EDA and DPL project blog by Aditi Dhavale (22070126006), Ananya Sachan (22070126010), Anshul Shinde (22070126015) AIML A1.

Introduction:

In today's world of modern technologies, Air travel is becoming more and more common amongst all sectors of people, all around the world. This brings upon the common and publicly despised problem of air traffic which causes multiple flight delays each day, all around the globe. According to data from the Bureau of Transportation Statistics, 20.8 percent of flights were delayed so far in 2023, compared to 18.8 percent in 2019. This number might seem small but when compared to the total number of flights departed, the severity is easily noticed. These flight delays disrupt travel plans and increase operational costs. Predicting and preventing them can enhance passenger experience and operational efficiency. This is our goal for this project.

Machine learning has become a household term amongst people nowadays. This is due to the surge of solutions developed using machine learning in recent years. It has become an important tool for making predictions, classifications, etc. Hence, we too have used machine learning to accurately predict Air plane flight delay prediction using factors such as: origin, destination, departure time, departure delay, the time elapsed between landing and arrival at the destination gate, actual arrival time and much more.

This project focuses on harnessing machine learning to predict and prevent airline delays, highlighting the various variables, strategies, and technologies that are reshaping the future of air travel. We will explore the key challenges faced by airlines, the various data sources and models employed to predict delays, and the proactive measures taken to mitigate disruptions before they snowball into major inconveniences. With a focus on passenger satisfaction and operational efficiency, this research seeks to shed light on how machine learning is transforming the aviation landscape and paving the way for a smoother and more enjoyable travel experience for all. For this project, we are using Flight data of 2019 from the official government aviation website as this was the earliest data available for now(https://www.faa.gov/air_traffic/by_the_numbers).

Problem Statement:

Unlocking the Skies: Harnessing Machine Learning to Predict and Prevent Airline Delays for Smoother Travel.

Dataset Description:

2019.csv[5] from U.S. Department of Transportation's (DOT).
The dataset consists of 5048576+ entries in 19 categories in total.

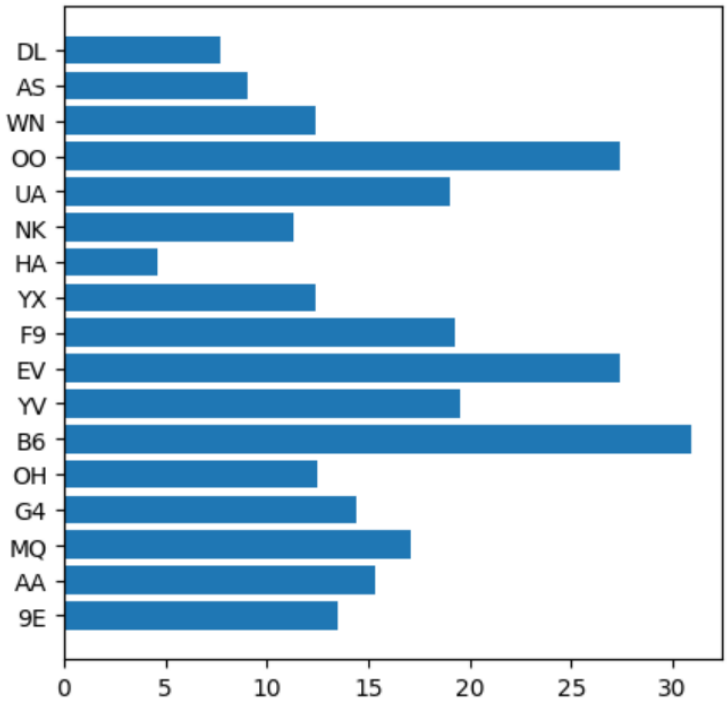
1. FL_DATE: The date of the flight.
2. OP_UNIQUE_CARRIER: The unique code assigned to the carrier operating the flight.
3. OP_CARRIER_FL_NUM: The flight number assigned by the operating carrier.
4. ORIGIN: The origin airport of the flight.
5. DEST: The destination airport of the flight.
6. DEP_TIME: The actual departure time of the flight.
7. DEP_DELAY: The departure delay of the flight, calculated as the difference between the scheduled and actual departure times.
8. TAXI_OUT: The time elapsed between departure from the origin gate and takeoff.
9. WHEELS_OFF: The time at which the aircraft's wheels leave the ground during takeoff.
10. WHEELS_ON: The time at which the aircraft's wheels touch the ground during landing.
11. TAXI_IN: The time elapsed between landing and arrival at the destination gate.
12. ARR_TIME: The actual arrival time of the flight.
13. ARR_DELAY: The arrival delay of the flight, calculated as the difference between the scheduled and actual arrival times.
14. AIR_TIME: The time spent in the air during the flight, calculated as the difference between wheels-off and wheels-on times.
15. DISTANCE: The distance flown by the aircraft during the flight, measured in miles.
16. CARRIER_DELAY: The delay caused by circumstances within the airline's control, such as maintenance or crew problems.
17. WEATHER_DELAY: The delay caused by weather conditions.
18. NAS_DELAY: The delay caused by circumstances within the National Airspace System, such as air traffic control or airport operations.
19. SECURITY_DELAY: The delay caused by security-related issues, such as passenger screening of baggage inspection.
20. LATE_AIRCRAFT_DELAY: The delay caused by a previous flight using the same aircraft arriving late, causing a delay in the subsequent flight.

	FL_DATE	OP_UNIQUE_CARRIER	ORIGIN	DEST	DEP_TIME	DEP_DELAY	ARR_TIME	ARR_DELAY	DISTANCE	WDAY	DAY	MONTH	YEAR
0	2019-01-01	9E	GNV	ATL	601.0	1.0	722.0	-1.0	300.0	1	1	1	2019
1	2019-01-01	9E	MSP	CVG	1359.0	-5.0	1633.0	-36.0	596.0	1	1	1	2019
2	2019-01-01	9E	DTW	CVG	1215.0	-5.0	1329.0	-16.0	229.0	1	1	1	2019
3	2019-01-01	9E	TLH	ATL	1521.0	-6.0	1625.0	-14.0	223.0	1	1	1	2019
4	2019-01-01	9E	ATL	FSM	1847.0	-15.0	1940.0	-25.0	579.0	1	1	1	2019

Flight delay Dataset(Image created by the author)

Airline Code	Name	Airline Code	Name
9E	Endeavor Air	F9	Frontier Airlines
AA	American Airlines	YX	Republic Airways
MQ	Envoy Air	HA	Hawaiian Airlines
G4	Allegiant Air	NK	Spirit Airlines
OH	PSA Airlines	UA	United Airlines
B6	JetBlue Airways	OO	SkyWest Airlines
YV	Mesa Airlines	WN	Southwest Airlines
EV	ExpressJet Airlines	AS	Alaska Airlines
		DL	Delta Air Lines

actual names of respective airlines



Average delay per airline

Finding the most occurring origin-destination pair

```
Route: JFK -> LAX | Occurrences: 3188
Route: LAX -> JFK | Occurrences: 3190
Route: LAX -> SFO | Occurrences: 3743
Route: LGA -> ORD | Occurrences: 3776
Route: ORD -> LGA | Occurrences: 3776
Route: SFO -> LAX | Occurrences: 3752
```

the most occurring route is of LGA: LaGuardia Airport & ORD: O'Hare International Airport

Final delay time for each carrier

```
Final delay time for each carrier:
ORIGIN  DEST  OP_UNIQUE_CARRIER
LGA     ORD   AA                 20.276085
              DL                 68.573171
              NK                 47.140351
              OO                 58.068712
              UA                 35.483960
              YX                 36.284848
```

Delayed flights in each time interval of the day:

INTERVAL	early morning	morning	afternoon	night
OP_UNIQUE_CARRIER				
AA	2	67	101	95
DL	1	18	36	34
NK	7	16	0	29
OO	2	68	146	96
UA	4	31	96	106
YX	3	6	21	19

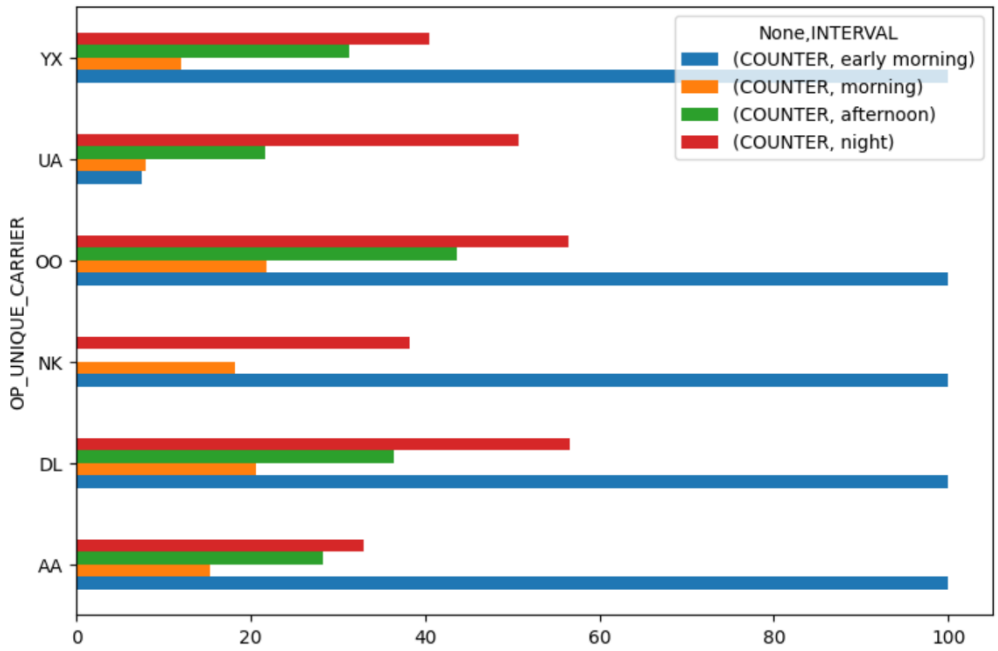
Flights in each time interval of the day:

INTERVAL	early morning	morning	afternoon	night
OP_UNIQUE_CARRIER				
AA	2	438	357	288
DL	1	87	99	60
NK	7	88	0	76
OO	2	311	335	170
UA	53	386	444	209
YX	3	50	67	47

Percentage of flights delayed in each interval of the day:

INTERVAL	early morning	morning	afternoon	night
OP_UNIQUE_CARRIER				
AA	100.00000	15.296804	28.291317	32.986111
DL	100.00000	20.689655	36.363636	56.666667
NK	100.00000	18.181818	0.000000	38.157895
OO	100.00000	21.864952	43.582090	56.470588
UA	7.54717	8.031088	21.621622	50.717703
YX	100.00000	12.000000	31.343284	40.425532

Percentage plot of flight delayed in each interval of day:



Carrier with highest number of flights:

OP_UNIQUE_CARRIER	
9E	60880
AA	228103
AS	61466
B6	72768
DL	225381
EV	37584
F9	28242
G4	24294
HA	19689
MQ	75751
NK	46030
OH	69080
OO	194934
UA	144288
WN	330225
YV	53701
YX	76818

WN carrier has the highest number of flights.

Our project consists of 3 basic stages: Data preprocessing, Exploratory data analysis & flight delay prediction. In the first stage, we dropped unnecessary columns and due to the extremely large data and system limitations, we only considered 3 months data in the dataset. i.e. January, February and March data. The rest of the data preprocessing such as calculating final delay and gain as well as normalizing the distance and time data were all done on this edited dataset. Further EDA was done using the preprocessed dataset to find the most used carrier, busiest times, carrier with maximum delay, delay distribution during days as well as months and carriers too, etc. Along with this, we also observed the correlation between multiple factors and finally decided to apply the Linear regression model for the flight delay prediction. Our r2 score was a good score of 0.9772 which tells us that this model works best for this dataset.

Github link: https://github.com/Aditi-Dhavale/EDA_DPL_flightDataset/tree/main