

MGMT 467 | Team 8 | Assignment 2 Team Ops Brief

Decision rule:

We optimize lives saved under limited capacity by maximizing expected utility at an operating threshold of 0.60 (favoring recall for true survivors over precision), with a fallback to 0.50 when capacity tightens further.

Evidence (baseline → engineered → cross-model)

Model A (engineered BQML LOGISTIC_REG + TRANSFORM)

- **Feature set:** Canonical Titanic fields (pclass, sex, age, sibsp, parch, fare, embarked) with imputation; engineered interactions/normalization; train/eval split in SQL.
- **Why it works:** Survival signal in sex × pclass, age, and fare is strong and fairly monotonic; linear log-odds capture most of the separation without overfitting.
- **Headline metrics:** Evaluation shows high separability (ROC AUC ≈ 0.86) and stable calibration around the mid-range of predicted probabilities. Confusion analysis at $\tau = 0.60$ raises TP / recall with a modest precision trade-off—appropriate for a rescue-style objective.

Model B (engineered BQML)

- **Feature set:** Adds family_size = sibsp + parch + 1, fare_bucket (low/mid/high), and sex_pclass interaction directly in SQL, then trains LOGISTIC_REG.
- **Headline metrics:** ROC AUC ≈ 0.85 , log-loss comparable to Model A. Calibration is similar; confusion at default $\tau = 0.50$ is balanced, but expected-cost is slightly worse than Model A at our preferred operating point ($\tau \approx 0.60$).

Takeaways across models

- Both models generalize well and beat a manifest-only baseline.
- Marginal lift from extra features (Model B) shows up more in ranking within pclass=3 than in global AUC; however, under a recall-favoring threshold, Model A edges out on expected-cost because it's a touch better calibrated in the 0.55–0.70 band where our policy operates.
- **Practical implication:** If we must choose one global model today, we ship Model A at $\tau = 0.60$. If we support segmented deployment later, we will revisit Model B's strengths on large families and low-fare cohorts.

Policy (deployment, cost, fairness)

- **Deploy now:** single global Model A @ $\tau = 0.60$ (recall-leaning; lowest expected cost with $C_{FN}=4$, $C_{FP}=1$).
- **Segment later:** shadow test a $pclass=3$ / $family_size>1$ specialization (Model-B style) only if it reduces expected cost without >5 pp parity gaps.
- **Fairness guardrail:** keep precision gaps < 5 pp across sex, pclass, embarked; investigate segmented thresholds if exceeded.

Monitoring

- **Weekly (owner: Artemii):**
 - **ECE** (calibration) and expected-cost@ τ on closed cohorts.
 - Precision@ τ by group + parity gap; PSI on age, fare, sex
- **Auto-actions:**
 - If $ECE > 0.05$ or parity ≥ 5 pp for 2 weeks \rightarrow run τ sweep 0.50–0.65 and recalibrate.
 - If capacity $>95\%$ for 3 days \rightarrow temporary $\tau = 0.50$ (ops override), log & review.

Risks & mitigations

- **Label delay / mismatch** \rightarrow evaluate on delayed cohorts; keep shadow logs.
- **Calibration drift** \rightarrow temperature scaling monthly or when ECE trigger fires.
- **Segment bias** \rightarrow parity alerts; try segmented thresholds or group-aware calibration.
- **Linear limits** \rightarrow keep Model-B features staged for segment pilot.

Repository hygiene & repeatability

- **SQL-first layout:** clean views \rightarrow CREATE MODEL \rightarrow ML.EVALUATE \rightarrow confusion@ τ \rightarrow expected-cost \rightarrow group cuts.
- **Params:** project/dataset names; fixed imputations; log τ + cost matrix each run.

Model governance

- **Assumptions:** manifest features stable; FN harm \gg FP; transforms frozen pre-train.
- **Limits:** linear boundary; sensitive to sex/pclass coding; no post-split feature leakage.

- **Controls:** access-limited model dataset; PR + Ops/Compliance sign-off for any τ /feature/cost change.

Bottom line

- **Ship:** Model A @ $\tau = 0.60$ (global).
- **Keep warm:** Model-B feature pipeline for a targeted pclass=3 / family_size>1 experiment.
- **Watch:** weekly calibration, expected-cost, and <5 pp precision parity.