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 ($\tau \approx 0.60$).

Model C (subgroup specialization – pclass = 3, BQML LOGISTIC_REG)

- **Feature set:** Same engineered features as Model B (family_size, fare_bucket, sex_pclass, plus pclass, sex, age, fare, embarked), but trained only on third-class passengers (pclass = 3).
- **Why it works:** Third-class passengers have different survival patterns with lower fares, bigger families, and lower survival odds. Training only on this group helps the model focus on the patterns that matter most for them.
- **Headline metrics:** On the pclass = 3 group, Model C shows similar or slightly better AUC than Model B and better calibration in the middle probability range. Confusion at $\tau = 0.50$ gives higher recall with a small drop in precision.

Takeaways across models

- All 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.
- **Model C**, trained only on **pclass = 3**, fits that group's patterns better and gives **higher recall** and slightly better calibration for that subgroup.
- **Practical implication:** If we must choose one global model today, we ship Model A at τ = 0.60. If we support segmented deployment later, we will revisit Model B's strengths on large families and low-fare cohorts.
- **Segment decision**: If we ever use different models for different groups, Model C would be the best option for third-class passengers, especially large families and low-fare travelers.

Policy (deployment, cost, fairness)

Monitoring

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

Risks s mitigations

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

Repository hygiene s repeatability

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

Model governance

- **Assumptions:** manifest features stable; FN harm ≫ 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 @ τ = 0.60 (global).
- **Keep warm:** Model-B and Model C feature pipeline for a targeted pclass=3 / family_size>1 experiment.
- **Watch:** weekly calibration, expected-cost, and <5 pp precision parity.