AirFly-Insights

Infosys SpringBoard Internship

Dataset:

Kaggle: Flight Delay and Causes

https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes

Insights

- Dataset contains flight schedule, delay reasons, cancellations, and airport information.
- Key delay-related columns: ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay.
- Origin and Destination columns with airport details
- Delay values are often right-skewed with extreme outliers.
- Missing values present in elapsed times, delays, and airport codes.
- Time columns (DepTime, ArrTime, CRSArrTime) are stored as integers instead of proper time formats.
- There are totally 29 Columns,500k rows.

Key Performance Metrics:

- Average departure and arrival delays per airline or per airport.
- Top origin and destination airports with most delays.
- Morning flights vs evening flights delay analysis.
- Delays caused by weather, carrier, NAS, security, or late aircraft.

Preprocessing Technique:

a) **Date Conversion**

- Typecast Date from integer format into standard datetime format**.**
- Further extract Day, Month, and Weekday to enable time-series analysis and seasonal trend detection.

b) **Time Standardization**

- Convert flight time columns (DepTime, ArrTime, CRSArrTime) from integer format to HH:MM string**.**
- This ensures consistent representation and enables calculation of time differences .

c) **Handling Missing Values in Numeric Columns**

- Impute missing values in numeric features (DayOfWeek, ActualElapsedTime, CRSElapsedTime, AirTime, ArrDelay, DepDelay, Distance, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay) using mean values**.**

d) **Handling Missing Values in Categorical/Time Columns**

- For categorical and time-related columns (Org_Airport, Dest_Airport, DepTime, ArrTime), use grouped imputation based on FlightNum or Carrier.
- This ensures that imputed values are contextually accurate, preserving route-specific and airline-specific operational patterns.