



Flight Delay Prediction

Project 4 - KU Data Analytics Bootcamp

**Presented by: Peiwen Chiu, Levi Fahring,
and Jen Zapata**

Purpose



Building on the foundation of Project 1, which centered on analyzing historical flight delays, we enhanced our approach by integrating machine learning techniques to predict the probability of future delays.



This evolution allowed us to transition from simply understanding patterns in past delay data to proactively forecasting potential disruptions, helping airlines and passengers make more informed decisions.





Data Source

Flight Delay Data:

- A cleaned dataset that includes flight details such as carrier, airport name, arr_delay, and weather_delay.
- https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Weather Data:

- Weather-related data such as temperature, wind speed, and precipitation to help predict weather-related delays.
- <https://open-meteo.com/>

Data Description

Year: The year in which the flight data was recorded.

Month: The month in which the flight data was recorded.

Carrier: The airline code representing the carrier operating the flight. (Ex: “AA” for American Airlines)

Carrier_Name: The full name of the airline carrier. (Ex: “American Airlines”, “Delta”, “United Airlines”)

Airport: The airport code where the data was recorded (origin or destination airport). (Ex: “LAX” for Los Angeles International Airport).

Airport_Name: The full name of the airport.

Arr_Flights: The total number of arrival flights at the given airport for the specified time period.

Arr_Del15: The number of arrival flights that were delayed by 15 minutes or more.

Carrier_CT (Carrier Count): The total number of delays caused by the airline carrier itself.

Weather_CT (Weather Count): The total number of delays caused by adverse weather conditions.

NAS_CT (National Aviation System Count): The number of delays caused by air traffic control issues or airport capacity limitations.

Security_CT (Security Count): The number of delays caused by security-related incidents.





Data Description (Continued)

Late_Aircraft_CT (Late Aircraft Count): The number of delays caused by the late arrival of the aircraft from a previous flight.

Arr_Cancelled: The number of arrival flights that were canceled.

Arr_Diverted: The number of arrival flights that were diverted to a different airport due to unforeseen circumstances.

Arr_Delay: The total delay (in minutes) for all arrival flights during the specified time period.

Carrier_Delay: The total delay time (in minutes) caused by the airline carrier itself.

Weather_Delay: The total delay time (in minutes) caused by adverse weather conditions.

NAS_Delay: The total delay time (in minutes) caused by National Aviation System (NAS) issues, such as air traffic control or airport capacity problems.

Security_Delay: The total delay time (in minutes) caused by security-related incidents.

Late_Aircraft_Delay: The total delay time (in minutes) caused by the late arrival of the aircraft from a previous flight.

Problem Statement

Flight delays are a major inconvenience for passengers and airlines. The goal of this project is to:

1. To analyze historical flight and weather data.
2. To build machine learning models for predicting overall flight delays and weather-specific delays.
3. To create an interactive dashboard that visualizes delay predictions and key metrics.





Machine Learning Models

Model Objective:

- We implemented multiple machine learning models to predict flight delays based on historical flight and weather data.
- These models help identify patterns and make accurate predictions to improve operational planning.

Algorithms Used:

- **Random Forest Classifier:** Predict flight delays by capturing complex, non-linear relationships in the data.
- **Random Forest Regressor:** Like the classifier, it combines multiple decision trees to improve accuracy and prevent overfitting, making it suitable for regression tasks like predicting the exact delay duration.

Insights & Findings

Highlights the most important features that the models used to make predictions.

- **Random Forest** model provided the most reliable predictions for both overall delays and weather delays.
- **Weather Variables Improve Prediction Accuracy:** Adding temperature, wind speed, and precipitation as features improved the model accuracy significantly.
- **Late Aircraft Delays Have a Ripple Effect:** One of the most significant predictors of delays is the late arrival of the previous flight, which often causes a chain reaction of delays.





Features and Target Used in the Weather-Related Delay Model

Feature Name	Description
Origin	Airport code of the departure airport
Dest	Airport code of the destination airport
Is_Weekend	Whether the flight is on a weekend
Weather Delay History	Rolling average of weather delays caused by various weather conditions from Jan. 2019 to Sept 2024.
Delayed Departure	Whether the departure was delayed
Delayed Arrival	Whether the arrival was delayed
Total Weather Delays	Sum of weather delays and cancellations

Target Variables:

- **Delayed Departure**
 - Departure delays have a direct impact on the entire flight schedule and can cause a ripple effect.
 - Weather impact is more significant on Departure.

Model Performance

Weather-Related Random Forest Classifier

Classification Report				
	precision	recall	f1-score	support
No Delay (0)	0.9993	0.9908	0.9950	346083
Delay (1)	0.9139	0.9927	0.9517	33959
accuracy	0.9910			
macro avg	0.9566	0.9918	0.9734	380042
weighted avg	0.9916	0.9910	0.9912	380042

True Negative = 342,907

False Positive = 3,176

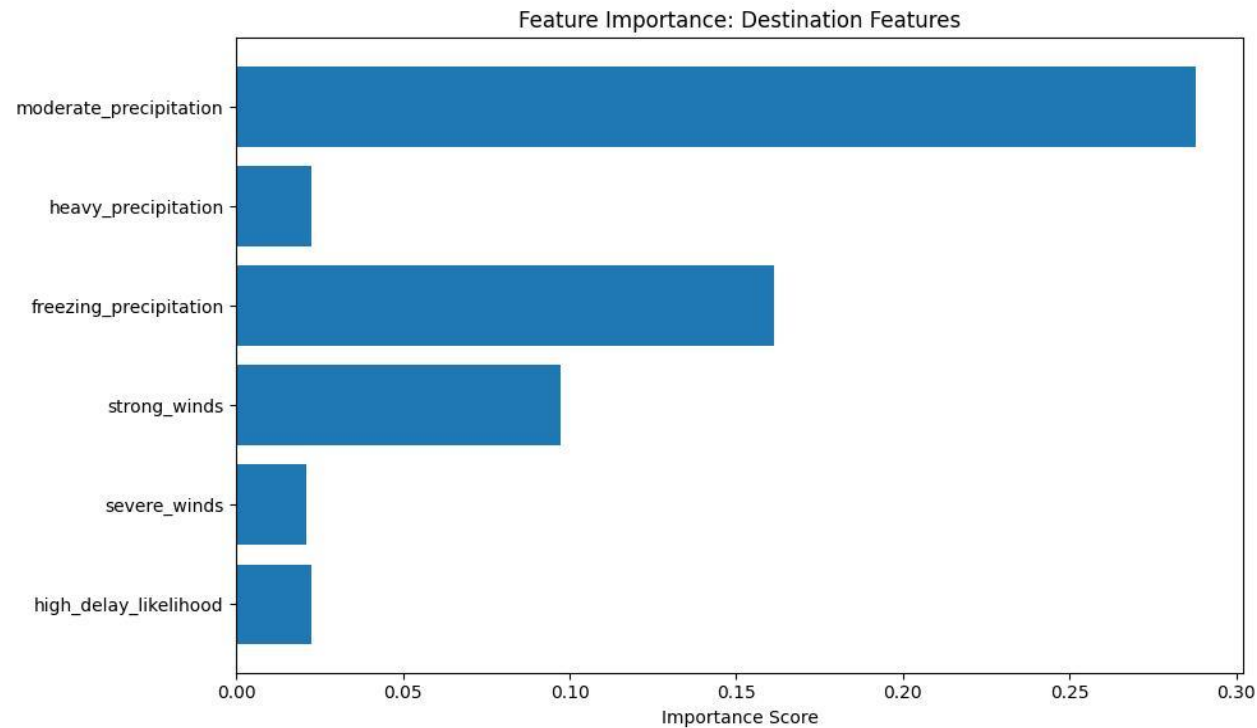
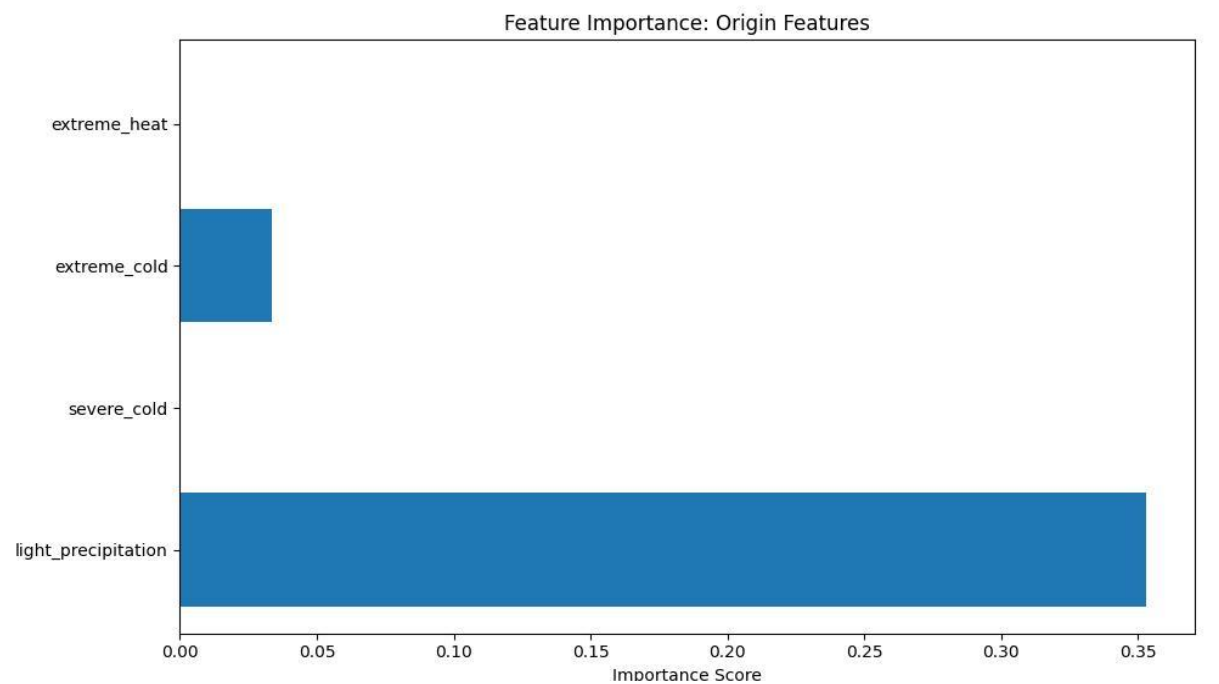
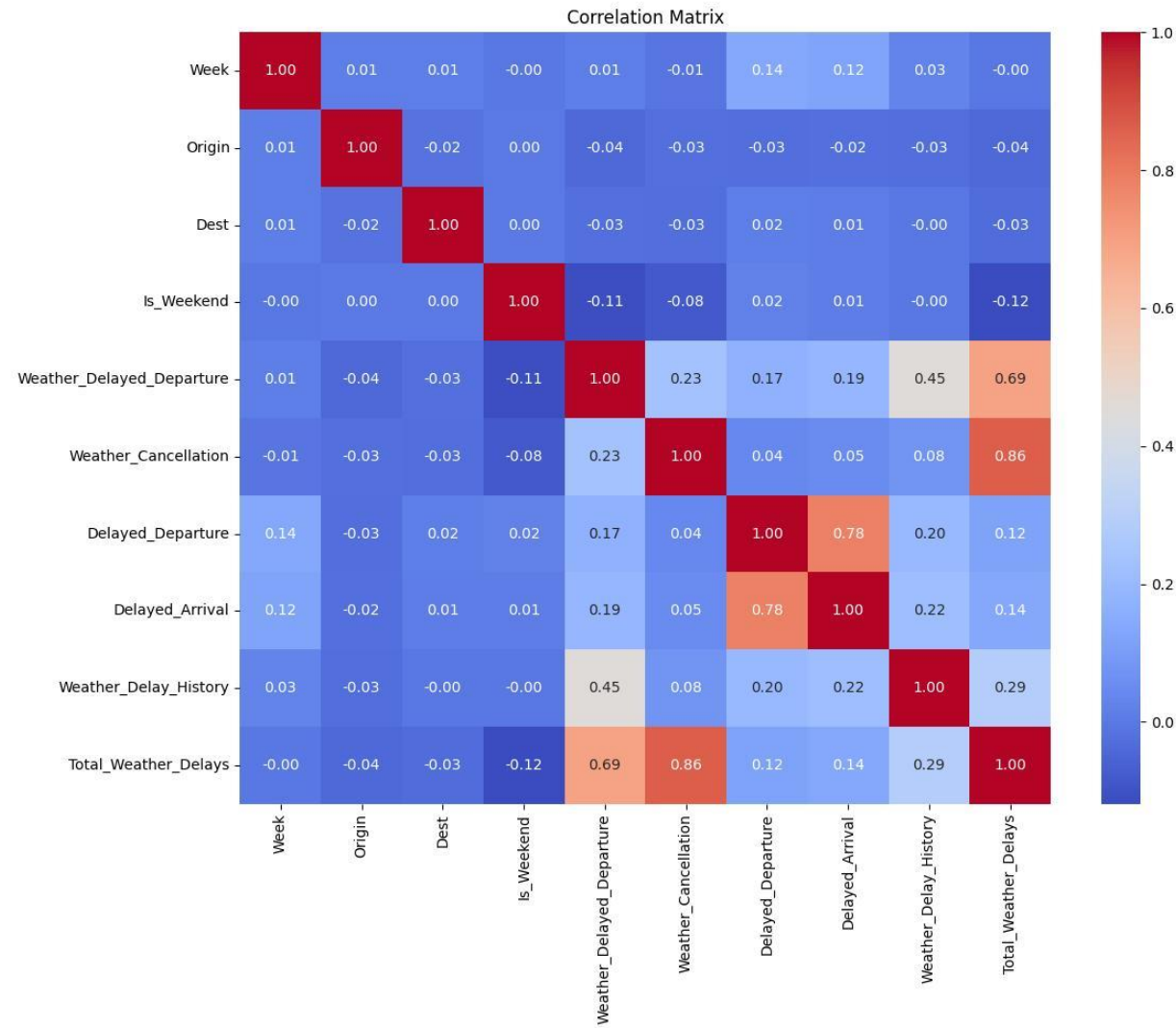
False Negative = 248

True Positive = 33,711

- High True Negative Count
- Relative Low False Positive and False Negative
- Although accuracy is high, but there are significantly more No Delay (0) than Delay (1). This could be indicating the data set is imbalance.
 - Looking at more years of data

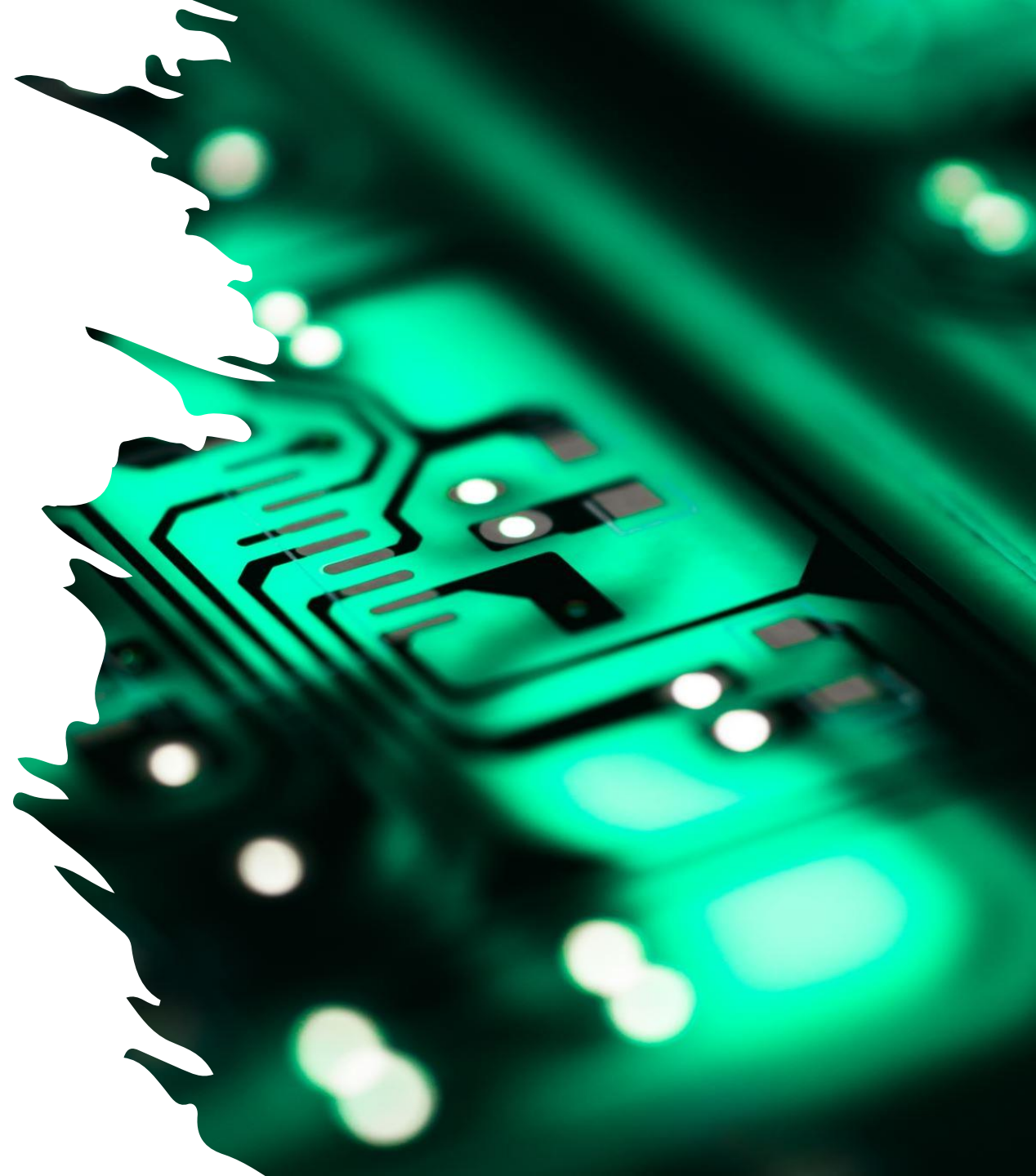


Findings



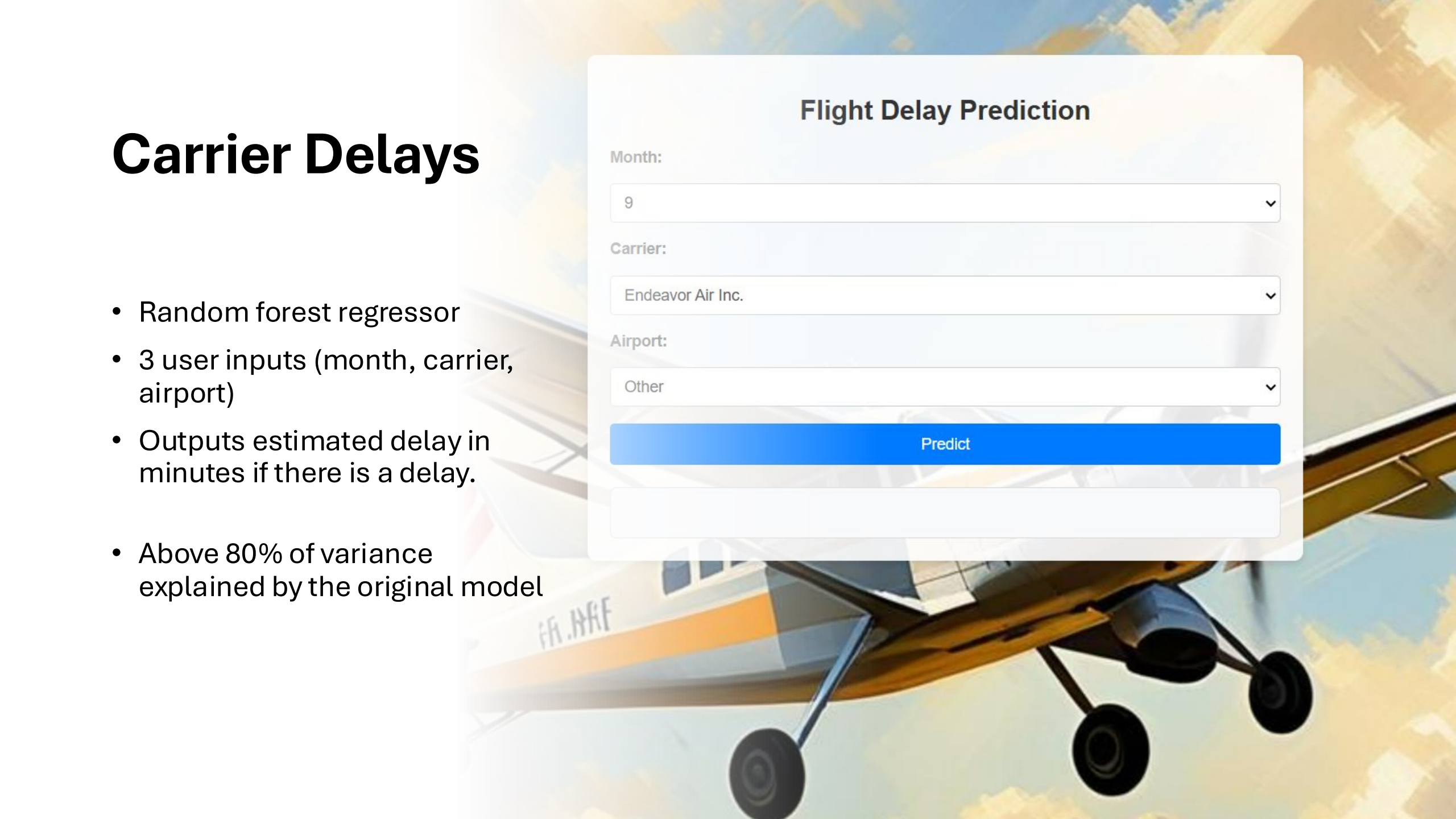
Arrival Delays

- Random forest regressor
- 3 user inputs (month, carrier, airport)
- Outputs estimated delay in minutes if there is a delay.
- Above 90% of variance explained by the original model



Carrier Delays

- Random forest regressor
- 3 user inputs (month, carrier, airport)
- Outputs estimated delay in minutes if there is a delay.
- Above 80% of variance explained by the original model



Flight Delay Prediction

Month:

9

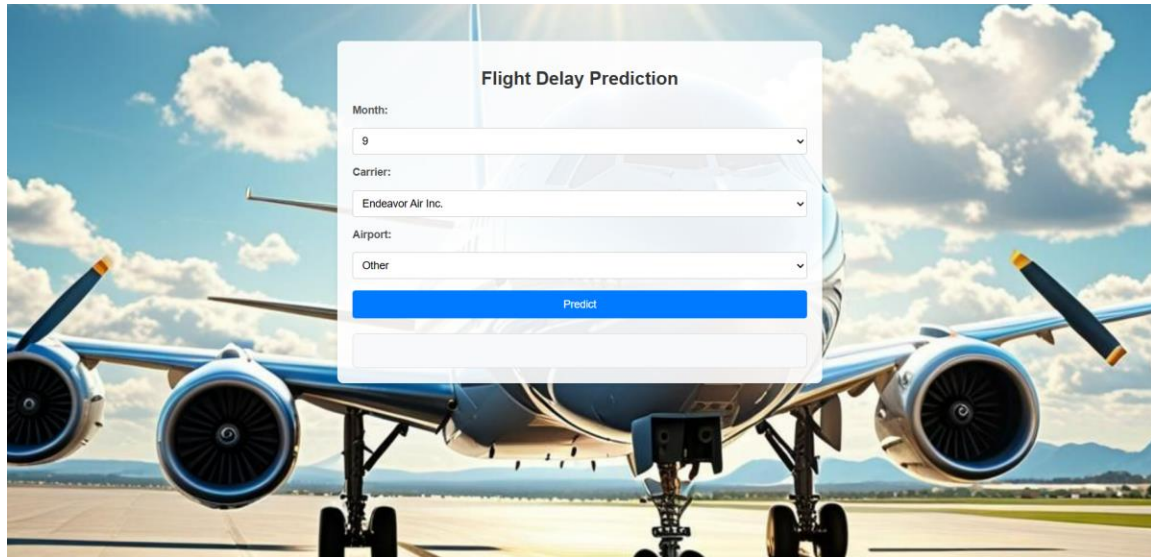
Carrier:

Endeavor Air Inc.

Airport:

Other

Predict



Weather Delays

- Random forest regressor
- 3 user inputs (month, carrier, airport)
- Outputs estimated delay in minutes if there is a delay.
- Above 70% of variance explained by the original model

THREE OVERVIEW MODELS ONE DATASET

- All three models were trained on the same data set
- Some have vastly better outcomes than others
- Some scored above 90, others below 70.





HOW THE MODEL WORKS

- First the data was cleaned (removing empty rows/outliers)
- Trained on random forest regression model.
- The model predictions were then saved to a csv
- A javascript/html/css script was used to create a user interface
- The interface used the previously mentioned prediction csv
- The output is the delay in minutes based on the users input.

Model Results

- A couple of issues
- Although, the original model r^2 was relatively good (over 90, over 80, and over 70) for the 3 models, the score is not fully representative of what goes on in the dash board.
- There are a number of important features that the user would not have knowledge about.
- To get around this, we aggregated the mean of these values into the model based on their respective months.
- This caused the 'month' input to be weighted too heavily, and the other two inputs to make no difference.
- So given that, the model is more of an educated guesstimate of expected delay than an actual prediction.





Conclusion

- The Flight Delay Prediction project builds on our initial Flight Delay Analysis by integrating machine learning to predict future delays.
- While the first project focused on identifying delay patterns from factors like weather, travel peaks, and airline performance, this project transitions to predictive modeling.
- By leveraging machine learning, we developed a tool that forecasts delay probabilities using historical data.
- The project demonstrates the power of combining data analytics with machine learning to solve complex challenges and paves the way for advancements in predictive capabilities across various domains.

Dashboard Demo

1. The Problem

1. Frequent traveler Ashley struggles with unpredictable flight delays that disrupt her plans.
2. Airline manager Marcus faces inefficiencies due to a lack of tools to predict and manage delays.

2. The Solution

1. The **Flight Delay Prediction Dashboard** uses historical flight and weather data combined with machine learning to predict delays.
2. It provides actionable insights for travelers like Ashley and managers like Marcus.

3. How It Works

1. Users can input details (e.g., travel month, carrier, airport) to see delay predictions.
2. Visualizations like graphs help identify patterns and improve decision-making.

4. The Impact

1. Travelers plan proactively with delay forecasts.
2. Airlines optimize resources and schedules.
3. Authorities improve infrastructure and response strategies.

This project transforms flight delays from a frustration into an opportunity for smarter planning, empowering everyone in the aviation ecosystem.





Q&A/ DISCUSSION