

PREDICTIVE MODELING.

Predictive models turn telemetry into probabilities — that a sign-in is hijacked, that a CVE will be exploited, that a host is beaconing. **A score is a forecast with an error rate, not a verdict; the program around it decides whether it helps.**

- CALIBRATED PROBABILITIES
- BASE RATES & FALSE ALARMS
- DRIFT & RETRAINING
- ADVERSARIAL INPUTS

01 THE MECHANISM

Labels Are the Product

A model learns from labeled outcomes — exploitation observed, an account confirmed compromised, malware confirmed on a host. **It can only predict what its labels captured.** Exploitation nobody observed is absent from EPSS training data by design. [01]

Features Feed a Score

EPSS derives about 2,850 features (exploit code, chatter, CVSS metrics, age) and gradient-boosted trees emit a calibrated 0–1 probability of exploitation in the next 30 days. **A score of 0.05 means 5%, not “low.”** [01]

Base Rates Set the Ceiling

When true events are rare, the false-alarm rate — not the detection rate — decides whether an alert deserves belief. Axelsson showed a useful intrusion detector needs a false-alarm rate on the order of 1×10^{-5} . **Volume decides credibility.** [05]

02 THE FLOW

From Signal to Score

BUILD

1

Frame: define the event (exploitation within 30 days, an unauthorized sign-in), the decision the score will drive, and who owns that decision. The NIST AI RMF MAP function puts context before metrics. [03]

2

Train: assemble labeled history and features; hold out data from a later period so the test resembles deployment. Document the limits of validity beyond the conditions the model was developed under. [03]

3

Calibrate: compare predicted probabilities with observed rates, then measure coverage, effort and efficiency at candidate thresholds instead of quoting one accuracy number. [01]

From Score to Action

OPERATE

1

Threshold: choose the operating point from coverage against effort. A workload the size of CVSS Critical maps to roughly EPSS ≥ 0.04 , the 90th percentile. [02]

2

Act: route scores into a control that can respond — risk-based Conditional Access that requires MFA or a secure password change at a chosen level, or a remediation queue ordered by probability. [06]

3

Monitor: watch inputs and outcomes in production and retrain when performance leaves the calibrated baseline. Confirmed exploitation or compromise overrides any score. [01] [02]

03 CORE PILLARS & COVERAGE

PILLAR	WHAT IT COVERS
Calibrated Output	A probability that matches observed frequencies. EPSS scores average about 2.8% with a median of 0.7% — a long tail, so thresholds come from the distribution, not from intuition.
Population vs. Local Context	EPSS scores the global population; it does not know presence, reachability or consequence in your environment. Localize every score with those three checks.
Real-Time and Offline Paths	Entra scores some risk at sign-in (5–10 minutes to report) and some offline with more context (up to 48 hours); a risk level can change after the fact.
Learning Periods	Behavioral baselines need history: unfamiliar-sign-in detection stays in learning mode for at least five days; atypical travel learns for 14 days or 10 logins.

04 SCORING APPROACHES COMPARED

APPROACH	SUFFICIENT ALONE?	WHAT IT MISSES
Static Severity (CVSS)	No	Ranks technical impact with no empirical calibration; roughly the top 10% is “Critical” whether or not anyone ever exploits it.
Confirmed Exploitation (KEV)	No	Records what already happened: complete evidence, zero foresight. A listing overrides a low forecast, never the reverse.
Rules and Signatures	No	Precise for known patterns and blind to new ones; every rule’s false-alarm rate compounds against rare true events.
Behavioral Baselines	No	Flags departure from a learned norm; needs a learning period and can baseline an attacker who was already present.
Calibrated Predictive Model	No	Best efficiency at a given coverage, yet population-level and label-bound; needs local context, drift monitoring and an override for confirmed events.

PREDICTIVE MODELING.

SCORE · CALIBRATE · ACT

05

PROGRAM OPERATIONS

Measure the operating point, not the model.

Report coverage (true events caught), effort (items acted on) and efficiency (actions that were real) at the threshold you actually run; a random baseline and the previous method are the honest comparators. [01]

Rank by probability, then override by evidence.

Order the queue by forecast, but treat anything on CISA KEV as exploited regardless of score; BOD 26-04 timelines run from KEV status, exposure, automatability and impact — three days at the top tier. [02] [09]

Close the loop on outcomes.

Feed confirmed compromises, dismissed false positives and “confirmed safe” decisions back as labels; NIST AI RMF MEASURE 2.4 expects production monitoring and MANAGE 4.1 a post-deployment monitoring plan. [03]

Budget the analyst's attention.

Count expected false alarms per day before enabling automated response: 10^5 events a day at a 10^{-5} false-alarm rate is one false alarm a day; at 10^{-3} it is a hundred. [05]

06

CONTROL MAPPING

VULNERABILITY TRIAGE

EPSS publishes a daily 0–1 probability and percentile for every CVE via CSV and API. Do not multiply EPSS by CVSS — the product has no interpretable meaning. Use EPSS for the long tail and KEV for confirmed activity. [02]

IDENTITY RISK

Entra ID Protection scores low, medium and high with machine-learning algorithms; risk-based Conditional Access requires MFA or a secure password change at a chosen level. Leaked credentials and verified threat-actor IP are always high: evidence, not inference. [06] [07]

CLOUD TELEMETRY

Amazon GuardDuty applies threat-intelligence feeds and ML models to CloudTrail management events, VPC flow logs and DNS logs; Extended Threat Detection correlates individually harmless events into attack sequences. [08]

MODEL GOVERNANCE

NIST AI RMF organizes the work as GOVERN, MAP, MEASURE and MANAGE, with valid and reliable as the first trustworthiness characteristic; NIST AI 100-2 catalogs the training-time poisoning and deployment-time evasion to test against. [03] [04]

07

WHAT THIS DOES NOT SOLVE

Novel Attacks Have No Labels

A forecast reflects observed history. **A technique never seen in training data scores like background noise.**

Adversaries Adapt

Evasion crafts inputs that look normal; poisoning corrupts the training set. **The model becomes a target.** [04]

Context Lives Outside the Model

Presence, reachability and blast radius are yours to supply. **A high score on an asset you do not have is nothing.** [02]

Rare Events Stay Hard

No calibration repairs a base rate of one in a million. **Evidence and human review still close the case.** [05]

08

MYTHS VS. REALITY

MYTH	REALITY
“A 5% EPSS score means ignore it.”	Across thousands of CVEs, 5% is a substantial number of exploitation events ; set thresholds from coverage against effort. [02]
MYTH	REALITY
“Multiply EPSS by CVSS to get a risk score.”	A probability times an ordinal rank has no interpretable meaning . Combine the two by policy, not by arithmetic. [02]
MYTH	REALITY
“A 99% accurate detector is right 99% of the time it fires.”	With rare intrusions the false-alarm rate dominates the Bayesian detection rate; most alerts can still be false. [05]
MYTH	REALITY
“Once trained, the model is finished.”	Inputs and attackers drift. EPSS retrains when it leaves its calibrated baseline; NIST AI RMF expects production monitoring. [01] [03]
MYTH	REALITY
“A low score clears a KEV entry.”	KEV is confirmed past exploitation; EPSS is a 30-day forecast. Confirmed activity wins. [02] [09]

09

SPEAK THE PROGRAM — KEY TERMS

CALIBRATION

Predicted probabilities **match observed frequencies**: of items scored 0.10, about one in ten experiences the event.

BASE RATE

The prior frequency of the event. **Rare events** make even accurate detectors produce mostly false alarms.

COVERAGE / EFFORT / EFFICIENCY

The operating measures: events caught, items acted on, and **actions that were real**.

CONCEPT DRIFT

The relationship between inputs and outcomes **changes over time**, degrading a fixed model.

EVASION VS. POISONING

Deployment-time inputs crafted to fool a model versus **training-time corruption** of its data.

REAL-TIME VS. OFFLINE RISK

Scored at sign-in for immediate policy, or **after the fact with more context**, the level can change.