Flight Delay Prediction and Rescheduling

[Github]

Devansh Sharma

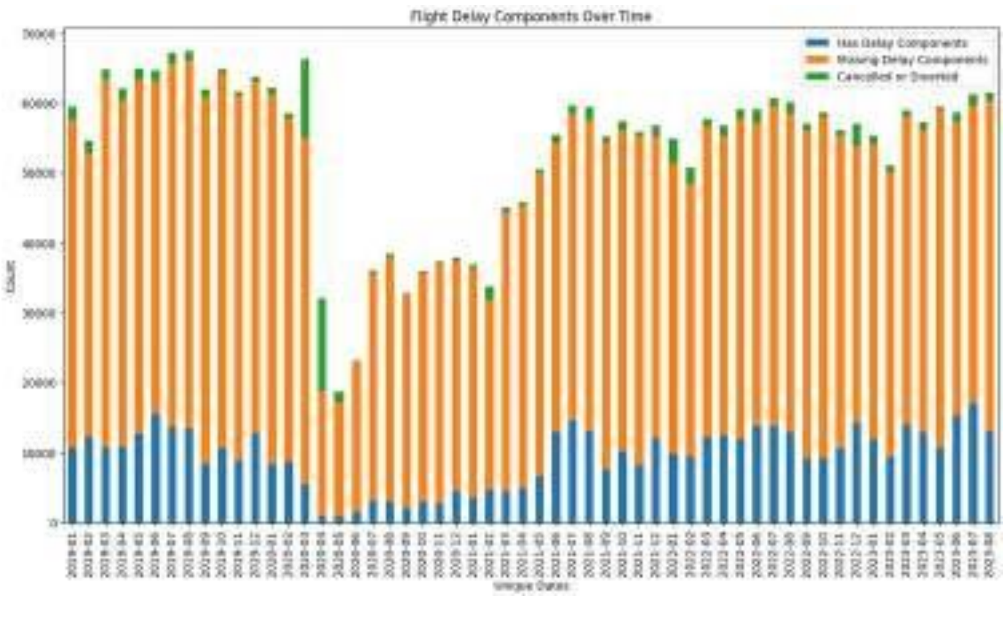
Table of Contents

1 PROBLEM STATEMENT

Airline operations are multifaceted, involving a delicate balance of flight scheduling, crew management, fuel optimization, and adherence to strict regulations. These operations face constant challenges from unpredictable variables like weather, maintenance issues, and crew availability, which can result in delays that disrupt schedules and increase operational costs. To address these challenges, it is essential to build efficient, data-driven solutions that enable airlines to predict and mitigate delays effectively while ensuring smooth and timely service. The FAA estimated that flight delays cost the aviation industry $33 billion annually in 2019, highlighting the substantial economic impact. These delays also contribute to environmental concerns through increased fuel emissions.

In this challenge, we aim to create a predictive solution that leverages historical and operational data to enhance decision-making in airline operations. The primary objectives are twofold: firstly, to develop a model that accurately forecasts flight delays based on historical data, helping airlines proactively allocate resources; and secondly, to design a system that can reschedule flights dynamically to minimize overall delays across routes, optimizing the utilization of fleet and crew.

This solution will be designed to generalize across various airports, routes, and aircraft types, reflecting the complex, real-world environment of airline operations. By implementing this data-driven model, airlines can improve operational efficiency, reduce unnecessary costs, and enhance overall customer satisfaction.



2 ANALYSIS OVERVIEW

1. **Dataset Acquisition and Purpose**
    - The dataset spans from **August 2019** to **August 2023** and was sourced from the **U.S. Department of Transportation's Bureau of Transportation Statistics.** Collected via **API** into an AWS EC2 instance, it covers **flight schedules**, **delays**, and **cancellation reasons**, providing the foundation for delay prediction and rescheduling analysis.
2. **Data Preparation and Transformation**
    - Since the initial dataset consisted of up to **3 million rows**, a **reduced sample of 0.5 million rows** was created, optimizing the workload without compromising accuracy.
3. **Feature Selection and Encoding**
    - Key features included **origin** and **destination**, **scheduled** and **actual times**, and **delay reasons. Redundant variables** were **pruned** using **Pearson correlation** and the **Kruskal-Wallis H-test** to identify the most influential factors in delay prediction. **Categorical variables** (e.g., airline codes, delay types) were **encoded**, and continuous **time-related features** were converted into a standardized format, such as **minutes past midnight**, to allow precise **temporal analysis.**
4. **Data Cleaning and Outlier Handling**
    - The dataset underwent rigorous cleaning to handle missing values, focusing on **removing records** where **delay reasons were missing. Outliers** were **managed using interquartile range (IQR) thresholds.**
5. **Predictive Modeling for Delay Prediction**
    - Benchmark regression models (e.g., **XGBoost**) and advanced time-series models (**LSTM**, **LSTM with CNN** architecture) were used to predict delays. LSTM models effectively captured temporal patterns, providing higher accuracy in predicting potential delays.
6. **Flight Rescheduling Optimization**
    - **Genetic algorithm** was developed to **reschedule flights** under constraints (e.g., crew schedules). This model minimized delays by prioritizing flights at risk of delay, demonstrating practical benefits for real-world applications.
7. **Evaluation and Results**
    - The **LSTM** models **outperformed baseline models**, achieving accurate delay predictions. The rescheduling approach effectively reduced cumulative delays, indicating its value in operational optimization.
8. **Challenges and Overview**
    - Challenges included **data sparsity from 2020** (COVID-19 impact). Future work may involve using ensemble techniques and **real-time data** (e.g., weather) to improve prediction accuracy.

3 DATA DESCRIPTION

- **Dataset Source and Scope**
  - The dataset is sourced from the U.S. Department of Transportation's Bureau of Transportation Statistics. It spans flights from August 2019 to August 2023, covering U.S. domestic airline operations. This timeframe provides a rich basis for analyzing delays, capturing seasonal patterns and operational disruptions over multiple years.
- **Purpose of Data Collection**
  - The data was gathered to support analysis of flight delays and cancellations, with the aim of building predictive models that help improve flight scheduling and reduce delays. By understanding the factors that contribute to delays, the dataset supports informed decision-making for operational efficiency.
- **Key Variables**
  - The dataset includes the following primary variables:
    - **Flight Details:** Fields such as origin and destination airports, and scheduled vs. actual departure/arrival times provide essential data points for evaluating punctuality.
    - **Delay Information:** Detailed delay information is available, with categorical fields indicating reasons (e.g., weather, crew, and maintenance) for each delay.
    - **Categorical Features:** Includes identifiers such as airline codes and airport codes, allowing analysis across different carriers and routes.
    - **Additional Attributes:** Unique identifiers, such as flight numbers and timestamps, are used to track individual flights over time.
- **Size and Structure of the Dataset**
  - Initially, the dataset contained approximately 29 million rows; for this analysis, it was reduced to a subset of 3 million rows, optimizing it for efficient processing and modeling. Data is structured chronologically by month, then aggregated by year for consistent temporal analysis.
- **Data Quality and Limitations**
  - The data underwent quality checks to address missing values and outliers. Fields with missing delay reasons were removed to maintain data integrity, while outliers were handled using interquartile range (IQR) thresholds to ensure accurate modeling. However, some limitations persist, such as sparse data from 2020 due to the impact of COVID-19 on flight schedules.
  - The dataset also has weather situations missing from both the origin and the destination, which could be very important given that weather situations often cause delay in flight situations.
- **Relevance of Data for Project Objectives**
  - The dataset's rich temporal and categorical features align well with our objectives to predict flight delays and reschedule flights. Time-based attributes support delay prediction by identifying trends, while categorical variables (such as airport and airline codes) enable model training across diverse routes and operational conditions, essential for real-world application.

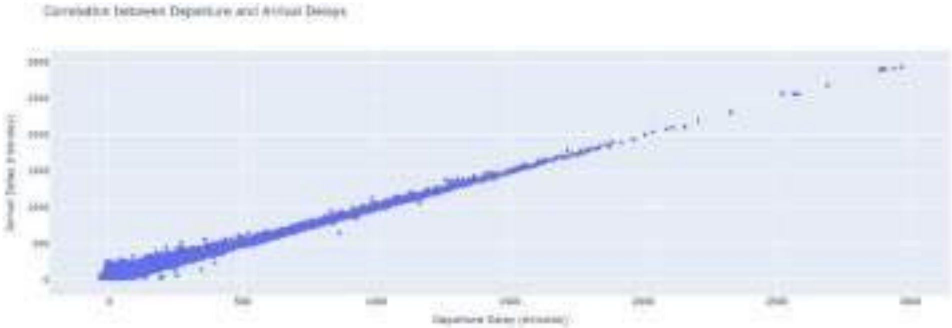
3 Data Description

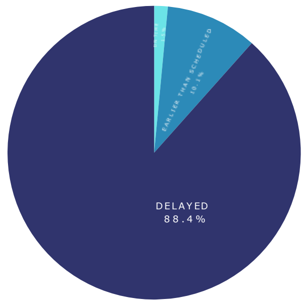
Name	Description
FL_DATE	Flight Date (yyyymmdd)
AIRLINE_CODE	Unique Carrier Code
DOT_CODE	An identification number assigned by US DOT to identify a unique airline (carrier)
FL_NUMBER	Flight Number
ORIGIN	Origin Airport
ORIGIN_CITY	Origin Airport, City Name
DEST	Destination Airport
DEST_CITY	Destination Airport, City Name
CRS_DEP_TIME	CRS Departure Time (local time: hhmm)
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY	Difference in minutes between scheduled and actual departure time
TAXI_OUT	Taxi Out Time, in Minutes
WHEELS_OFF	Wheels Off Time (local time: hhmm)
WHEELS_ON	Wheels On Time (local time: hhmm)
TAXI_IN	Taxi In Time, in Minutes
CRS_ARR_TIME	CRS Arrival Time (local time: hhmm)
ARR_TIME	Actual Arrival Time (local time: hhmm)
ARR_DELAY	Difference in minutes between scheduled and actual arrival time
CANCELLED	Canceled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies the Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
ELAPSED_TIME	Elapsed Time of Flight, in Minutes
AIR_TIME	Flight Time, in Minutes
DISTANCE	Distance between airport (miles)
DELAY_DUE_CARRIER	Carrier Delay, in Minutes
DELAY_DUE_WEATHER	Weather Delay, in Minutes
DEAY_DUE_NAS	National Air System Delay, in Minutes
DELAY_DUE_SECURITY	Security Delay, in Minutes
DELAY_DUE_LATE_AIRCRAFT	Late Aircraft Delay, in Minutes

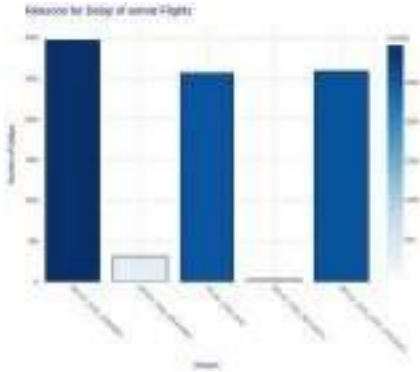
4 Data Preprocessing and EDA

- **Feature Selection**
  - **Pearson's Correlation** between 6 non-categorical independent attributes: CRS_DEP_TIME, TAXI_OUT, CRS_ARR_TIME, TAXI_IN, CRS_ELASPED_TIME DISTANCE and dependent variable: ARR_DELAY.
  - **PCA** on **normalized data did not help increase correlation.**
  - **Eliminated redundancies** in **categorical attributes which was verified using The Kruskal Wallis H-test.**

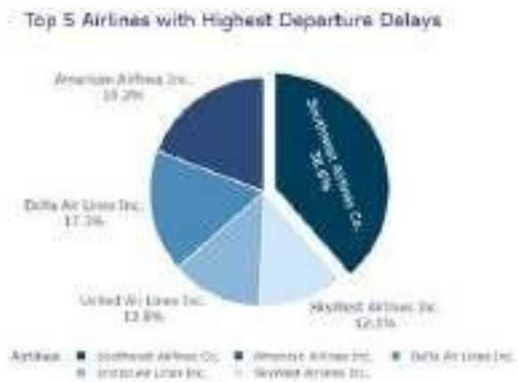
- **Data Summary Statistics and Insights**
  - **Most of the flights are delayed as is shown by the graph -** Over 85 percent of the flights already depart late which suggest need for optimization in their scheduling, and also the arrival delays show a very high correlation with departure delay, with their correlation being over 97 percent.



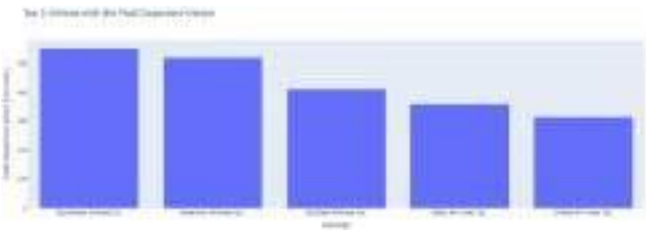




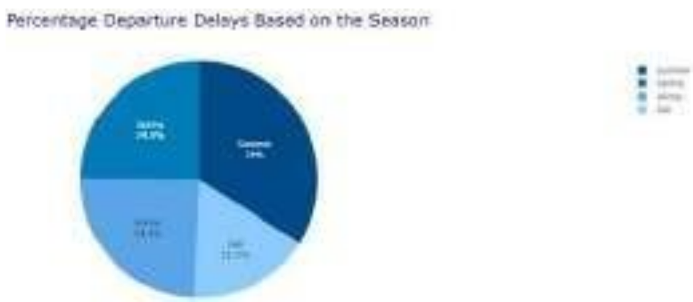
- **Most** of the **delay** is caused due to **Carrier,** followed by **late aircrafts. Security reasons** are the **least cause** of flight delays.



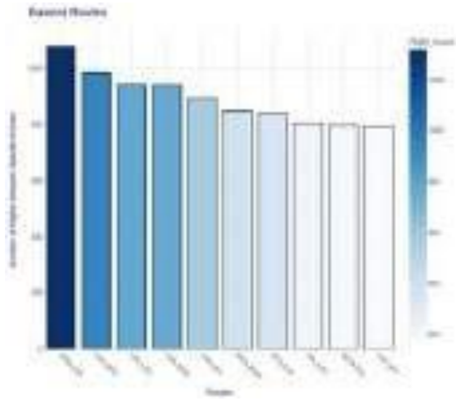
- Certain airline companies, more notably **Southwest Airlines** have an unusally **high departure delay** on an average, comprising of nearly **40** percent of the departure delays.



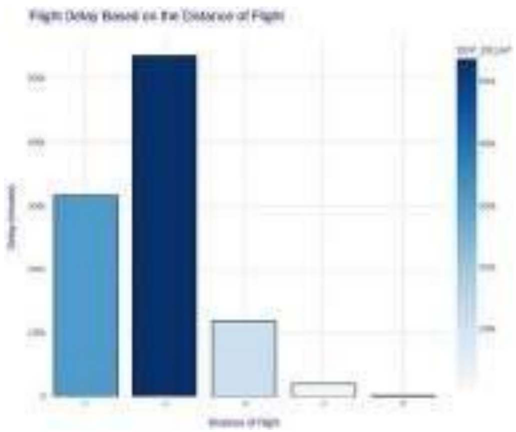
- Certain airline companies, more notably **South-west Airlines** have an unusally **high departure delay** on an average, comprising of nearly **40** percent of the departure delays.



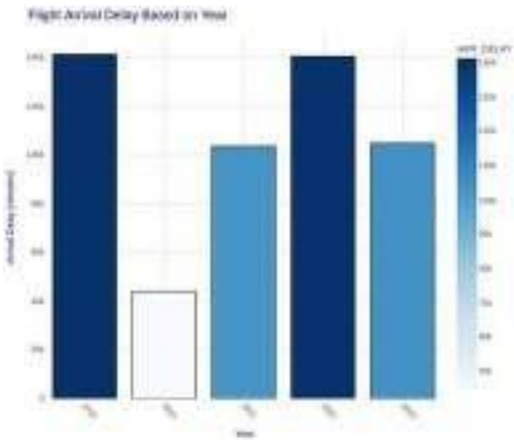
- **Temporal analysis** shows **seasonality trends. Summer** shows **34% departure delays** in flights, **twice as much as fall**.



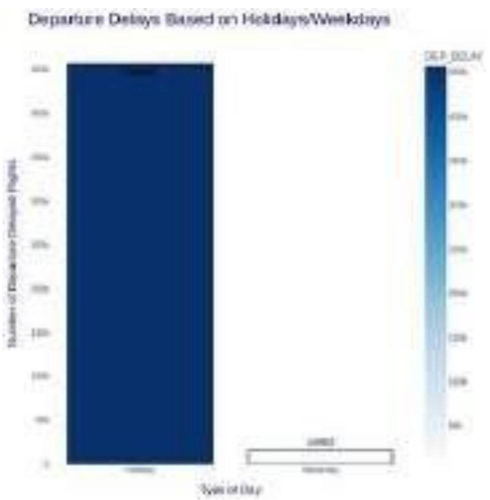
- **Certain routes are the busiest. ORD-LGA** being one of the busiest routes has the **highest number of departure delays,** which is quite obvious.



- **Long distance travelling flights** have **lesser average delay,** as compared to shorter flights, which suggest that airports prefer to let shorter distance flights go.



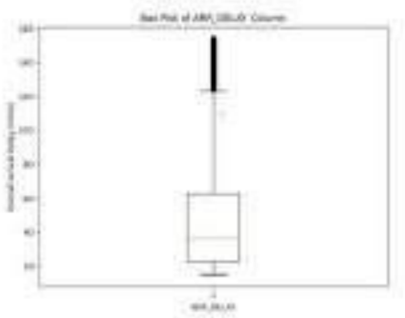
- Also**, year 2020 and 2021,** had the **least number of flights** with flight operations being dormant for most parts of the year, hence the least load on airports and thus, **less average delay.**



- **Number of delays** on **holidays** is nearly **30 times than on weekdays**. Airports experience more load on these days, as compared to weekdays which leads to more average delays.

- **OUTLIER PRUNING**

	Pre-Outlier Removal	Post-Outlier Removal
# of Records	533,863	489,828
Mean	67.526	47.828
Standard Dev.	93.909	32.869
Minimum	15	15
Maximum	2934	154



We incorporated all these insights into the dataset and performed feature engineering to create additional variables such as season, time of day, and route traffic levels. These enhancements significantly improved our data's quality and predictive power.

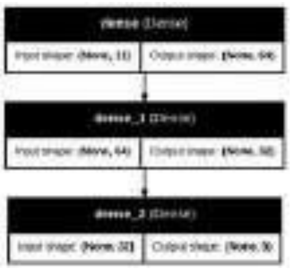
5 Predictive Modelling for Delay Prediction

- **Split and Normalization - Data** was **divided into a 75%/25% train/test split.**
- **Categorical attributes** were **encoded into integer labels.**
  - **ORIGIN** and **DEST** were **encoded together** to ensure consistent relationship.
- **Non-categorical attributes were Z-Score normalized.**
  - **Fitted on training dataset and then applied to the test data.**
- The **baseline machine learning algorithms** used in our model are:
  a. **XGBoost Regressor**
  b. **Artificial Neural Network (ANN)**
- The **time-series machine learning algorithms** used in our model are:
  a. **Long-term Short Memory (LSTM)**
  b. **LSTM + CNN Hybrid Model.**

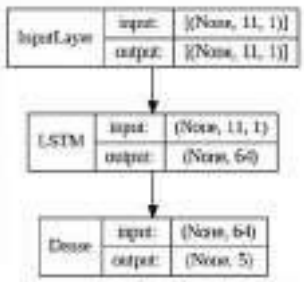
  Their **adeptness** in **handling vanishing gradient problems** and **capturing long-term dependencies** was found to be very **useful** for our dataset.

ARCHITECTURE OF NEURAL NETWORK AND TIME SERIES -

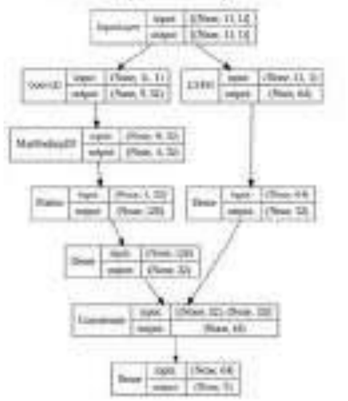
A N N



LSTM



LSTM/CNN Hybrid



Training and Evaluation Metrics-

1. **Mean Absolute Error (MAE) :**

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_{\text{pred},i} - y_{\text{true},i}|$$

- **MAE** quantifies the **average magnitude of errors between predicted and actual values.** It is **useful for interpreting our results because** it is **measured in minutes.**

2. **Mean Squared Error (MSE) :**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **MSE** quantifies **the average squared magnitude of errors between predicted and actual values.** It is **useful for training models because it gives higher weighting to lower frequency values, and we have a lot of zeros.**

Time Series Results -

Total Model Error

	Mean Squared Error	Mean Absolute Error
LSTM	367.868	10.022 9.93
XGBoost	345.02	10.296 9.53
LSTM + CNN Hybrid	367.868	9.684 10.09
*Neural Network	1118.49	10.240 10.02

*Best non-time-series model

- Expected **superiority of time series models, especially LSTM and its variants,** in predicting flight delays **due to their capacity to capture temporal dependencies,** offering valuable insights into machine learning approaches for flight delay prediction.

6 RE-SCHEDULE MODELLING

Genetic Algorithm Overview

A **Genetic Algorithm (GA)** is a **search heuristic** that **mimics the process of natural selection**. This heuristic is routinely used to **generate useful solutions to optimize and search problems.** The algorithm **reflects the process of natural evolution,** where the **fittest individuals** are **selected for reproduction** in order **to produce offspring of the next generation.**

Key Concepts -

Population: A set of potential solutions to the problem.

Chromosomes: A representation of a solution. Typically, this is a string of bits, but other representations are possible.

Genes: Parts of a chromosome, representing a specific trait of the solution.

Fitness Function: A function that evaluates how close a given solution is to the optimum solution of the problem.

Selection: The process of choosing the fittest individuals from the population to create offspring.

Crossover (Recombination): A genetic operator used to combine the genetic information of two parents to generate new offspring.

Mutation: A genetic operator used to maintain genetic diversity within the population by randomly tweaking the genes of individuals.

Genetic algorithms (GAs) are **effective for optimization** because they **use a natural selection-inspired approach** to **explore a vast solution space**, avoiding local optima and **finding near-optimal solutions**. By evolving a population of solutions over generations through selection, crossover, and mutation, GAs can handle complex, nonlinear, and multi-dimensional problems. They are particularly useful when the solution landscape is unknown, as they don't require gradient information, making them versatile for a variety of optimization challenges.

	FL_DATE	AIRLINE	FL_NUMBER	ORIGIN	DEST	CRS_DEP_TIME	Optimized_CRS_DEP_TIME
1	2023-08-31	Frontier Airlines Inc.	2479	IAH	PHX	1517	1639
2	2023-08-31	Republic Airline	5834	LGA	MSN	1993	1951
3	2023-08-31	Southwest Airlines Co.	3188	DEN	AUS	1078	1130
4	2023-08-31	Delta Air Lines Inc.	1850	CLT	ATL	0922	0926
5	2023-08-31	Delta Air Lines Inc.	531	LAX	DTW	2359	-

7 CHALLENGES AND LIMITATIONS

1. **Data Quality and Missing Values**
   - **Handling missing values**, especially in critical **delay-related fields**, posed a challenge. We had to **remove almost 80 percent** of the dataset, because there were **null values** present **in the delay reasons.**
   - Records with **missing delay reasons** or **incomplete flight details** had to be **removed**, which reduced the dataset and may have led to a loss of potentially valuable data.

2. **Outliers and Data Sparsity**
   - **Significant outliers,** particularly **extreme delay values,** affected the initial analysis. While these were **managed using the interquartile range (IQR) method, removing** these records might have also **excluded some real but rare events,** potentially **affecting the model's ability** to predict extreme delays.
   - **The COVID-19 pandemic** led to **sparse data in 2020** due to **reduced flight operations, impacting the continuity of trends and making it challenging to build a model that generalizes well over time.**

3. **Complexity of Delay Factors**
   - Flight delays are influenced by a combination of **external (weather, air traffic)** and **internal (crew availability, maintenance) factors. Capturing all relevant variables** in a predictive model proved challenging**, especially for rare or unpredictable delay reasons.**

4. **Model Generalization**
   - Although the models performed well on the test set, **their ability to generalize to completely new airports, airlines, or routes** (not present in the training data) remains uncertain. **Additional data and testing** would be **required to confirm broader applicability.**

5. **Computational Constraints**
   - Running time-series models like **LSTM on large datasets required considerable computational resources, especially for tuning hyperparameters.** Optimization for speed without compromising accuracy was a balancing act that limited the number of experiments we could perform.

6. **Limited Real-Time Integration**
   - Our models were built on **historical data without real-time inputs** like live weather conditions or airport congestion. This **limits** their **current applicability,** as real-world implementations would benefit from dynamic data inputs to improve prediction accuracy.

8 FUTURE SCOPE AND CONCLUSION

While our model provides significant insights and optimizations, several areas could enhance its effectiveness further:

1. **Integration of Real-Time Data:** Incorporating real-time data, such as **live weather updates and airport traffic,** could improve the accuracy of delay predictions by **accounting for sudden changes in operational conditions. This work has already begun from our side, that is we are trying to incorporate weather condition features like temperatures during time of departure of flight, and min. and max. temperatures in the departure city on that particular date. Factors like wind speed, wind direction, precipitation amount have also been added using OpenCage,** which further adds to the realism in data. Once real time data is integrated, the accuracy of delay predictions could improve.

2. **Advanced Ensemble Techniques:** Exploring **ensemble models** that combine multiple machine learning approaches may yield more robust results, especially for complex scenarios with multiple interacting delay factors.

3. **Broader Geographical and Operational Coverage:** Expanding the model to include **international flights and diverse airline operations** would **enhance its generalizability**, making it **applicable to global airline networks.**

4. **Optimization for Fuel and Cost Efficiency:** Incorporating **metrics related to fuel consumption and operational costs in the rescheduling model could lead to more comprehensive optimizations,** benefiting both airlines and environmental sustainability.

5. **Automated Decision Support System: Integrating the predictive and rescheduling models into an automated decision support system** could **enable real-time recommendations** for flight management teams.