via OpenML (fetch_openml('adult', version=2, as_frame=True), the binary >50K target), drop records with missing values, one-hot encode categoricals and standardise the six numeric attributes using training statistics only (d = 104 features), and take a 75/25 stratified split with seed 42 (N_train = 33,916, N_test = 11,306). We fix all pseudo-random seeds. The model families are implemented from scratch in numpy so that input gradients are exact: a logistic-regression linear baseline, a one-hidden-layer MLP of width 64 (the baseline), and a one-hidden-layer MLP of width 128 (ReLU activations, logistic/softmax output, binary cross-entropy, mini-batch gradient descent). The attacks are FGSM (Eq. 3) and PGD (Algorithm 1) with $T \in \{1, 5, 10, 20, 40\}$, $\alpha = \varepsilon/8$, and a random start, both masked to the standardised continuous columns; the defence is adversarial training (Eq. 5) with a PGD inner maximiser and $\varepsilon\_train \in \{0.05, 0.10, 0.20\}$; the evaluation grid is $\varepsilon \in \{0, 0.02, 0.05, 0.10, 0.20, 0.30\}$; and the transfer protocol crafts PGD on the baseline and evaluates on independently trained, distinctly seeded same- and cross-architecture targets. Software versions used for the reported run are Python 3.12.3, numpy 2.5.0, pandas 3.0.3, scikit-learn 1.9.0, and scipy 1.18.0. The script prints all metrics and writes experiment/results.json; every number in Tables 1–6 is taken directly from that output. No proprietary, licensed, or non-public data is used. Reference implementations of the attacks and the adversarial-training loop follow directly from the cited works [5, 9].

7. Conclusion

Adversarial robustness is a load-bearing property for AI systems that sit on the critical path of financial infrastructure, yet it is invisible to the benign-distribution metrics that currently govern model acceptance. In a fully reproducible experiment on the real, public Adult dataset [1] we showed that an accurate classifier loses a measurable share of accuracy under gradient-based evasion even when only six continuous attributes are perturbable (e.g. 85.3% → 79.4% at $\varepsilon = 0.10$ and → 64.8% at $\varepsilon = 0.30$ under PGD), that the degradation is at least as severe under PGD as under FGSM and

deepens with both the perturbation budget and the number of PGD iterations before saturating, that the fragility is shared across linear and neural model families, that adversarial training recovers nearly all of the lost robust accuracy along a budget-indexed trade-off at a clean-accuracy cost under half a point, and that adversarial examples transfer to independently trained models well enough to make black-box attacks credible. All of these findings are measurements on a public dataset via the included script, not measurements of any live system, and we have been explicit that even our defence numbers are empirical bounds against a specific masked attack rather than certified guarantees.

The practical implication is a shift in evaluation discipline: report adversarial accuracy alongside clean accuracy across a justified budget grid and against a saturated attack, tie the budget and the training budget to the economics of each attacker-controllable input channel, adopt defence in depth rather than trusting any single mechanism, and pursue certified robustness where an auditable guarantee is needed. As financial institutions absorb ever larger pre-trained models and as regulators formalise robustness expectations, treating the gap between benign and adversarial performance as a first-class risk metric is, we argue, the minimum responsible posture.

Compliance with ethical standards

Acknowledgement

The author thanks the maintainers of the UCI Machine Learning Repository and OpenML for providing public access to the Adult dataset used in this study.

Disclosure of conflict of interest

The author declares no financial or non-financial conflicts of interest in relation to this work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

The UCI Adult dataset is publicly available from the UCI Machine Learning Repository and OpenML. All code and the results file required to reproduce the reported numbers are available from the corresponding author on reasonable request.

References

[1] Becker B, Kohavi R. Adult. UCI Machine Learning Repository; 1996. DOI: 10.24432/C5XW20.

[2] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258; 2021.

[3] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union; 2024.

[4] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research. 2018;270(2):654–669.

[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR); 2015.

[6] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. The Review of Financial Studies. 2020;33(5):2223–2273.

[7] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Computing Surveys. 2023;55(12):1–38.

[8] López de Prado M. Advances in Financial Machine Learning. Hoboken (NJ): John Wiley & Sons; 2018.

[9] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR); 2018.

[10] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017. p. 3–18.

[11] Sirignano J, Cont R. Universal features of price formation in financial markets: perspectives from deep learning. Quantitative Finance. 2019;19(9):1449–1459.

[12] Zhang Z, Zohren S, Roberts S. DeepLOB: Deep convolutional neural networks for limit order books. IEEE Transactions on Signal Processing. 2019;67(11):3001–3012.

[13] Zou A, Wang Z, Carlini N, Nasr M, Kolter JZ, Fredrikson M. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043; 2023.