Flight Delay Predictions

Data Science Project Background and Scoping

Team members: Grace Pham, Heidi Tran, Jessica Jha, Khanh Tran,
Professor: Dr. Bhupesh Shetty

Goals

In scope

We will perform data analysis to answer the following questions:

- Which airlines are on time the most?
- Which airports have high numbers of delayed flights?
- Which are the busiest months/days to fly on average?
- Which airlines have bounced back the most/ the quickest based on the pre-pandemic and post-vaccine data (i.e. pre-2020 and post 2020)?

We will also use Machine Learning models to make prediction on:

- Which flights will have the highest tendency to be on time?
- Which flights will have the highest tendency to be delayed?
- What factors have the most impactful influences on delays in flights' departure and arrival?
- How long will a flight be delayed?

Out of scope

- Using the whole dataset (7.2 GB) for training and testing .

Actions

Data Reduction

Before jumping into the dirty work, there are some actions to be taken for the sake of simplifying the problem and reducing the necessary workload. Since the size of the original dataset (7.2 GB) is too big to be handled comfortably in a Jupyter notebook with our currently available local resources, plus it is also considered an overkill to utilize the whole dataset for the scope of this project, we will conduct some data reduction. The goal is to cut the dataset down to a more reasonable size that will fit our resources yet still allows us to achieve our predefined goals. Ideally, since the original dataset is a time-series one, we choose to keep the records from the most

recent years. We will be experimenting with different numbers of years to come up with the best option. An obvious trade-off of data reduction is the loss of meaningful analysis and insights as well as an expected drop in evaluation metrics during the model testing phase. As a result, it is utmost important to keep all aspects of the dataset in balance. This can also be considered a statistical problem where we will need to maintain the proportions between the number of values of different attributes, especially the ones that will be used in training the model. Skewed dataset needs to be avoided at all costs. A good ratio of number of attributes : number of records is also an action that will be taken to prevent underfitting models. In other words, the data reduction task will need to be monitored with great care both horizontally and vertically.

Approach

The original dataset will be imported and cut down using Spark, or PySpark to be specific. It is highly possible that the Spark cluster will be in Local Mode - not the best way to utilize Spark but it is convenient and suitable for the size of our dataset. This action will be followed by the reduced data being converted to a Pandas DataFrame and exported as a CSV file. After that, Pandas will be in heavy use for the data analysis and modeling tasks, along with Matplotlib or Seaborn for data visualization.

Data

Data Retrieval

The dataset was retrieved from IBM. The original dataset is 7.2 GB in size. This dataset provides information on roughly 200 million U.S. domestic flights on United States Bureau of Transportation Statistics along with the flights' information including flight date, place of origin, destination, delay time, flight time, etc.

Data Description

Details on all attributes of flights in this dataset will be demonstrated as below:

Feature	Description

Year	Year
Quarter	Quarter
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week (numeric)
FlightDate	Date of Flight
Reporting_Airline	Airline Unique Carrier Code
DOT_ID_Reporting_Airline	Number assigned by US DOT to identify a unique airline
IATA_CODE_Reporting_Airline	Airline Code assigned by IATA
Tail_Number	Aircraft tail number
Flight_Number_Reporting_Airline	Flight Number
OriginAirportID	Origin Airport ID
OriginAirportSeqID	Origin Airport Sequence ID
OriginCityMarketID	Origin City Market ID
Origin	Origin Airport Code
OriginCityName	Origin City Name
OriginState	Origin State
OriginStateFips	Origin State FIPS place code

OriginStateName	Origin State Name
OriginWac	Origin Airport World Area Code
DestAirportID	Destination Airport ID
DestAirportSeqID	Destination Airport Sequence ID
DestCityMarketID	Destination City Market ID
Dest	Destination Airport Code
DestCityName	Destination City Name
DestState	Destination State
DestStateFips	Destination State FIPS code
DestStateName	Destination State Name
DestWac	Destination Airport World Area Code
CRSDepTime	Computer Reservation System (scheduled) Departure Time
DepTime	Departure Time (hhmm)
DepDelay	Departure delay (minutes)
DepDelayMinutes	Absolute value of DepDelay
DepDel15	Departure Delay >15?
DepartureDelayGroups	Departure delay 15 minute interval group
DepTimeBlk	Computer Reservation System (scheduled) time block

TaxiOut	Taxi out time (minutes)
WheelsOff	Wheels off time (local time, hhmm)
WheelsOn	Wheels on time (local time hhmm)
TaxiIn	Taxi in time (minutes)
CRSArrTime	Computer Reservation System (scheduled) Arrival Time
ArrTime	Arrival time (local time, hhmm)
ArrDelay	Arrival delay (minutes)
ArrDelayMinutes	Absolute value of ArrDelay
ArrDel15	Arrival Delay >15?
ArrivalDelayGroups	Arrival delay 15 minute interval group
ArrTimeBlk	Computer Reservation System (scheduled) arrival time block
Cancelled	1 = canceled
CancellationCode	A = Carrier, B = Weather, C = National Air System, D = Security
Diverted	1 = diverted
CRSElapsedTime	Computer Reservation System (scheduled) elapsed time
ActualElapsedTime	Actual elapsed time
AirTime	Flight time (minutes)
Flights	Number of flights

Distance	Distance between airports (miles)
DistanceGroup	250 mile distance interval group
CarrierDelay	Carrier delay (minutes)
WeatherDelay	Weather delay (minutes)
NASDelay	National Air System delay (minutes)
SecurityDelay	Security delay (minutes)
LateAircraftDelay	Late aircraft delay (minutes)
FirstDepTime	First gate departure time at origin airport
TotalAddGTime	Total ground time away from gate
LongestAddGTime	Longest time away from gate
DivAirportLandings	Number of diverted airport landings
DivReachedDest	1 = diverted flight reached scheduled destination
DivActualElapsedTime	Elapsed time of diverted flight reaching scheduled destination
DivArrDelay	Difference in minutes between scheduled and actual arrival time
DivDistance	Distance between scheduled and diverted airport
Div1Airport	Diverted Airport 1
Div1AirportID	Diverted Airport 1 ID
Div1AirportSeqID	Diverted Airport 1 Sequence ID

Div1WheelsOn	Diverted Airport 1 wheels on time (local, hhmm)
Div1TotalGTime	Diverted Airport 1 total ground time away from gate
Div1LongestGTime	Diverted Airport 1 longest ground time away from gate
Div1WheelsOff	Diverted Airport 1 wheels off time (local, hhmm)
Div1TailNum	Diverted Airport 1 aircraft tail number
Div2Airport	Diverted Airport 2
Div2AirportID	Diverted Airport 2 ID
Div2AirportSeqID	Diverted Airport 2 Sequence ID
Div2WheelsOn	Diverted Airport 2 wheels on time (local, hhmm)
Div2TotalGTime	Diverted Airport 2 total ground time away from gate
Div2LongestGTime	Diverted Airport 2 longest ground time away from gate
Div2WheelsOff	Diverted Airport 2 wheels off time (local, hhmm)
Div2TailNum	Diverted Airport 2 aircraft tail number
Div3Airport	Diverted Airport 3
Div3AirportID	Diverted Airport 3 ID
Div3AirportSeqID	Diverted Airport 3 Sequence ID
Div3WheelsOn	Diverted Airport 3 wheels on time (local, hhmm)
Div3TotalGTime	Diverted Airport 3 total ground time away from gate

Div3LongestGTime	Diverted Airport 3 longest ground time away from gate
Div3WheelsOff	Diverted Airport 3 wheels off time (local, hhmm)
Div3TailNum	Diverted Airport 3 aircraft tail number
Div4Airport	Diverted Airport 4
Div4AirportID	Diverted Airport 4 ID
Div4AirportSeqID	Diverted Airport 4 Sequence ID
Div4WheelsOn	Diverted Airport 4 wheels on time (local, hhmm)
Div4TotalGTime	Diverted Airport 4 total ground time away from gate
Div4LongestGTime	Diverted Airport 4 longest ground time away from gate
Div4WheelsOff	Diverted Airport 4 wheels off time (local, hhmm)
Div4TailNum	Diverted Airport 4 aircraft tail number
Div5Airport	Diverted Airport 5
Div5AirportID	Diverted Airport 5 ID
Div5AirportSeqID	Diverted Airport 5 Sequence ID
Div5WheelsOn	Diverted Airport 5 wheels on time (local, hhmm)
Div5TotalGTime	Diverted Airport 5 total ground time away from gate
Div5LongestGTime	Diverted Airport 5 longest ground time away from gate
Div5WheelsOff	Diverted Airport 5 wheels off time (local, hhmm)

Div5TailNum	Diverted Airport 5 aircraft tail number

Analysis

Exploratory Analysis

- We will use various libraries in Python to answer questions proposed in the Goals section, for quantifiable and tangible results that are less advanced and don't need models to be answered.
- We will look into Tableau if that helps us visualize geospatial data.

Modeling Analysis

Some of the possible models we are planning to use:

- Linear Regression
- Logistic Regression
- Random Forest
- Artificial Neural Network

Following, we will evaluate the accuracy of our models using python libraries such as:

- Cross-validation
- Accuracy
- Precision
- Recall & F1-Score ratings
- Root mean square error scores

Considerations

Data Quality Considerations

- Reducing the size of the dataset both in the number of rows and columns. Since stale data can harm the model's performance, we would consider only keeping recent data, for example the last 5 years of data.

- After brief analysis of the data, there are values that are not numbers, some empty values, we will have to deal with these by either dropping the columns, calculating means of values and replacing the NaN/0s, or some other method of tidying the empty values.

Ethical considerations:

- Regarding privacy, transparency, discrimination, equity, and accountability issues for this dataset, there is no personal identifiable information linked to passengers or pilots nor their demographic or monetary information, but solely time and departures of airline companies and its metadata with other factors.

Website:
https://developer.ibm.com/exchanges/data/all/airline/