## Airline Delays Prediction

## **INTRODUCTION**

Passenger inconvenience is often caused by airline cancellations and delays, making it an important issue to address in the airline industry. As someone who works in this field, I have always been fascinated by market prediction and flight performance measures. By utilizing publicly available datasets on airline performance and applying learned techniques, it is possible to accurately predict an airline's performance.

For my final project, I plan to analyze airline data to identify various factors that influence a carrier's performance. With different models, I hope to determine the best approach to predict the likelihood of a flight being either on time or delayed.

**Understanding airline delays**:

Airline delays reported are categorized into 5 causes.

* 1. Carrier Delays

Some reasons for a carrier delay include reasons such as late arrival of an aircraft due to a previous delay, maintenance or crew issues, baggage loading/unloading, fueling, aircraft cleaning, etc.

* 1. Late Aircraft Delays

A previous flight was delayed causing the current flight to be delayed.

* 1. NAS Delays

Delays and cancellations caused by the national aviation system refer to conditions, such as airport operations, heavy traffic volume, and air traffic control.

* 1. Weather Delays

Delays caused by adverse weather conditions at the origin or destination causing difficult flying conditions.

* 1. Security Delays

Delays at the terminal or concourse, due to re-boarding of aircraft because of a security breach, inoperative screening equipment, and/or long lines over 30 minutes at screening areas.

**Problem Statement**

This study aims to provide valuable insights to both customers and airlines by predicting a flight's performance. Customers can access information on the likelihood of their flight arriving on time during the booking process or prior to departure. Airlines can also benefit from this by analyzing their performance and anticipating potential delays based on aircraft, origin, and destination. This allows them to implement corrective measures to minimize cancellations and delays and improve their on-time record. Using data science, we can gain a better understanding of the top-performing airlines and the reasons behind delays, diversions, and cancellations across multiple carriers.

**Data Source**

The Excel data has airline performance factors such as canceled, diverted, delayed, and on-time data. The downloaded raw data has up to 34 columns.

<https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E> (Download Raw Data link for data).

Due to the execution size limitation, this study is restricted to one month of airline data, May’2022.

## **Summary of work done**

The data was loaded to conduct a swift analysis of the features and the following steps were performed.

**Data Transformation:**

The data was effectively transformed through the following steps:

* 1. Imputing missing values for time-related columns
  2. Creating essential columns (such as Status)
  3. Adding features that evaluate the reason for delays and cancellations to ensure accuracy.

**Data Visualization:**

1. Overall airline performance in May’22

A picture containing text, screenshot, diagram, plot

Description automatically generated

HA- Hawaiian Airlines had the best on-time performance in May’22.

1. Number of Delays by Delay Reason

A picture containing text, screenshot, diagram, plot

Description automatically generated

From the above chart, we can see most airline delays in May’22 was due to carrier delays.

1. Airline with the most delays

A picture containing text, screenshot, line, plot

Description automatically generated

F9 – Frontier Airlines had the most delays in May’22 followed by G4 - Allegiant Air

1. Following is a chart of the top 10 airports with the most delays.

Origin Airport

A picture containing text, screenshot, font, parallel

Description automatically generated

Destination Airport

A picture containing text, screenshot, font, diagram

Description automatically generated

From the charts, we can see flights to and from Pago Pago International Airport had the most delays.

**Data Preparation**

During the data preparation phase, a series of steps were taken, including dropping unnecessary columns or features, adding new relevant features, and addressing any missing data.

## **Model building and evaluation**

Before building models, the following steps were performed:

1. We split the data into categorical and numerical features.

2. Created dummy variables for the categorical features.

3. Standardized the numerical features using the StandardScaler.

4. As the number of delays in the dataset was significantly less compared to on-time performance, we ran a test to check the balance of test and train data.

5. The test showed that the on-time performance to delayed ratio was 80:20, which could lead to inaccurate results if models were built on this data. To build accurate models on this dataset, we used the Synthetic Minority Oversampling Technique (SMOTE), a statistical technique that increases the number of cases in the dataset in a balanced way.

Upon completing the preparation of the data for modeling, we conducted testing on three distinct models: RandomForest Classifier, DecisionTree Classifier, and LogisticRegression. The accuracy score of each of the models is as follows:

- RandomForest Classifier: 91.92 %

- DecisionTree Classifier: 64.84 %

- LogisticRegression: 88.03 %

## **Conclusion**

From the models, we can say that there is a significant difference between accuracy scores across all models and RandomForest Classifier has the best score. Although the model seems to have a good accuracy score, I believe there could be some more fine-tuning before the model can be deployed.

The data was restricted to a month due to size limitations. There could be more delays across other months that are not considered. For example, there could be more delays in the winter due to frequent storms. I would recommend evaluating the results with more data (a year’s worth).

Understanding the right steps to perform in modeling is essential. In my experience, not selecting the right features caused incorrect model outcomes. Apart from this, I believe data preparation also plays a vital role.

There are other models that can be built and evaluated to explore which model suits best for this dataset.