# Optimizing Airline Operations: Predicting Delays for Smarter Routes and Efficient Resource Management

Dipti Aswath

## Executive Summary

### Problem Statement:

Airlines and airports face significant operational challenges due to flight delays, which can be caused by a variety of factors including flight status, weather conditions, air traffic congestion, aircraft specifics, and inefficiencies in ground and passenger handling. The objective is to predict flight delays by developing a multi-class classification model that considers both departure and arrival delays, helping improve operational planning and customer satisfaction.

### Rationale:

Flight delays can have widespread consequences for airlines, from passenger dissatisfaction to operational disruptions. Developing a predictive model for flight delays not only addresses the core issue of minimizing delays but also enhances decision-making processes across various facets of airline operations.

#### Business Use Case: Airline Route Optimization:

* **Proactive Rerouting**: Predicting delays enables airlines to reroute flights, helping avoid congested airspace or airports experiencing delays, thereby reducing the overall impact on flight schedules and minimizing delay time.
* **Weather Avoidance**: Forecasting weather-related delays allows for dynamic adjustments to flight routes, ensuring safer and more efficient travel by avoiding adverse weather conditions.

#### Business Use Case: Airline Operations Management:

* **Scheduling Adjustments**: Delay predictions empower airlines to adjust schedules in real time, reallocating resources like gates, ground crew, and equipment to manage disruptions effectively.
* **Resource Allocation**: Advanced knowledge of delays facilitates more efficient allocation of resources, such as repositioning aircraft and crew, reducing the cascading impact of delays on subsequent flights.
* **Passenger Communication**: With accurate delay predictions, airlines can inform passengers in advance, manage expectations, and provide better customer service, such as offering rebooking options or compensation for delays.

#### Example Usage:

#### An AI system that predicts flight delays could:

1. Suggest alternate flight paths that are less likely to experience delays.
2. Provide passengers with timely updates and rebooking options.
3. Dynamically adjust flight schedules to manage disruptions effectively.
4. Allocate resources efficiently to minimize the impact on subsequent flights.

### Research Question:

How can a multi-class classification model be developed to accurately predict flight delays by assessing multiple factors, including departure and arrival delays, using data related to flight status, weather conditions, air traffic, aircraft specifics, and ground operations?

### Key Findings from Exploratory Data Analysis:

1. Highest 2019 departure and arrival delays by Carriers

A graph showing the average departure of a flight

Description automatically generated

A graph showing the average arrival of passengers

Description automatically generated

1. Top 20 congested airports in 2019 with flight delays

A graph of the number of concordments

Description automatically generated

A map of the united states with different colored spots

Description automatically generated

1. SMOTE Resampling on Training Data

A white background with black text

Description automatically generated

### By examining segment number trends across Distance Group descriptors, we gain valuable insights into how aircraft operational schedules and the number of daily flights contributed to 2019 delays.

**Segment Number Decreases with Distance**: As flight distance increases, the number of segments (flights) decreases. Aircraft flying longer routes complete fewer flights in a day due to time constraints.

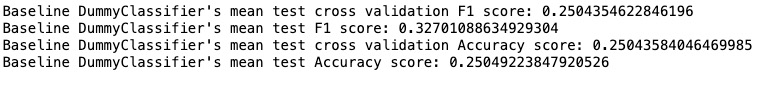
**Delays Correlate with Higher Segment Numbers**: Flights scheduled for more segments in a day are more prone to delays, regardless of distance. These delays are likely due to operational factors, such as shorter turnaround times, leading to delayed departures and arrivals.

A graph of different colored lines

Description automatically generated with medium confidence

### Model Evaluation Summary and Performance Metrics:

**Baseline Model with Dummy Classifier:**



**Multinomial Logistic Regression Classifier:**

A screenshot of a computer

Description automatically generated

A close-up of a graph

Description automatically generated

**Decision Tree Classifier:**

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Random Forest Classifier:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Data Sources:

Kaggle Dataset from [here](https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/data), that is comprised of multiple csv's listed below.

1. Air Carrier Summary
2. Aircraft Inventory
3. Air Carrier employee support (Ground Crew, Flight Attendants)
4. Flight On Time Reporting Status with Air Carrier info for 2019-2020
5. Airport Weather
6. Airport and Carrier look-up codes

## Methodology Used:

1. Data Preparation

Involved cleaning and merging multiple CSV files to create a unified data-set with ~4M entries (for training) and ~2M entries (for testing). Feature engineering resulted in approximately 35 features from the original raw data.

1. Feature Engineering

* Delay Categories: Classified delays into four classes/categories. a) **On-time Departure and Arrival:** Flights that depart and arrive within their scheduled times. b) **Delayed Departure, On-time Arrival:** Flights that experience delays during departure but still arrive on time or within a minimal delay window. c) **On-time Departure, Delayed Arrival**. d) **Delayed Departure and Arrival:** Flights that experience delays both in departure and arrival times.

A graph showing different types of classes

Description automatically generated with medium confidence

* Aggregation Features: Created historical delay averages such as,

CARRIER\_HISTORICAL = captures the historical average delay rate of each carrier per month

DEP\_AIRPORT\_HIST = captures historical average delay rates for flights departing from specific airports per month

PREV\_AIRPORT\_HIST = captures historical average delay rate for the airport from which the aircraft arrived before the current departure

DAY\_HISTORICAL = captures historical average delays associated with each day of the week, adjusted monthly

DEP\_BLOCK\_HIST = captures historical average delay rate for different departure time blocks, aggregated by month

* Time Based Features: Extracted season from month and part of the day from departure and arrival time blocks

A comparison of different colored bars

Description automatically generated

A group of bars with numbers

Description automatically generated with medium confidence

* Distance Based Features: Mapped distance groups to descriptive text

A close-up of a graph

Description automatically generated

* Delay Based Features: Created new features by combining actual departure and arrival times with planned times to create new delay features, ELAPSED\_TIME\_DIFF, DEP\_DELAY, ARR\_DELAY
* Employee Features: Created employee statistics features FLT\_ATTENDANTS\_PER\_PASS and PASSENGER\_HANDLING for analyzing airline and carrier operations
* Removed highly correlated features with VIF

A close-up of a document

Description automatically generatedA close-up of a number

Description automatically generated

* Engineered features with their descriptions can be found here

1. Data Pre-Processing: Missing values and outliers detected were removed. Categorical features were target encoded and Numerical features were scaled.
2. Modeling and Evaluation: Classification algorithms used were Decision Trees, Random Forest, and multi-nomial Logistic Regression, with evaluation metrics: F1 Score, PR AUC, ROC AUC and Accuracy scores. Sequential Feature Selection was used to select 5 features from among the 34 predictor variables.

## Outline of Project:

1. Engineered Features and description here
2. Merged Data for train and test is here
3. Raw Data used for creating merged data is here
4. Exploratory Data Analysis plots from Auto Viz Visualization Library is here (credits: )
5. Images captured in this README are located here
6. Jupyter notebooks, a) Notebook for data preparation, merges and initial exploratory analysis is here, b) Notebook for data pre-processing, additional exploratory analysis, modeling and evaluation is here

## Next Steps:

1. Use Dimensionality Reduction and Clustering to reduce dimensions, and cluster features together to reduce the count of 34 predictors. Relying on Sequential Feature Selection alone, takes a while to train any of the classification models.
2. Use the SVC classifier to compare its performance with the others
3. Use StreamLit and Fast API to serve the flight prediction delays via an application interface