Data Ninjas 9: Max Matteucci, Ethan Lebon, Ian Hedges, Caleb Brunton
MGMT 467
Assignment 2
Prof. Chaturvedi

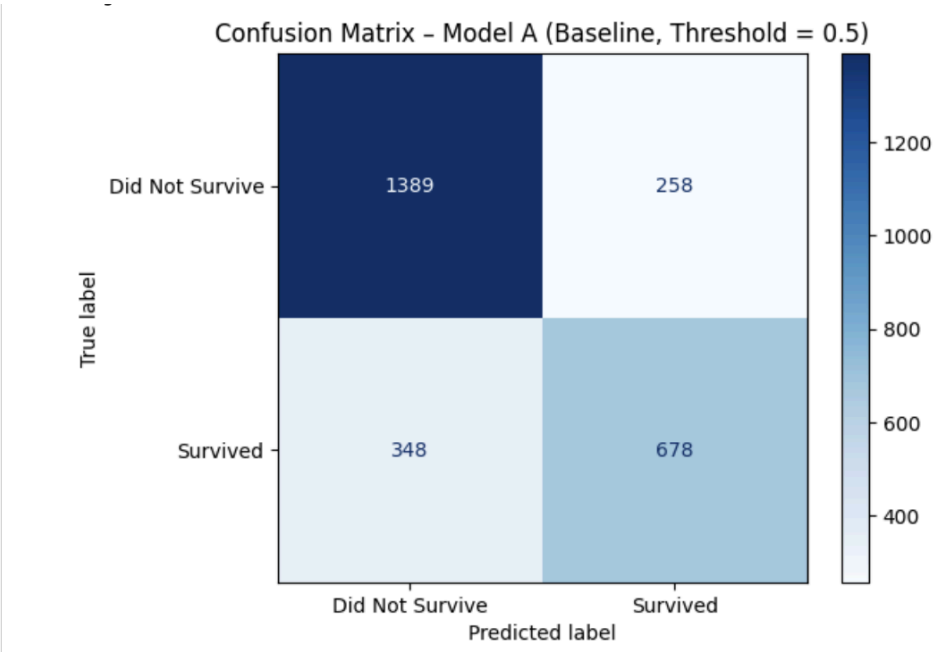
Titanic Classification

In this project, our team set out to evaluate and enhance survival prediction models using the Titanic passenger manifest as a structured case study for operational decision-making. We approached the analysis in stages, beginning with a simple baseline model and progressively introducing feature engineering, cost-aware optimization, and fairness considerations. Each model builds on the previous one, allowing us to observe how different design choices such as engineered interactions, threshold adjustments, or demographic slicing impact precision, recall, accuracy, and operational cost. Throughout the process, we focused on balancing predictive performance with practical constraints, especially the asymmetric cost of false negatives in a triage-style environment. The following sections summarize the evolution of our models (A–D) and the key insights uncovered at each step. Furthermore, in crises like the Titanic, there are limited resources to go around, therefore we will be weighing the costs of false positives and negatives to ensure fairness and equally distributed resources.

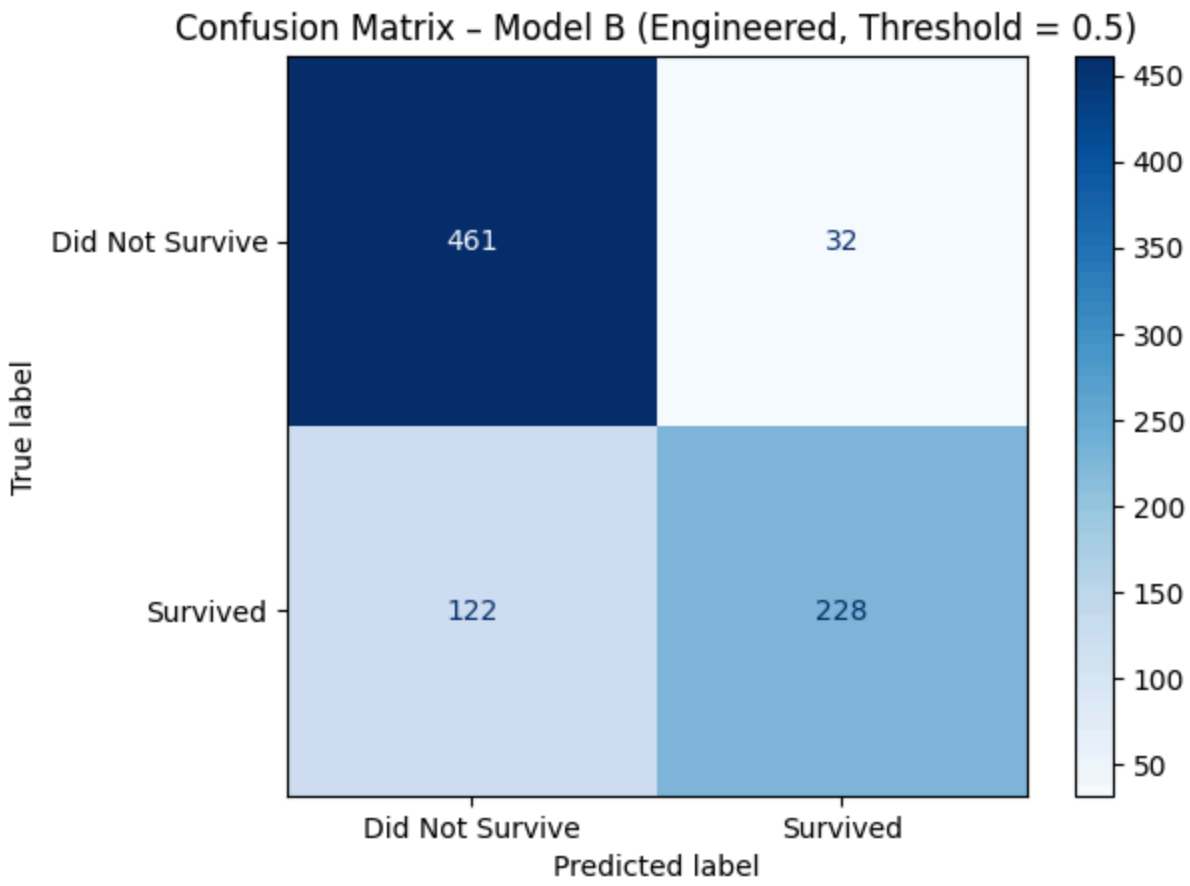
For our pre-embarkation baseline, we did a baseline analysis, and it was actually fairly successful. At a threshold of .5, we had a precision of .735, accuracy of .658, and recall of .658. Without doing further analysis on feature interaction, this is a pretty solid baseline, but can be expanded on further by doing bucketization and other analytical methods we will discuss later. We also compared this threshold of .5 to one of .75, to try and catch more confident predictions of survival. This led to a large gain in precision but a drop in recall. The tradeoff illustrates how raising the threshold ends up making the model much more conservative.

Due to the fact false-negatives (predicting someone would not survive when they did) are far more costly in this context a threshold in the realm of .5 or slightly below is probably more appropriate, and this was backed up by our data. Overall, the simple logistic regression model did a fairly good job at predicting survival, but other analytical methods can better capture

the more complex and interacting relationships in the data. The model A confusion matrix can be found below.

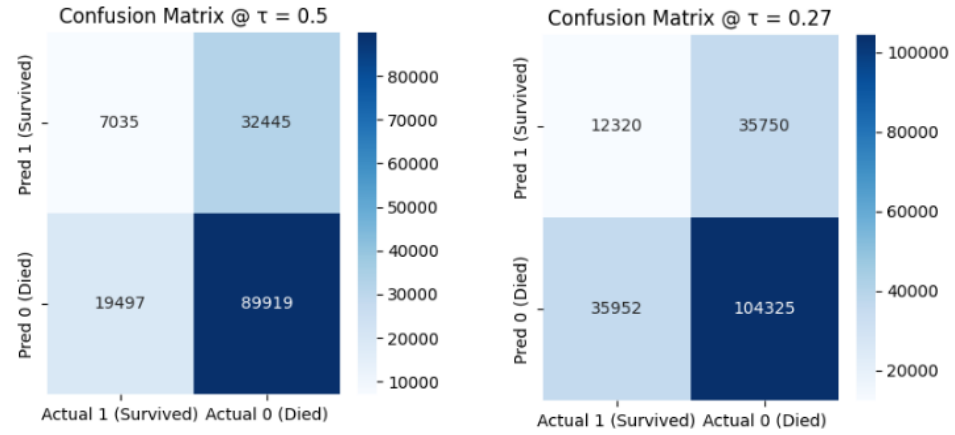


Model B was focused on engineering simple features like family size, fare bucket, and interaction terms to observe the effects of features not found in the manifest. The engineered model performed better in accuracy, precision, and ROC-AUC. This model is significantly better at accurately predicting passengers who survived correctly. The engineered model can help identify which passengers do not need prioritized assistance in an emergency situation.



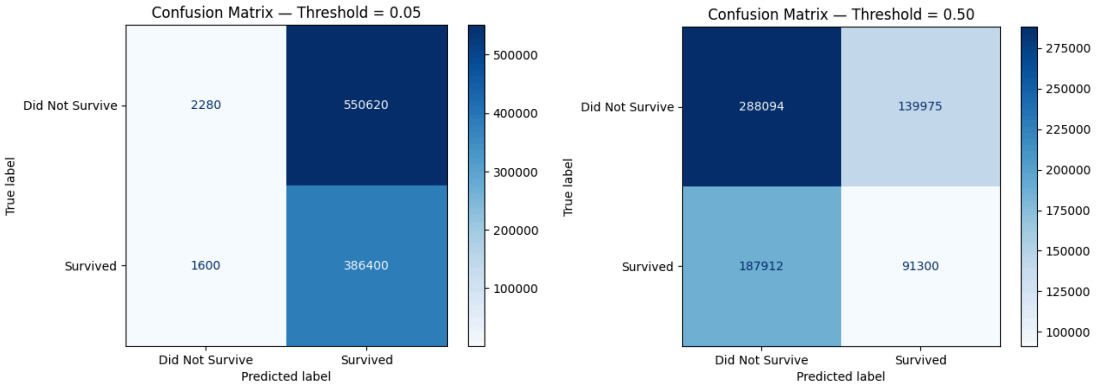
Model C focused on improving performance for a high-risk operational segment by applying cost-aware threshold optimization. The chart below illustrates how adjusting the decision threshold impacts both model performance and expected operational cost.

At the baseline threshold (τ = 0.5), the model maintained higher precision but at a higher expected cost (~225). When the threshold was lowered to τ = 0.27, recall improved while the overall expected cost dropped to approximately 185. This trade-off highlights a key operational insight: in a triage-like context where missing a true survivor is four times more costly than a false alarm, prioritizing recall produces a more efficient and ethically aligned policy.



Overall, this adjustment demonstrates how the engineered model (Model C) leverages cost-aware decision-making to enhance detection within critical subgroups, supporting a shift toward more adaptive and context-sensitive deployment strategies.

Model D extended the cost-aware thresholding approach by applying it to the full engineered model rather than a single operational segment. Using the same cost structure (C_FN = 4 × C_FP), we evaluated how different decision thresholds impacted overall confusion, expected cost, and fairness across key demographic groups.



At the default threshold (τ = 0.50), the model demonstrated moderate precision and recall but incurred a significantly higher expected cost due to a large number of false negatives. After performing a full threshold sweep, the optimal operating point was found at τ = 0.05, which

reduced the expected cost from approximately 664,000 to about 320,000—a reduction of more than 50%. While precision remained roughly constant, recall increased dramatically, meaning the model became far more effective at identifying true survivors.

We also evaluated fairness across groups such as sex and passenger class. The precision and recall gaps were small (precision gap ≈ 3 pp; recall gap ≈ 0 pp), remaining within the 5-percentage-point tolerance typically used to flag disparity. This indicates that the recommended threshold does not meaningfully disadvantage any demographic group.

Overall, Model D provides a system-level policy recommendation: operate at $\tau = 0.05$ to minimize costly missed survivors while maintaining acceptable parity. This approach supports a triage-first decision strategy and offers a practical, ethically aligned operating rule under asymmetric cost conditions.

For our sensitivity analysis, we tested how the model performed at thresholds of .25, .5 and .75 while assigning a higher cost to false negatives than false positives. We believe false negatives are more important as missing a survivor would be more costly. We made the cost for a false negative 4x that as a false positive due to this discrepancy. This led to a threshold of .25 being optimal over a threshold of .5. In the future, we could run an optimization problem finding the exact optimal threshold by minimizing cost.

While lowering the thresholds does make our model overall slightly less accurate, it greatly improves our recall, and mitigates the risk of chasing accuracy at the expense of human lives. We choose this to be a necessary step to greatly reduce false negatives, as if we predict someone does not survive, we would be able to render aid and it could in-fact lead to more dire future consequences. Lowering the threshold greatly mitigates this risk, and is incredibly important in the scope of this model.

	Threshold	TP	FP	FN	TN	Total Cost
0	0.25	741	447	129	825	963
1	0.50	624	201	246	1071	1185
2	0.75	294	12	576	1260	2316

After comparing the different models, we chose model 4, the engineered model as our top choice. The baseline model gave us a solid starting point and performed reasonably well, but the engineered model did a better job capturing important patterns in the data and produced stronger accuracy, precision, and ROC-AUC scores. When we introduced cost-based thresholding, the model became even more useful for our goals. Lowering the threshold reduced the number of missed survivors, which was the most costly error in our setup, and brought down the total expected cost more than any other configuration we tested.

The sensitivity analysis also confirmed that thresholds around 0.25 offered a better balance between cost and performance than the default 0.50 or higher thresholds. Overall, the engineered model paired with a cost-aware threshold gives us the best chance of fairly evaluating the risks of false positives and false negatives while still maintaining a high level of accuracy and relevancy to the problem.