Flight Delay Prediction

MACHINE LEARNING PROJECT

Nicholas Kuok Jin Shung
FCS12124
7 January, 2026

PROBLEM STATMENT

Flight delays often happen without warning, making it hard for passengers and airlines to plan their schedules.

Stakeholders :

- Passengers
- Airlines
- Airport Operators

Impact :

Unpredicted delays cause long waiting times for passengers and increase operating costs for airlines and airport.

Data Overview

Source :

Kaggle flight delay dataset (2024)

Granularity :

One row represents one flight with details

Size :

7,079,081 rows and 35 columns

Target Variable:

delayed = (arr_delay > 15)

Objectives & Key Questions

Objectives :

- Build a machine learning to predict flight delays
- Use flight information to classify flights as delayed or on time

Key Questions :

- Can flights data predict delays ?
- How well does the model identify delayed flights ?

Methodology

Data Preprocessing

- Dropped irrelevent columns
- Filtered canceled flight (cancelled == 0)
- Create delay label using [‘arr_delay’] > 15 minutes

Feature Selection

- Input : month, day of week, scheduled departure hour, distance, airline, origin, destination
- Target : Binary all the delay (1 = Delayed , 0 = On-Time)

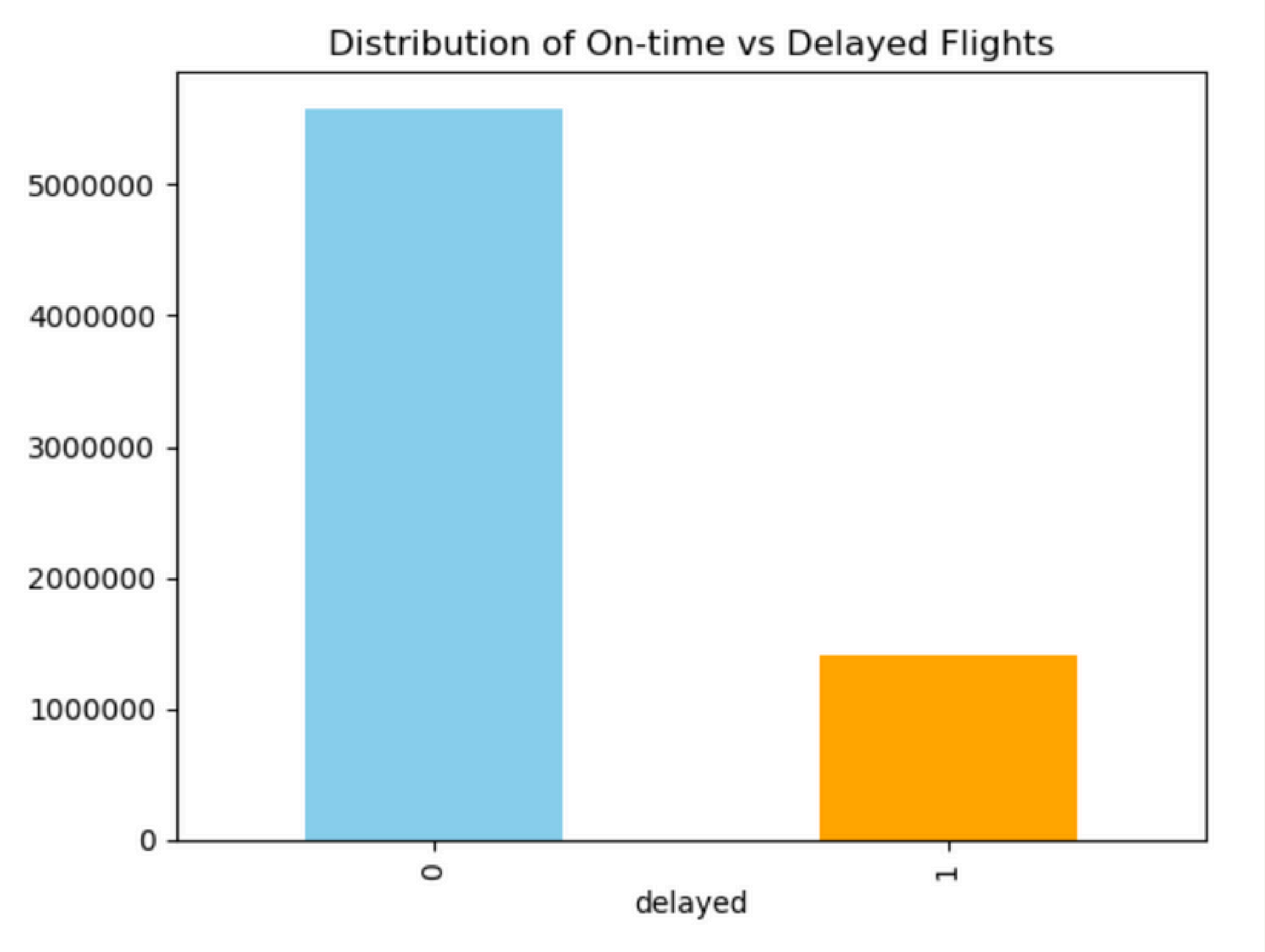
Modeling

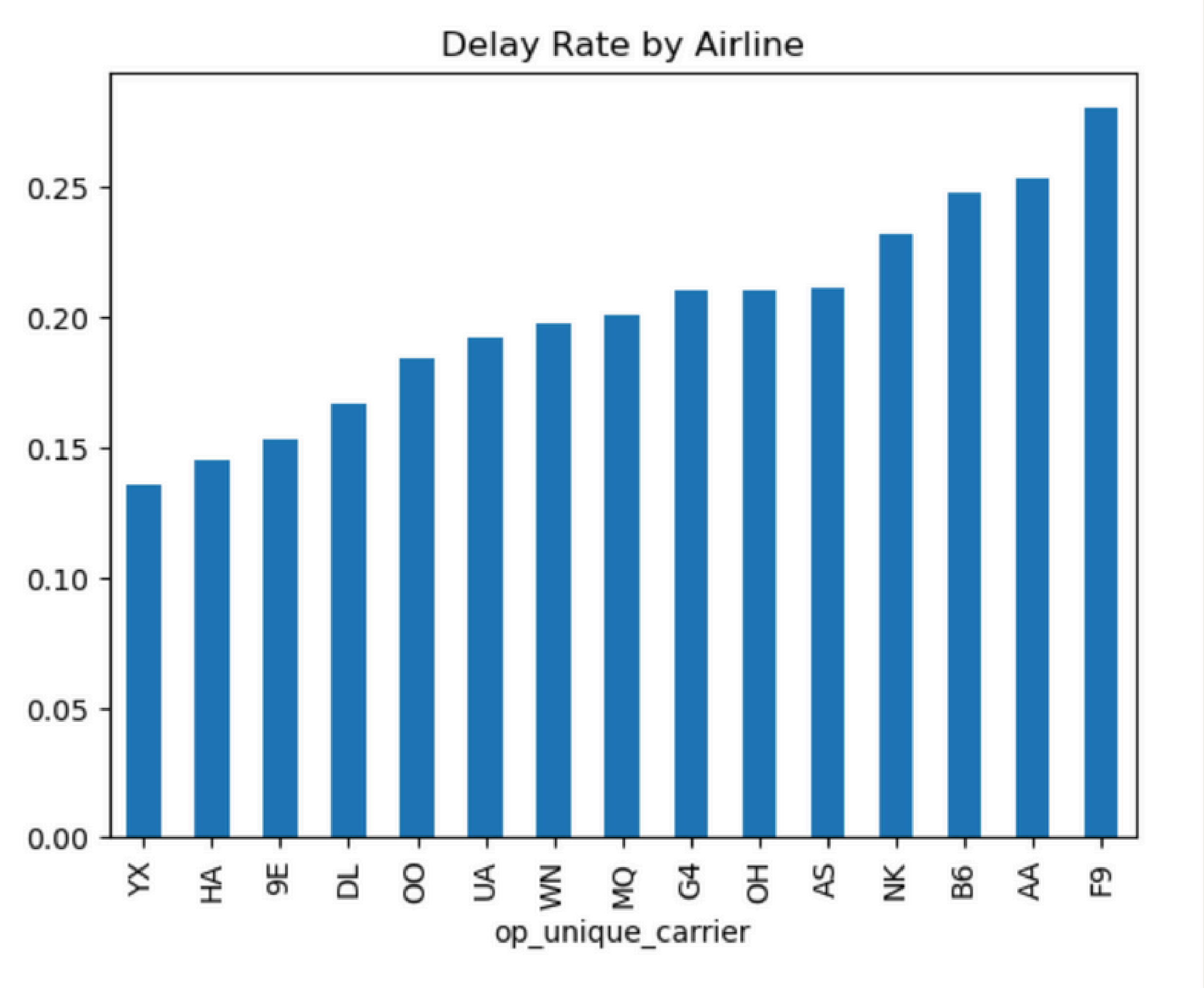
- Split dataset :
  80% train / 20% test
- Algorithm :
  Logistic Regression

Evaluation & Deployment

- Metrics : Accuracy , Confusion Matrix , F1 Score
- Deployment : Gradio App

Eda Key Finding





Evidence : Most flights are on time , fewer are delayed .

Interpretation : The dataset is unbalanced

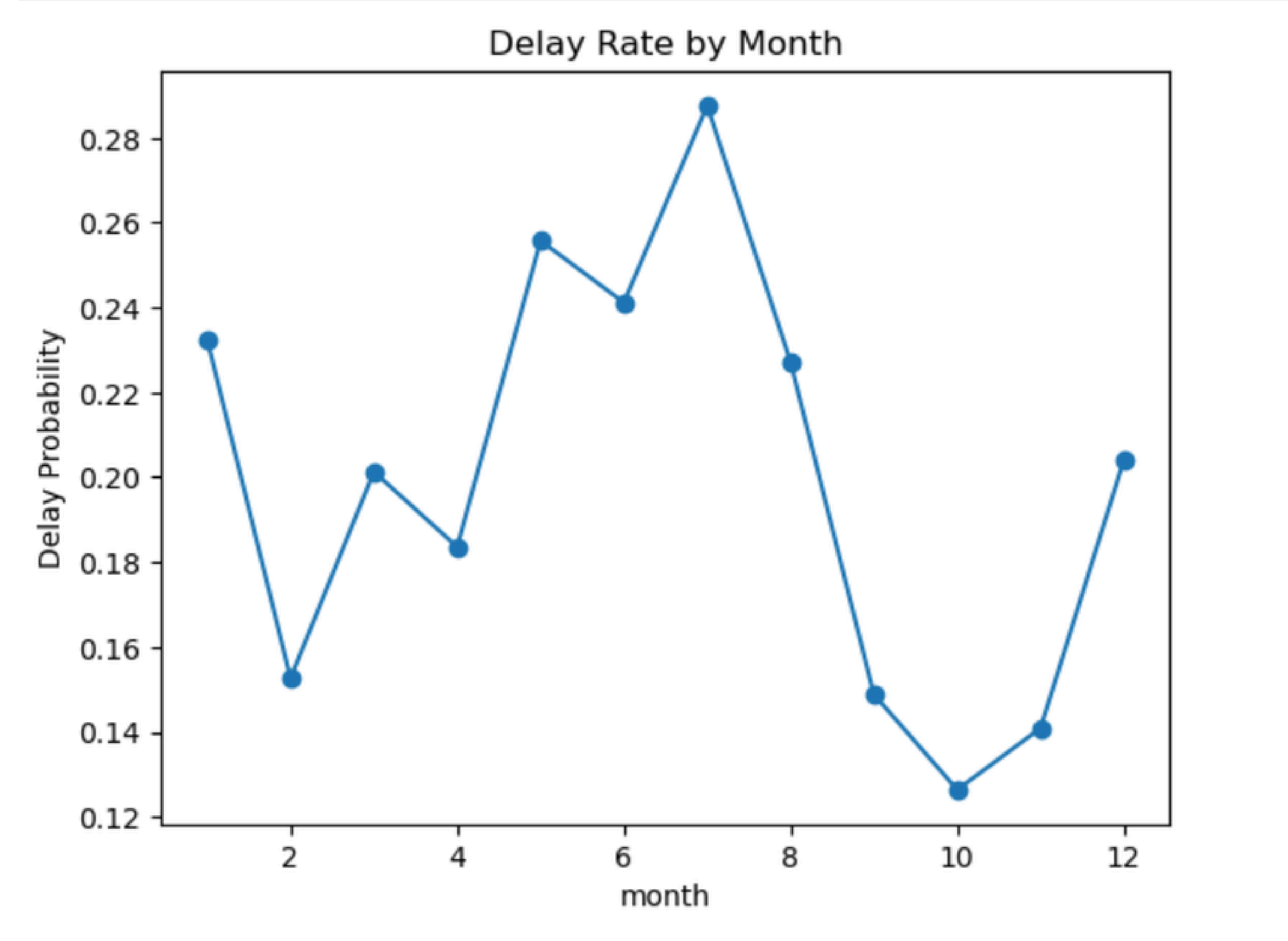
Action : Use F1-score and confusion matrix for model evaluation instead of accuracy alone.

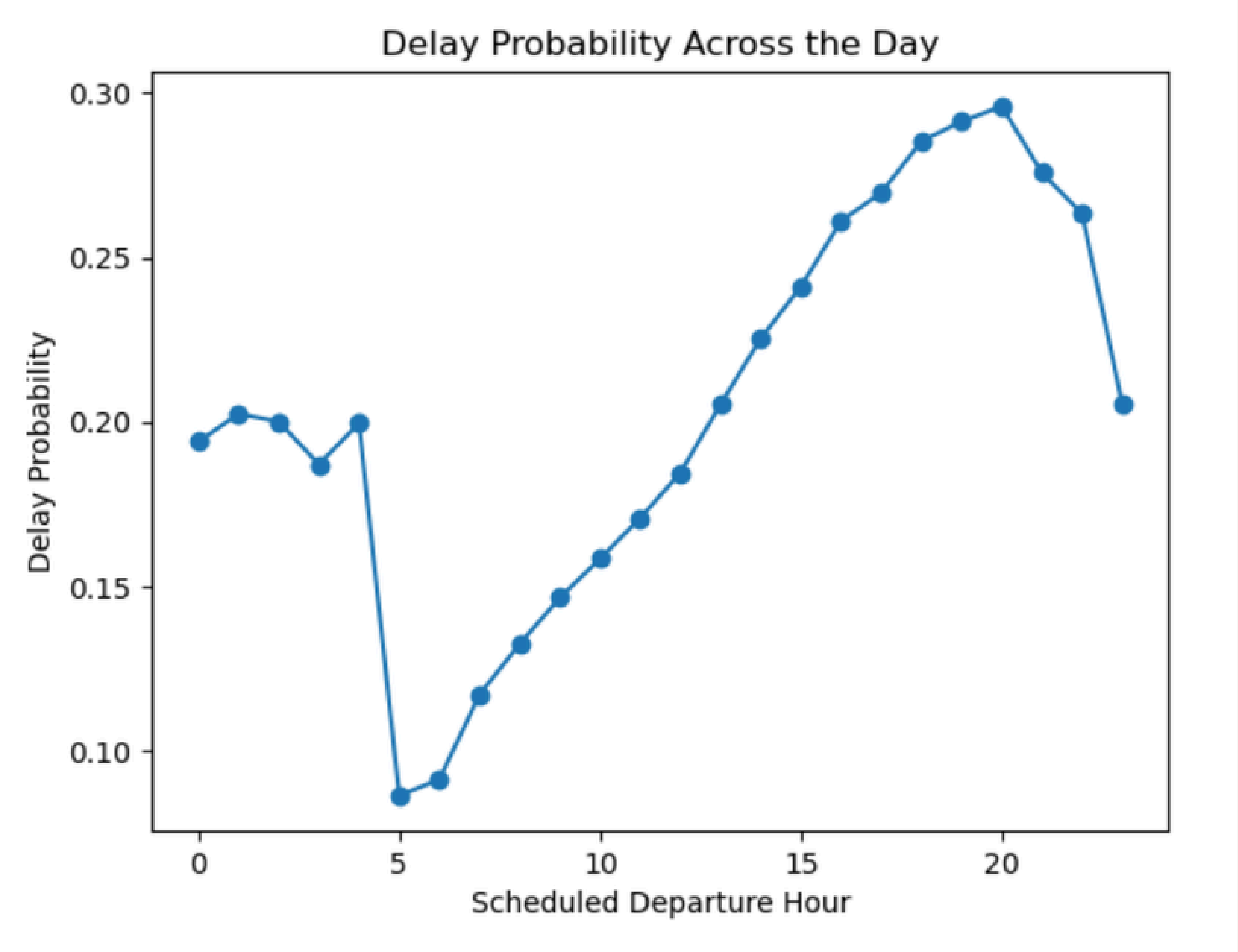
Evidence : Different airlines show different delay rates.

Interpretation : Airline choice affects the likelihood of flight delays.

Action : Include airline information as an feature in the model.

Eda Key Finding





Evidence : Delay rates change across different months, with some months having higher delays than others.

Interpretation : Seasonal factors affect the likelihood of flight delays.

Action : Include **month** as a feature in the model

Evidence : Flights departing later in the day show a higher probability of delay compared to early morning flights.

Interpretation : Flights departing later in the day show a higher probability of delay compared to early morning flights.

Action : Include SDH as a feature in the model

Modeling Approch

Algorithms :

Used Logistic Regression because it is suitable for binary classification problems such as predicting whether a flight is delayed or not.

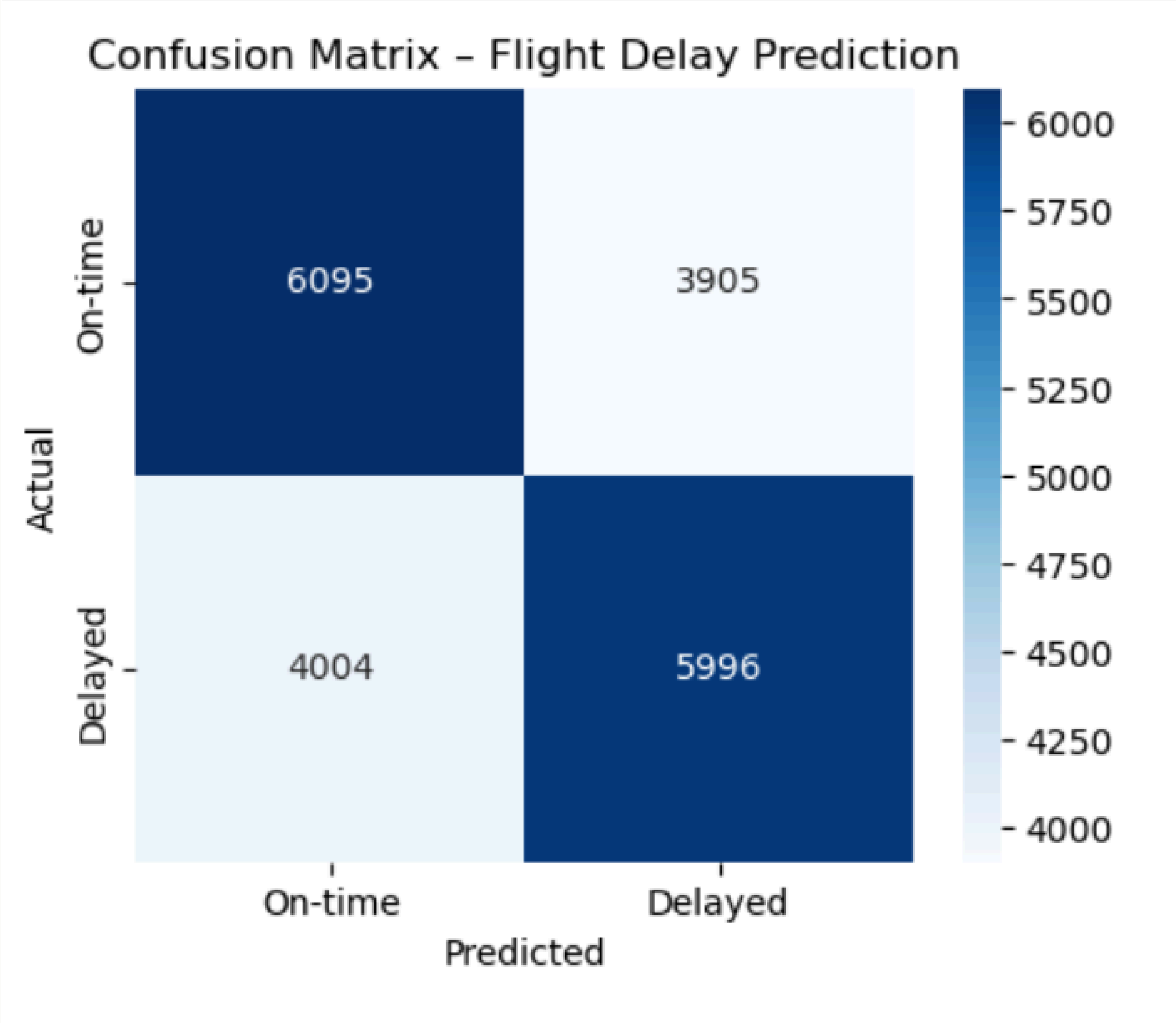
Validation :

Split data into 80% traning and 20% testing

Feature Engineering :

- Converted scheduled departure time (HHMM) into hour format
- Filled missing numerical values using the mean
- Categorical features were encoded using a pipeline
- Only pre-departure features were used to avoid data leakage

Result & Evaluation



Primary Metrics :

Accuracy : 60%

F1 Score : 0.60

“So What” ? :

- The model performs better than random guessing and can provide a useful early warning for possible delays.

Project Demo

Measure of Success

The model achieved an F1 Score of about 0.60, which means it can reasonably predict whether a flight will be delayed or on time.

- The model can identify flights likely to be delayed
- This helps airlines plan schedules better and reduce passenger waiting time

Challanges & Limitations

- **The dataset has fewer delayes flights compared to on-Time flights**
  - F1-score and confusion matrix were used instead of accuracy alone..

- **Large Dataset**
  - The original dataset was very large (7millions of rows).

Future Work & Recommendations

Add Weather Information

- Weather conditions such as rain, snow, or storms are major causes of delays and could improve model performance.

Tech Stack

Language : Python

Libraries : Pandas , Scikit-Learn , Joblib , Gradio , Matplotlib/Seaborn

Infastructure : Github , Gradio

Thank you!