Flight Delay Prediction For Aviation Industry Using Machine Learning

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 Following Figure, 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 Notebookestablishments on our PCs.

The data that I used comes from Kaggle and it consists of a multi-year dataset ranging from 2009 to 2018 separated by year, so one file per year. Each one of these files contains an average of 28 categories with a few million rows. Because of the size of each file I chose to work only with the one corresponding to 2018 which consists of over 7.2 million rows.

As I mentioned above, I will only be considering those categories that you are aware of before the planes takes off. This way my predictions are before the delay is

announce on the departure boards and obviously, before you board the plane. To give you an idea of some of the categories that I dropped, here is a list of them:

- Taxi Out
- Wheels Off
- Wheels On
- Taxi In
- Arrival Delay
- Actual Elapsed Time

All of the above will help the models increase their predictive power and therefore, their accuracy. Now, the category that I will be adding to biassed my models and prove my point will be the "Departure Delay" (not listed above), which is totally predictable that if your plane leaves late, chances are it will be arriving late at it's destination, but remember, airlines try to account for these delays by reducing their elapse time, and in some cases, as Figure 1 suggests, they succeed and the planes end up arriving on time or even earlier.

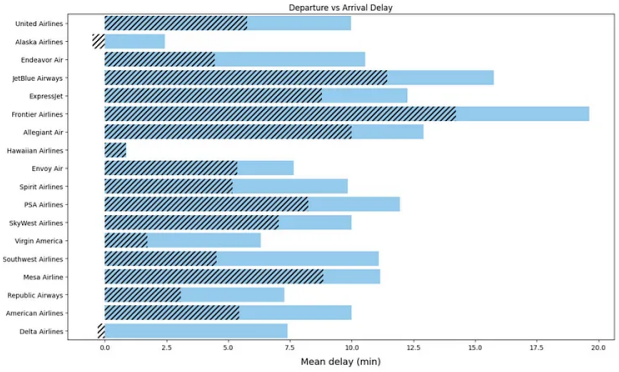

Figure 1. Comparison between "Departure Delays" (light blue) and "Arrival Delays" by airline (area covered by the oblique lines)

LIBRARIES :

Not many libraries are needed for this notebook. I will import Pandas and numpy but I might just end up using pandas as this notebook is basically modifying dataframes to end up with the desire model input

```
import pandas as pd
```

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set_style('darkgrid')

pd.set_option('display.max_columns', None)
import datetime, warnings
warnings.filterwarnings("ignore")
```

Data Pre-processing/Cleaning

The data pre-processing and cleaning was done in two separate stages, a first one were I dealt exclusively with cleaning and a second focused on feature engineering.

At this point I had to make a pause and decide what the definition of a delayed flight would be for the project because this is what would be determining if I could drop or not any other columns and/or rows. So, for a flight to be considered delayed it had to meet only one criteria:

“ARRIVE LATE AT ITS DESTINATION”

Quite simple, and this means that even if a flight has a delay from its departure, but still arrives on time, it will not be considered to be delayed. The same goes for a canceled flight, which in theory never arrived to its destination.

This also implies that this will be a binary classification problem where a “0” means that the flight arrives on time, and “1” that the flight will be delayed. This takes us to the next question, is the dataset balanced? Figure 2 illustrates how imbalanced this dataset is with an almost 2:1 ratio, so this needs to be taken into account when the models performance are evaluated, as accuracy won’t be enough and therefore I will be looking at Precision and Recall as well.

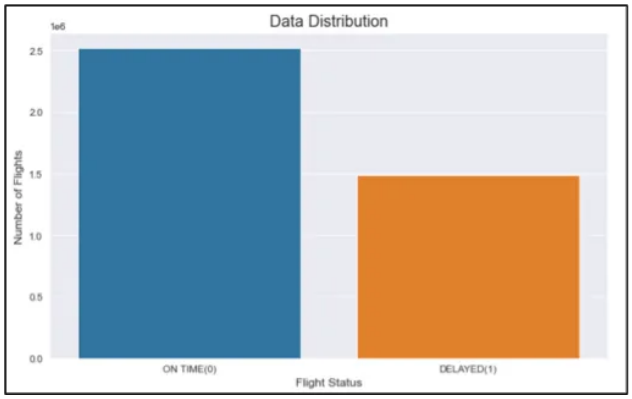

Figure 2. Data distribution showing highly imbalance categories favoring the arrivals on time

	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