Predicting Flight Delays

How Weather & the Taylor Swift Era’s Tour Impacts Flight Delays

Hope Husemann
Data Science Capstone Project, Nov 3, 2024
Springboard

The Problem: Predict the likelihood or extent of flight delays based on historical flight data, weather conditions, airline/airport characteristics, and potentially external factors such as NFL games. Develop a machine learning model to predict flight delays to provide valuable insights and decision-making support for airlines, airports, and passengers to better manage delays.

Predicting flight delays has broad applications beyond the immediate scope of the airline industry including, but not limited to the following:

- Logistics and Supply Chain Management: Freight companies can use delay predictions to manage logistics and adjust delivery schedules based on anticipated disruptions in air cargo transport. By anticipating delays, logistics companies can optimize routing and scheduling to minimize the impact on their supply chains.
- Large Businesses and Corporate Travel: Businesses may need to partner with airlines for corporate travel needs. Predicting flight delays helps businesses choose airlines that offer more reliable service, improving the efficiency of corporate travel and reducing risk of disruptions..
- Travel Insurance: Insurance companies can use delay predictions to better manage claims related to flight delays, helping to streamline the claims process and improve customer service. They can also use predictive insights to price travel insurance policies more accurately, taking into account the likelihood of flight delays.
- Travel and Hospitality Industry: Travel agencies and booking platforms can use delay predictions to offer more accurate travel itineraries and recommendations. By understanding patterns in flight delays, these platforms could consider implementation of dynamic pricing and availability in real time to mitigate the impact of delays.

Context:
Flight timelines are often unpredictable to the average person, because delays can be caused by so many different factors. Our goal is to review historical flight data, weather conditions, airline/airport characteristics, and potentially external factors such as NFL games to provide meaningful data in regards to their upcoming flights to predict likelihood or extent of delays.

Goal:
Develop a machine learning model to predict flight delays to provide valuable insights and decision-making support for airlines, airports, and passengers to better manage delays**.**

The submission shows

- All data wrangling steps were applied where necessary.
- EDA figures and visualizations were created and reviewed for feature relationship understanding.
- the create dummy features steps were completed.
- the features were magnitude standardized.
- the split into train and test data subsets was completed. ○
- three different models were built.
    - Linear
    - Regression
    - Other
    - Other
- the model performance comparison table is filled in.

Upload datasets:

```
flights_data = pd.read_csv(r"C:\Users\hopeh\Desktop\data_science_bootcamp\flight_times_capstone\flights_airport_iata.csv", low_memory=False)
weather_data = pd.read_csv(r"C:\Users\hopeh\Desktop\data_science_bootcamp\flight_times_capstone\weather_iata.csv", low_memory=False)
```

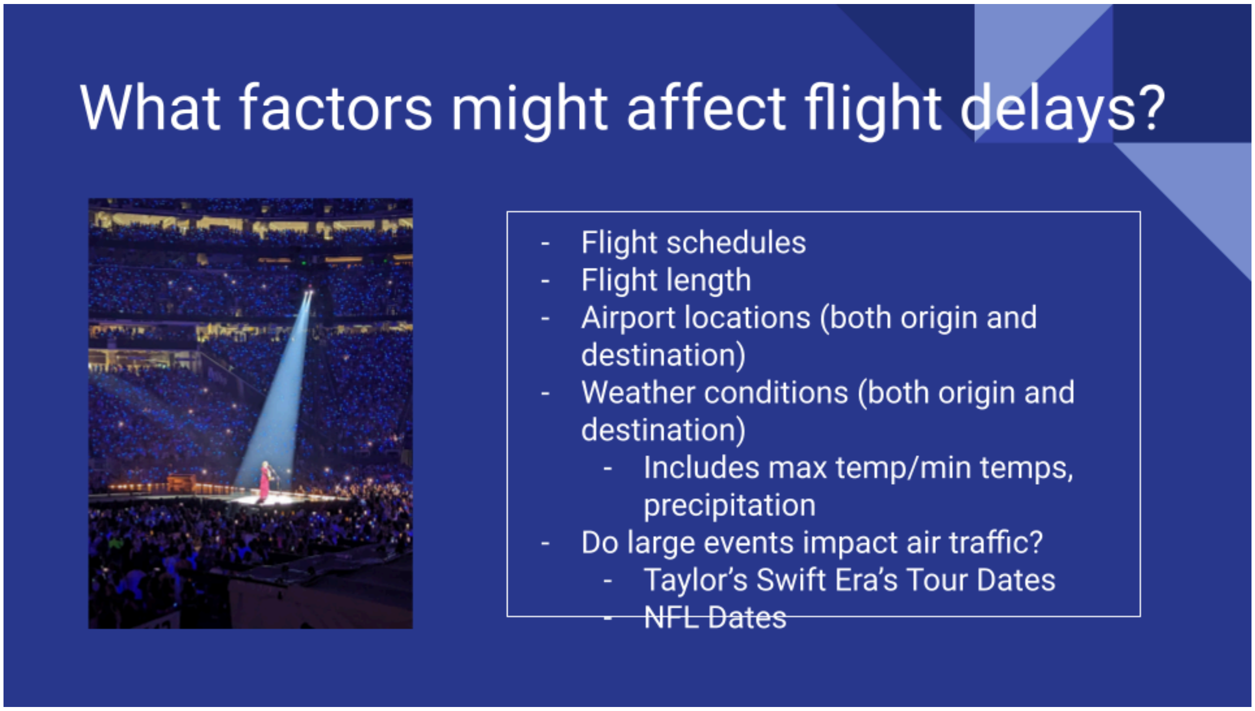
- Merged airport data, flight data, airport data based on the date, location, and nearest medium/large airport nearby (via IATA Codes.)
- Merged flights data and weather data.
    - At origin airport and destination airport
- Merged Era’s Tour Dates by nearest IATA Code

```
print(weather_data.columns)
'latitude', 'longitude', 'elevation', 'date', 'prcp', 'snow', 'snwd',
    'tmax', 'tmin', 'tobs', 'city', 'state_abbr_x', 'iata', 'airport_name',
    'state_abbr_y'
```

```
print(flights_data.columns)
'day_of_week', 'date', 'op_unique_carrier', 'tail_num',  'op_carrier_fl_num', 'origin_iata', 'origin_city',
'dest_iata', 'dest_city', 'crs_dep_time', 'dep_time', 'taxi_out', 'wheels_off',  'wheels_on', 'taxi_in',
'crs_arr_time', 'arr_time', 'cancelled', 'diverted', 'crs_elapsed_time', 'actual_elapsed_time', 'air_time',
        'flights', 'distance', 'distance_group', 'carrier_delay', 'weather_delay', 'nas_delay', 'security_delay',
'late_aircraft_delay', 'origin_state', 'dest_state', 'iata', 'origin_latitude', 'origin_longitude', 'iata_dest',
'airport_name_dest', 'dest_latitude', 'dest_longitude', 'state_abbr_dest', 'airport_ref', 'airport_ident',
        'type_of_airport', 'airport_name', 'elevation_ft', 'origin_state.1',  'municipality', 'scheduled_service',
'unique_id', 'id', 'length_ft', 'width_ft', 'surface', 'lighted', 'closed', 'le_ident', 'le_displaced_threshold_ft',
'he_ident', 'he_displaced_threshold_ft'], dtype='object
```

- Standardized dataframe formatting
- Handle duplicate columns
- Handle duplicate rows
- Investigate and handle null values
- Investigate and remove outliers, removing most outliers from early departures as this is very rare due to nature of operations.
  - Number of rows with 'tour_date' set: 14642 flights in USA surrounding the Era's Tour

Investigate which features are correlated with arrival delays, departure delays and total elapsed delay time.



What factors might affect flight delays?

- Flight schedules
- Flight length
- Airport locations (both origin and destination)
- Weather conditions (both origin and destination)
  - Includes max temp/min temps, precipitation
- Do large events impact air traffic?
  - Taylor's Swift Era's Tour Dates
  - ~~NFL Dates~~