# Team 1 Ops Brief

## Model A (Pre-Embarkation Baseline, Global) - *Arman Hyder*

We address how a maritime operator should prioritize limited staff attention during an emergency to maximize survival odds without systemic unfairness. Our baseline uses a global logistic-regression model with manifest-only features (Pclass, Sex, Age, Fare, Embarked) and a standard operating threshold of 0.50 to keep the rule simple, auditable, and ready for immediate use. If operations later impose a recall floor or a fixed capacity, the threshold—or a top-K policy—can be adjusted accordingly.

### Evidence

The model is trained and evaluated in BigQuery ML with an 80/20 split. ML.EVALUATE provides AUC (ranking quality) and log_loss (calibration), while ML.CONFUSION_MATRIX at 0.5 clarifies the balance between missed rescues (FN) and false alarms (FP). Coefficients from ML.WEIGHTS are interpretable: being female and paying higher Fare increase survival likelihood; lower class number (1st class) correlates with higher survival; Age effects are monotone on average but can be non-linear. These patterns are pretty intuitive.

### Policy

We recommend a global deployment at 0.50 for the initial hand-off. Using a simple expected-cost framework with FP cost = 1 (wasted staff cycle) and FN cost = 5 (missed rescue), the operating loss is FP×1 + FN×5 computed from the confusion matrix; if FN dominate under real constraints, lower the threshold or switch to a top-K policy sized to staff capacity. Because historical outcomes differ by Sex and Pclass, fairness is monitored, not hard segmented at baseline: track precision and recall by subgroup and trigger review if any gap exceeds five percentage points.

### Monitoring

Monitor log_loss (or Brier) and global recall weekly, along with subgroup parity; re-check threshold monthly and recalibrate if drift increases error or parity gaps persist. Model A is a compact, auditable starting point that delivers immediate triage value and establishes clear procedures for threshold or top-K tuning and fairness governance as operational objectives and capacity limits are finalized.

## Model B – Daniel Gallagher

We address how a maritime operator should prioritize limited staff attention during an emergency to maximize survival odds without systemic unfairness. Our engineered model (Model B) builds upon a baseline by incorporating features intended to capture family and class interactions, using a global logistic-regression model with features including Pclass, Sex, Age, Fare, Embarked, family_size, fare_bucket, and sex_pclass interaction. A standard operating threshold of 0.50 is used initially, keeping the rule relatively simple. If operations later impose a recall floor or a fixed capacity, the threshold—or a top-K policy—can be adjusted accordingly.

### Evidence

The model is trained and evaluated in BigQuery ML. ML.EVALUATE provides AUC (ranking quality) of 0.844109 and log_loss (calibration) of 0.461262, indicating good discriminatory power and calibration. ML.CONFUSION_MATRIX at 0.5 clarifies the balance between missed rescues (FN) and false alarms (FP). From the confusion matrix (TN=202, FP=24, FN=60, TP=72), we see the model is better at identifying non-survivors than survivors at this threshold, with a notable number of false negatives. Coefficients from ML.TRAINING_INFO (not explicitly shown in the provided cells but available from the model) would show the impact of features like Sex, Pclass, and engineered features on survival likelihood. The engineered features like family_size, fare_bucket, and sex_pclass_interaction are intended to capture more nuanced patterns than the baseline.

### Policy

We could recommend a global deployment at 0.50 for an initial assessment. Using a simple expected-cost framework with FP cost = 1 (wasted staff cycle) and FN cost = 5 (missed rescue), the operating loss is FP×1 + FN×5 computed from the confusion matrix (24*1* + *60*5 = 24 + 300 = 324). If FN dominate under real constraints, lowering the threshold or switching to a top-K policy sized to staff capacity could be considered. Because historical outcomes differ by Sex and Pclass, fairness should be monitored, not hard segmented at baseline: track precision and recall by subgroup and trigger review if any gap exceeds five percentage points.

### Monitoring

Monitor log_loss (or Brier) and global recall weekly, along with subgroup parity; re-check threshold monthly and recalibrate if drift increases error or parity gaps persist. Model B, with its engineered features, is a step towards capturing more complex relationships in the data, aiming for improved predictive performance while maintaining a degree of interpretability.

## Model C – Ethan Garcia

### Decision Rule

Prioritize maximizing the identification of potential survivors within the Pclass 3 subgroup using a policy informed by Model C, potentially adjusting the operational threshold from 0.5 based on the trade-off between false positives and false negatives and available resources for this group.

### Evidence

Model C, a logistic regression, was specialized for Pclass 3 passengers. Key evaluation metrics for this subgroup are AUC 0.751, log loss 0.467, and accuracy 0.788. The confusion matrix at a 0.5 threshold for Pclass 3 is: TN=352, FP=35, FN=71, TP=43. This shows strong performance in identifying non-survivors but a significant number of false negatives (71) among actual survivors in Pclass 3 at this threshold.

### Policy

Model C supports a specialized policy for Pclass 3. The 0.5 threshold is conservative, minimizing false alarms but missing survivors. Given the high number of false negatives, a lower threshold or a top K approach tailored to Pclass 3 capacity could improve survivor identification. An expected-cost analysis (FP cost=1, FN cost=5) on the Pclass 3 confusion matrix yields a loss of 390, highlighting the impact of false negatives. This specialization allows for targeted resource allocation where risk is concentrated.

### Monitoring

Monitor Model C's performance on incoming Pclass 3 data weekly. Track AUC, log loss, precision, recall, and confusion matrix components specifically for this subgroup. This helps detect data drift within Pclass 3, ensuring the policy remains effective. Ownership for monitoring should be assigned to ensure consistent oversight.

## Model D – Aditya Ghorpade

### Decision Rule

We optimize life-safety first, efficiency second by minimizing an expected-cost objective where false negatives are 1.7× more costly than false positives (C_FN = 1.7 · C_FP). Operationally, the rule is: *"Prioritize passengers whose predicted probability of survival from the engineered Model B exceeds ~0.25."* This global threshold (~0.25 instead of the default 0.50) is chosen from a sweep over thresholds on the EVAL split to reduce missed at-risk passengers (FN) while keeping the number of "unnecessary" interventions (FP) operationally manageable.

## Evidence

The engineered Model B improves AUC and log_loss over the manifest-only baseline and yields better recall for survivors at 0.5. When evaluating Model B on EVAL data across thresholds, the ~0.25 operating point provides the lowest expected cost under the 1.7:1 penalty, with a favorable trade-off between precision and recall. However, fairness analysis at this threshold shows substantial sex-based disparities exceeding the 5-percentage-point tolerance and therefore must be flagged.

## Policy

We recommend deploying Model B as a global scoring engine with a single operating threshold of around 0.25 for triage decisions, rather than separate models or thresholds by segment, to keep the policy simple and operationally robust. This threshold minimizes expected cost under the C_FN = 1.7·C_FP assumption and materially improves survivor recall relative to a 0.5 cutoff. However, the fairness analysis shows substantial sex-based parity gaps, so the policy should be treated as provisional: acceptable as a first-pass scoring tool, but not as an unmonitored, fully automated allocator of scarce staff attention. In practice, we would combine the global threshold with additional procedural checks (e.g., manual review lanes or softer prioritization rules for under-served groups) and revisit the possibility of segment-specific calibration if fairness gaps persist.

## Monitoring

Monitor weekly the model's recall and precision at the deployed threshold, expected cost, and fairness parity gaps (target < 5 pp). A designated operations or analytics owner should review these metrics and adjust the threshold or escalate fairness issues as needed.