Operations Brief

Decision Rule

Prioritize minimizing False Positives (FP) to reduce operational inefficiencies, employing a diversion prediction threshold of 0.75.

Evidence

- Baseline Model Performance: The confusion matrix with the default cutoff of 0.5 shows 0 true positives and false positives but very high false negatives and true negatives. The number of false negatives is 3134 and the number of true negatives is 1140896. The custom 0.75 cut off also shows 0 true positives and false positives but very high false negatives and true negatives. The number of false negatives is 3020 and the number of true negatives is 1142862.

	TP	FP	FN	TN
0	0	0	3134	1140896

	TP	FP	FN	TN
0	0	0	3020	1142862

Baseline Model Evaluation:

	precision	recall	accuracy	f1_score	log_loss	roc_auc
0	0.0	0.0	0.997338	0.0	0.017975	0.713678

- Engineered Model Performance: This model performs better than the baseline model and has a perfect precision value. It also has a higher AUC value, making it more reliable.

Engineered Model Evaluation:

	precision	recall	accuracy	f1_score	log_loss	roc_auc
0	1.0	0.002006	0.997394	0.004004	0.017207	0.7663

- Interpretation: While the engineered model offers better discriminative power (higher AUC), both models exhibit very low recall due to the data imbalance, making direct classification of diversions challenging at current thresholds. However, the engineered model provides a more reliable signal when it does predict a diversion, justifying the chosen precision-focused threshold.

Model B Evidence

- Engineered Model Performance**:** The engineered day-of-ops model shows perfect precision (1.0), meaning all predicted diversions are correct, though recall remains extremely low (0.0019) due to the class imbalance. Accuracy is high but not informative given the dominance of non-diversion

events. The ROC AUC improves to 0.750, indicating better discriminative power than Model A, and log loss is slightly lower, showing more reliable probability estimates.

	precision	recall	accuracy	f1_score	log_loss	roc_auc
0	1.0	0.001917	0.997377	0.003827	0.01729	0.750426

- Interpretation: Compared to Model A's pre-departure baseline, Model B provides a clear uplift in reliability and discriminative performance, primarily driven by the inclusion of near-departure features such as departure delay and delay buckets. While both models suffer very low recall due to extreme data imbalance, Model B gives more trustworthy signals when it predicts a diversion, justifying the precision-focused approach. The improvement in AUC confirms that real-time operational signals add measurable value over baseline schedule-level predictions.

Model C Evidence

- Engineered Model Performance: The model does have a very high precision (1.0) and accuracy (0.998) meaning that the model is always correct when predicting a diversion, and the accuracy can be misleading due to the imbalanced nature of the dataset. But with a very low recall score (0.0063) and a low f1-score (0.012), it means that the model is very bad at balancing precision and recall, and it actually identifies only a small amount of actual diverted flights.

	model_version	precision	recall	accuracy	f1_score	log_loss	roc_auc
0	localized	1.0	0.00627	0.997778	0.012461	0.015474	0.663949

	expected_label	FALSE	TRUE
0	False	142346	0
1	True	317	2

- Interpretation: Compared to model B, Model C does show a higher f1 score, meaning that it is slightly better at balancing precision and recall. With a lower log_loss score, it indicates that the predicted probabilities are closer to the true labels. However, model C has a lower ROC compared to model B which shows a poor performance in distinguishing positive and negative classes. With an insignificant increase in the model performance, the specialization in origins of 'ATL', 'ORD', and 'JFK' for the localized segment, may not be the features that provide important predictive information.

Model D Evidence

The Model D framework extends the engineered Big Query ML model by introducing a cost-based threshold optimization policy to determine the most efficient diversion-alert setting. Using the same engineered features—route, day_of_week, and dep_delay_bucket—the model applies a cost matrix where False Positives (C_FP = $1,000) represent wasted operational effort and False Negatives (C_FN = $6,000) represent the cost of a missed diversion. Iterating thresholds from 0 to 1 identified the optimal threshold = 0.55, minimizing expected cost.

At this point, the model achieved:
True Positives (TP): 24
False Positives (FP): 49
False Negatives (FN): 2,969
True Negatives (TN): 1,143,762
Expected Cost: $17.87 million (≈ $8,000 savings vs default 0.5 threshold)

The ROC AUC of 0.759 confirms improved discriminative power and more reliable probability estimates compared with earlier models.

Interpretation: Model D represents a shift from pure statistical tuning to operational cost minimization. By explicitly quantifying FP vs FN impacts, it delivers a policy-aligned decision rule that balances safety and efficiency. Compared with Model C's fixed precision-heavy threshold, Model D dynamically chooses the cutoff that minimizes total financial loss while maintaining high precision (~1.0). This makes it the most deployable and economically grounded version—transforming small performance gains into measurable cost savings and aligning model behavior with real airline risk-management priorities.

Policy

- Global Deployment and Cost Analysis: The current model is trained for global deployment according to Model D evidence. The expected cost analysis ($17.87 million) would amount to approximately $8000 in savings when compared to using the baseline model with a 0.5 threshold.
- Hub Precision: Although fairness across different groups (e.g. airlines, routes) or specific hub precision was not explicitly evaluated in this iteration, it would be a crucial consideration for future enhancements or segmented deployments.
    - While the Model B iteration improves overall predictive power, fairness across airlines, hubs, or specific routes was not explicitly measured. Future deployment could include segmented evaluation to ensure equitable performance and avoid bias toward certain carriers or routes.
    - While Model C is localized for the specialization in origins 'ATL', 'ORD', and 'JFK', it did not capture the full picture of the diversion risk. From the confusion matrix we can see that it only captured 2 True Positive. The limitation on the model can result from a limited number of origins and focus on the subset of risk. Since it only gives the result of 3 mixed hub results, it may eliminate the possibility of diversion risk happening in other airports. Diversion risk may be high in other hubs too, and it is important to study from the previous data of diverted flights, causes of the diversion to happen, and seasonality of diversion risk.

Monitoring

- Key Metrics: The primary metrics to monitor would be ROC AUC, Precision, and Recall for diverted flights.

- Thresholds: The thresholds for acceptable performance would be when the AUC remains above 0.75 and precision above 0.95 when the model predicts a diversion. Additionally, we might also consider monitoring data drift in input features such as departure delay, distance, and carrier.
- Cadence: The key metrics and thresholds will be reviewed weekly for model performance, and monthly for data drift.
- Owner: We would assign responsibility for monitoring and model recalibration to flight-specific authorities, such as the Data Science Team, and Flight Operations Analytics, after the final model is ready for deployment.