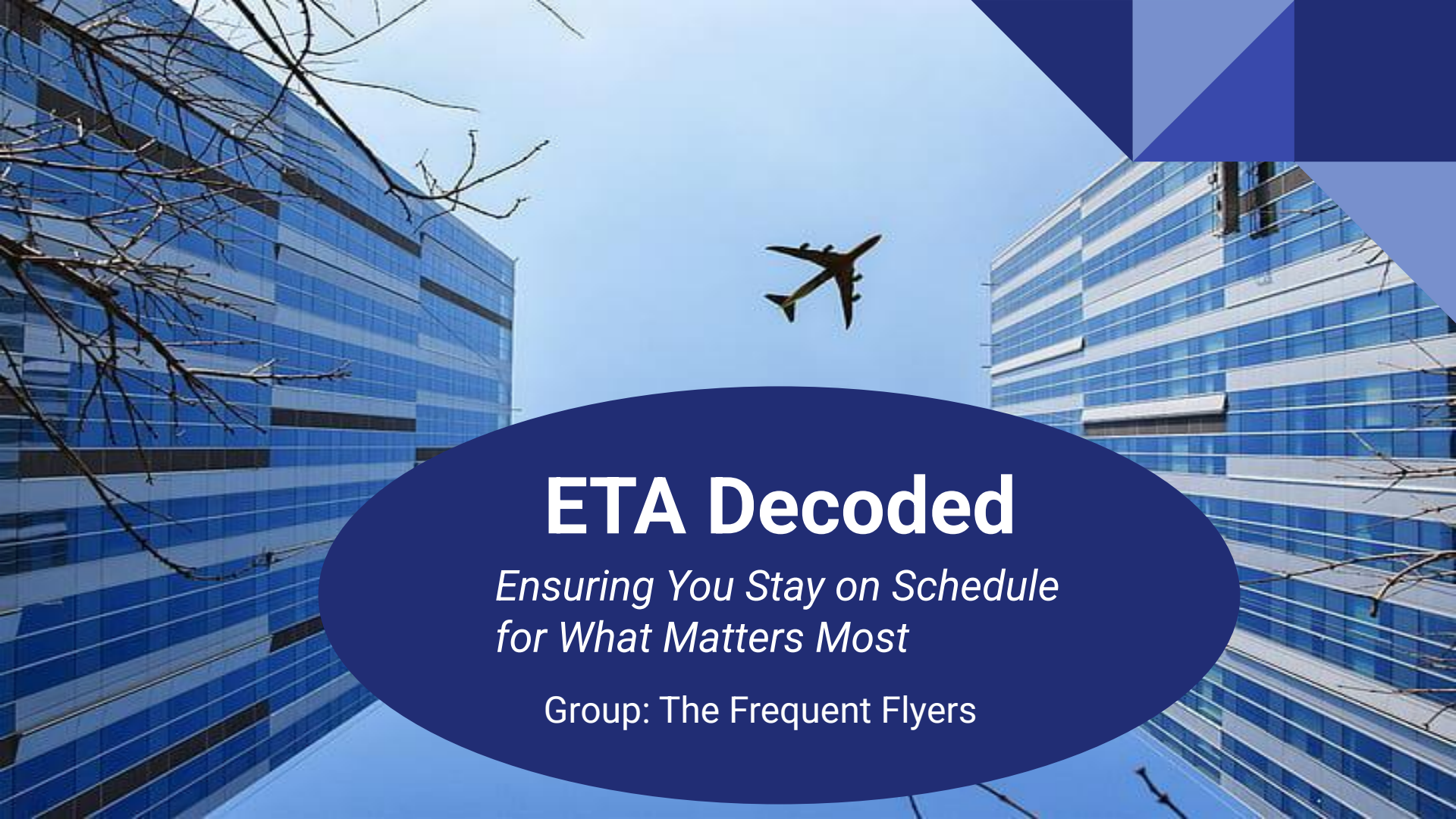


ETA Decoded

Ensuring You Stay on Schedule for What Matters Most

Group: The Frequent Flyers

The Team

Data Cleaning, EDA, Relevant Features
Lei

Baseline Model 1 & Fine Tuning	Feature Engineering/ Baseline Model 2
Linda	Audrey

Linear Regression	Logistic Regression	Decision Tree	Random Forest
Elsie	Naima	Diana	Maya

OUTLINE

1. Project Introduction
2. The Dataset
3. Our Methods
4. Results

Introduction

Research Questions

How can we predict flight arrival delays accurately using available flight and operational data?

Motivation

- Improving Airline Operations
- Enhancing the passenger experience

Stakeholders

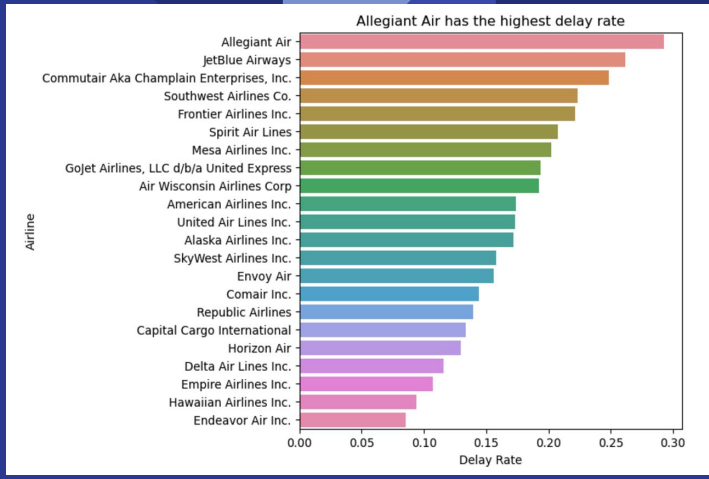
- Airlines
- Passengers
- Airport Employees

The Dataset

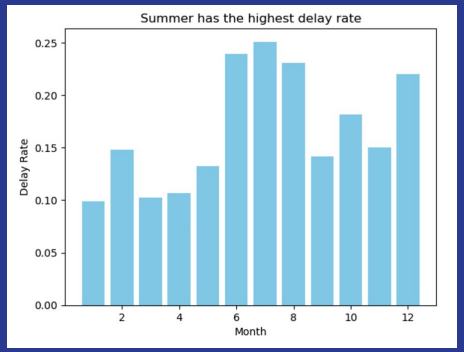
Flight Status Prediction

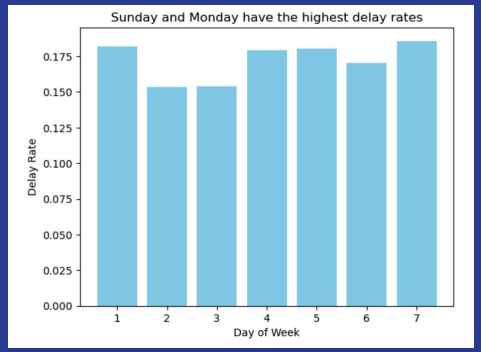
Can you predict which flights will be delayed or cancelled in 5 years of data?

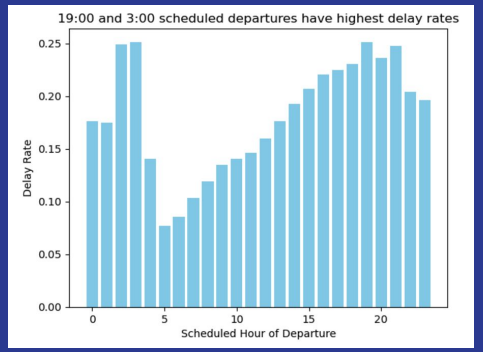
kaggle

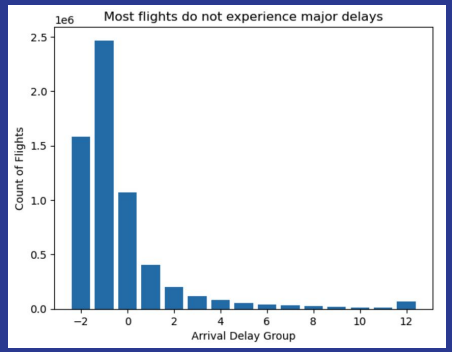


- **Dataset:** Combined_Flights_2021.csv
- **Original Shape:** 6,311,871 instances, 61 columns
- **Class Imbalance:** ~82.7% of instances belong to the negative class









PRE-PROCESSING

Relevant Feature Selection	Data Cleaning	EDA	Feature Engineering
Selecting relevant or potentially important features for our modeling objectives	Handle null values, data types, invalid entries, and duplicated records	Examine summary statistics, distributions, outliers, correlations, unique values, and create visualizations	Creating new variables to improve predictive power and ensure better insights

Relevant Feature Selection

From the 61 columns, we chose 21 columns as relevant or potentially important features for our modeling objectives, categorized into 7 groups.

We plan to choose at most 1 feature from each group to avoid redundancy and multicollinearity.

The 7 feature groups include:

1. **Flight Date:** *When is the scheduled flight date?*
2. **Flight Time:** *When is the scheduled departure time or arrival time?*
3. **Airline:** *Which airline will operate the flight?*
4. **Flight Number & Aircraft Number:** *What's the flight number & which aircraft was used?*
5. **Origin Location:** *Where is the flight planned to take off from?*
6. **Destination Location:** *Where is the flight planned to land?*
7. **Distance:** *What's the distance between planned origin and destination airports?*

Target Variable Selection

Target variables for each type of prediction:

- Continuous Prediction: `ArrDelayMinutes`
- Binary Prediction: `ArrDel15`
- Categorical Prediction: `ArrivalDelayGroups`

Feature Engineering

- Started with **21** columns after initial selection.
- Expanded to **57** columns after tuning and feature engineering.

Engineered variables to capture relationships:

Weekend & Holiday Indicators: Flagged travel patterns for non-working days.

Seasonal Labels: Highlighted weather impacts on delays.

Airport Capacity: Quantified origin/destination operational load.

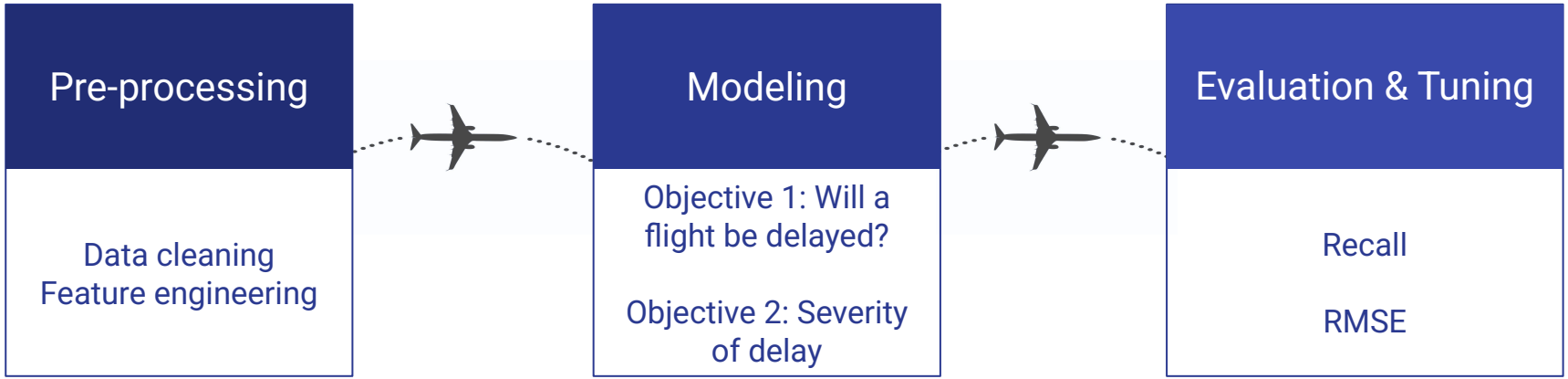
Delay Trends by Aircraft: Captured per-aircraft delay histories.

Balanced dimensionality using one-hot encoding and scaled numeric features to improve model performance.

*** Observed Potential Data leakages

Methods

Methods Overview



Model 1

Predict whether or not a flight arrival will be delayed

Target Variable: ArrDelay15

Approach:

- Logistic Regression
- Decision Trees
- Random Forests

Metrics: Recall, Precision, Accuracy, F-1, AUC

Evaluation & Tuning: Threshold optimization, hyperparameter tuning, class adjustments, cross-validation, confusion matrix, classification report

Model 2

Predict how much (in minutes) the flight will be delayed

Target Variable: ArrDelayMinutes

Approach:
Built a regression model to predict how long in minutes will a flight be delayed to arrival at destination.

Metrics: RMSE, MAE, R-squared

Evaluation & Tuning: Feature selection, hyperparameter tuning, ridge/lasso regression, cross-validation

Model 1 Performance and Metrics

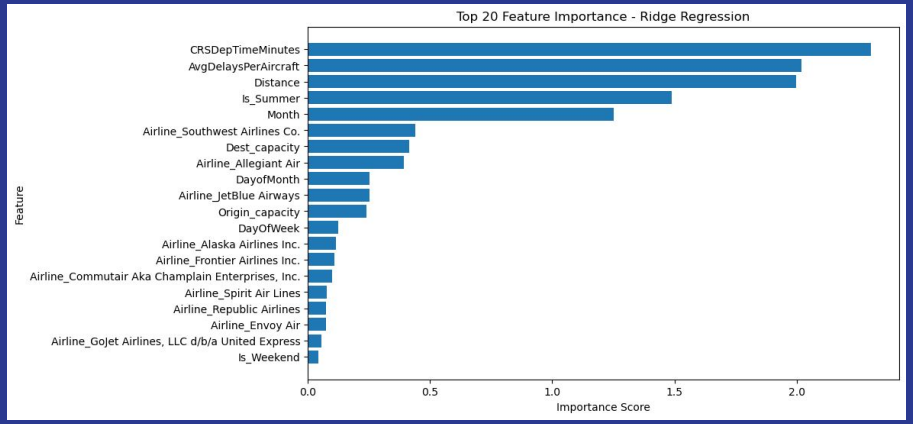
Model	Accuracy	Recall*	Precision	F-1
Logistic Regression	62.54	62.67	25.86	0.3662
Decision Tree	64.43	63.50	27.22	0.3811
Random Forest	69.58	79.10	33.75	0.4731

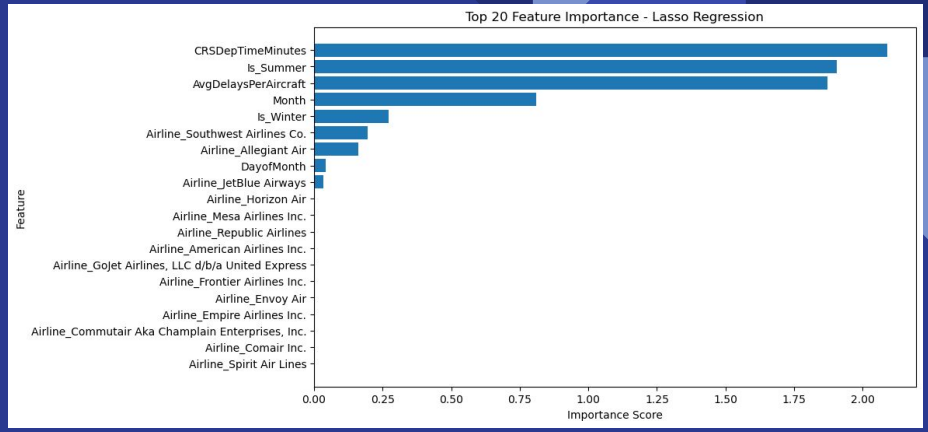
Model 2 Performance and Metrics

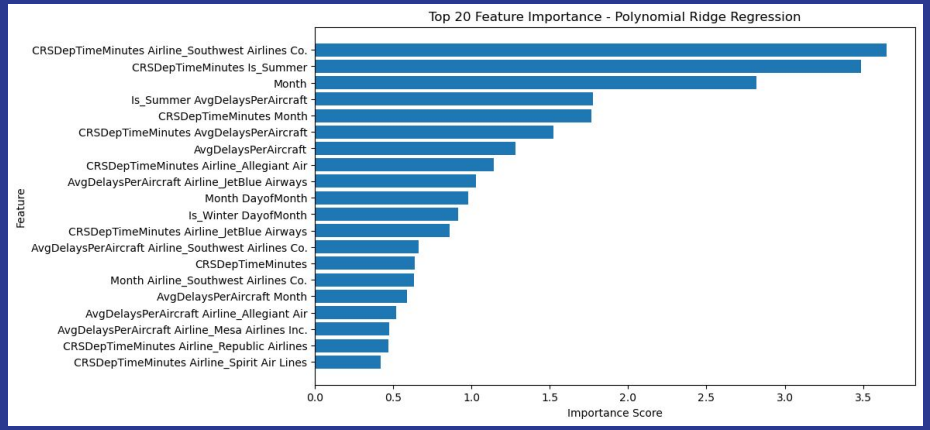
Metric	Ridge Regression	Lasso Regression	Polynomial Ridge
RMSE	17.03	17.02	16.87
MAE	11.27	11.35	11.1
R^2	0.06	0.06	0.07

- The Polynomial Ridge model performed best among the evaluated approaches, demonstrating slight improvements in predictive accuracy.
- The models explained only 6%–7% of the variance in arrival delays, indicating that a linear model does not capture

Model 2 Feature Importance







Real-World Implications

- Average of **over 100,000** commercial flights operating per day globally
- Average of **30,000** delayed flights per day globally
- Benefit of accurate predictions:
  - True positives/True negatives
- Cost of inaccurate predictions:
  - False positives/False negatives
- Ready for the real-world?
  - **Real-time/live data:** weather, maintenance issues, air traffic, global events, etc.
  - **Adaptive modeling:** responsive to real-time changes

Lessons Learned

- High correlation between features
- Leakages
- Very large dataset