PITON Global

EXECUTIVE WHITE PAPER · WP-45

DATA SCIENCE · PHILIPPINES

# The model-reliability standard: the executive economics of data science outsourcing to the Philippines.

An analysis of why models built is a volume vanity metric, how model reliability and production impact — never model throughput — decide the true cost of a data-science operation once models that never ship, undetected drift, misleading validation, and rework are counted, and the vendor-selection discipline that gets a model into production and keeps it honest.


AUTHORS

**John Maczynski** — Chief Executive Officer
**Ralf Ellspermann** — Chief Strategy Officer



AUGUST 2026
MANILA · ESTABLISHED 2001

# Contents



**ABOUT THIS SERIES — VOLUME 45**

Every data-science proposal is priced per model built, because model count is easy to show. This volume is about the only data-science outcome a business actually depends on: the model that is honestly validated, shipped to production, and monitored so it stays reliable as the world moves. The gap between building a model and running a reliable one in production is where notebook-ware and silent drift live. The full series lives at piton-global.com.

THE PITON-GLOBAL LEADERSHIP TEAM

# About the authors

## John Maczynski

Chief Executive Officer

John has bought data-science capacity for three decades, and learned that the fast team that builds every model and ships none is the most expensive one a business can hire — because a model that never reaches production returns nothing, a model validated dishonestly misleads with confidence, and a model nobody monitors drifts until it quietly costs money. His standing question ignores the models-built report: of the models you built, how many were honestly validated, shipped to production, and still reliable a quarter later? Applied at PITON-Global, that question is what clients get on every data-science sourcing decision — personally, with no account layer in between.

j.maczynski@piton-global.com
+1 402-598-8740

## Ralf Ellspermann

Chief Strategy Officer

Ralf has spent twenty-five years running Philippine data-science floors, and he judges a team by its model reliability and production impact, not its models-per-quarter: whether the model was honestly validated and shipped, and whether it stayed reliable under monitoring. The data-science contract in Section 3 codifies what he learned rebuilding teams that built models fast and left them in notebooks. He now reads data-science providers the way he once ran them — from the production-deployment rate and the drift log backward.

r.ellspermann@piton-global.com
Manila, Philippines

PITON Global

ABOUT PITON-GLOBAL

PITON-Global is a buyer-side advisory practice, Manila-based since 2001, that has never operated a delivery floor or taken a placement commission. Its stock-in-trade is knowing the Philippine provider market at the operating level — 1,000+ mapped operations, re-verified continuously, including each data-science team's real model reliability and production-deployment record rather than its models-per-quarter

headline — and converting that knowledge into a shortlist of six to ten genuinely qualified partners per engagement. Buyers from Tripadvisor, Garmin, TSYS, HealthPartners, Yetipay, Workrise, and Norman Disney & Young have run their searches through the practice.

SECTION 01

# Executive summary

**-56%**
Fully-loaded cost per production-reliable model vs. US onshore

**88%+**
Production-deployment rate on vetted teams, up from 30–50%

**-64%**
Reduction in undetected drift and model rework on vetted teams

**6.2×**
First-year ROI, reconstructed case (Section 6)

A data-science engagement is bought on the price per model and paid for on the models that never earn. The vendor reports dozens of models built at a fraction of the onshore cost; the business finds the models stuck in notebooks that never shipped, the model validated on leaked data that looked brilliant and failed in production, the deployed model nobody monitored that drifted into losing money, and the data scientists rebuilding what should have worked. The operation hit its models-built number and the business inherited notebook-ware, a misleading validation, and a drifting model in production. The gap between building a model and running a reliable one in production is where the real cost of cheap data science hides, because a model that does not ship or does not hold up is not throughput delivered — it is a promise that never paid.

The Philippines — English-fluent, quantitatively deep, with a maturing data-science and MLOps talent pool — is where data science is delivered with the most model-reliability discipline, because that pool can run the honest validation, production deployment, and drift monitoring that reliability demands. Five findings:

1 **Models built is the industry's most flattering number.** Any team can train models in a notebook. Vetted teams are measured on outcome: honest validation, production deployment, and sustained reliability. Weak providers quote a low per-model price the business repays in models that never ship, misleading validation, and drift. Section 2 shows the gap.

2 **The production-reliable model, not the trained notebook, is the product.** A model that returns comes from honest validation, a path to production, and monitoring that keeps it reliable — not from a high training metric on a slide. Teams that deliver reliability ship models that earn; teams that build loosely leave notebook-ware and drift that cost more than the model.

3 **A bad model cuts both ways — never ships, or ships and drifts.** A model stuck in a notebook returns nothing on the spend; a model shipped on dishonest validation drifts and loses money with false confidence. Vetted teams hold production-deployment above 88% and monitor for drift precisely because both tails cost — the model that never earns and the one that quietly stops.

4 **The unshipped model and the silent drift, not the training run, are what cost.** A model done wrong costs the spend that never returned, the wrong decisions a drifting model drives, and the rework to fix it — many times the per-model saving. Teams that deliver reliability ship models that earn and catch drift early, and that production impact — not the offshore rate alone — is where the case study's return is built.

5 **Contract on the production-reliable model, not the built one.** The contracting failure of data science is paying per model and hoping it ships and holds up. Vetted deals price against production-deployment floors, validation-honesty standards, and drift-monitoring SLAs, making reliable models the team's problem, which is where it belongs.

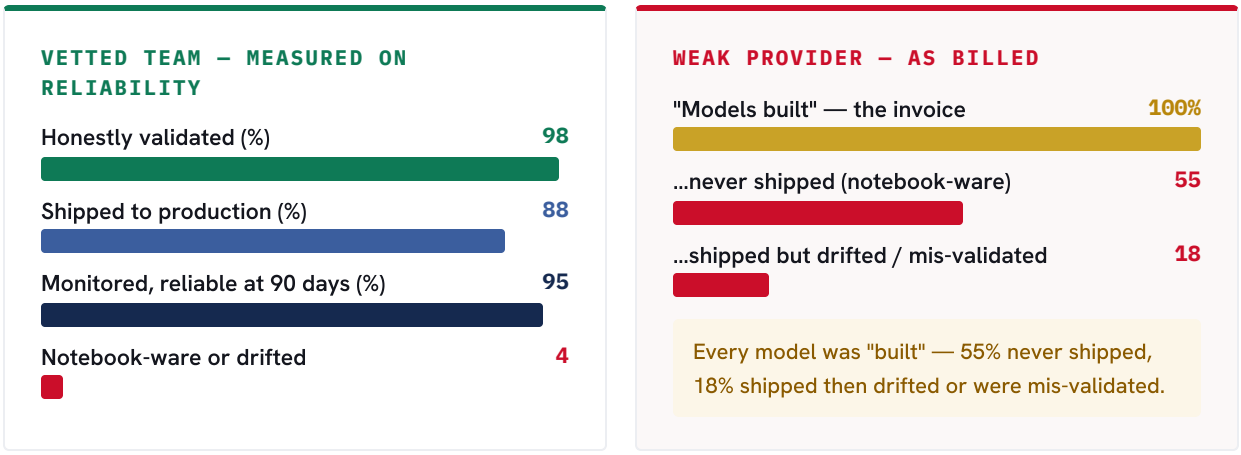
The intended reader is a chief data or AI officer, head of data science, or COO buying data science per model and paying for it in notebook-ware, misleading validation, and drift. Sections 2–4 separate the teams that deliver reliability from the teams that build models; Sections 5–7, the method and the proof.

SECTION 02

# The volume mirage: models built versus production-reliable models

Figure 1 shows one year of a data-science team's output two ways: as the models-built report a weak vendor bills, and as the model-reliability reality production and the drift monitor experience. The models-built number is identical in spirit to a chatbot's containment rate — a figure engineered to look like output while notebook-ware, misleading validation, and silent drift accumulate underneath.

FIG. 1 — ONE YEAR OF MODELS: MODEL-RELIABILITY QUALITY VS. REPORTED VOLUME




Left: PITON-Global engagement instrumentation, 2025–26, model reliability audited. Right: composite of pre-engagement vendor reporting reviewed in diligence.

The economics follow the model reliability, not the models count. On the left, eighty-eight percent of models ship to production and ninety-five percent are still reliable at ninety days, so the spend earns; on the right, a cheap per-model price leaves fifty-five percent in notebooks and ships eighteen that drift or were mis-validated, paying in dead spend, wrong decisions, and rework at once. A low per-model price that never ships or silently drifts is not a saving; it is dead spend and false confidence, delayed and compounding. The model-reliability audit — validation honesty, production deployment, drift monitoring — is the cheapest due-diligence instrument in this paper.

SECTION 03

## The data-science contract: validate it honestly, ship it to production, monitor it for drift

Figure 2 lays out the three clauses of a working data-science contract — the machinery a floor read verifies before any per-model price is believed.

FIG. 2 — THE DATA-SCIENCE CONTRACT, AS OPERATED ON VETTED TEAMS

**VALIDATE IT HONESTLY**

The model is validated with no leakage, an honest hold-out, and metrics tied to the business objective — not a flattering number on training data, and a weak result reported rather than hidden. Honest validation is what stops a model that looks good from failing the moment it meets real data.

**SHIP IT TO PRODUCTION**

The model is engineered to deploy — packaged, integrated, and put behind a real prediction path — so it earns in production rather than dying in a notebook. Shipping is the step most model factories skip, and the only one where a model starts returning on its cost.

**MONITOR IT FOR DRIFT**

Live performance, data drift, and concept drift are monitored with alerting and a retraining trigger, so a model that starts to slip is caught and refreshed rather than quietly losing money — with a reliability report the business can read. Monitoring is what keeps a shipped model reliable as the world it models moves.

↓ yields ↓

**THE PRODUCTION-RELIABLE MODEL**

HONESTLY VALIDATED · 88%+ IN PRODUCTION · DRIFT-MONITORED

Models are validated honestly, shipped to production, and monitored for drift — so notebook-ware disappears, deployed models earn, and a slipping model is caught before it costs. The operation gets cheaper per production-reliable model and more impactful at the same time, which is the entire promise of outsourcing data science to people who measure model reliability, not models built.

Operating parameters from PITON-Global data-science engagements, 2025–26. Model development is governed by the client's data-science leadership and MLOps standards.

In diligence, the contract is tested from production's side: pull built models and re-validate a sample for leakage and honest metrics; check the production-deployment rate against models built; audit the drift-monitoring and retraining record. Perhaps one team in ten survives it cleanly. Steps 3 and 5 of the framework exist to find that one before the contract does — because everything in Table 3 of Section 6 depends on model reliability, not models built.

SECTION 04

## Cost and performance: cost-per-production-

# reliable-model economics and benchmarks

**TABLE 1 — FULLY-LOADED COST PER PRODUCTION-RELIABLE MODEL: THREE OPERATING MODELS (2026)**

| OPERATING MODEL | COST / MODEL | VS. US BASELINE |
|---|---|---|
| US onshore in-house data-science team | $47.5k | baseline |
| Low-cost offshore model shop (per-model) | $32.0k | −33% |
| **Vetted Manila team — validated, shipped, monitored** | $21.0k | −56% |

Basis: PITON-Global engagement pricing 2025–26; cost per production-reliable model = fully-loaded operating cost divided by models honestly validated, shipped to production, and reliable at 90 days (notebook-ware and drifted/mis-validated models excluded); the model shop looks cheap per model but a low deployment rate, misleading validation, and drift lift its true cost per reliable model; US comparator fully-loaded in-house cost.

**TABLE 2 — DATA-SCIENCE BENCHMARKS, VETTED VS. BASELINE (2025–26)**

| METRIC | VETTED TOP TIER | BASELINE | EXECUTIVE READING |
|---|---|---|---|
| **Production-deployment rate** | 88%+ | 30–50% | The only model number returns follow |
| **Validation honesty (holds in production)** | 96%+ | 70–84% | Real, or a leaked-data mirage |
| **Undetected drift vs. baseline** | −64% | baseline | "Monitor it for drift" — held or not |
| **Time-to-production vs. baseline** | −47% | baseline | Earning sooner, or stuck in review |
| **Model rework rate vs. baseline** | −61% | baseline | Built once, or rebuilt to ship |

Source: PITON-Global data-science engagement data 2025–26; baseline = typical pre-engagement operations billed on models built.

The strategic read: the model shop's −33% per model is a false economy — a low deployment rate, misleading validation, and drift lift its true cost per reliable model far above the per-model gap, and a model that never ships or silently drifts is a cost no rate recovers. The vetted Manila team's −56% is real because it is measured where the value is: the model honestly validated, shipped to production, and

monitored. Buy on models built and you buy the right column of Figure 1; buy on cost-per-production-reliable-model and you buy model reliability — the same discipline every volume in this series keeps arriving at from a different door.

SECTION 05

## Sourcing discipline: the 7-Step Vendor Vetting Framework, applied to data science

Data-science diligence adds one hard requirement to every step: a per-model price is real only when the model reliability behind it survives a re-validation for leakage and a check of the production-deployment record. The funnel below is representative of a 20–50-seat data-science search.

**FIG. 3 — THE 7-STEP FRAMEWORK AS A DATA-SCIENCE FUNNEL**

| | STEP | WHAT IT ELIMINATES IN A DATA-SCIENCE SEARCH | FIELD REMAINING |
|---|---|---|---|
| 1 | **Requirements definition** | Throughput-first scoping: fixes the production-deployment floor, the validation-honesty standard, the drift-monitoring SLA, and the time-to-production target the operation must meet by use case | 1,000+ |
| 2 | **Market mapping** | Model shops: operations that bill models built and treat notebook-ware and drift as the client's problem, and generalist floors with no MLOps or production-ML depth | ~38 |
| 3 | **Capability screening** | Teams without the machinery: production-ML engineers, audited deployment history, an honest-validation and MLOps workflow, drift monitoring, and coverage across your model use cases | ~11 |
| 4 | **Compliance & security audit** | Weak governance: data protection and residency, model governance and reproducibility, access controls on training data and the ML platform, SOC 2 scope covering the data-science stack | ~7 |
| 5 | **Operational diligence** | The re-validation: re-validate built models for leakage and honest metrics; check production-deployment rate against models built; audit the drift-monitoring record; interview on how they handle a model that fails in production | 4–6 |
| 6 | **Competitive RFP & negotiation** | Per-model pricing: finalists bid against production-deployment floors, validation-honesty standards, and drift-monitoring SLAs — all carrying credits for models that never ship or drift | 2–3 |
| 7 | **Transition & launch governance** | Cold start: the team calibrates against your ML platform, use cases, and deployment path on a pilot model before it owns delivery, with the deployment floor certified before steady state, PITON-Global embedded buyer-side | 1 |

Field-remaining counts represent a 20–50-seat data-science search, PITON-Global engagements 2024–26.

The standing guarantees hold: **vendor neutrality** — PITON-Global operates no data-science floors, builds no models, and takes no placement economics, so the shortlist reflects audited model-reliability performance, not inventory; **right-sizing** — teams for whom your operation is a flagship account. Six to ten genuinely qualified providers, delivered within 48–72 hours, free to the buyer, competing through a fully transparent, fair, comprehensive, and highly competitive RFP.

SECTION 06 · ANONYMIZED CASE STUDY

# Re-basing the operation on model reliability: a 40-seat data-science rebuild, reconstructed

Engagement DZ-059 is a US enterprise — commissioning roughly 90 models a year across forecasting, propensity, and risk use cases — that had offshored data science to a per-model shop. The model price was excellent; the reliability was not: barely a third of models reached production, several had been validated on leaked data and failed live, deployed models were drifting undetected, and the business had stopped believing the roadmap. The chief AI officer wanted the capacity without the notebook-ware and the drift. Identifying details are anonymized under NDA.

### SELECTION (WEEKS 0–8, RE-VALIDATIONS)

Requirements fixed an 88% production-deployment floor, a validation-honesty standard, and a drift-monitoring SLA, with a mandatory honest-validation and MLOps workflow. Mapping produced 38 claimants; the machinery screen — production-ML engineers, deployment history, honest validation, drift monitoring — left 11; the model-governance audit left 7. Blind re-validations of each finalist's built models for leakage and a production-rate check shortlisted 5. A Manila team won priced against production deployment and drift monitoring, with a re-validation written into the reporting spec.

### TRANSITION (MONTHS 2–7, CALIBRATED FIRST)

The team calibrated against the enterprise's ML platform, use cases, and deployment path on a pilot model before owning delivery, and stood up drift monitoring on the existing deployed models; the deployment floor was certified in month 4. The honest-validation-and-MLOps loop fed feedback from day one. By month 7 the production-deployment rate reached 88%, validation honesty hit 96%, undetected drift fell 64%, and time-to-production dropped 47%. PITON-Global chaired the model-reliability review board to steady state.

| TABLE 3 — DZ-059 FIRST-YEAR VALUE BUILD-UP (US$) | | |
|---|---|---|
| VALUE COMPONENT | YEAR 1 | BASIS |
| Labor arbitrage on the 40-seat data-science operation | $1.62M | 75.0k productive hrs × ($31.50 − $9.90) |
| Production impact captured (models that ship) | $0.88M | Return on models now earning in production |
| Drift & mis-validation loss avoided | $0.34M | Losses from drifting/mis-validated models removed |
| Model-rework reduction | $0.16M | Rebuild-to-ship rework removed |
| **Gross value created** | **$3.00M** | |
| Less: engagement & transition costs | ($0.48M) | Calibration, MLOps & drift-monitoring build, platform integration; advisory fee $0 (provider-paid model) |
| **Net value ÷ cost = first-year ROI** | **6.2×** | $3.00M ÷ $0.48M |

Figures rounded; deployment, validation, and drift effects audit-verified. Method in Methodology & notes.

Fifty-four percent arbitrage, forty-six percent from production impact, avoided drift losses, and less rework — a typical split for the series, earned because the operation was re-based on model reliability rather than model volume. Every line traces to a diligence step: the deployment floor to Step 1, the reliability-based pricing to Step 6, the honest-validation workflow to Step 3's contract. The chief AI officer's line at the year-one review: "The model price barely moved. What moved is that models actually ship now — and the ones in production don't quietly stop working."

SECTION 07

## The executive playbook: governance for the first 180 days

A data-science transition succeeds when the team is measured and paid on production deployment and model reliability — client in command throughout.

**DAYS 0–30**

**Define reliability, contract on it**

Set the production-deployment floor, the validation-honesty standard, and the drift-monitoring and time-to-production thresholds with data-science leadership before launch, and document the honest-validation and MLOps workflow by use case. Contract against production deployment and reliability — never per model alone — with floors and SLAs carrying credits. Stand up the re-validation as the program's arbiter of truth.

**DAYS 30–90**

**Calibrate, ship a pilot, monitor**

Calibrate the team against your ML platform, use cases, and deployment path on a pilot model before it owns delivery, and stand up drift monitoring on existing deployed models; certify the deployment floor and honest-validation loop. Switch on the MLOps-and-monitoring loop from day one. Publish the model dashboard: production-deployment rate, validation honesty, undetected drift, time-to-production, rework rate.

**DAYS 90–180**

**Own delivery, certify steady state**

Hand delivery to the team only after it holds the deployment floor and validation-honesty standard. Review the model board monthly with the drift and deployment-rate trends on the table. Certify steady state on two quarters at floor including a retraining cycle; hand over on documented criteria; keep the quarterly re-validation standing.

**THE FIVE FAILURE MODES WE ARE HIRED TO PREVENT**

Paying per model while models never ship · building notebooks instead of shipping to production · validating on leaked data and reporting a mirage · deploying models with no drift monitoring · owning delivery before production reliability is proven. All five are settled at selection and contract.

The data-science argument, in one sentence: a built model is activity and a production-reliable model is honest validation, a live deployment, and monitoring that keeps it earning, and a vetted Philippine team delivers –56% cost per production-reliable model with an 88%+ deployment rate, honest validation, and undetected drift cut by nearly two-thirds — for buyers who re-validate the reliability instead of counting models built, and compete the finalists through a fully transparent, fair, comprehensive, and highly competitive RFP. Anyone can build a model. We make sure you buy the team whose models ship and stay reliable.

METHODOLOGY & NOTES

# How the numbers were built

**Reliability and deployment figures** (Fig. 1, Table 2) derive from engagement instrumentation in PITON-Global data-science engagements, 2025–26. Production deployment and validation honesty are measured by re-validating built models for leakage and checking deployment against models built; undetected drift, time-to-production, and rework are measured against the client's pre-engagement baseline. The weak-provider column is a composite of pre-engagement vendor reporting reviewed in diligence, shown for contrast.

**Cost-per-production-reliable-model figures** (Table 1) are fully loaded — wages, benefits and statutory charges, facilities, compute and ML-platform tooling, management and QA overhead including MLOps engineers and validation reviewers, attrition-driven recruiting — divided by models honestly validated, shipped to production, and reliable at 90 days, with notebook-ware or drifted/mis-validated models excluded from the denominator; US comparators reflect fully-loaded in-house cost on the same basis. A production-reliable model is honestly validated, deployed, and monitored.

**The case study** (DZ-059) is anonymized under NDA; volumes and dates are generalized, figures rounded. Deployment, validation, and drift effects are audit-verified; the production-impact line reflects return on models now earning, and the drift-loss line is a conservative estimate. ROI = net first-year value ÷ all engagement-related client costs. PITON-Global's advisory fee is provider-paid and appears at $0 in the client cost base; neutrality safeguards are documented at piton-global.com.

**Market context** draws on IBPAP reporting and PITON-Global's proprietary provider mapping (1,000+ operations; model reliability and production deployment re-verified May 2026). Model development is governed by the client's data-science leadership and MLOps standards. Estimates and projections are PITON-Global analysis and should be cited as such. The full series is available at piton-global.com.