



# **Empowering Air Travelers: Real-time Flight Delay Prediction and Optimal Flight Selection with Ensign and Machine Learning**

The Three Musketeers



# Our team



Toby Chiu



Kevin Sianto



Jason Chandra

# Agenda

- Introduction to the Project
- Dataset
- Project Setup
- Machine Learning Model
- Demo
- Results
- Conclusion and Recommendation

# Problem Statement:

The aviation industry faces the task of efficiently managing flight delays, which can result in various consequences such as passenger inconveniences, heightened operational expenses, and scheduling disruptions. According to FAA/Nextor, the annual costs of delays (direct cost to airlines and passengers, lost demand, and indirect costs) in 2018 amounted to a staggering **\$28 billion**.

# Value Proposition:

Our primary objective is to create a real-time flight delay prediction system tailored for **consultants**, helping them identify flights with expected delays, providing **confidence intervals** for predictions, and offering **specific departure and arrival airport pair predictions** to make informed travel decisions.

# Use Case:

Date	Aircraft	Origin	Destination	Departure	Arrival	Duration
<a href="#">17-Oct-2023</a>	A321	John F Kennedy Intl ( <a href="#">KJFK</a> )	Los Angeles Intl ( <a href="#">KLAX</a> )	08:10PM EDT	11:07PM PDT	Scheduled
<a href="#">16-Oct-2023</a>	A321	John F Kennedy Intl ( <a href="#">KJFK</a> )	Los Angeles Intl ( <a href="#">KLAX</a> )	08:49PM EDT	10:41PM PDT	En Route
<a href="#">15-Oct-2023</a>	A321	John F Kennedy Intl ( <a href="#">KJFK</a> )	Los Angeles Intl ( <a href="#">KLAX</a> )	08:37PM EDT	10:37PM PDT	4:59
<a href="#">14-Oct-2023</a>	A321	John F Kennedy Intl ( <a href="#">KJFK</a> )	Los Angeles Intl ( <a href="#">KLAX</a> )	08:26PM EDT	10:46PM PDT	5:19

The departure and arrival times for *most* frequent flights are the same every day, we can use historical data and flight path to predict the likelihood of delay.

# Dataset/ Data Collection

**Data Source:** OpenSky API and FlightRadar24 API

**Description:** Each sample in the dataset represents a specific flight with various attributes related to that flight at one point in time

**Unit of Analysis:** One unit of aircraft flight

**Frequency of data updates:** every 15 seconds

**Model Engineering:** Datasets contain airline information, altitude, velocity, geographical coordinates, scheduled departure/arrival, which can be used as inputs in the ML model

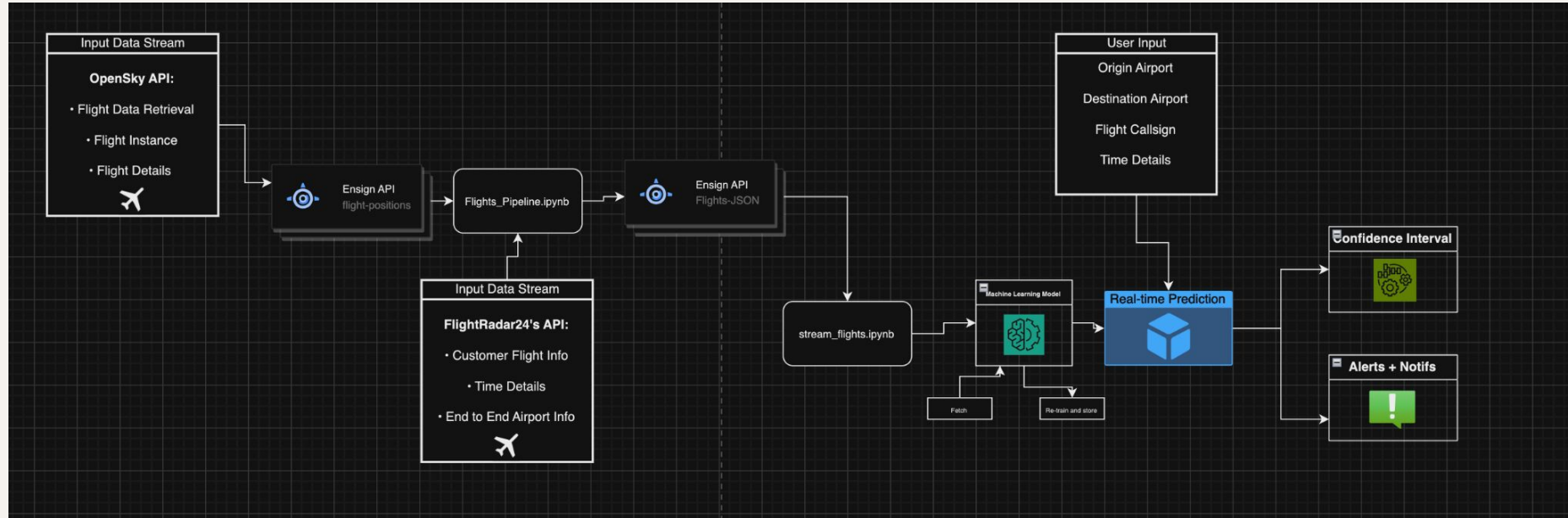
## Limitations:

Unreliability of Opensky/FlightRadar24 APIs (slow queries/timeout due to rate limit) leads to decreased data quality - missing/incorrect rows

Imperfect fit between APIs - creative workarounds required

```
{
  "actual_departure": 1697522102,
  "airline": "Delta Air Lines",
  "airline_icao": "DAL",
  "barometric_altitude": 10668,
  "callsign": "DAL1140 ",
  "category": 0,
  "destination_airport_icao": "KMCO",
  "estimated_arrival": 1697535616,
  "geo_altitude": 11148.06,
  "historical_delay": "-1284",
  "historical_flight_time": "13465",
  "icao24": "a057cc",
  "last_contact": 1697526279,
  "latitude": 36.6649,
  "longitude": -102.6205,
  "on_ground": false,
  "origin_airport_icao": "KSLC",
  "origin_country": "United States",
  "position_source": 0,
  "scheduled_arrival": 1697536800,
  "scheduled_departure": 1697521500,
  "sensors": null,
  "special_purpose_indicator": false,
  "time_position": 1697526279,
  "transponder_code": "6027",
  "true_track": 117.11,
  "velocity": 242.73,
  "vertical_rate": -0.33
}
```

# Project Setup





# Methodology: Pipeline

**Live Demo**

# Machine Learning Model

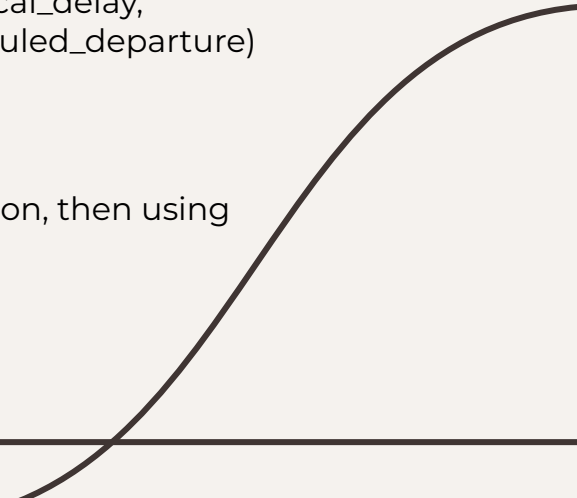
The model that we are going to use is **Linear Regression**. The output that we are trying to predict is the **arrival time**, which is a continuous variable. We chose linear regression because it is **highly interpretable** to figure out which features affect our prediction and its **simplicity to implement**.

## Feature Engineering:

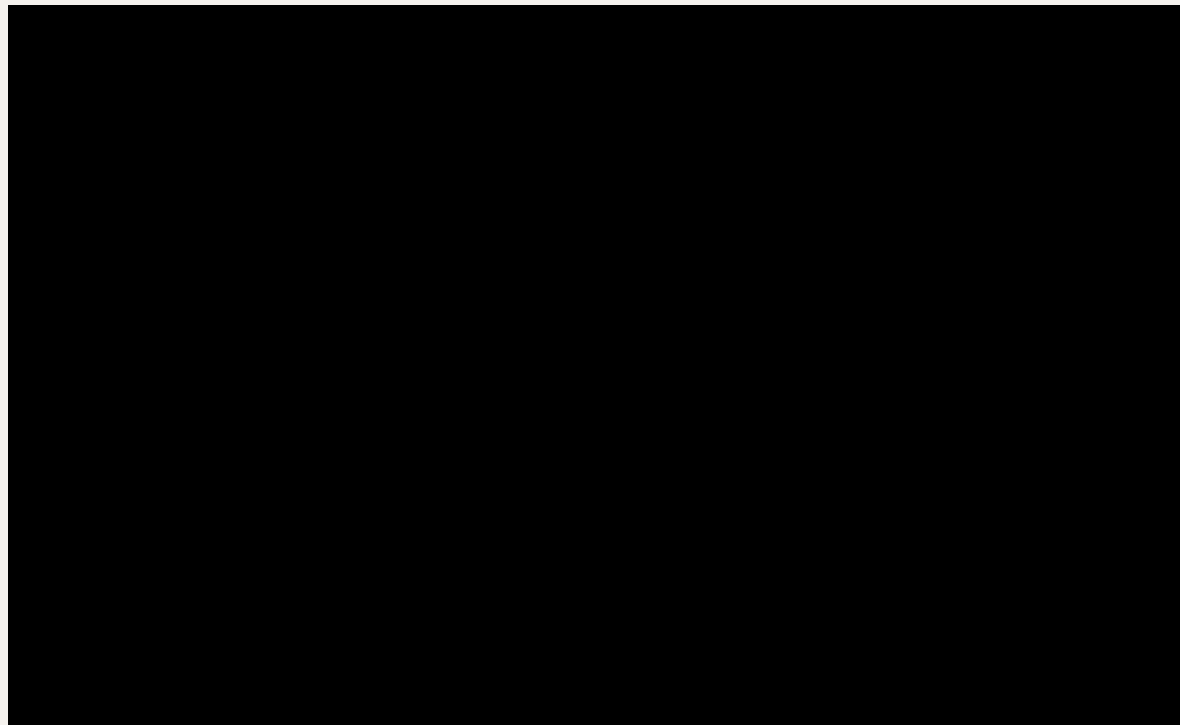
- **Selection:** Time-based features, airport pairs, and historical flight data are used (Fields: actual\_departure, callsign, destination\_airport\_icao, historical\_delay, historical\_flight\_time, origin\_airport\_icao, scheduled\_arrival, scheduled\_departure)
- **Scaling** - `model |= preprocessing.StandardScaler()`

## Validation:

Model performance is validated through data splitting and cross-validation, then using evaluation metrics such as RMSE, MAE and R-squared.



# Demo of Output



# Modelling Results

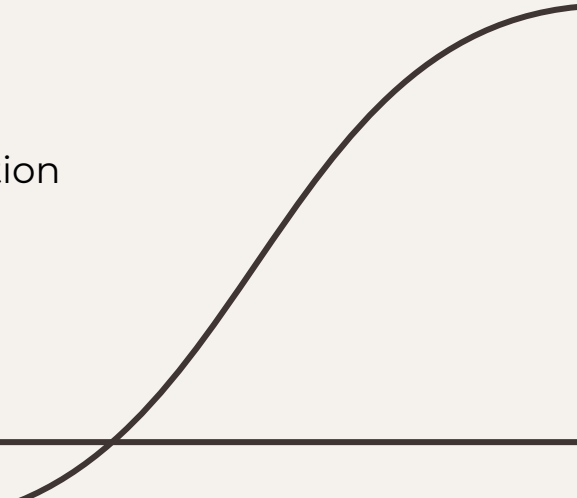
We didn't get to this stage yet but if we did...

Evaluate key metrics such as *RMSE*, *MAE*, *R-squared* to minimize errors.

- With more time, we could consider other models such as polynomial regression, decision trees, random forests, etc. and do cross validation to see which model performs the best.

Validation:

Perform k-cross validation to estimate the model's generalization performance



# Conclusion and Recommendation

- The project provides actionable recommendations for consultants to make informed travel decisions, offering insights into which flights are likely to experience delays and provides confidence intervals to guide consultants and travelers.
  - Limitations: Data quality, coverage of API, privacy, query limit.
  - Further steps could involve real-time implementation and integration with travel platforms, enhancing the user experience for consultants and travelers.
  - Future extensions: expanding the system to cover a broader range of flights and regions. Or, we could integrate additional features, such as weather data, to enhance prediction accuracy and further empower air travelers.
- 