— PORTFOLIO CASE STUDY

SCM

Supplier Performance & Risk Scorecard

Forward-looking supplier evaluation for aerospace and defense sourcing
Kenyon Woodley | Western Washington University | 2026 | Stack: Python + Streamlit + Claude API

01 — PROBLEM

Most supplier evaluations are report cards. They describe the past. They don't surface the risk that's coming.

In aerospace and defense, many late deliveries are not complete surprises. The signals were there: a supplier running single-site, a sole-source relationship with no qualified alternate, a geography with increasing geopolitical friction. Sourcing teams see those signals in isolation, in scattered spreadsheets and quarterly reviews, but rarely in a system that aggregates them into a coherent risk picture before the problem lands on the production line.

The default approach: tracking OTD and quality, escalating when something breaks: is reactive by design. It treats every supplier as a known quantity until it isn't. In a supply chain where lead times are measured in months and re-sourcing is measured in years, reactive is expensive.

> "How healthy is this supplier, really?" A question that deserves a structured answer, not a gut check: and one that should surface risk before it materializes, not after.

02 — INSIGHT

Performance tells you what happened. Risk tells you what's coming. A useful scorecard needs both layers.

Standard supplier scorecards measure delivery, quality, and responsiveness. Those dimensions matter, but they are backward-looking by nature. A supplier with 97% OTD and ITAR registration missing is not a safe supplier for defense work. A supplier with strong quality metrics and a single manufacturing site in a geopolitically exposed region carries asymmetric downside risk that the performance score doesn't capture.

The key architectural decision was to split the engine into two explicit layers: a Performance layer that scores what the supplier has done, and a Risk and Resilience layer that scores what could go wrong. Both layers contribute to the composite score, but they are computed independently and displayed separately. A sourcing manager can see in a single view whether a supplier's composite is being held up by strong performance despite high risk, or vice versa: and act accordingly.

Compliance is handled outside both layers entirely. A missing ITAR registration or AS9100 gap is not a scoring input: it is a hard flag that renders independently and can cap the recommendation regardless of composite score.

03 — ENGINE ARCHITECTURE

A deterministic two-layer scoring engine with ordinal recommendation logic

The engine is built in Python across five modules with strict separation of concerns. Scoring, compliance, and recommendation are fully deterministic. Claude generates the plain-language narrative and improvement actions from the computed scores: it never generates or modifies a number.

COMPOSITE SCORE FORMULA

Composite = (Performance Score × Layer Weight) + (Risk Score × Layer Weight)

Three scoring modes shift layer emphasis without changing dimension weights within each layer:
Balanced: Performance 50% / Risk 50% (default)
Performance Recovery: Performance 70% / Risk 30% (use when rehabilitating a supplier)
Risk Reduction: Performance 30% / Risk 70% (use when evaluating geopolitical or single-source exposure)

Dimension weights within each layer are commodity-profile-aware: Machined Parts, Electronics, Raw Material, Castings/Forgings, Standard
All lookup tables, band definitions, and mode weights are centralized in models.py — single source of truth

The recommendation engine produces one of four outcomes in fixed order of severity: Expand relationship, Maintain with monitoring, Issue corrective action plan, Initiate re-sourcing evaluation. Compliance caps are applied after the score-based recommendation. Caps can only increase severity, never reduce it. A supplier with a high composite score and a critical compliance gap will receive a more severe recommendation than the score alone would produce. The AI narrative receives the final recommendation as a fixed input and explains it. It cannot change it.

04 — SCORING DIMENSIONS

Nine dimensions across two layers, each tied to a specific input and an explicit scoring function

Performance Layer: scores historical supplier behavior across four dimensions.

P1

ON-TIME DELIVERY

OTD % trailing 12 months. Piecewise linear scoring; baseline expectation 95%+. Weight shifts upward under Performance Recovery mode for commodity profiles where schedule is paramount.

P2

QUALITY

Defect rate in PPM or % (whole percent convention: 1 = 1%). Normalized to PPM internally via half-open band lookup, which eliminates boundary ambiguity. Castings/Forgings profile weights quality at 40%, the highest of any dimension.

P3

RESPONSIVENESS

Average business days to respond to commercial inquiries: PO acknowledgment, change requests, RFQ responses. One metric, one input. No blended averages that obscure the source.

P4

COST STABILITY

Purchase price variance % trailing 12 months. Positive PPV = cost creep. Piecewise linear; negative PPV (savings) scores 100. Raw Material profile weights this at 30%, reflecting commodity price sensitivity.

Risk & Resilience Layer: scores forward-looking exposure across five dimensions.

R1

SINGLE-SOURCE EXPOSURE

Nested lookup: single_source status sets base severity, backup source count refines within that condition. Annual spend drives an optional criticality multiplier: $1M+ single-source triggers an additional 20% penalty. Highest-weight risk dimension across all commodity profiles.

R2

GEOGRAPHIC CONCENTRATION

Four-tier risk system. Tier 4 (China, Taiwan) reflects geopolitical tension, export control friction, and strategic competition risk. Taiwan designated separately given semiconductor concentration and cross-strait exposure. Electronics profile weights this at 35%: highest geo weight of any profile.

R3

FINANCIAL FRAGILITY

Explicit deduction model: base score by size (Small 50, Mid 70, Large 85), minus financial concern flag (−25), minus price volatility band (Medium −10, High −25). Floor at 0. Max intentionally capped at 85: proxy, not verified financials.

R4

LEAD TIME VOLATILITY

Quoted vs. actual lead time variance %. Piecewise linear interpolation: eliminates the 20-point band cliffs that occur in step-function scoring. Signals capacity or production planning instability.

R5

OPERATIONAL REDUNDANCY

Single manufacturing site vs. multiple. Binary input, explicit score (90 / 20). Single-site suppliers carry disproportionate disruption risk for precision parts with no domestic backup manufacturing.

05 — RATING BANDS & RECOMMENDATION LOGIC

Four bands. One deterministic recommendation. Compliance caps the ceiling.

PREFERRED

85: 100

Expand relationship. Consider for additional scope and strategic sourcing agreements.

ACCEPTABLE

70: 84

Maintain with monitoring. Targeted improvement plan for lowest-scoring dimensions.

AT RISK

55: 69

Issue corrective action plan. Evaluate dual-sourcing for critical parts.

CRITICAL

0: 54

Initiate re-sourcing evaluation. Immediate escalation to leadership.

Compliance caps are applied after the score-based recommendation using ordinal index comparison: never string matching. Caps only increase recommendation severity, never reduce it. A supplier scoring 91 (Expand) with missing ITAR registration on applicable work is capped at Issue Corrective Action Plan. The cap reason is passed directly to the AI narrative, which must acknowledge it explicitly.

06 — COMPLIANCE HARD FLAGS

Compliance is not a scoring dimension. It is a hard gate that operates independently of composite score.

Each certification is evaluated against an applicability toggle: ITAR, CMMC, and NADCAP flags only apply when the scope of work makes them relevant. A supplier without ITAR registration on commercial-only work receives no flag. A supplier without ITAR registration on defense-applicable work receives a critical flag and a recommendation cap.

AS9100	WARNING	Aerospace quality management system standard. Always evaluated regardless of program type. Missing AS9100 caps recommendation at Maintain with monitoring: treated as a baseline expectation gap, not a disqualifier.
ITAR	CRITICAL	U.S. State Department registration for defense-related articles and services. Evaluated only when ITAR-applicable work is confirmed. Missing registration means the supplier should not be awarded ITAR-relevant work until registration is resolved. Caps recommendation at Issue corrective action plan and renders a critical flag regardless of composite score.
CMMC	WARNING	Cybersecurity Maturity Model Certification: DoD supply chain requirement. Evaluated only when DoD work is applicable. Level 2 required for controlled unclassified information (CUI). Level 3 called out distinctly for highest-sensitivity programs. Below Level 2 caps recommendation at Maintain with monitoring.
NADCAP	WARNING	Special process accreditation for heat treat, NDT, plating, and similar operations. Evaluated only when applicable processes are confirmed. Missing accreditation caps recommendation at Issue corrective action plan. Suppliers holding NADCAP without it being required receive a positive "Additional Capability" signal.

07 — AI NARRATIVE LAYER

Deterministic first. AI explains the result — it does not generate it.

The AI layer serves one purpose: translate a structured score profile into a readable supplier evaluation brief and a specific, actionable set of improvement recommendations. Every score and recommendation is computed before the AI is called. The prompt passes computed scores as fixed facts and instructs the model that implying a different recommendation is a defect.

When compliance caps are triggered: whether or not they escalated the recommendation: the prompt explicitly surfaces the gap and requires the narrative to address it. If ITAR is missing on applicable work, the brief must state "this supplier is disqualified from ITAR-relevant work until registration is obtained." Softened compliance language is treated as a prompt failure.

WHAT AI GENERATES

Supplier summary (2–3 sentences), key strengths grounded in high-scoring dimensions, primary risk factors in order of severity, suggested improvement actions specific enough for a supplier review meeting. One comparative paragraph when 2+ suppliers are loaded into the session table.

WHAT AI NEVER TOUCHES

Numerical scores, rating band assignment, recommendation selection, compliance flag status, cap logic. All of these are computed deterministically before the AI call. The model string and max_tokens are fixed in code: the AI layer is boxed, not open-ended.

08 — ADDITIONAL FEATURES

Data confidence scoring, what-if sensitivity, and multi-supplier comparison

FEATURE	WHAT IT DOES	WHY IT MATTERS
Data Confidence Badge	Scores data quality 0–100 from months of history, transaction volume, data source, and audit recency. Renders as High / Medium / Low badge alongside composite score.	A 92 composite on 18 months of audited data is not the same as a 92 on three POs and a vendor self-report. The badge makes that distinction visible.
What-If Sensitivity	Computes composite lift from realistic improvements on the two lowest-scoring performance dimensions. Filters zero and negative deltas: only genuine improvements are shown.	Converts the scorecard from a snapshot into a planning tool. Tells the sourcing manager which dimension improvement has the most leverage before the supplier review meeting.
Commodity Profile Weights	Five profiles (Standard, Machined Parts, Electronics, Raw Material, Castings/Forgings) shift dimension weights within each layer. Layer weights remain user-controlled via scoring mode.	Electronics suppliers should be weighted heavily on geographic risk. Castings should be weighted heavily on quality. A single fixed weight table treats all commodities identically: this doesn't.
Spend Criticality Multiplier	Annual spend field applies a penalty to single-source dimension score. $250K+ triggers 10% reduction; $1M+ triggers 20% reduction.	Single-source on a $2K/year part is not the same risk as single-source on a $2M/year part. Consequence scales with exposure.
Multi-Supplier Comparison	Session-state comparison table accumulates up to 5 suppliers. Plotly heatmap with green-to-red diverging scale across all 9 dimensions. Best fit by scoring mode shown for all three modes simultaneously.	Supplier decisions are rarely made in isolation. The comparison view surfaces which supplier wins under each evaluation lens: and lets the sourcing manager see the tradeoff before committing.

09 — SCORE LINEAGE

Every composite score is fully auditable from input to output

The Score Lineage expander in the UI renders a Plotly Sankey diagram showing exactly how weighted dimension scores flow into layer scores, and how layer scores combine into the composite. Each link's width is proportional to its weighted contribution. A sourcing manager being challenged on a score can open the lineage view and walk any reviewer through the exact math: dimension by dimension, weight by weight.

This is not decoration. In aerospace and defense sourcing, every recommendation is potentially reviewed by program managers, contracts, and legal. A scorecard that can't explain itself in a challenge setting is a liability. Score lineage makes the deterministic math fully transparent without requiring the reviewer to read code.

10 — TECHNICAL DESIGN

Five modules, 92 passing tests, one centralized source of truth

MODULE STRUCTURE

models.py → scoring.py + compliance.py → recommendation.py → ai_narrative.py → streamlit_app.py

```
models.py: all constants, lookup tables, enums, validated input schema: single source of truth
scoring.py: two-layer deterministic engine, confidence score, sensitivity analysis
compliance.py: flag evaluation, applicability-gated cap logic
recommendation.py: ordinal recommendation + compliance cap resolution
ai_narrative.py: Anthropic SDK call, prompt construction, comparison narrative
streamlit_app.py: full UI layer — imports from all five backend modules
```

The test suite enforces 92 assertions including: PPM band boundary correctness, lead time piecewise interpolation (no band cliffs), validation bounds across all input fields, cap logic separation (cap_triggered vs cap_escalated), same-severity compliance triggering, sensitivity filtering, weight table integrity across all commodity profiles, and confidence score label assignment.

11 — LIMITATIONS & SCOPE

What this tool does not do — and why those are explicit decisions

KNOWN **Manual inputs only.** No ERP integration, no supplier portal, no automated data pull. All inputs are entered by the sourcing analyst from memory, PO records, or a supplier review. This is a deliberate scope decision: integration dependencies would make the tool brittle and harder to demo in any environment.

KNOWN **No persistent supplier database.** The comparison table accumulates within a session and resets on refresh. There is no saved state, no supplier history, and no longitudinal tracking. Trend analysis across quarters would require a database layer that is explicitly out of scope.

KNOWN **Financial fragility is a proxy, not verified financials.** The financial dimension uses supplier size, a concern flag, and price volatility as inputs: not Dun & Bradstreet data, not audited financials. The 85-point maximum for the most favorable inputs is intentionally conservative to reflect this uncertainty.

KNOWN **Geographic risk is tier-based, not real-time.** Country tiers are static and do not reflect breaking news or real-time geopolitical events. The tier system is auditable and explainable: that's the tradeoff for not having a live data dependency.

KNOWN **No export.** The Streamlit screen is the artifact. There is no PDF or Excel export. Score lineage, compliance flags, and AI narrative are designed to be presented from the live tool, not forwarded as a document.

12 — BROADER CONTEXT

Fourth tool in the sourcing operations portfolio

The Supplier Performance & Risk Scorecard is the deepest engine in the portfolio so far, with the most scoring dimensions and the most complex recommendation logic. It is designed to sit downstream of the Part Prioritization Framework (which surfaces which suppliers need attention first) and inform the Make vs. Buy Decision Framework (which decides whether to source at all).

SUITE PROGRESS

→ Part Prioritization Framework: Complete

→ RFQ Lifecycle Tracker: Complete

→ Parametric Should-Cost Model: Complete

→ Supplier Performance & Risk Scorecard: Complete

→ Make vs. Buy Decision Framework: In Development

DESIGN PHILOSOPHY

Deterministic scoring. AI explains outputs, never generates them. Every number traces to an auditable input. Compliance is a hard gate, not a weighted variable. The tool produces the assessment: the sourcing engineer makes the call. Built for the sourcing professional who needs to defend a recommendation in a leadership setting, not for the analyst building dashboards.

13 — CLOSING NOTE

Built to close the gap between "we track OTD in a spreadsheet" and "we have a structured, defensible view of supplier health across performance, risk, and compliance." Deterministic by design. Explainable by requirement. Aerospace-specific by intent. **The engine produces the assessment. The sourcing engineer makes the call.**

Python engine + Streamlit UI + Claude API. | Two-layer scoring: 4 performance dimensions, 5 risk dimensions. | Three scoring modes, five commodity profiles. | Four compliance hard flags with applicability toggles and ordinal cap logic. | Data confidence scoring, what-if sensitivity, spend criticality multiplier. | Multi-supplier comparison heatmap with session state. | Score Lineage Sankey diagram. | 92 passing tests across 18 assertion groups. | Geo risk tiering across 40+ countries.