Predicting Departure Delay with Weather and Flight Data

Spring 2022

Group 6 - ***Zephyr***

The Team

Gauri Ganjoo

Gauri previously worked at tech startup focused on predicting athlete's success and automating data retrieval and cleaning.

Matt Lyons

Matt works in clinical quality improvement. He studies both clinical outcomes and customer satisfaction/experience.

Rumi Nakagawa

Rumi has experience in consulting/ in-house market researcher to support CXOs. Recent years shifting her career to data science.

Anand Patel

Anand works as a VR software engineer, and has experience in research & airplane development. Shifting his career to data science.

Outline

Project Overview

Background, Problem Statement, Objective, Client, Project Scope, Metrics, Project Workflow, Timeline

Data Preprocessing

Dataset, Exploratory Data Analysis, Join Multiple Datasets, Data Pipeline, Data Cleaning

Feature Engineering

Feature Selection, Creating New Features, Dimensionality Reduction

Modeling

Overview of Baseline Model, Features, Secondary-Third Model, Hyper Parameter Tuning

Result

Result, Conclusion, Next Steps

1. Project Overview

Background

30% of flights delayed by more than 15 minutes in the United States

Federal Aviation Administration estimated the annual cost of flight delays to be $26.6 billion.

Flight Delay Hits Customer Satisfaction of Airline Companies

Components of Customer Satisfaction in Airline Industry(Source: WSJ)

WSJ 2019 Airline Scorecard: Rankings of major carriers in key operational areas, best to worst

Companies that shine in customer satisfaction are also successful in business

Rankings of Customer Satisfaction and Business Performance

	Customer Satisfaction	Revenue Available / Seat*Mile	Operating Income 2019	Passengers 2019
Delta	1	15.4 cent	$6.6 B	204 M (+6%)
United	8	13.9 cent	$4.3 B	162 M (+2.6%)
American	9	14.7 cent	$3.1 B	215 M (+5.6%)

souce: https://simpleflying.com/2019-us-airlines-performance/
https://www.cnbc.com/2020/01/16/best-and-worst-airlines-for-2019.html
https://graphics.wsj.com/dynamic-inset-iframer/?url=https://asset.wsj.net/wsjnewsgraphics/data-tables/3c928058-b857-43d3-87db-6c7bd82ecc1e.json

Our Client

https://www.priestmangoode.com/project/united-airlines/

Project objective:

We develop a flight delay prediction app for airline companies that enables them to minimize the loss due to delay and allocate their resources effectively.

2. Data Preprocessing

- Datasets
- Exploratory Data Analysis
- Data cleaning
- Data pipeline
- Join

Four datasets are joined to link flight and timely weather in airports

Datasets

Bureau of Transportation Statistics

- Included flight numbers, plane tail numbers, and information about departures and delays for a flight

National Centers for Environmental Information

- Included observations of temperature, wind and extreme weather conditions

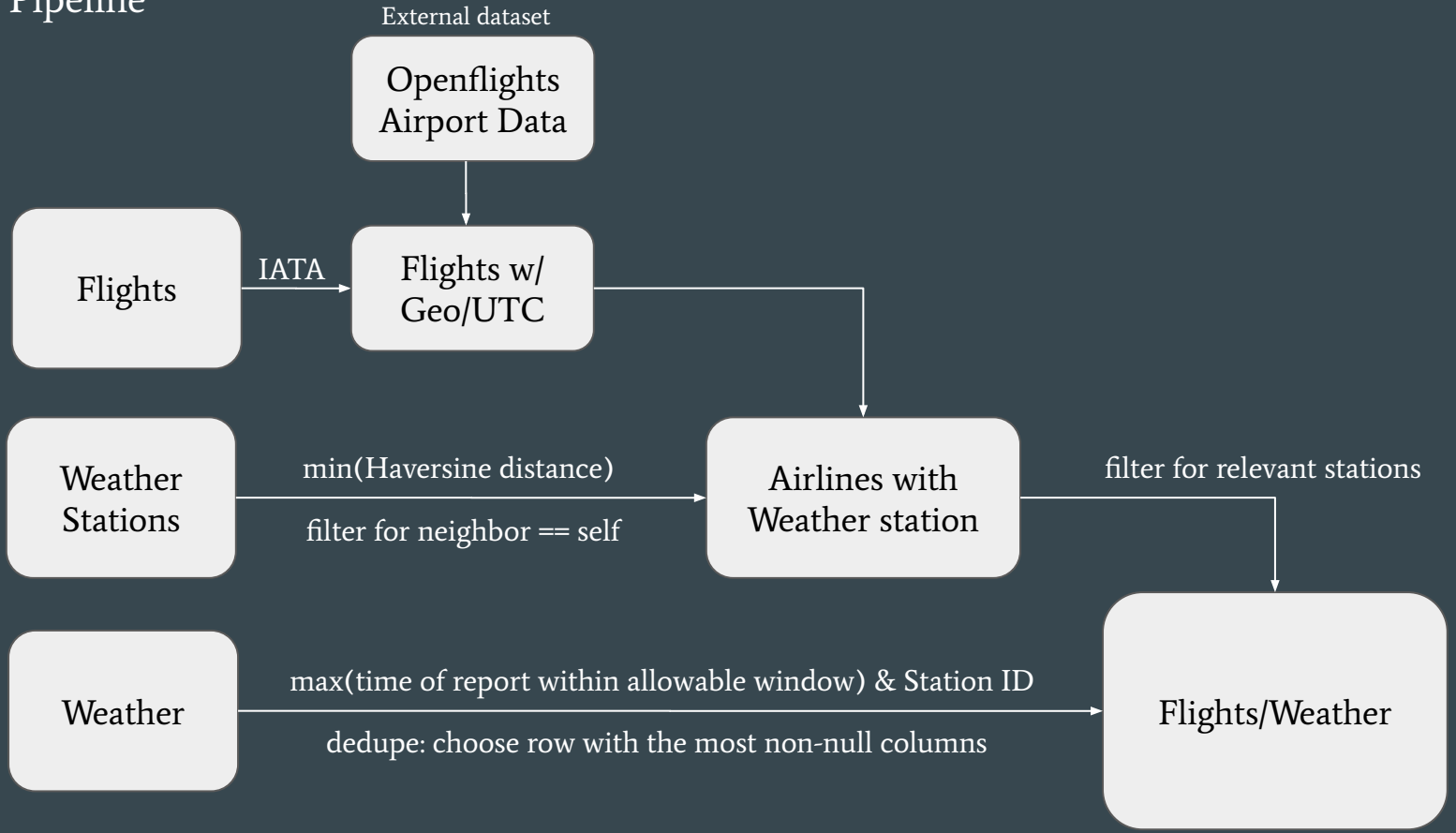
- Included weather stations and lat-long coordinates

OpenFlights

- Included airport IDs and lat-long coordinates

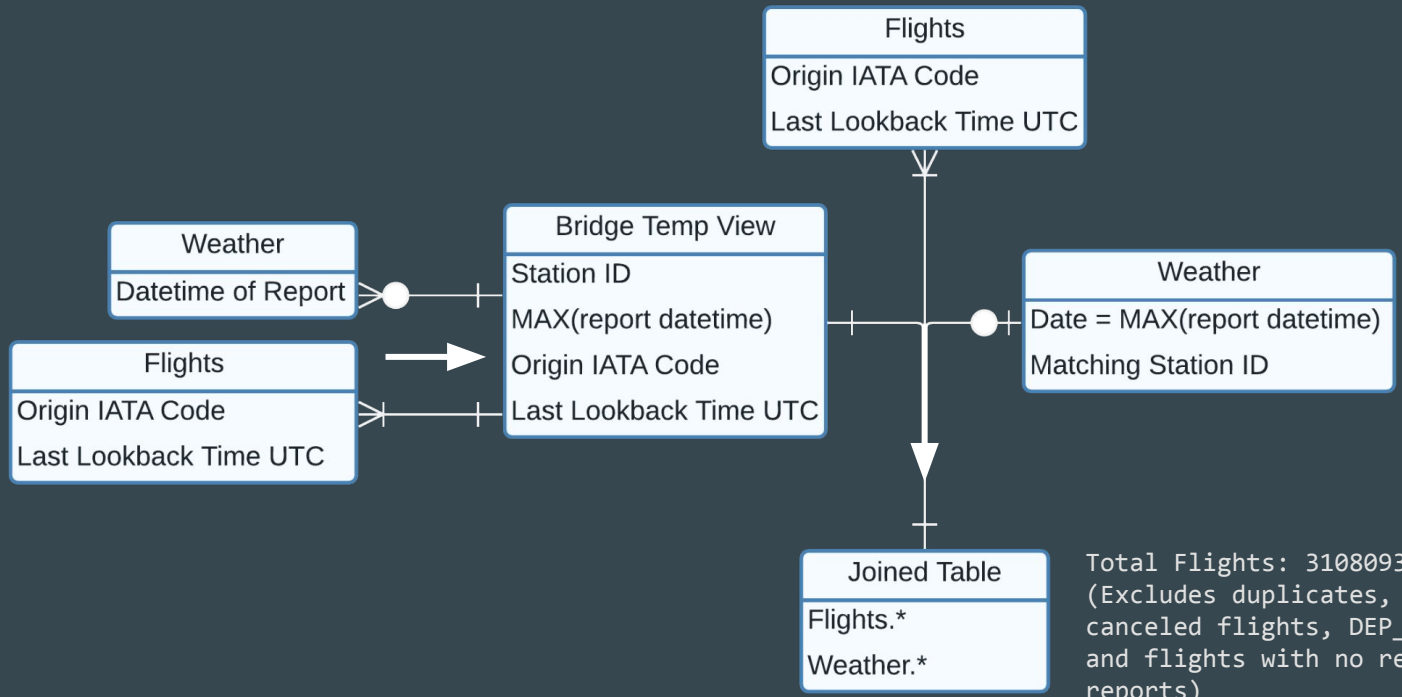
Import, Filter, Join, and Dedupe

Data Pipeline



Filter as much as possible, then join using “bridge”

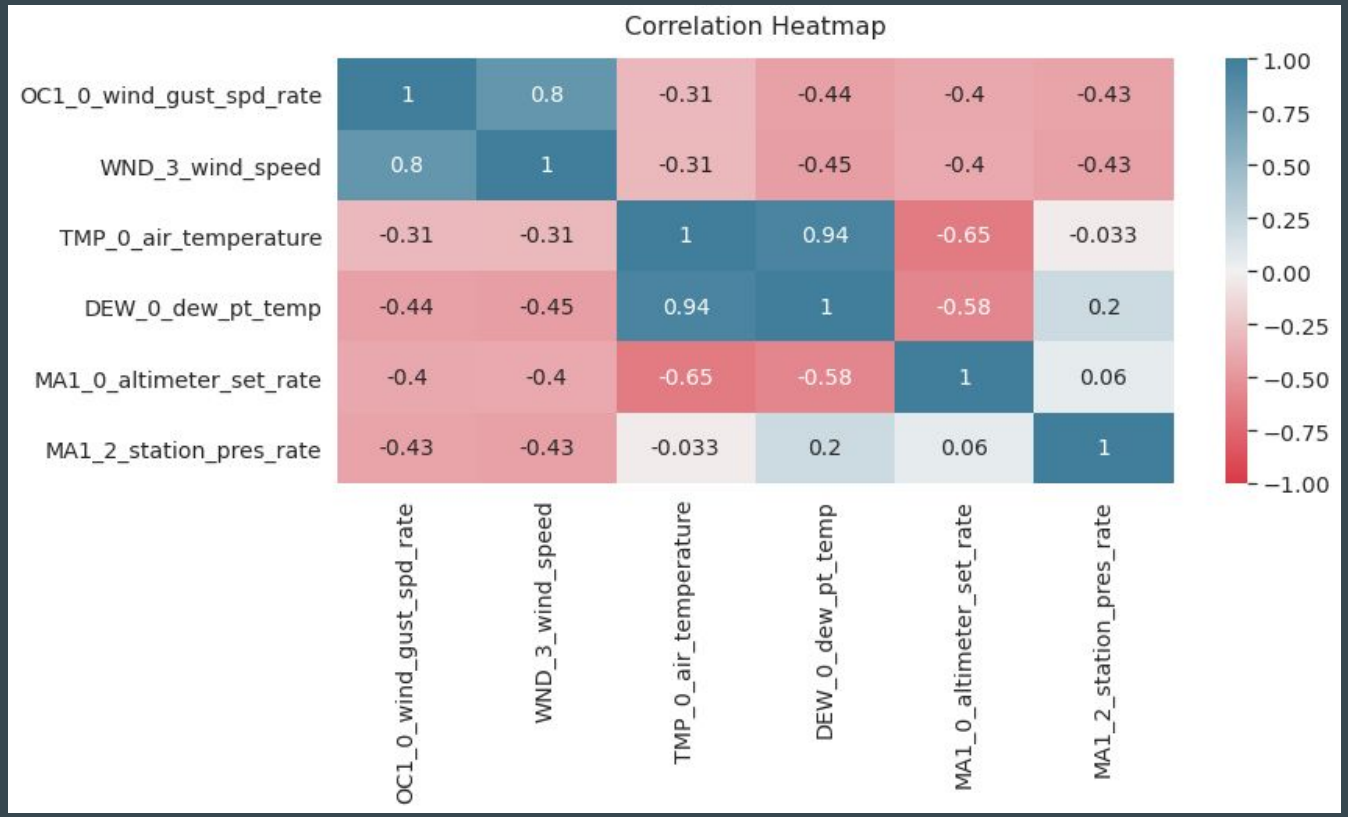
Approach & strategy to join datasets




```
Total Flights: 31080934
(Excludes duplicates,
canceled flights, DEP_DEL15 nulls,
and flights with no recent weather
reports)
```

Some weather features are unexpectedly collinear

EDA - Correlation Table of Weather Data(Post Join)



Having weather stations nearby the majority of the airports unlock the use of weather data in origin/destination airports for prediction

EDA - Distribution of Airport in Airline Data / Weather Stations in weather data

Airport

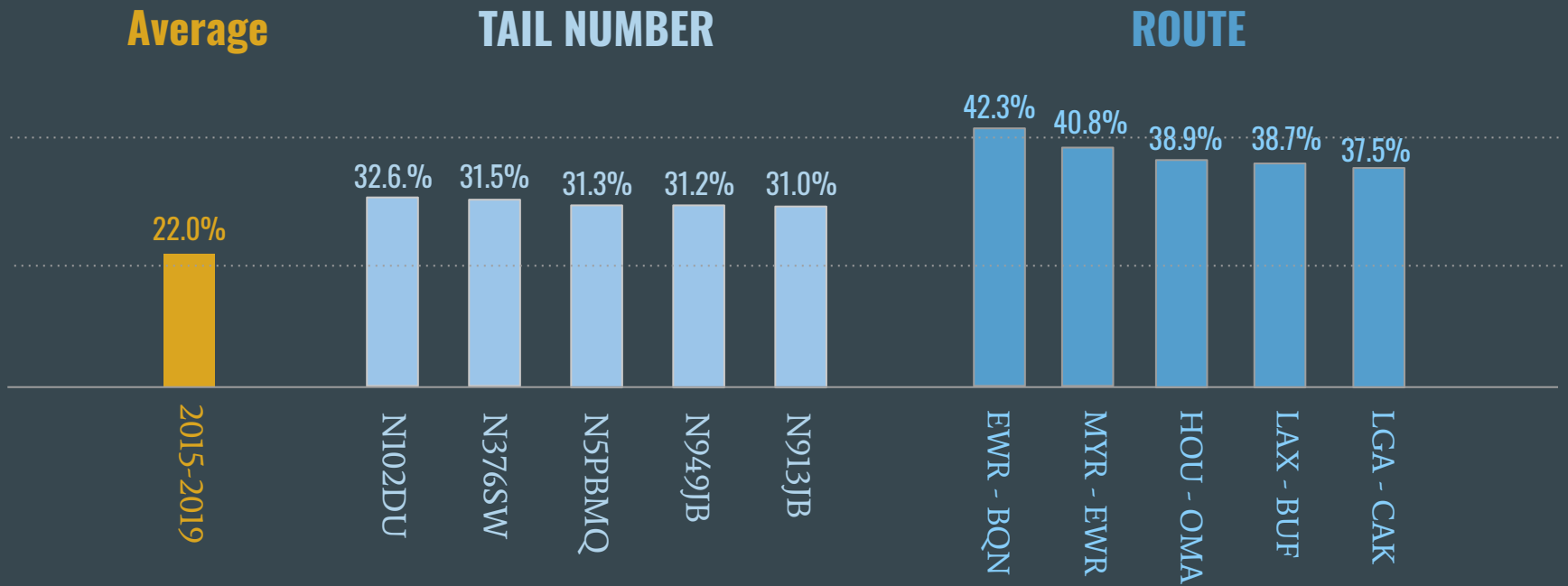
Weather Station

Airport that has a weather station nearby

*Flights and weather stations also cover islands such as Hawaii and Guam

Specific tail numbers/routes are 10-20% more likely to delay

EDA - Top 5 Delayed Flights (*filtered with > 1,000 count of flights in 5 years)



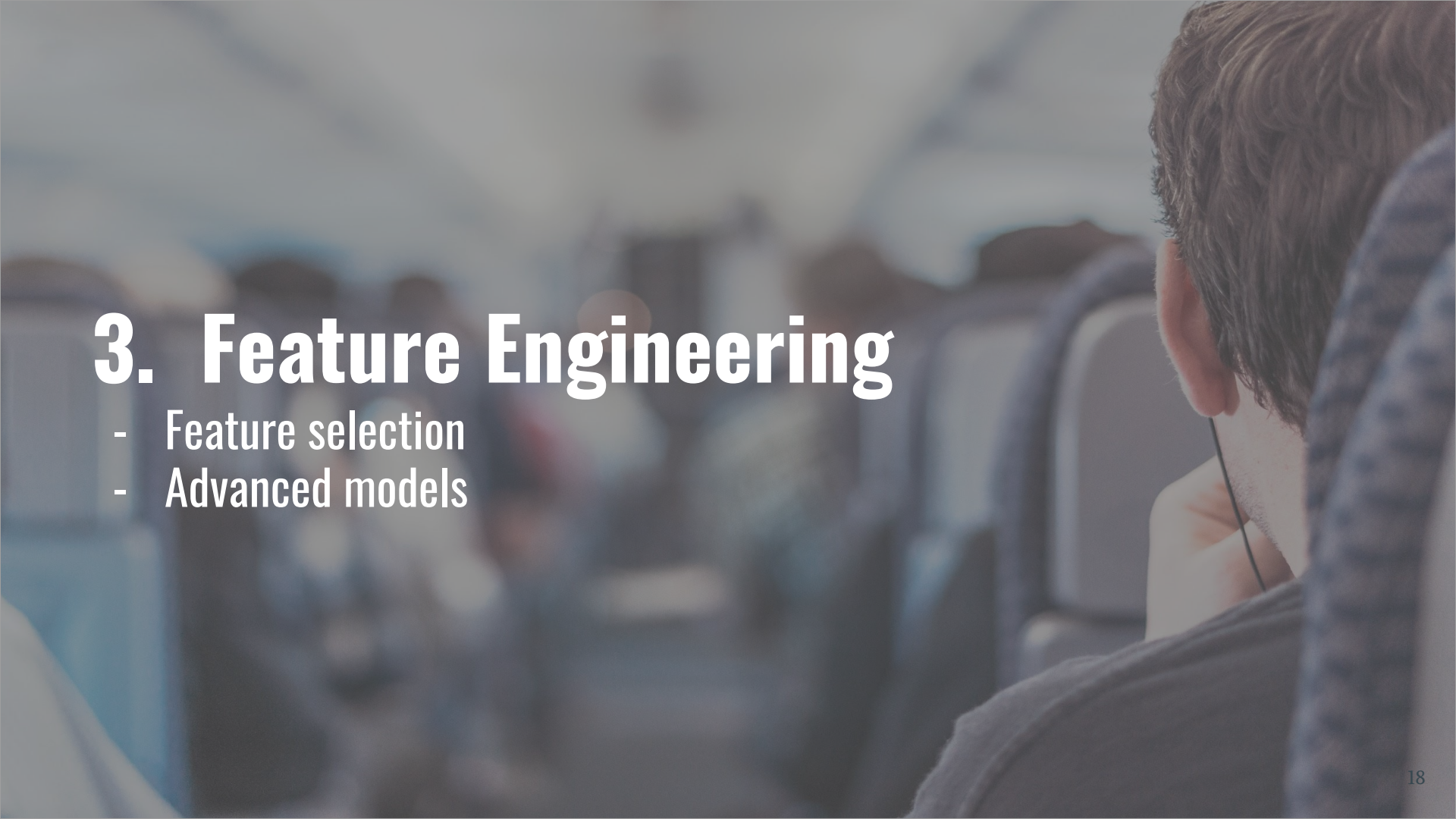
Cancellations, Duplicates, and Nested Features

Data cleaning

Flights	Weather

• Removed canceled flights and null DEP_DEL15
• Encountered some (true) duplicate records post join
	• Imputed numeric weather features using average at the nearest level
	• “Duplicates” (same reporting minute):
	  ○ Choose row with the greatest non-blank columns
	  ○ Broke ties arbitrarily
	• Filtered stations before joining

Imputed missing numericals using past data only

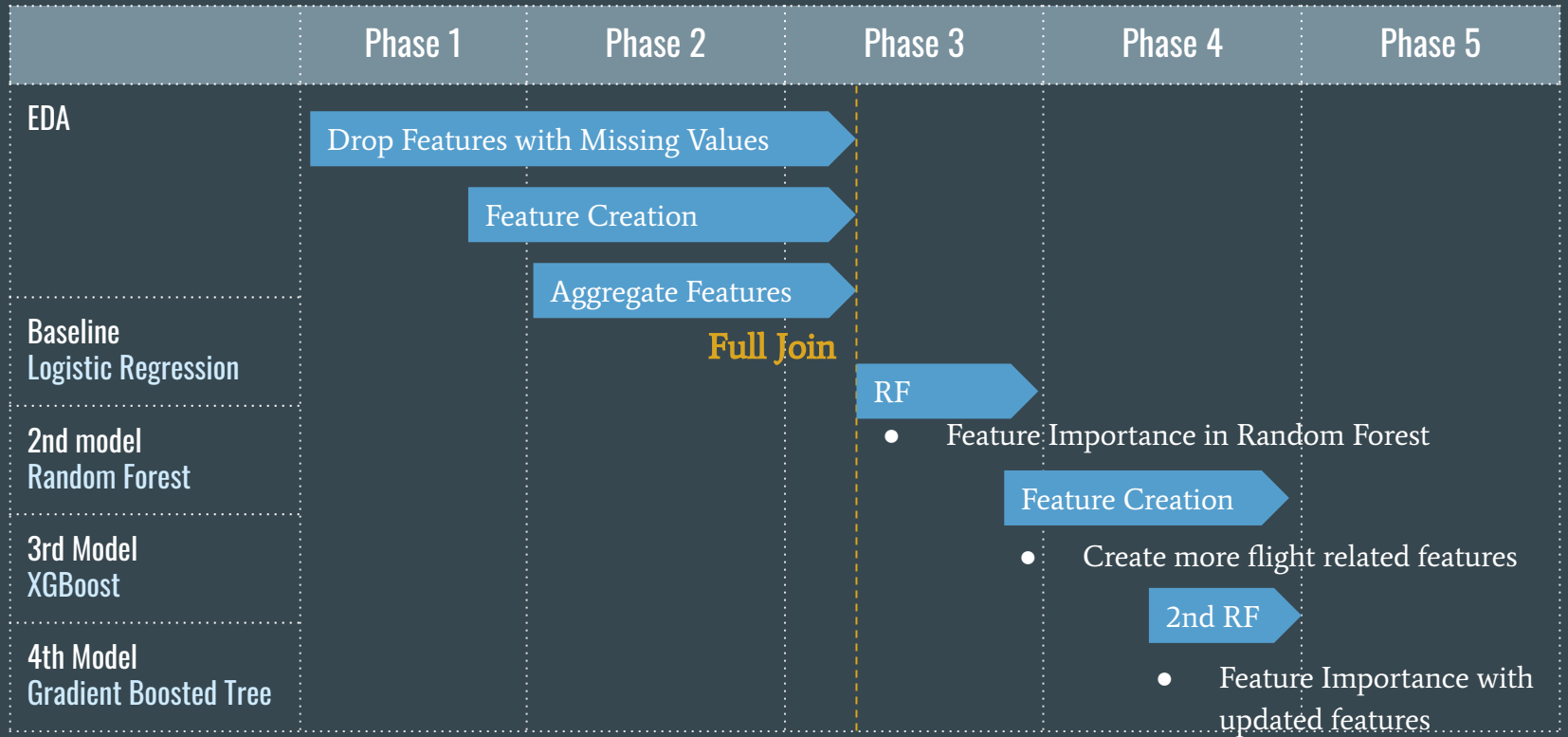


3. Feature Engineering

- Feature selection
- Advanced models

Features were created / selected in agile process

Journey of Feature Engineering



Features from Flight Data

Creating new features (Flights)

Season of Year

Time of Day

Region of Country

Weekend/Holiday

<u>Tracking Plane via Tail #:</u>

- Prior Dep Delay: [0/1]
- Prior Dep Delay Min: [0, inf]
- Prior Arr Delay: [0/1]
- Prior Arr Delay Min: [0, inf]
- Plane is Here?: [0/1]

<u>Carrier Info:</u>

- Avg carrier delay (24 hrs): [0, 1]

<u>Airport Congestion:</u>

- # flights scheduled today @Origin: [0, inf]
- # flights scheduled today @Dest: [0, inf]
- Avg delay at Origin (24 hrs): [0, 1]

<u>Geography</u>

- Elevation @Origin: [0, inf]
- Elevation @Dest: [0, inf]
- Flight Distance: [0, inf]

3 Plane Arrival Scenarios

Tracking Plane via Tail #:

- Plane on its way?
- Time In Between Plane's Last Flights
- Time Inb Flights Rolling Avg by Carrier, by Origin, by Destination

Scenario 1	Recent Dep & Arrival before T-2		Scenario 2	Last Departure before T-2, plane is on its way but not here currently		Scenario 3	Arrival and Dep before T-2 are at another airport. Plane has not left yet.	
2nd Most Recent Arrival	At Airport A					3rd Most Recent Arrival	At Airport A	
			2nd Most Recent Departure	To Airport A				
						2nd Most Recent Departure	From Airport A to Airport B	
Most Recent Departure	From Airport A to Current							
			2nd Most Recent Arrival	At Airport A		2nd Most Recent Arrival	At Airport B	
Most Recent Arrival	At current airport		Most Recent Departure	From Airport A to Current				
T-2 hours	Prediction Time		T-2 hours	Prediction Time		T-2 hours	Prediction Time	
			Most Recent Arrival	At current airport		Most Recent Departure	From Airport B to Current	
						Most Recent Arrival	At current airport	
Scheduled Dep	From current airport		Scheduled Dep	From current airport		Scheduled Dep	From current airport	
Recent arrival is before prediction	1			0			1	
plane_is_here	1			0			0	
recent departure is before prediction	1			1			0	
(if recent departure's dest == current airport origin)								
departed_for_current_aiport	1		departed_for_current_aiport	1		departed_for_current_aiport	0	
time_inb_flight_min	scheduled_dep_time - recent_arr_time			recent_dep_time - 2nd recent_arr_time			2nd recent_dep_time - 3rd_recent_arrival_time	

More Features from Flight Data

Creating new features (Flights)

Carrier Info:

- Carrier delay (%, min) last quarter
- Carrier as ordinal (Brieman's method)

Airport/Route Congestion:

- Delay Network PageRanks (arr & dep)

- Hourly Delay Rolling Avg 24 hrs
- Route Delay Rolling Avg 24 hrs

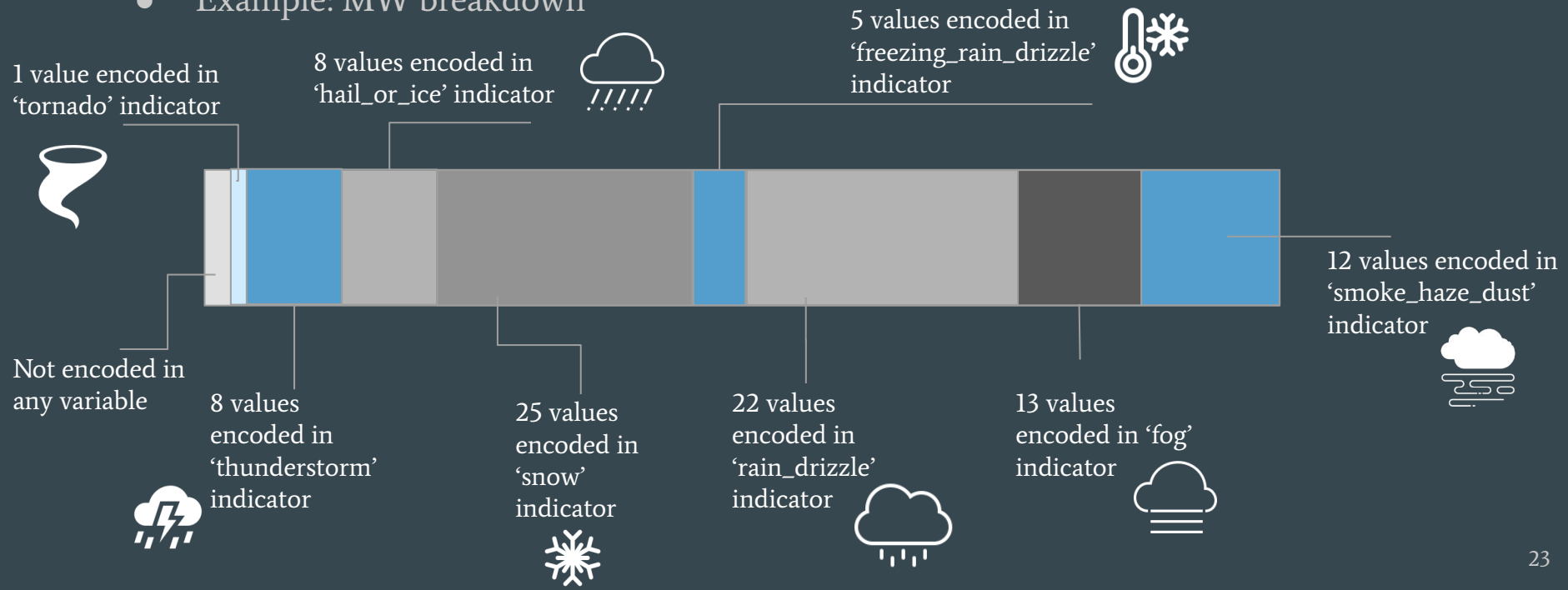
- Origin Airport Rolling Avg 24 hrs
- Origin delay (%, min) last quarter
- Origin as ordinal (Brieman's method)

- Route delay (%, min) last quarter
- Route as ordinal (Brieman's method)

Sparse categorical features to extreme weather indicators

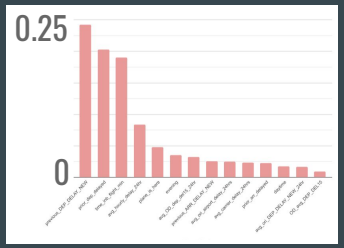
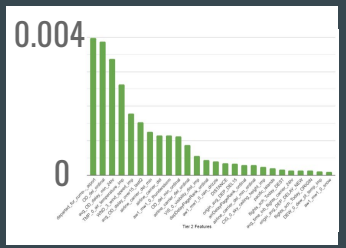
Creating new features (weather)

- Combined AW, MW into categories (e.g. “rain/drizzle”)
- Example: MW breakdown

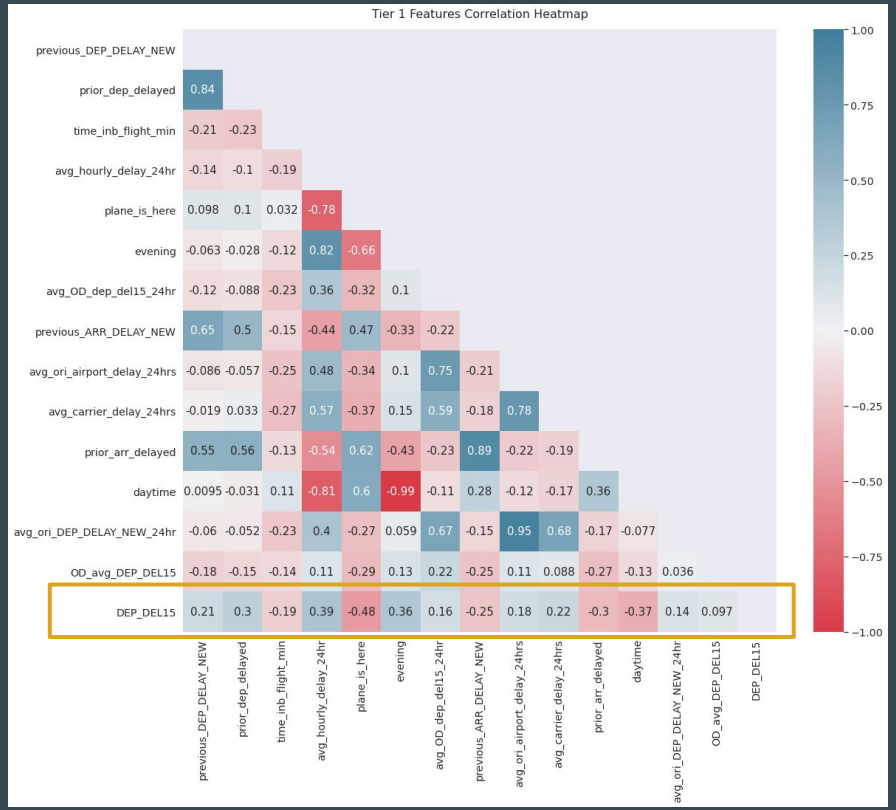


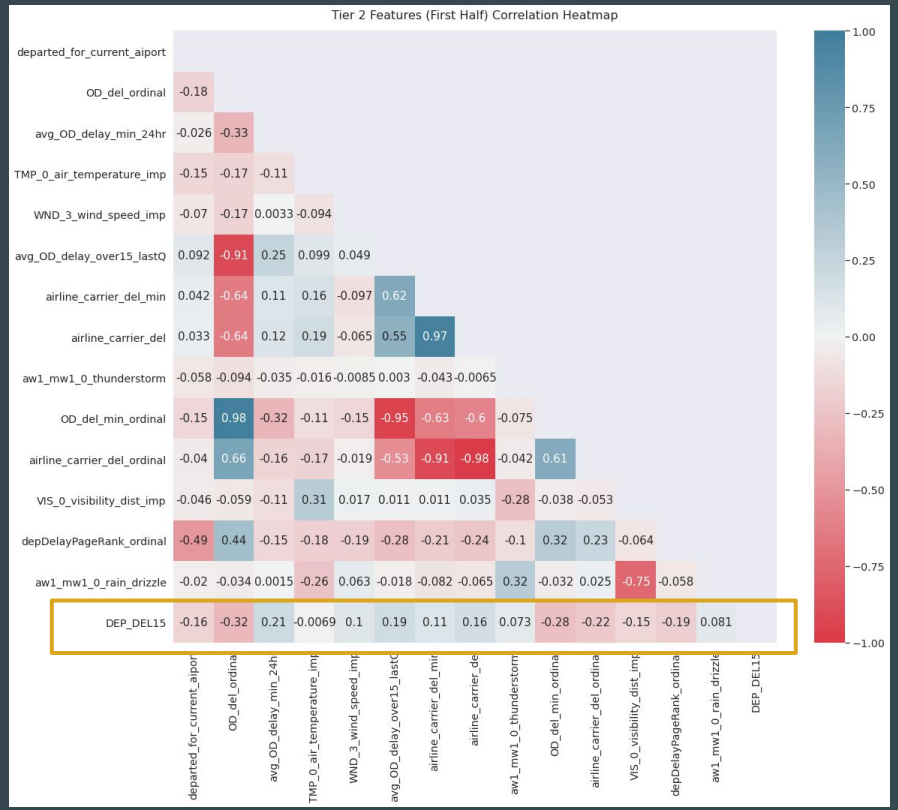
After the second trial focus was shifted to Tier 1 vs Tier 1 + Tier 2

Feature Importance in Random Forest(2nd trial)

	Range	Count	Example	Distribution
Tier 1	> 0.005	13	- prior_dep_delayed - plane_is_here - evening ...	
Tier 2	<= 0.005 > 0.0001	33	- pacific_islands - thunderstorm - depDelayPageRank ...	
Others	<= 0.0001	31	(Rest of the all)	-

Correlation Table: Tier 1 & Top Tier 2 Features



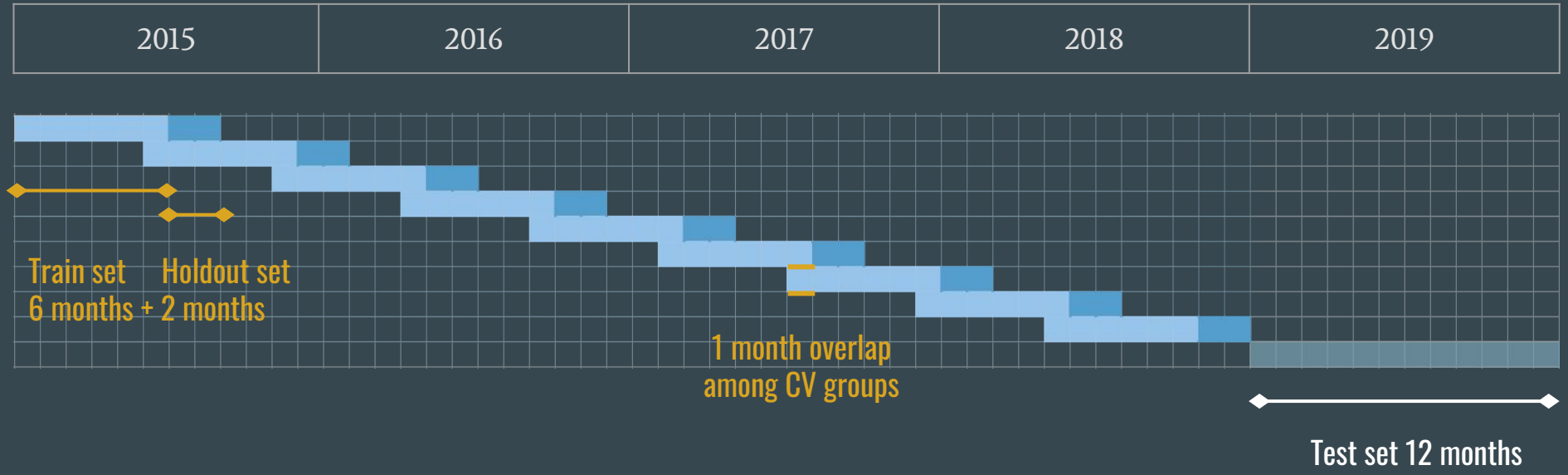


4. Modeling

- Model Selection
- Cross Validation Strategy
- Hyperparameter Tuning

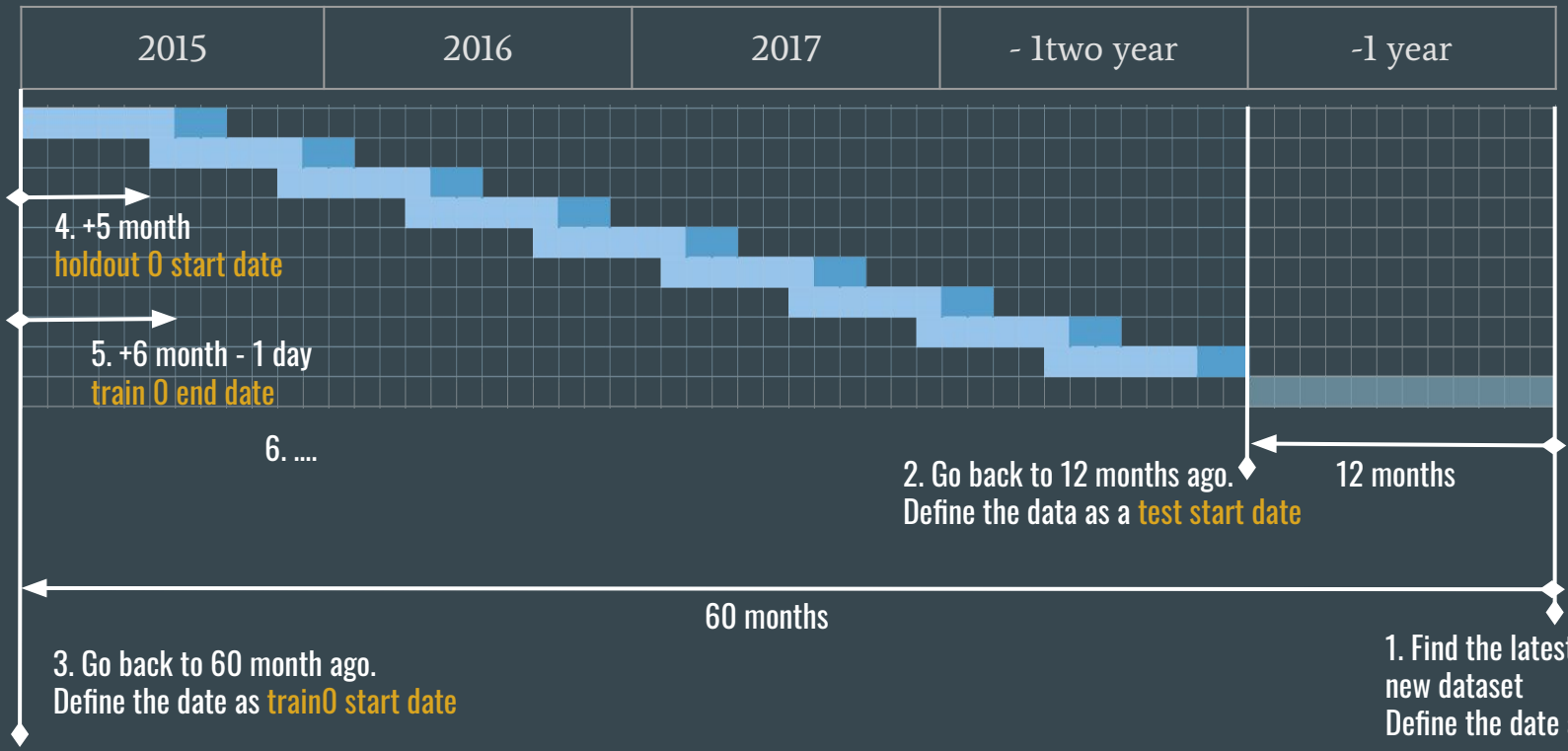
1 month overlap avoids overfitting due to the period of each group

Cross Validation - Block Time Series



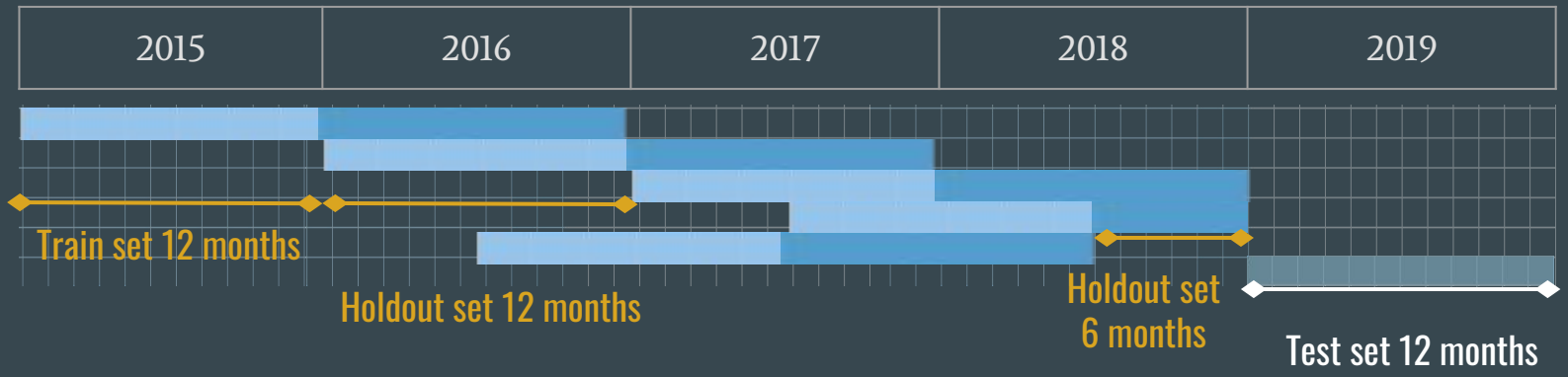
CV also works with new dataset by finding the latest five-years record

Cross Validation - Block Time Series



Custom CV split enables train data to reflect seasonality feature and latest trend better

Custom CV Strategy

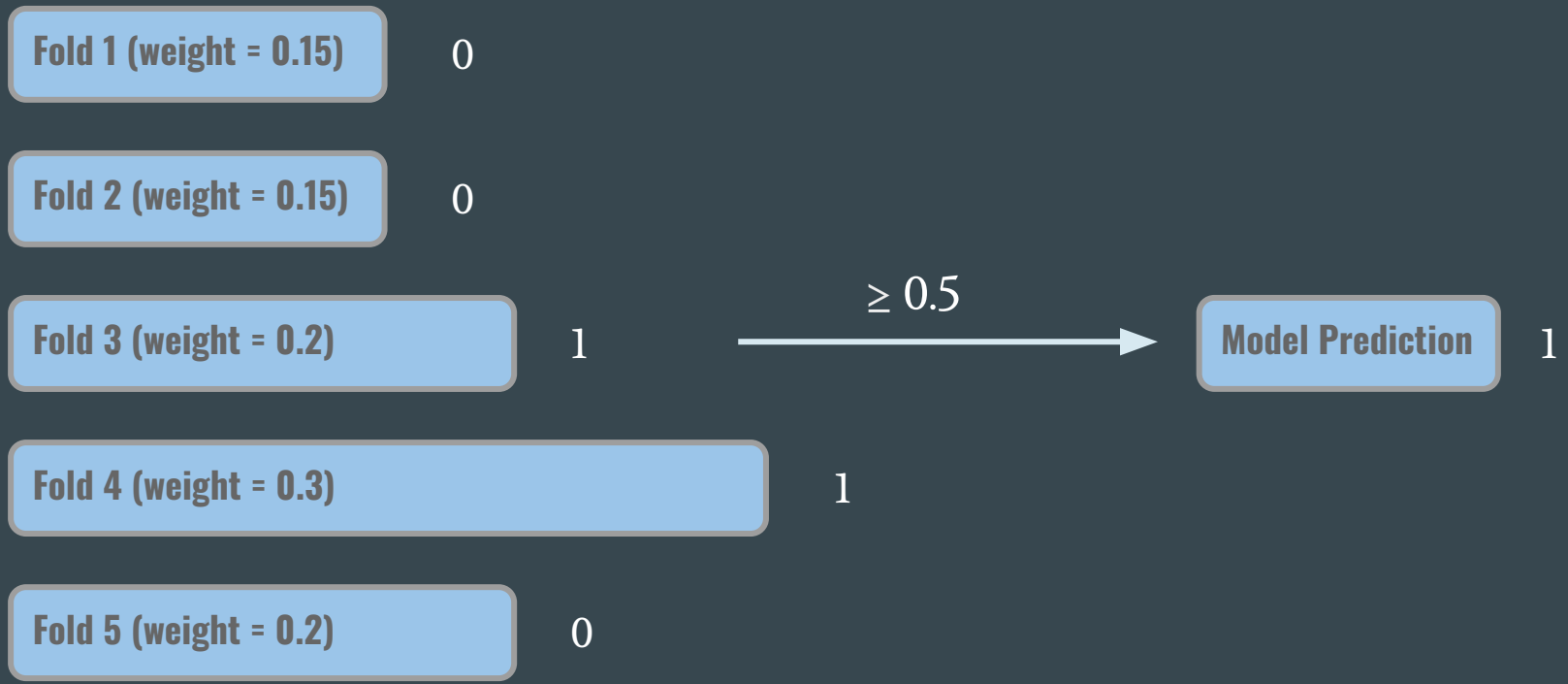


Why use 6 month holdout for Fold 4?

- We wanted to capture the most recent data available while keeping 2019 held out

For every input row (flight), we take a weighted vote between each fold by recency

Voting between folds



Established tree models are implemented to tackle classification problem

Model Selection / Characteristics and Strength

Outcome Variable: DEP_DEL15

1 -> Delay
0 -> No Delay

1. Logistic Regression

- Outcome is interpretable as probability
- Algorithm can be regularized to avoid overfitting
- Fast to train

2. Random Forest

- One of the established algorithm in the past literature of flight delay
- One of the most successful practicalized model in classification

3. XGBoost

- Builds deep trees well-suited to extract meaning from features
- Many regularization options to prevent overfitting (which it is prone to do)

4. Gradient Boosted Trees

- Builds sequential trees well-suited to boost performance on misclassifications

Adding to general tuning, multiple advanced techniques were applied to develop models

Four Approaches for Enhanced Model

Undersampling

- used to re-balance data
- best ratio found: 1.6:1

Hyperparameter Tuning

- used a combination of grid search and random search for tuning

Voting

- Originally applied to create manual forest with time series data in Random Forest
- Expanded to boosting models

Threshold Optimization

- optimization are applied to each CV model
- tuned to maximize f1 score of each fold

Best Hyperparameters

Applied Hyperparameters in Each Machine Learning Algorithms

Logistic Regression

- Max Iterations=10, Regularization Parameter=0.001, Elastic Net Parameter=0.5, Thresholding

Random Forest

- MaxDepth=8, MaxBins=64, minInstancesPerNode=10, numTrees=100, Thresholding

XGBoost

- Eta, Gamma, Lambda, Max Depth, Subsample, Base Score, # Parallel Trees (varied based on fold)

Gradient Boosted Trees

- Max Depth = 5, # of Trees = 100, minInstancesPerNode = 10, maxBins = 64, Thresholds

5. Result

Summary Table of Models Investigated

Our best performing model was **XGBoost**, with an F1 Score of 0.55 (T1, T2 features).

- Best models on leaderboard achieved ~0.8 F1 Score using Boosting, better models also used Boosting or RF.
- They trained longer and spent more time hyperparameter tuning. Some had 10-30% more features.

Model Name	Section	Number of Features	Internal Weighted Voting	F1 Score
Logistic Regression (Best Baseline)	5.2.2	49	No	0.473
Logistic Regression Final	5.2.4	78	Yes	0.501
Random Forest	5.3.1	78	Yes	0.522
XGBoost	5.4.1	78	Yes	0.547
XGBoost	5.4.1	46	Yes	0.550
Gradient Boosted Trees	5.5.1	78	Yes	0.544
4 Model Voting Ensemble	5.6	78	Yes	0.546

Scalability Concerns

- 5 CVs & 4 Model Voting requires tracking and running operations on multiple prediction dataframes.
- More training time may needed, assuming that 2020 or 2021 data would not contribute to prediction due to irregular circumstances. Older records may need to incorporated to additionally

Limitations & Future Directions

- Time-Series Data
  - Limits our Cross Validation Split to rolling windows
  - Requires past data to make features (such as Brieman's method) and impute values, which may make predictions worse due to changes over time.
- Cluster
  - Sharing a cluster and investigating many different models limited in-depth tuning to Logistic Regression and XGBoost. More time spent on Tuning RF, GBT might have yielded better F1 Scores according to Gap Analysis.
- Next Steps
  - **Longer random search hyperparameter tuning** for GBT, RF, XGBoost using full data
  - Try random search hyperparameter tuning for GBT, RF, XGBoost using **only most important features.**
  - Remodel RF, XGBoost, GBT final model using best parameter **on all of the training data, instead of voting between models.**
  - Voting system in the combination of regression models

Questions