

Predicting Flight Cancellations

Katia Lopes-Gilbert

Springboard Capstone Project



Project Overview

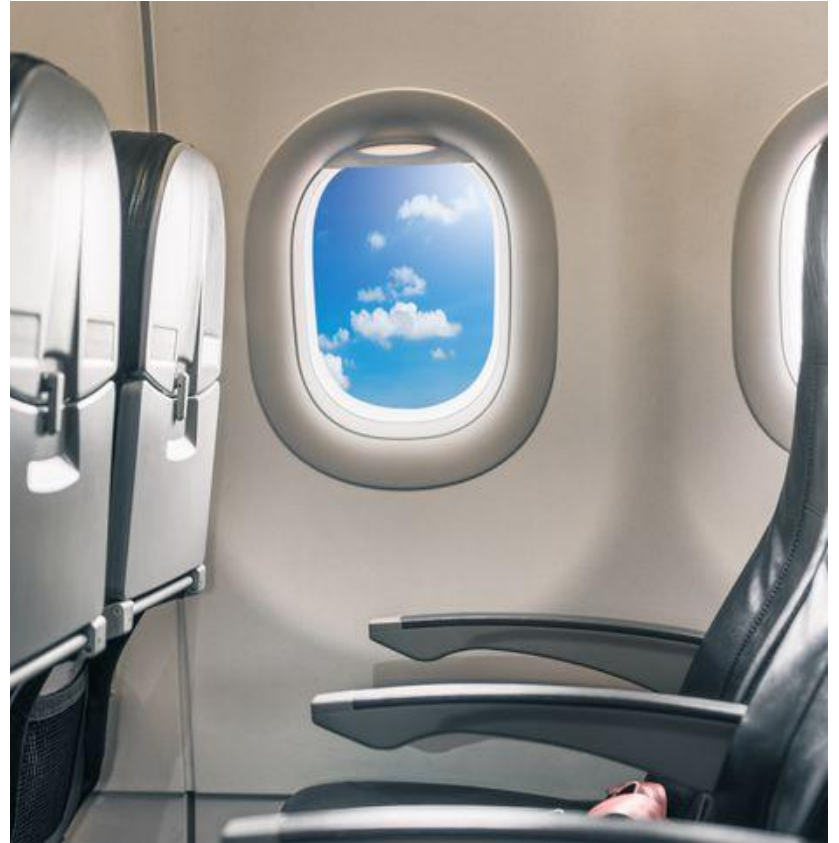
Flight **cancellations** have **significant economic and logistical impacts**, costing airlines and passengers **billions** of dollars **annually**.

How might we improve airline operations and passenger experiences by correctly predicting flight cancellations?



Data Sources

- 2023 Airline Performance Data
- IATA Codes
- Airport Details
- Hourly Weather Data
- Aircraft Registration, Engine, and Aircraft Details



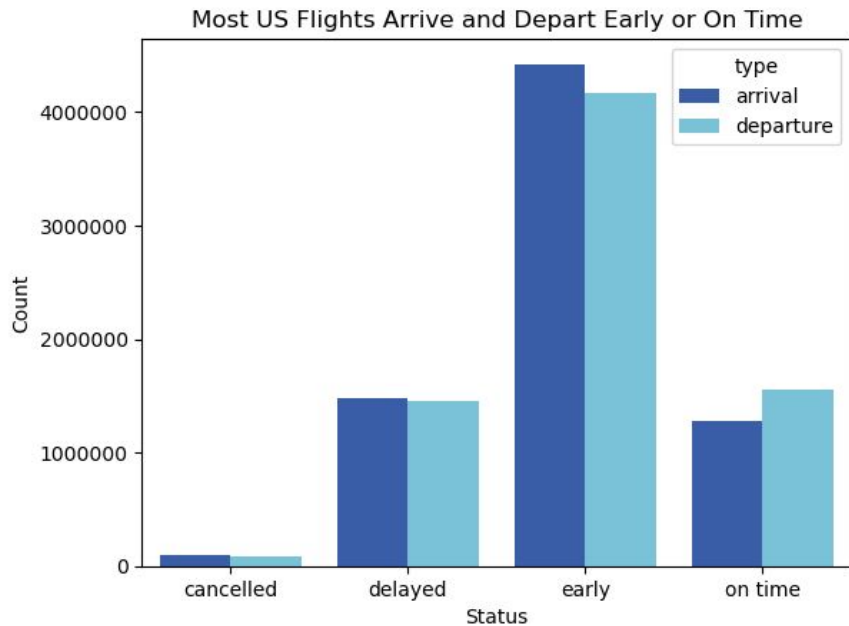
Data Cleaning & Wrangling

1. Investigated and dealt with null values
2. Data type transformations
3. New datetime features
4. Merging airports
5. Selecting top airports
6. Merging weather data



Data Summary

- **7.2M** flights in 2023
- 93.8K cancellations (**1.29%**)
- 78% depart & arrive early or on time
- Most flights arrive between 15 min early to 9 min late
- Departure delays over 31.5 min and arrival delays over 45 min are considered outliers



Exploratory Data Analysis

Investigated cancellations rates and performance for:

1. Historical Patterns
2. Airlines
3. Geographical Factors
4. Temporal Analysis
5. Temporal & Geographical Factors
6. Weather Patterns

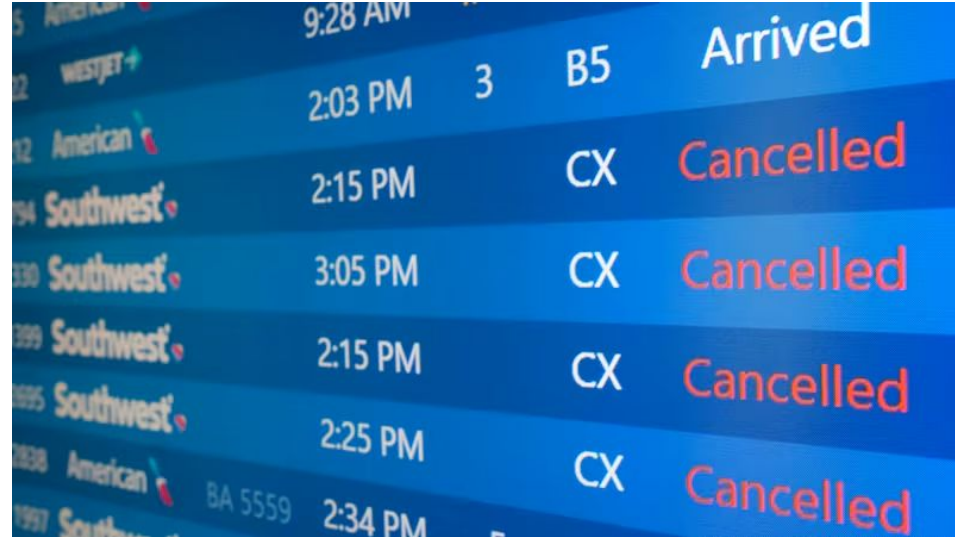


Feature Engineering: Historical Performance

Unique Flights: Flights grouped by airline, route and departure window

Unique fights that have **cancellations** within a **10 day window** are **considered outliers**.

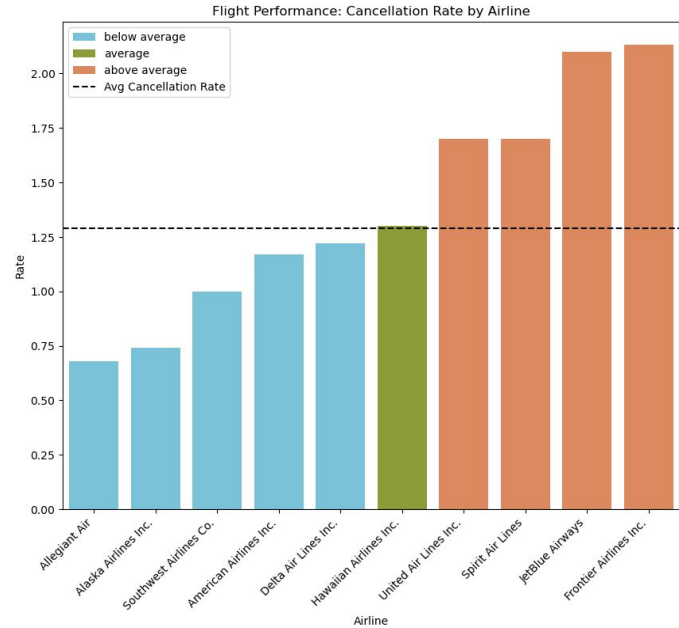
Unique flights that did experience at least one **cancellation** in the previous 10 days **experienced much higher cancellation rates** compared to unique flights that did not experience cancellations over a 10 day period. Some were between **2-3x more likely**.



WESTJET	9:28 AM			Arrived
American	2:03 PM	3	B5	
Southwest	2:15 PM		CX	Cancelled
Southwest	3:05 PM		CX	Cancelled
Southwest	2:15 PM		CX	Cancelled
Southwest	2:25 PM		CX	Cancelled
American	2:34 PM			
Southwest				

Airline Performance

- Cancellation rates varied across airline partners.
- Most airlines had cancellation rates that differed in a statistically significant way compared to the average cancellation rate.



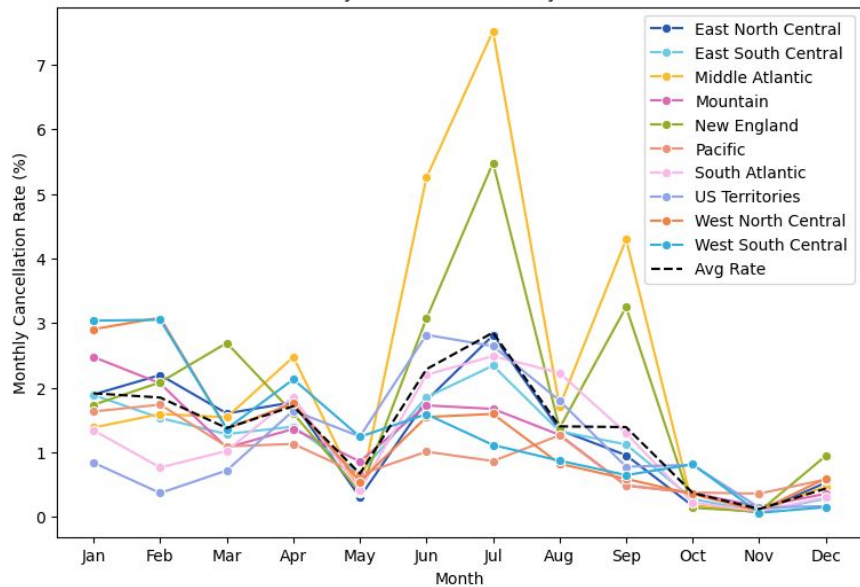
Temporal Analysis

- Month was a significant predictor of cancellations
- Busiest travel months not correlated with peak cancellations
- Weather is leading cause of cancellations most months of the year
- Carrier cancellations increase over time and peak when weather cancellations are at their lowest

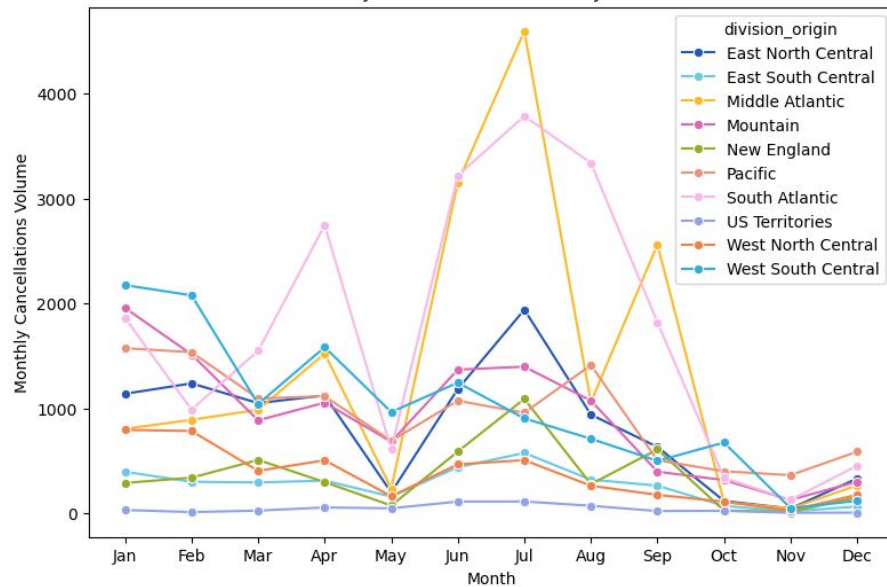


Geographical & Temporal Effects on Cancellations

Monthly Cancellation Rate by Division



Monthly Cancellation Volume by Division

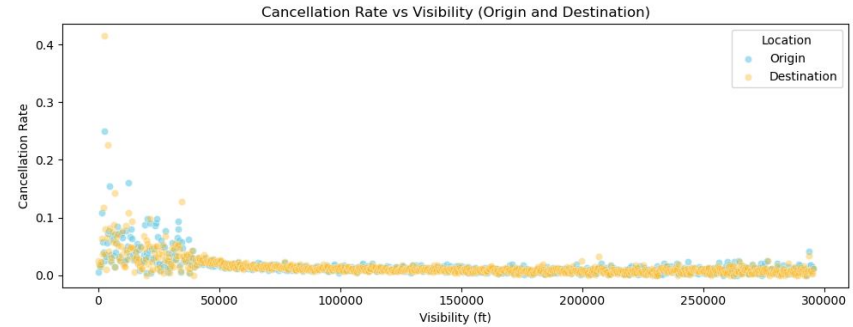
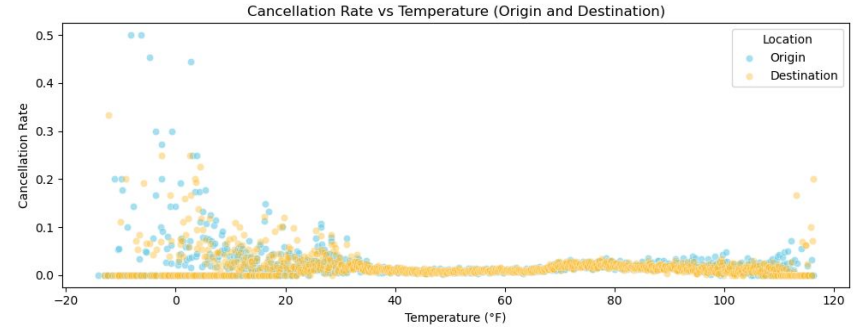


Origin and Destination Weather Patterns

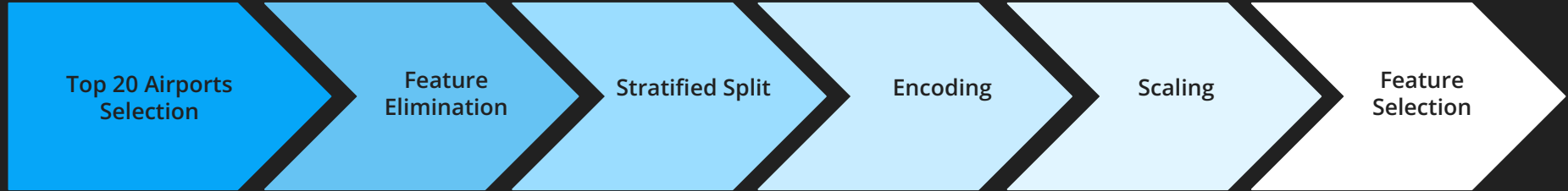
Flight cancellations tend to increase at temperature extremes, particularly in very cold weather (< 15 degrees F)

Flight cancellations are sensitive to very low visibility conditions (< 50,000 ft.)

Both origin and destination cancellation rates rise as weather impact severity increases.

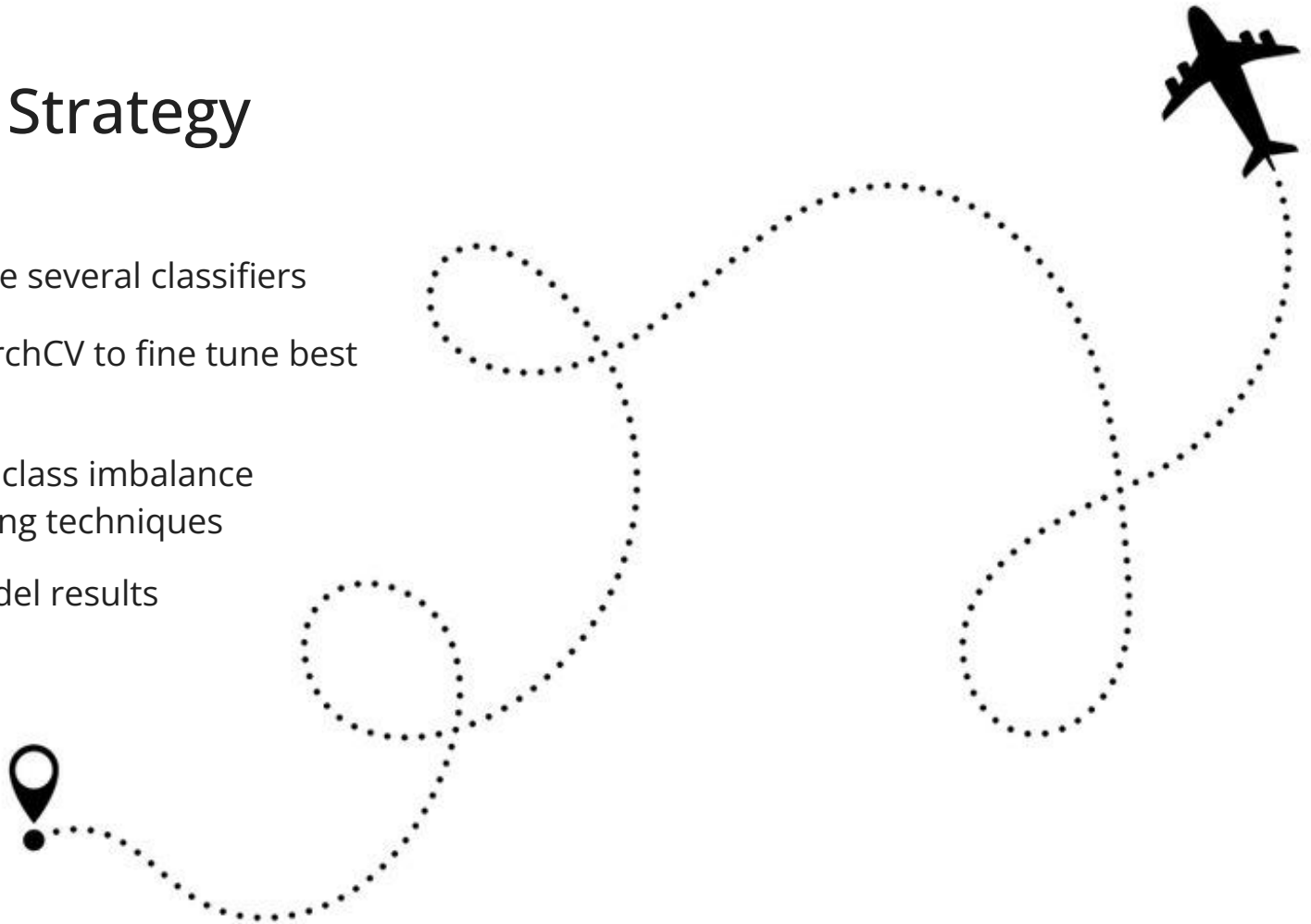


Preprocessing

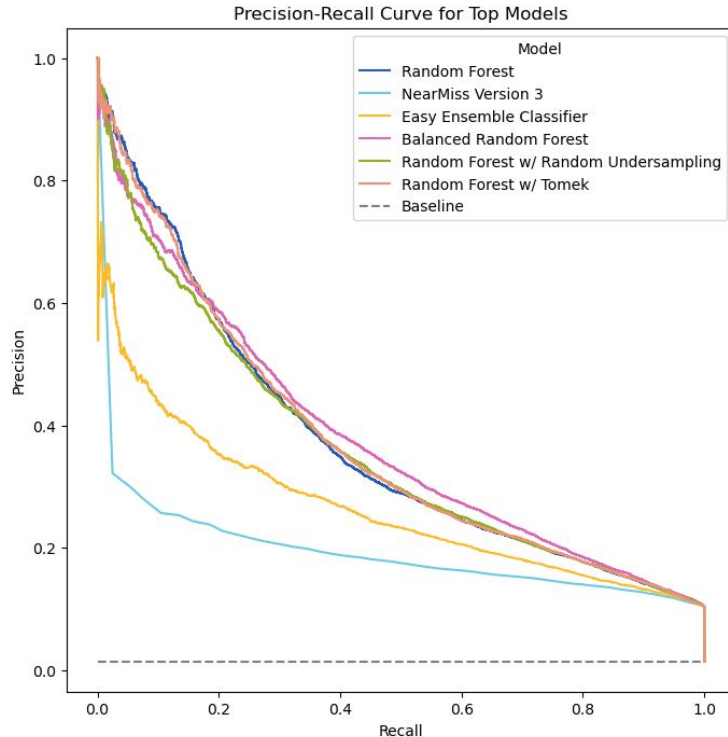


Modeling Strategy

- Cross validate several classifiers
- Use GridSearchCV to fine tune best classifiers
- Try different class imbalance undersampling techniques
- Evaluate model results



Model Results and Evaluation



- Prioritized models based on Recall Score and Area Under Precision-Recall Curve (AUPRC)
- Tried different feature handling (ex: month as categorical value instead of numerical)
- Investigated data distributions between True Positives and False Positives

Feature Importance

1. Cancelled Sum Over 10 Day Window
2. Number of Flights Over 10 Day Window
3. Month
4. Quarter
5. Visibility Destination
6. Visibility Origin
7. Temperature Origin
8. Temperature Destination
9. Day of Month
10. Departure Delay Max Over 10 Day Window

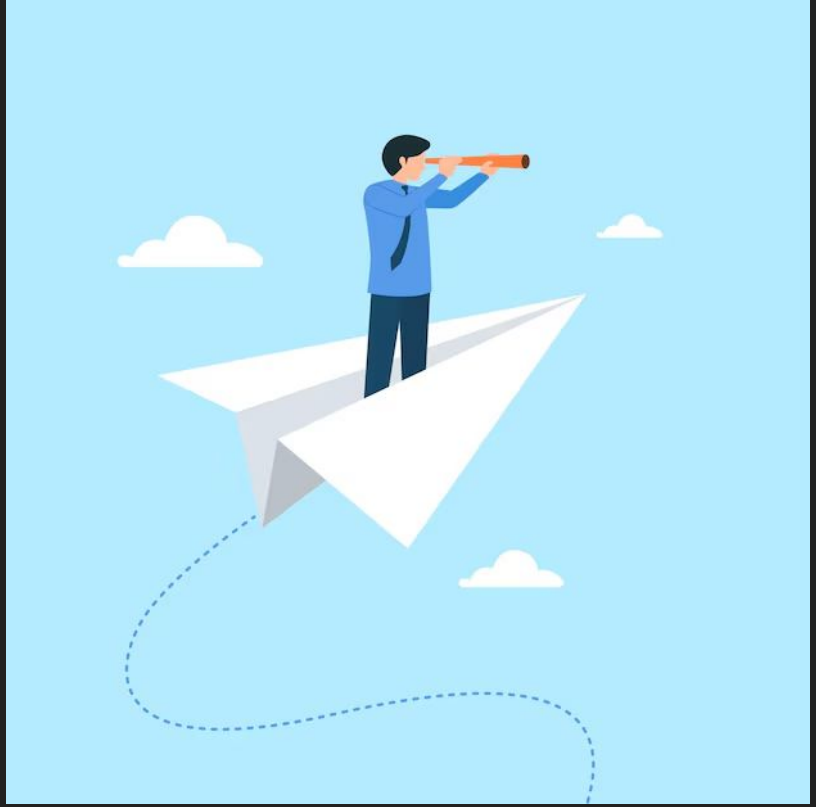
Challenges and Limitations

- Limited number of airports used for modeling due to volume of weather data that would have been required
- Complex relationships between weather patterns and geography not necessarily captured
- Low precision could be costly, but we need cost-benefit analysis



Future Improvements

1. Create different feature interactions
2. Incorporate aircraft and engine data
3. Predict delays in addition to cancellations
4. Expand to all airports



Acknowledgements

Thank you to my mentor, Jyant, for your help and guidance!



Questions ?