Predict Domestic (US and US territories) Flight Delay

Project Phase III Update

Team Members: Grace Lee, Shivangi Pandey, Sybil Santos-Burgan, Beijing Wu

W261 Summer 2022 Section 4
FP_Section4_Group1

Berkeley
UNIVERSITY OF CALIFORNIA

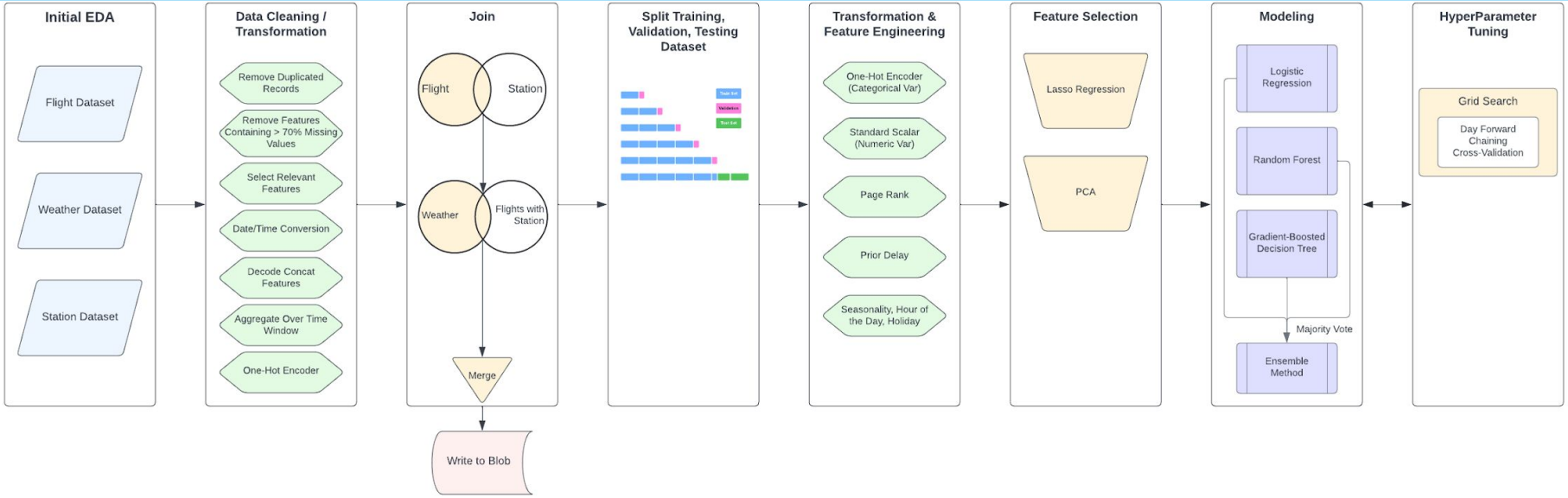
Goal

Delays are annoying.

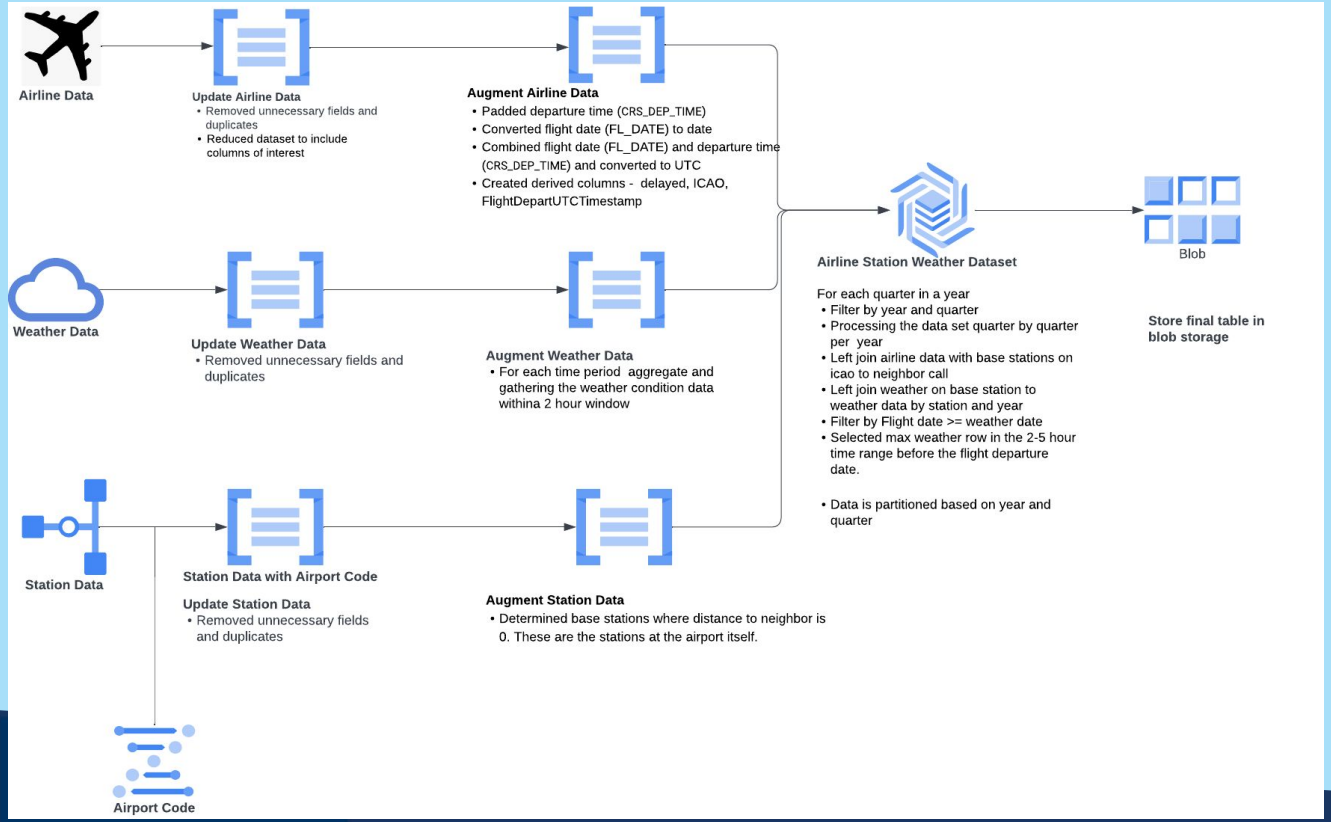
We believe predicting delays will help airports with planning. We will help them by predicting delays 2 hours before scheduled CRS departure time.

- **Null hypothesis**: The addition of prior flight information to current flight and weather data has no impact on predicting departure delays
- **Alternate hypothesis**: Adding information about tracking prior flights improves prediction of departure delays.

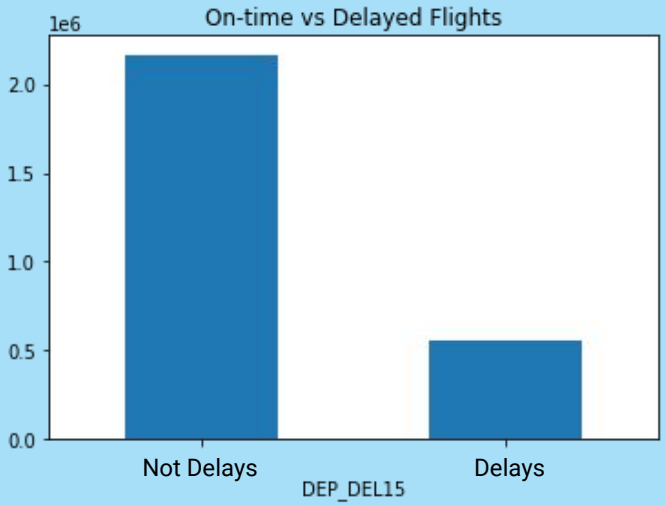
Model Pipeline



Model Pipeline



Airline EDA

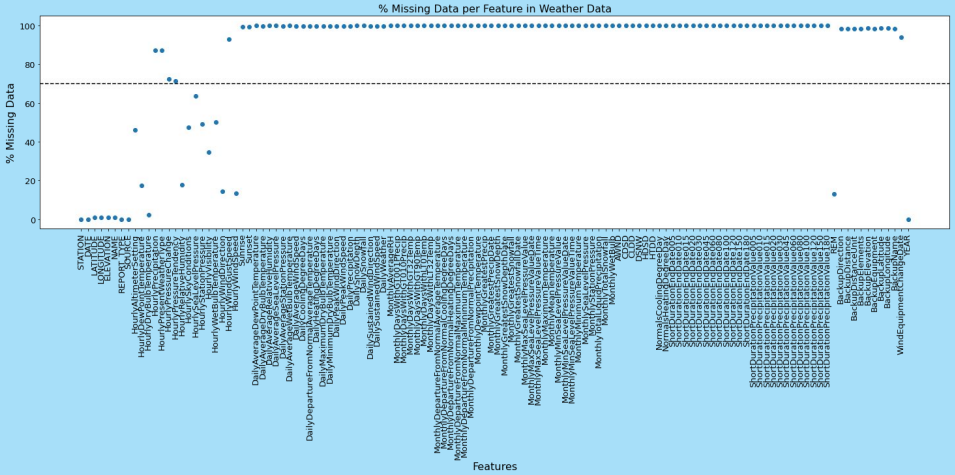


EDA on Missing Data:

Feature	Notes about Missing	Accounting for Missing
DEP_DELAY15	All missing values are due to cancelled flights	Consider cancels as extended delay: recode nulls as 1
ARR_DELAY	missing when cancelled or diverted	* We need this column for information about prior flight delay, currently we are undecided on whether to remove missing or fill
AIR_TIME	Null for diverted and cancelled flights	won't use this column because highly correlated with distance
ACTUAL_ELAPSED_TIME	Null for diverted and cancelled flights	won't use this column because highly correlated with distance
FIRST_DEP_TIME	mostly missing	won't use this column
TOTAL_ADD_GTIME	mostly missing	won't use this column
LONGEST_ADD_GTIME	mostly missing	won't use this column
DIV_REACHED_DEST	mostly missing	won't use this column
DIV_ACTUAL_ELAPSED_TIME	mostly missing	won't use this column
DIV_DISTANCE	mostly missing	won't use this column
TAIL_NUM	Only missing when flights cancelled	* We need this column for information about prior flight delay, currently we are undecided on whether to remove missing or fill
CARRIER_DELAY	this column, as well as weather, nas, security, late aircraft delays are missing	won't use this column

Weather EDA

Step 1: Removed columns with > 70% missing data



124 → 21 Features

Step 2: Select features that are relevant to flight delay

According to the Federal Aviation Administration, ***inclement weather****, including* ***thunderstorms****,* ***snowstorms****,* ***wind shear****,* ***icing*** *and* ***fog****, creates potentially hazardous conditions in the nation's airspace system. These conditions are, by far, the largest cause of flight delays.*

- ❑ Hourly Dew Point Temp
- ❑ Hourly Sky Conditions
- ❑ Hourly Visibility
- ❑ Hourly Wind Speed

21 → 10 Features

Berkeley
UNIVERSITY OF CALIFORNIA

Initial Features Selected

Feature	Description	Data Type
DEP_DEL15	Indicator of departure delay 15 minutes or more. (1 = Yes)	float64
YEAR, QUARTER, MONTH, DAY_OF_WEEK	Year, Quarter, Month, Day of the Week	int32
FL_DATE	Flight Date yyyy-mm-dd	string
OP_CARRIER	Airline	string
ORIGIN, DEST	Origin Airport, Destination Airport	string
CRS_DEP_TIME	Computer Reservation System scheduled departure time	int32
ARR_DELAY	Difference between scheduled and actual arrival time (in minutes)	float64
DIVERTED	Diverted Flight Indicator (1 = Yes)	float64
TAIL_NUM	Plane tail numbers	string
DISTANCE	Distance between airports (miles)	float64
CANCELLED	Cancelled Flight Indicator (1 = Yes)	float64
STATION	Weather station code	string
DATE	Weather reading date yyyy-mm-dd hh:mm:ss	string
NAME	Name of weather station	string
REPORT_TYPE	Code that denotes the type of geophysical surface observation	string
OvercastIndex	Decode from HourlySkyConditions; overcast weather indicator (1 = Yes)	float64
WindSpeedAvg	Wind speed (m/s) avg'ed over 2-hr window prior to the time indicated	float64
DewPointTempAvg	Dew point temp (°C) avg'ed over 2-hr window prior to the time indicated	float64
VisibilityAvg	Visibility (m) avg'ed over 2-hr window prior to the time indicated	float64

Observations

> Duplicates , Null values

> Airlines dataset => IATA codes (in origin) , datetime in local time zone

> Weather dataset does not have ICAO code and datetime in UTC time zone

> Weather station records information at random frequency of datetime

> Not all IATA codes have ICAO code

> Base station contains correlated data of stations and its neighboring station information

> Base station => ICAO code

> Base stations with zero distance to neighboring stations represents stations at airport

> No direct joining keys available between airline and weather

Data Transformation (For Join)

AIRLINE DATASET

> Filter Duplicates , NaNs & Nulls based on EDA criteria

> UTC Time Conversion

- Convert flight departure datetime (local timestamp) to UTC
- Convert UTC timestamp into unix timestamp
- Preserve datetime before 2 hrs of departure

> Add ICAO code

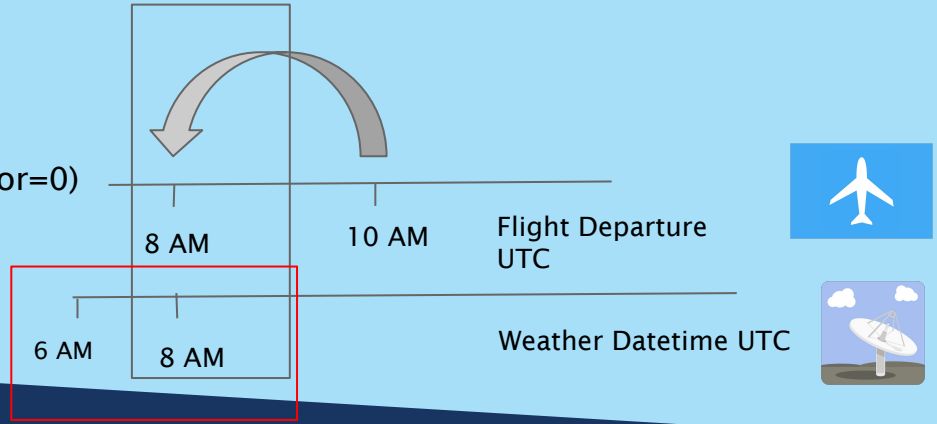
Python Libraries used:

Airporttime : for IATA-> ICAO code conversion & local to UTC timestamp convertor

BASE STATIONS

> Filter stations at airport (Criteria: distance to neighbor=0)

> Get station ids per ICAO code



Berkeley
UNIVERSITY OF CALIFORNIA

Data Transformation (For Join)

WEATHER DATASET

> Convert datetimes to rounded nearest hour

> Aggregate features for last two hour window

JOIN CRITERIA :

LEFT JOIN on Airline and Base Stations
on ICAO code
Further, LEFT JOIN on Weather
on two hour prior departure timestamp with rounded nearest hour weather AND
on station ids from base station and weather

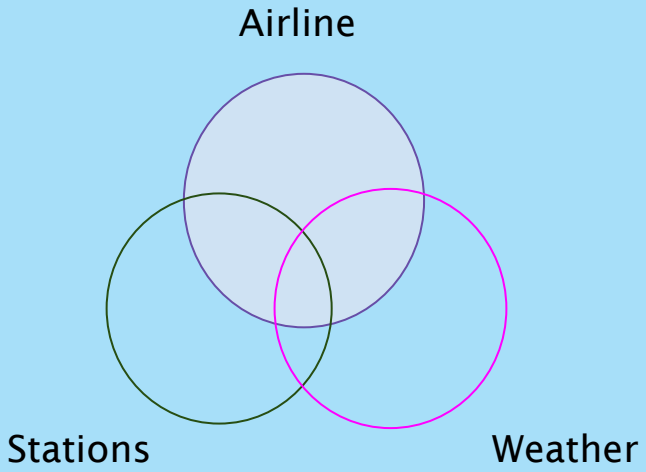
Lesson Learnt :
> Persist data in Blob storage for intermittent transformed dataset
> Read back the dataset from blob before join
> Slowness reasons : cyclic memory overwrites, garbage collection
> Avoid conditional join (ex >= ,< etc
> Check execution plan
> Cluster resource occupied with other executions

Join Statistics

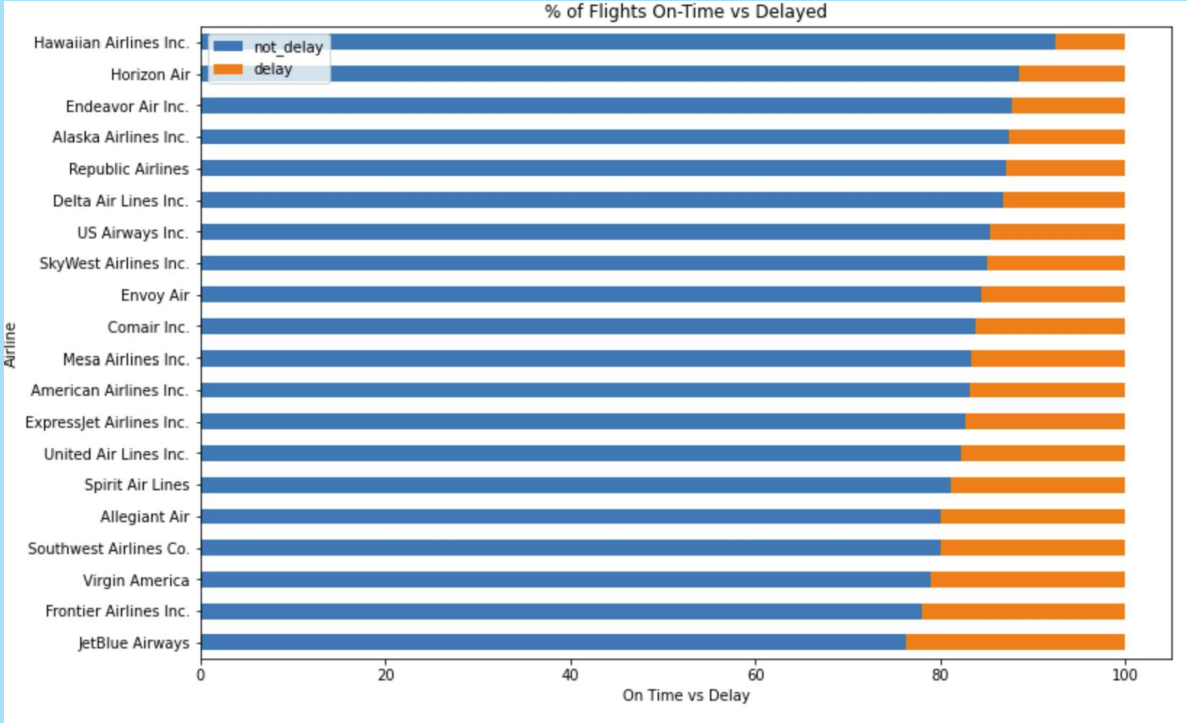
Original Dataset (Compressed PARQUET Files):
Airline : Record Size= ~ 74 million
Weather : Record Size = ~898 million
Base Station : Record Size =~ 5 million

Transformed Dataset (Compressed PARQUET Files):
Airline : Record Size= ~ 42 million
Weather : Record Size = ~ 122 million
Base Station : Record Size = ~ 2k
(ran on 28 cores, 112 GB memory capacity)

Join Time: 5.62 minutes

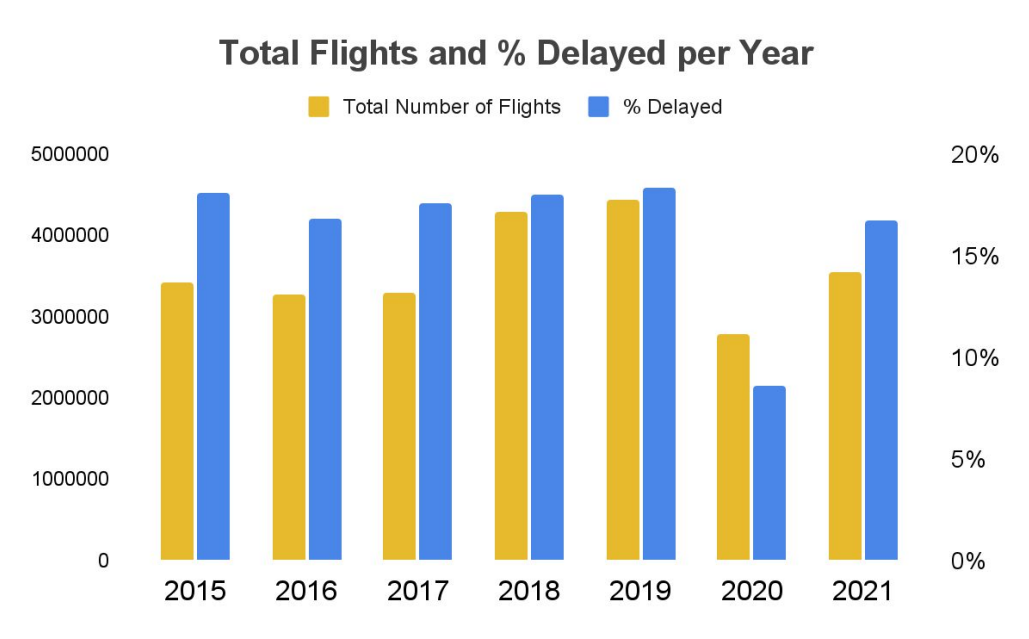


EDA on Joined Dataset



EDA on Joined Dataset - % Delayed

COVID vs PRE-COVID



Train/Test Split

Split

- Split Train/Test by Year
- Train = 2015 - 2020
- Test = 2021

Imbalanced Data

- Undersampling
  - Only on the training data after the split
  - Random sample the “not delays” to match similar number of records with the “delayed”

Evaluation Metrics

- **F2 Score**
  - Weighted harmonic mean of precision and recall
  - False positives are considered better than False negatives

$$F_2Score = 5(\frac{Precision * Recall}{4Precision + Recall})$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- **MCC (Matthews Correlation Coefficient)**
  - Represents confusion matrix as a single number
  - Takes into account the proportion of each class → **Helps with imbalanced data**
  - Range: [-1, 1] (-1 = opposite prediction, 0 = random guess, 1 = perfect)

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **Balanced Accuracy**
  - Takes into account of different class sizes

$$Balanced\ Accuracy = \frac{sensitivity + specificity}{2}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

- **AUC (Area Under the ROC Curve)**
  - Evaluates how well the model can distinguish between TP and FP
  - AUC = 1: perfect model
  - AUC = 0.5: random guessing

Baseline Model & Results

Logistic Regression

- Carrier
- Hour
- Month
- Visibility Avg

	F2	MCC	Balanced Accuracy	AUC
Train	0.669	0.214	0.606	0.648
Test	0.492	0.161	0.607	0.657

Next Steps

Planned Feature Engineering:

- Normalization
- Track previous flight delay
  - track tail numbers and arrival delay
- Holidays
- Page rank to understand centrality of an airport within the flight network
  - “Busy Airport” variable
- Type of airline
  - Major vs. regional, ultra low cost vs. low cost

Features Selection:

- Lasso regression to minimize cost function

Parameter Tuning:

- Forward Chaining Cross Validation