Flight Delay Prediction for aviation Industry using Machine Learning

INTRODUCTION

1.1 Overview

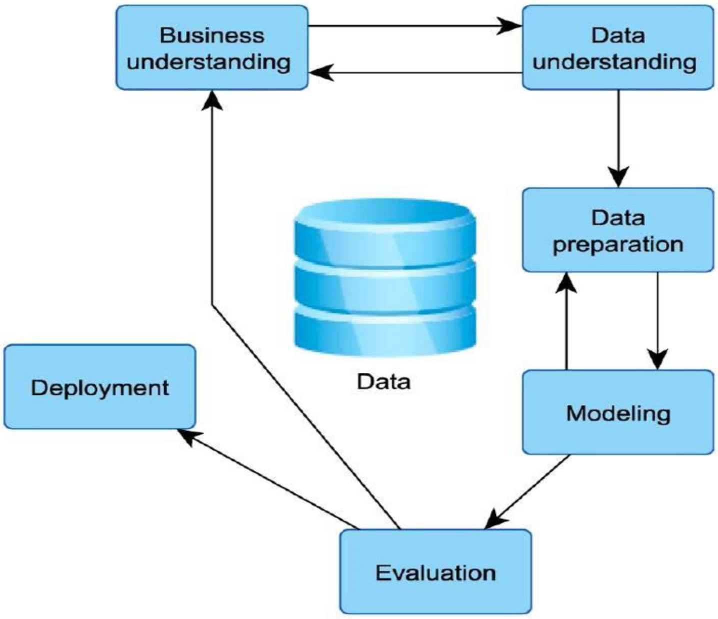
OVER the last twenty years, air travel has been increasingly preferred among travelers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in air traffic and on the ground. An increase in air traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air. These delays are responsible for large economic and environmental losses. According to, taxi-out operations are responsible for 4,000 tons of hydrocarbons, 8,000 tons of nitrogen oxides and 45,000 tons of carbon monoxide emissions in the United States in 2007. Moreover, the economic impact of flight delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy In response to growing concerns of fuel emissions and their negative impact on health, there is active research in the aviation industry for finding techniques to predict flight delays accurately in order to optimize flight operations and minimize delays.

1.2 Purpose

Using a machine learning model, we can predict flight arrival delays. The input to our algorithm is rows of feature vector like departure date, departure delay, distance between the two airports, scheduled arrival time etc. We then use decision tree classifier to predict if the flight arrival will be delayed or not. A flight is delayed when difference between scheduled and actual arrival times is greater than 15 minutes. Furthermore, we compare decision tree classifier with logistic regression and a simple neural network for various figures of merit. Finally, it will be integrated to web based application

2. PROJECT DEVELOPMENT

In this section, there is an overview of the process of data mining and data modeling, from collecting the data, through the data preparation and finally the data modeling. Data cleaning and formatting can be considered as the most critical part of the whole project according to several data scientists [10]. Figure 1 shows how the process of data mining works to extract knowledge using algorithms [11].

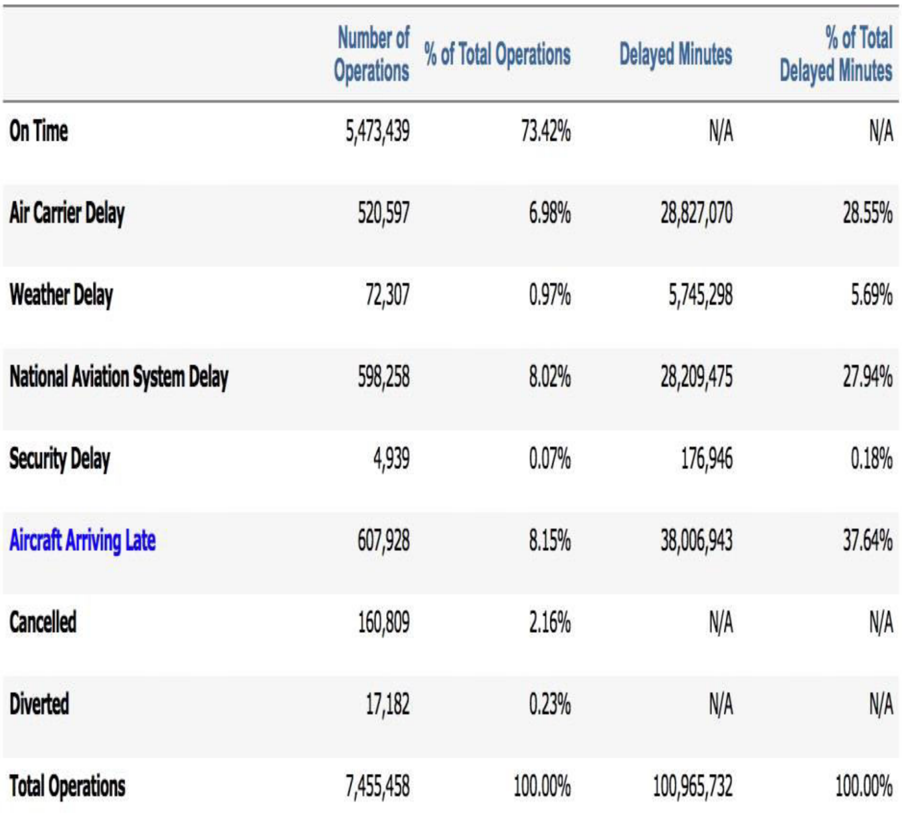


4 DATA COLLECTION

Once the project undertaking is completely comprehended, our subsequent stage is to gather the information that is required for future model building. The information accumulation was an issue as data was not situated at a single source. The data was kept in unique information design. To achieve the end goal, it requires a clear understanding of the correct location of the data.

As we can see in Figure 2 [12], the US Bureau of Transportation Statistics gives detailed information on every single household flight, which incorporates their booking and take off circumstances and real takeoff, origin, destination, date, and carrier. We consolidated a portion of the information properties with Local Climatological Data from National Oceanic and Atmospheric Administration (NOAA) to shape a join data set. Since the datasets for every year are very massive, we decrease our concentration to one-year, i.e., 2008, which as of now contains 1 million records for the most significant airplane terminals. In this venture, we have taken 2007 as our preparation set and 2008 as our test set.

Handling speed is a noteworthy thought since the machine learning methodology that functions admirably on smaller datasets cause issues with the Jupyter Notebook establishments on our PCs.



	Number of Operations	% of Total Operations	Delayed Minutes	% of Total Delayed Minutes
On Time	5,473,439	73.42%	N/A	N/A
Air Carrier Delay	520,597	6.98%	28,827,070	28.55%
Weather Delay	72,307	0.97%	5,745,298	5.69%
National Aviation System Delay	598,258	8.02%	28,209,475	27.94%
Security Delay	4,939	0.07%	176,946	0.18%
Aircraft Arriving Late	607,928	8.15%	38,006,943	37.64%
Cancelled	160,809	2.16%	N/A	N/A
Diverted	17,182	0.23%	N/A	N/A
Total Operations	7,455,458	100.00%	100,965,732	100.00%

Figure 2: Data Collection

5 DATA EXPLORATION FOR FLIGHT DATA

As a first exploratory examination, we consider the watched likelihood of deferral in minutes on the whole dataset. The best path is through a histogram, taking a departure at flight and arrival delays independently.

- Departure & Arrival Delay Distribution

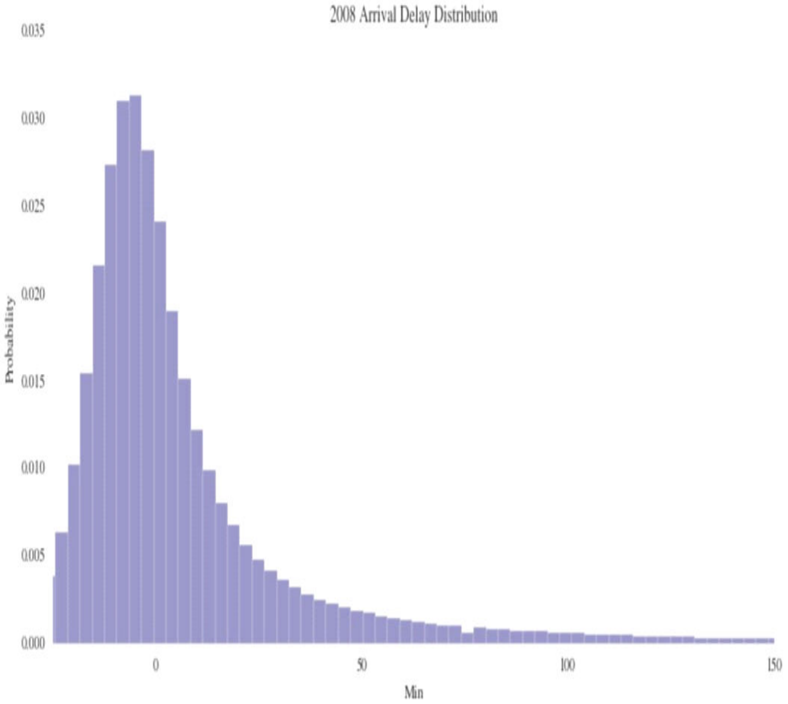


Figure 3: Arrival Distribution

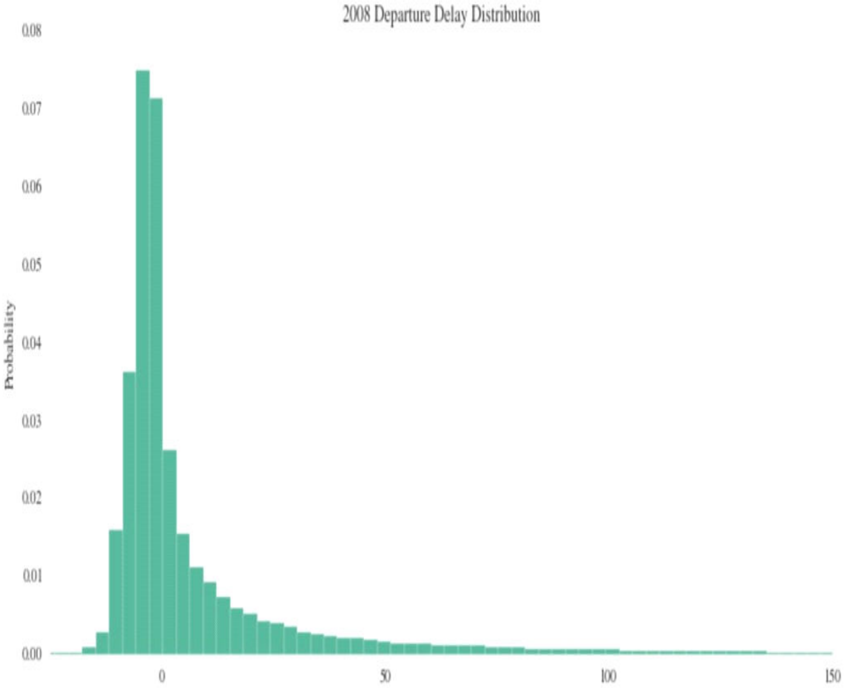


Figure 4: Delay Distribution

In Figure 3 and Figure 4, we notice a much higher probability of short delays. Notice the long right-hand tails from Figure 4. Some flights are delayed for very long time, e.g., over two hours. On the other hand, the delays were centered around just below zero. In both cases, the mode of the distribution is less than zero, meaning most of the flights leave from the gate and arrive at the gate even before the published schedule time of departure and arrival. As we will show below, the more extended delays cancel out, the shorter negative delays (advances), leading to average delays that are above zero.

The x-axis for the two plots is to scale. As a result, we can see that the arrival delay distribution, compared with the departure delay distribution, leans toward left. The scheduled time of an event defines a flight delay compared to the actual time of the event. Airlines usually put extra buffer time on a flight to ensure on-time arrival. Therefore, the departure delay and arrival delay distributions difference indicate that some departure delays were recovered during the flights due to the extra amount of time embedded in the flight time between two airports.

- Average Departure & Arrival by Month

Next, we consider the impact of the months on the delays. We would expect that winter months have the most extended delay. A column chart with departure and arrival delay in minutes plotted by month is the most effective way to see the potential effects of the months.

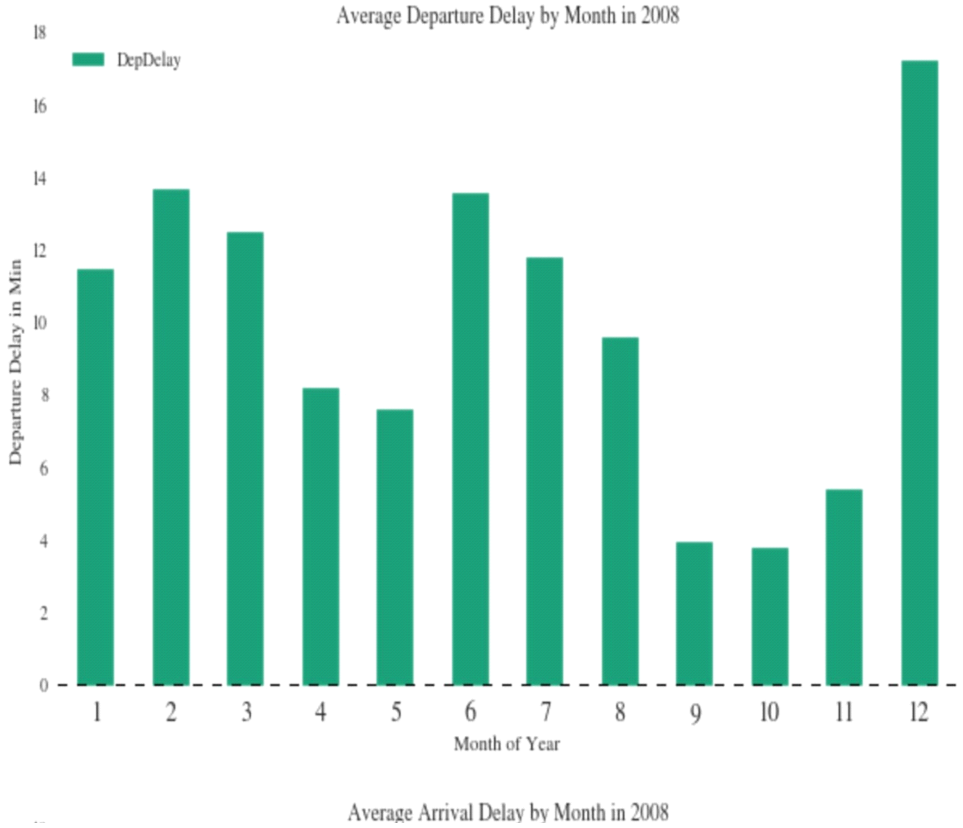


Figure 5: Average Departure Distribution

Figure 5 and 6 clearly show that for both departures and arrivals, the impact of December is clear - the highest delays are in that month. On the other hand, September, October, and November are the months with the least amount of delay. For the summer, June and July are marked by higher delays. Also, February posts high delay values as well. The reason for winter's high delay values is probably because of snowstorms in the northeast of the US.

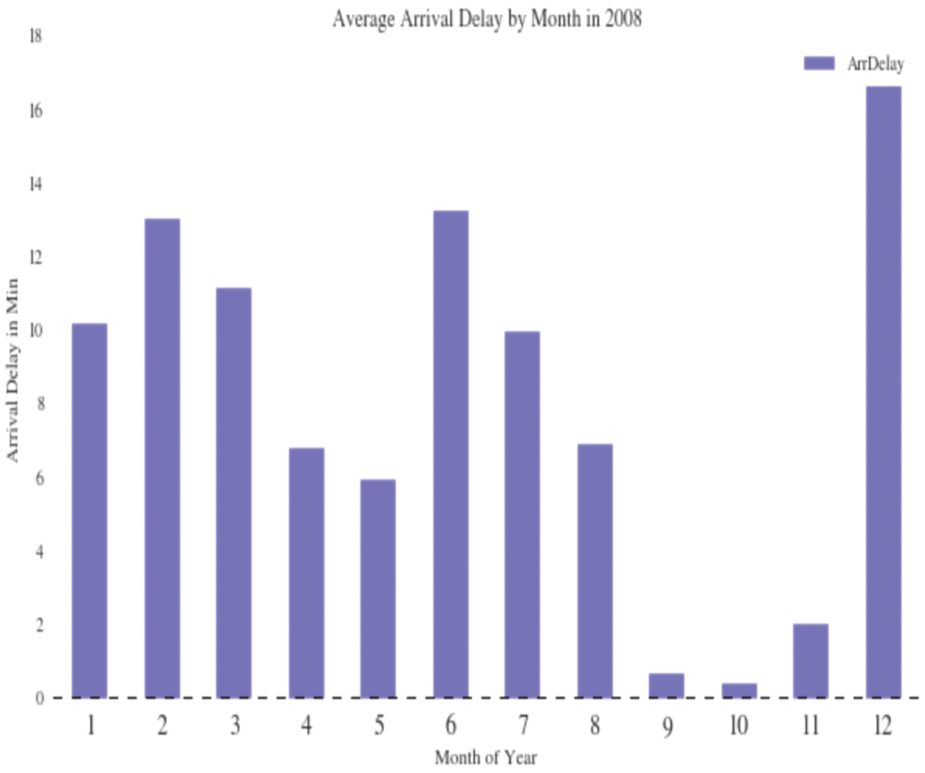


Figure 6: Average Arrival Distribution

Also, in Summer, thunderstorms in Chicago areas can cause high delay impact to the rest of the country. A snowstorm/storm may only affect operations at an airport or two. However, delay propagation, which marks as the significant contributor to the flight delay, can cause ripple effects on delay to downstream flight operations. The time of day should have an impact. Normally, flight delays cumulate throughout the day through a knock-on effect, where delayed flights provoke other delays because of tight schedules and runway congestion. We plot the mean delay by an hour of the day in a column chart.

- Mean Delays across 4 Different Airports

For the next step, we have picked four significant airports and compare these distributions across these airports to see whether the impacts of month and hour of the day are similar at different airports. This analysis could be easily extended to all airports, but for the purposes here, it is sufficient to consider just four.

We subset the dataset into flights departing or arriving at Chicago O'Hare (ORD), Boston Logan (BOS), San Francisco (SFO) and New York LaGuardia (LGA).

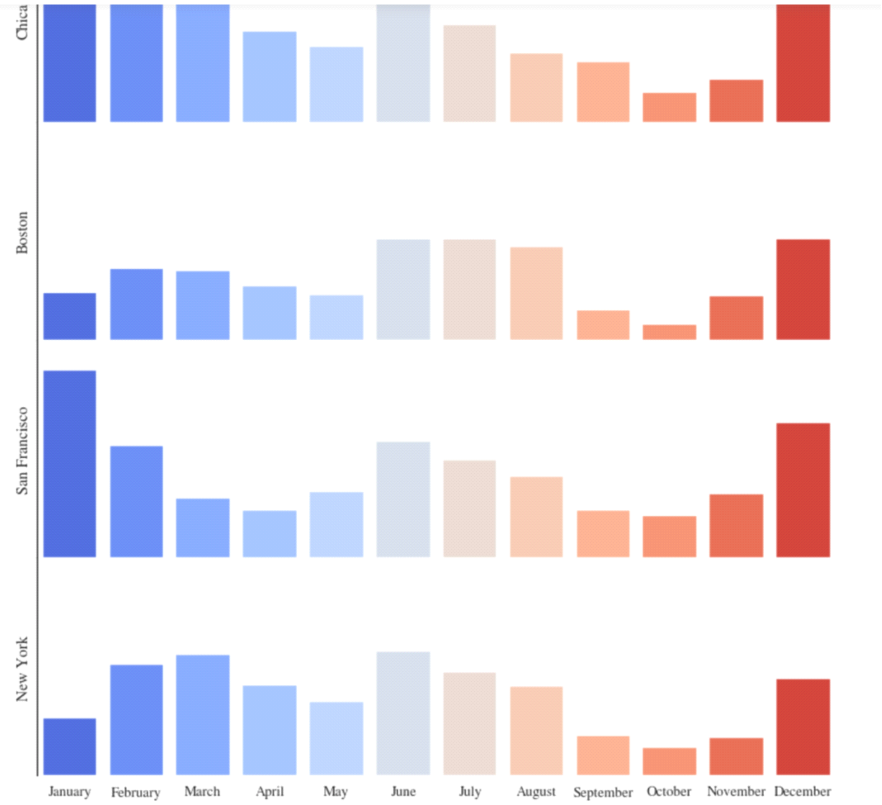


Figure 7: Departure Delays at Airport

First, we look into the departure delays from Chicago O'Hare (ORD), Boston Logan (BOS), San Francisco (SFO) and New York LaGuardia (LGA). As we can see in the Figure 7, the differences between Chicago O'Hare and San Francisco are similar to the overall profile for the mean delay at all airports, with higher delays in December and January and a midsummer bump. On the other hand, Boston Logan shows lower mean delays at the beginning of the year and more delays in all three summer holiday months. New York LaGuardia has many delays in the springtime in February, March, and June having a higher mean delay compared to December. In all locations, December is a month with higher than average delays.

Given that Northeast Corridor does receive a significant amount of rain/snow, further analysis is required for examining the cause and distributions (temporal and spatial relationship) of the flight departure delay.

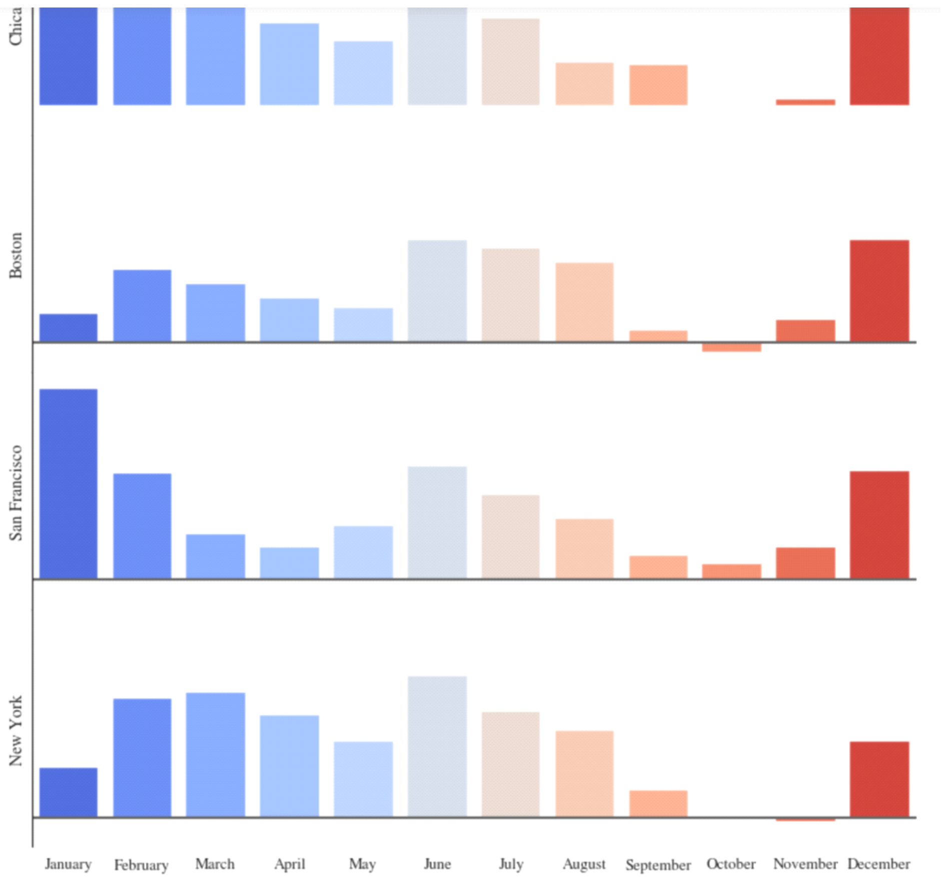


Figure 8: Arrival Delays at Airport

For arrival delays, we see four distinct peak months for Chicago O'Hare: December, January, February, and June in Figure 8. The latter may be because of holiday travel. The former three may be entirely weather related since these three months have the worst climatic conditions. San Francisco is similar to Chicago again. The peak delay is in January, possibly weather-related (bad winter conditions in January 2008). Still, the springtime peak in New York shows the most extended delays from February to

June. In all locations, the end of the year only has minimal delays in October and November, with delays rising back up in December.

NM2023TMID32012 - Flight delay prediction

Milestone 2 : Data collection & Preparation

Activity 1.1 : Importing the libraries

￼

Activity 1.2 : Read the dadaset

￼

Activity 2.1 : Handling missing values

￼

￼

Activity 2.2: Handling Categorical Values

￼

￼

￼

NM2023TMID32012 - Flight delay prediction

Milestone 3 :Exploratory Data Analysis

Activity 1: Descriptive statistical

Activity 2.1: Univariate analysis

Activity 2.2: Bivariate analysis

Activity 2.3: Multivariate analysis

Splitting data into dependent and independent variables

Splitting data into train and test

NM2023TMID32012 - Flight Delay Prediction

Milestone 4: Model Building

Activity 1: Training the model in multiple algorithms

Activity 1.1: Decision tree model

Activity 1.2: Random forest model

Activity 2: Test the model

NM2023TMID32012 - Flight Delay Prediction

Milestone 6: Model Deployment

Activity 1: Save the best model

Activity 2: Integrate with Web Framework