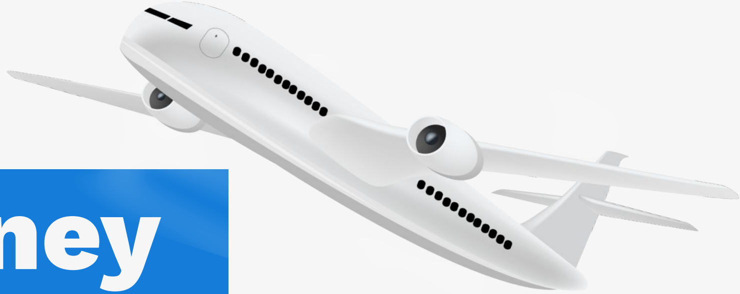


Take Off with Data:

Spark Airlines' Journey to Outsmart Flight Delays

261 Final Project - Team 2-1

Heesuk Jang

Karsyn Lee

Stephanie Cabanela

Raymond Tang

Overview

01

Abstract & Project Description

02

EDA & Feature Engineering

03

Modeling Pipeline

04

Results & Next Steps

01
Abstract & Project Description

Abstract

> *At Spark Airlines, we believe in the transformative power of travel in connecting people to the world.*
>
> *Our mission is to do the right thing for our* ***customers, communities, and planet****.*

Business Use Case

Business Problem: Predicting departure flight delays

Why do we care?

> *At Spark Airlines, our mission is to do the right thing for our* ***customers, communities, and planet****.*

Impact of flight delays:

- ***Customers*** → passenger dissatisfaction
- ***Communities*** → financial time and economic losses
- ***Planet*** → increase carbon emissions

Business Use Case

Business Problem: Predicting departure flight delays

Business Solution: Binary Classification Model

- *My flight is scheduled to departure in 2 hours. Will there be a delay?*
- Departure Delay >= 15 minutes

Business Metrics

Primary Metric:

- F2 Score (beta = 2.0)

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Secondary Metrics for Results based analysis:

- Recall
- Precision

Dataset

The Full Dataset

OTPW – joined dataset of ontime flight performance and weather data from 2015 – 2019

- 216 columns, 31.7 million rows
- data types: int, float, string, date, time features exist
- target variable = DEP_DEL15

Dataset

What did we focus on for Phase 2?

- **3 month OTPW: Jan 2015 – March 2015** – faster iterative development
    - 216 columns, 1.4 million rows
- **1 year OTPW: Jan 2015 – December 2015** – evaluation on baseline model pipeline
    - 216 columns, 11.6 million rows

02

EDA & Feature Engineering

Summary Statistics on the Selected Columns

column	description	count	mean	stddev	min	max
MONTH int	Time-related features that capture the temporal context of flights. Seasonality and specific days or periods (like weekends, holidays, or vacation seasons) can significantly influence flight schedules and the likelihood of delays.	11623708	6.524	3.405	1	12
QUARTER int		11623708	2.508	1.107	1	4
YEAR int		11623708	2015.000	0.000	2015	2015
DAY_OF_MONTH int		11623708	15.703	8.783	1	31
DAY_OF_WEEK int		11623708	3.927	1.989	1	7
DISTANCE float	The length of the flight can impact the likelihood of delays due to factors like fueling time and air traffic.	11623708	821.542	607.271	21	4983
DEP_DEL15 float	Target Variable	11451590	0.184	0.388	0	1
ELEVATION float	Airport elevation can affect aircraft performance, possibly influencing departure times.	11623708	251.928	398.849	0.3	2353.1
CRS_DEP_TIME int	The scheduled departure time and expected flight duration can be significant predictors of delays, especially during peak hours or longer flights.	11623708	1329.606	483.246	1	2359
CRS_ELAPSED_TIME flt		11623696	141.594	75.172	18	718
DIVERTED float	Flights that have been diverted in the past may have a higher chance of future delays.	11623708	0.003	0.051	0	1
FLIGHTS float	The number of flights (frequency) could indicate busier periods more prone to delays.	11623708	1.000	0.000	1	1
DISTANCE_GROUP int	Grouping flights by distance can help identify delay patterns for short, medium, and long-haul flights.	11623708	3.759	2.392	1	11
DEST str	The departure and arrival airports can be crucial factors as some airports might have higher instances of delays due to factors like traffic, weather, or operational issues.	11623708	n/a	n/a	ABE	YUM
ORIGIN str		11623708	n/a	n/a	ABE	YUM
OP_UNIQUE_CARRIER str	The airline operating the flight often influences delay patterns due to varying operational efficiencies and policies.	11623708	n/a	n/a	AA	WN
TAIL_NUM str	Specific aircraft might have different reliability or maintenance records, affecting departure punctuality.	11594302	n/a	n/a	7819A	N9EAMQ
HourlyWindSpeed int	Weather conditions at the time of the flight can significantly impact flight schedules, with adverse weather often leading to delays.	11589326	8.893	5.344	0	67
HourlyPrecipitation float		9695354	0.003	0.031	0	5.76
HourlyRelativeHumidity int		11591308	60.385	21.465	1	100
HourlyVisibility float		11583804	9.390	1.881	0	99.42

Summary Statistics on the Selected Columns

column		count	mean	stddev	min	max
MONTH	int	11623708	6.524	3.405	1	12
QUARTER	int	11623708	2.508	1.107	1	4
YEAR	int	11623708	2015.000	0.000	2015	2015
DAY_OF_MONTH	int	11623708	15.703	8.783	1	31
DAY_OF_WEEK	int	11623708	3.927	1.989	1	7
DISTANCE	float	11623708	821.542	607.271	21	4983
DEP_DEL15	float	11451590	0.184	0.388	0	1
ELEVATION	float	11623708	251.928	398.849	0.3	2353.1
CRS_DEP_TIME	int	11623708	1329.606	483.246	1	2359
CRS_ELAPSED_TIME	flt	11623696	141.594	75.172	18	718
DIVERTED	float	11623708	0.003	0.051	0	1
FLIGHTS	float	11623708	1.000	0.000	1	1
DISTANCE_GROUP	int	11623708	3.759	2.392	1	11
DEST	str	11623708	n/a	n/a	ABE	YUM
ORIGIN	str	11623708	n/a	n/a	ABE	YUM
OP_UNIQUE_CARRIER	str	11623708	n/a	n/a	AA	WN
TAIL_NUM	str	11594302	n/a	n/a	7819A	N9EAMQ
HourlyWindSpeed	int	11589326	8.893	5.344	0	67
HourlyPrecipitation	float	9695354	0.003	0.031	0	5.76
HourlyRelativeHumidity	int	11591308	60.385	21.465	1	100
HourlyVisibility	float	11583804	9.390	1.881	0	99.42

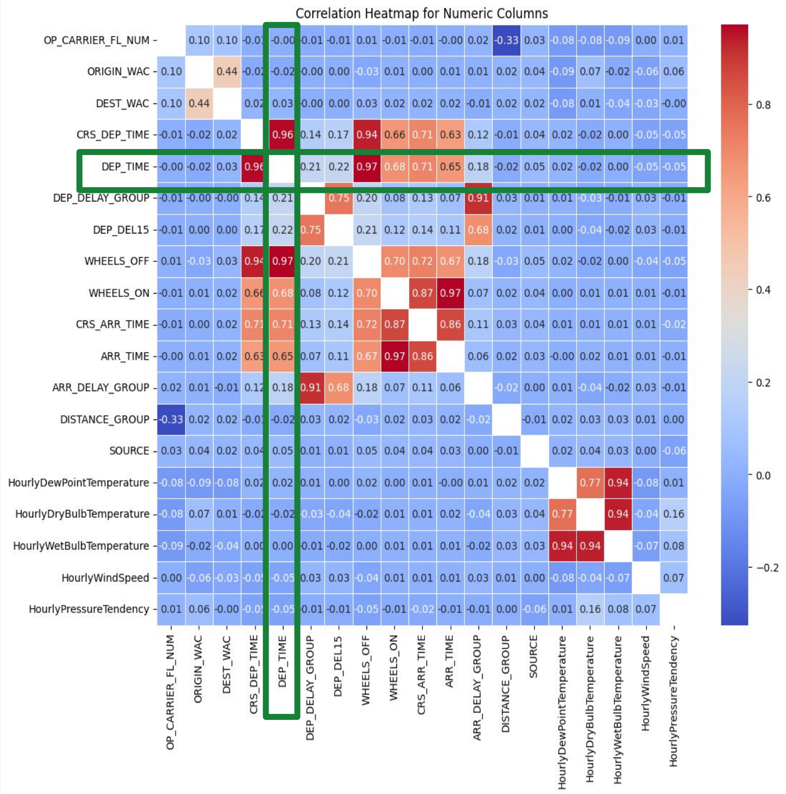
Summary Statistics on the Selected Columns

Column Name	count	mean	stddev	min	max
CRS_DEP_TIME	6019064	1398.150132	478.7325967	1	2359
DISTANCE	6019064	855.1018351	619.6686513	25	4983
ELEVATION	6019064	250.9033566	405.2289741	0.3	2353.1
DAY_OF_WEEK_sin	6019064	0.009237530767	0.7077602777	-0.9749279122	0.9749279122
DAY_OF_WEEK_cos	6019064	-0.02663869855	0.705889937	-0.9009688679	1
CRS_DEP_HOUR_sin	6019064	-0.1910342823	0.7309022564	-1	1
CRS_DEP_HOUR_cos	6019064	-0.3496396667	0.5541119361	-1	1
DAY_HOUR_interaction	6019064	107.9863007	47.92791738	24	191
TIME_INTERVAL_OF_DAY	6019064	0.590492143	0.9047150268	0	3
DAY_OF_MONTH	6019064	15.71623445	8.757503842	1	31
MONTH	6019064	6.485082564	3.375486249	1	12
HourlyPrecipitation	5023508	0.004005035907	0.0358555068	0	10.14
HourlyRelativeHumidity	6001313	60.64908229	21.85803803	1	100
HourlyVisibility	5997882	9.429484471	1.814999919	0	99.42
HourlyWindSpeed	6000947	9.179732632	5.621403246	0	2237
CRS_ELAPSED_TIME	6019063	146.2590604	76.67994459	4	718
OP_UNIQUE_CARRIER	6019064	None	None	AA-US	WN
ORIGIN	6019064	None	None	ABE	YUM
DEST	6019064	None	None	ABE	YUM
TAIL_NUM	6019064	None	None	7819A	PLANET
CARRIER_SIZE	6019064	None	None	CARRIER_LARGE	CARRIER_SMALL
DAY_TYPE	6019064	None	None	weekday	weekend
DEGREE_VISIBILITY	6019064	None	None	HIGH_VISIBILITY	LOW_VISIBILITY
FL_DISTANCE_GROUP	6019064	None	None	DIST_LONG	DIST_SHORTEST
5_DAYS_DIST_FROM_Independence	6019064	967460.5849	177425.0239	-1	999999
7_DAYS_DIST_FROM_Christmas	6019064	956906.5988	203065.1122	-1	999999
7_DAYS_DIST_FROM_NewYear	6019064	974644.3416	157199.5208	0	999999
DIVERTED	6019064	0.003215948526	0.0566180778	0	1
dep_del15_2hr_before	6019064	0.8829175101	0.3215185805	0	1

EDA

LASSO Feature Selection

Features	Coefficients
CRS_DEP_TIME	0.035599072873419756
HourlyRelativeHumidity_imputed	0.018687284077983014
HourlyWindSpeed_imputed	0.009424083809546286
6_DAYS_DIST_FROM_NewYear_idx	0.007825974350104068
CRS_ELAPSED_TIME_imputed	0.003791426948673926
HourlyPrecipitation_imputed	0.002668667189096292
6_DAYS_DIST_FROM_Christmas_idx	0.0008842165651147551
3_DAYS_DIST_FROM_Independence_idx	0
FL_DISTANCE_GROUP_idx	0
DEGREE_VISIBILITY_idx	0
DAY_TYPE_idx	0
CARRIER_SIZE_idx	0
TAIL_NUM_idx	0
OP_UNIQUE_CARRIER_idx	0
DIVERTED	0
MONTH	0
DAY_OF_MONTH	0
DAY_HOUR_interaction	0
CRS_DEP_HOUR_cos	0
DAY_OF_WEEK_cos	0
DAY_OF_WEEK_sin	0
ELEVATION	0
DISTANCE	0
DEST_idx	-0.00007749139001021719
ORIGIN_idx	-0.0001549738987923603
HourlyVisibility_imputed	-0.009176450353314968
TIME_INTERVAL_OF_DAY	-0.009557272962663378
CRS_DEP_HOUR_sin	-0.06712880652735072

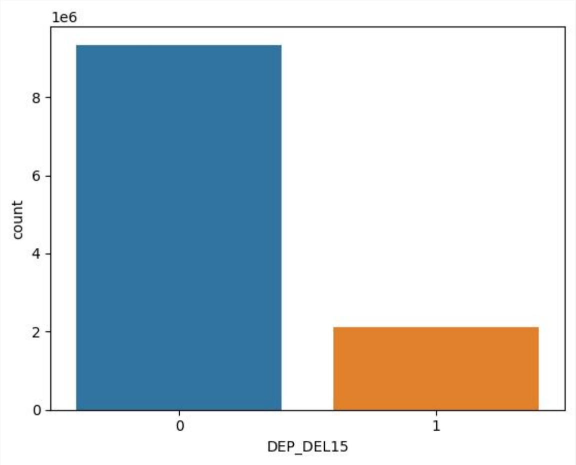


Pearson Correlation Analysis

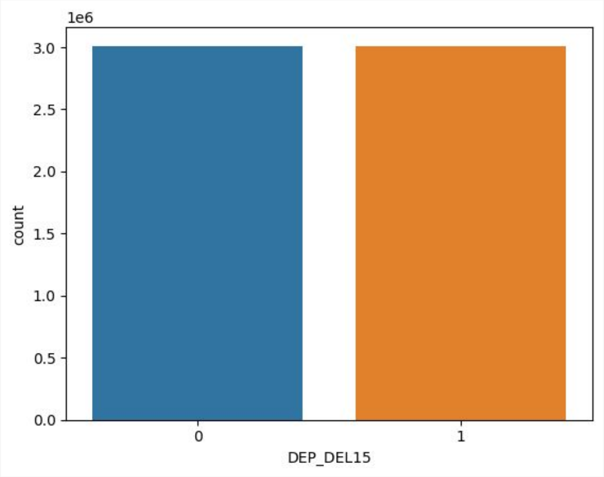
EDA

Under-Sampling

- Imbalanced data - significantly more observations were recorded where the flight was on time
- Non-delayed is 5x more than Delayed flights



Under-sampling



Data Pre-Processing

Numeric & Categorical Variables

Df_cleaned → Numeric Data

Df_cleaned → Categorical Data → String Index & One-Hot Encoding

This is the string index shown below.

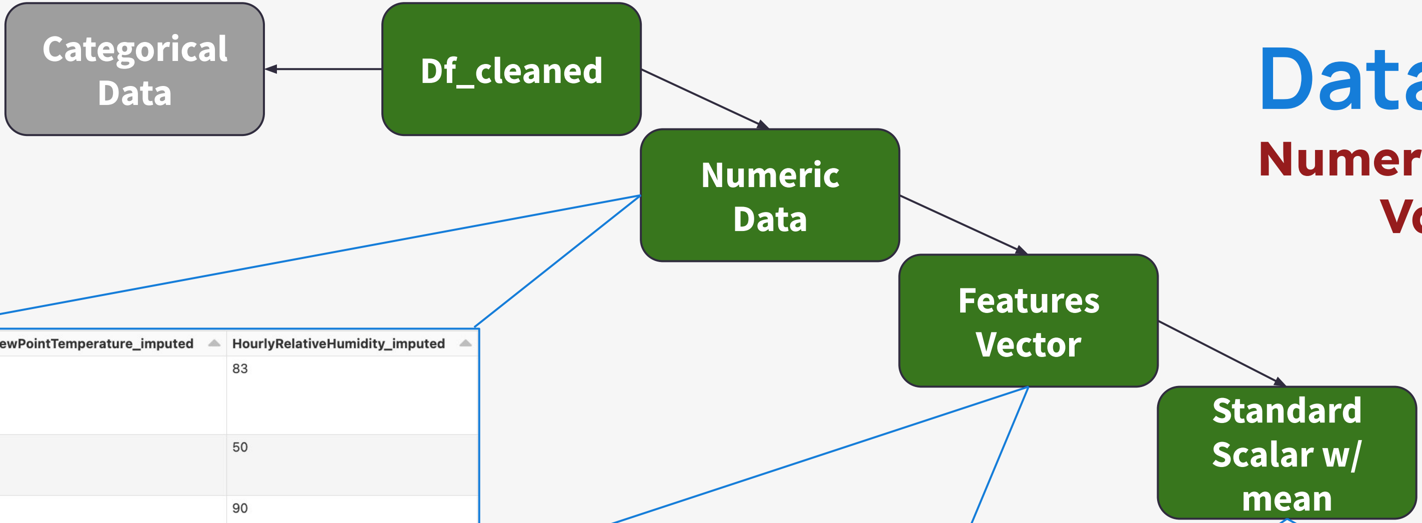
One-Hot Encoding:
A B C D => 1 2 3 4 => [0 0 0 1, 0 1 0 0, ...]

OP_UNIQUE_CARRIER_idx	ORIGIN_idx	DEST_idx
2	4	16
2	2	38
2	4	16

OP_UNIQUE_CARRIER	ORIGIN	DEST
AA	LAX	JFK
AA	DFW	HNL
AA	LAX	JFK

Data Pre-Proce

Numeric & Categorical Variables



HourlyDewPointTemperature_imputed	HourlyRelativeHumidity_imputed
61	83
67	50
65	90

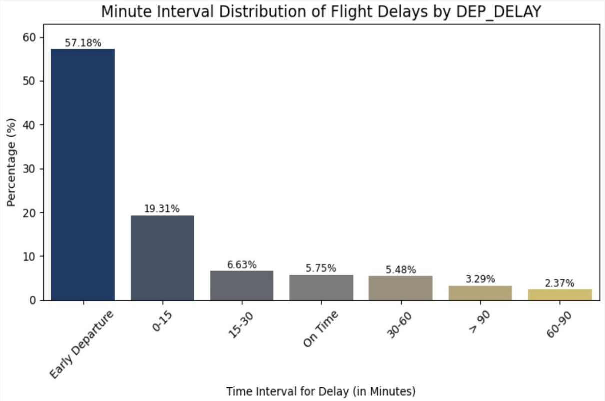
features_vec
{"vectorType": "dense", "length": 12, "values": [344, 1, 2475, 10, 29.600000381469727, 0.0032276585698127747, 10, 16, 250, 66, 61, 83]}
{"vectorType": "dense", "length": 12, "values": [497, 1, 3784, 11, 170.6999969482422, 0, 10, 11, 180, 88, 67, 50]}
{"vectorType": "dense", "length": 12, "values": [334, 1, 2475, 10, 29.600000381469727, 0, 10, 6, 260, 68, 65, 90]}

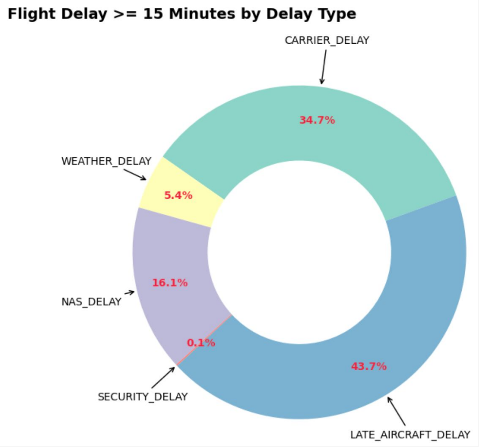
features_scaled_vec
{"vectorType": "dense", "length": 12, "values": [2.662575352023657, 0, 2.6992888838030313, 2.588925194713191, -0.552448397918632, -0.023629425322748643, 0.3346435361931144, 1.3109259201013694, 0.7029364936294197, 0.11849734007089043, 0.6933330726211824, 1.0097615122254673]}
{"vectorType": "dense", "length": 12, "values": [4.690888818343632, 0, 4.852065203810319, 3.006917192128604, -0.20084269638136729, -0.12262072042093952, 0.3346435361931144, 0.37855390081465584, 0.03989675387061614, 1.255086515114932, 1.0088661288600662, -0.5250308051127458]}
{"vectorType": "dense", "length": 12, "values": [2.530005844421044, 0, 2.6992888838030313, 2.588925194713191, -0.552448397918632, -0.12262072042093952, 0.3346435361931144, -0.5538181184720579, 0.7976564564521058, 0.22182362871125785, 0.903688443447105, 1.335323518933573]}

EDA

Departure Delay Fact

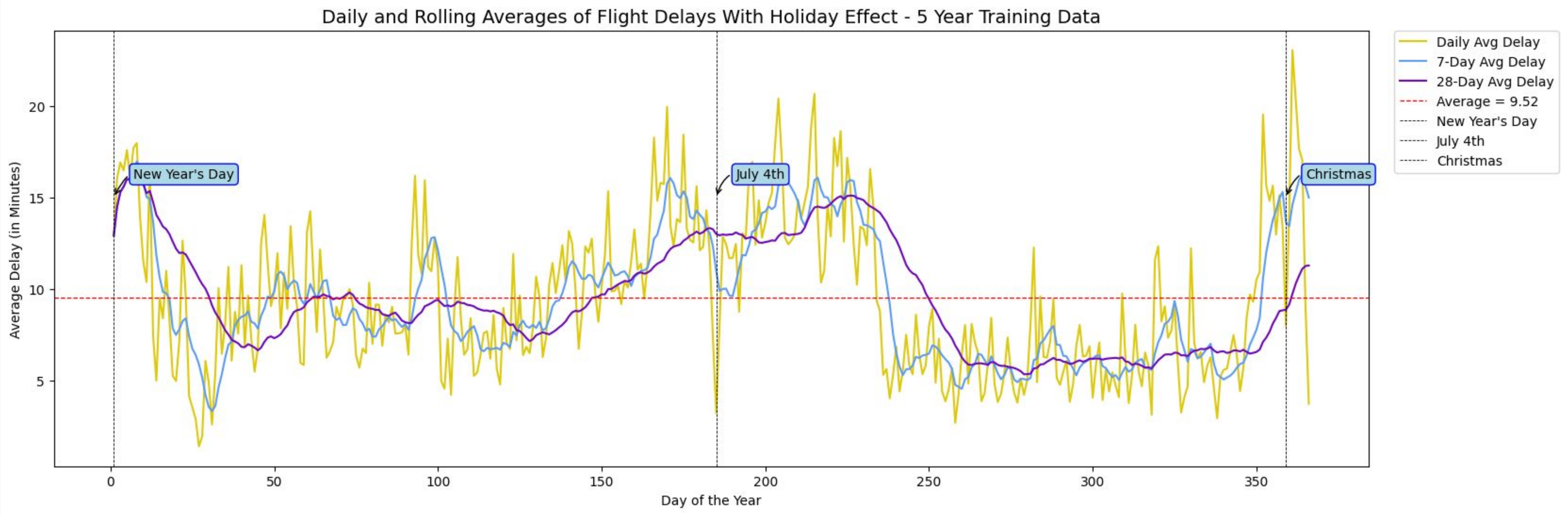
- **Delay Interval Distribution**
  - Early Departure: more than 50%
  - 0-15 Minute Delay: 19%
- **Cause of Delay (in Minutes)**
  - Late Aircraft Delay: 44%
  - Carrier Delay: 35%
  - National Air System Delay: 16%
  - Weather Delay: 5%





EDA Seasonality Analysis

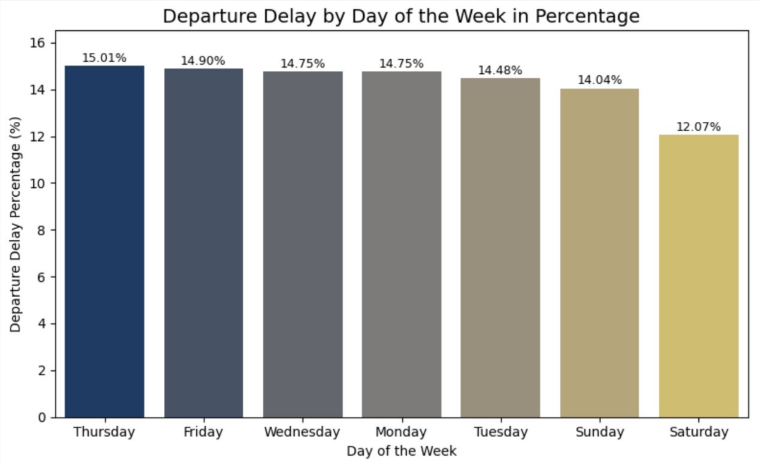
- **Delay** by **Day of Year**
  - Holiday Effect:
    - July 4th
    - New Year’s day
    - Christmas

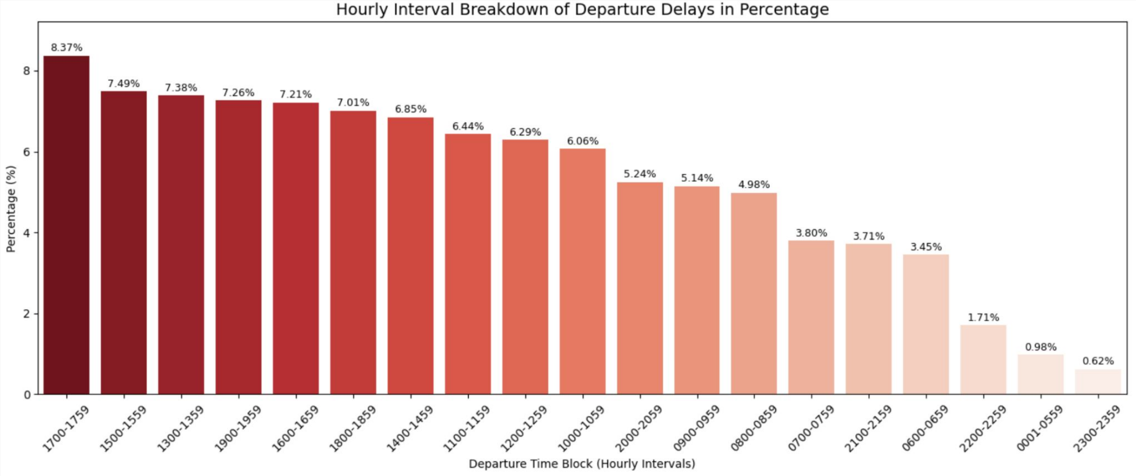


EDA

Trend Analysis

- **Delay** by **Day of Week**
  - Days with the highest delays:
    - Thursday: 15%
    - Friday: 15%
  - Days with the lowest delays:
    - Sunday: 14%
    - Saturday: 12%
- **Delay** by **Hour of Day**
  - Time interval pattern:
    - Afternoon & evening (1 PM - 7 PM) with high delays
    - Early morning & night time with low delays
- **Delay** by **Day of Month**
- **Delay** by **Quarter**
- **Delay** by **Month**

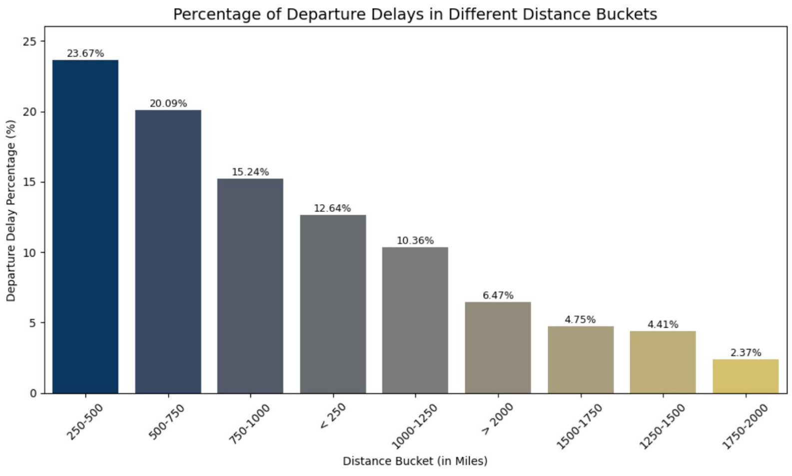


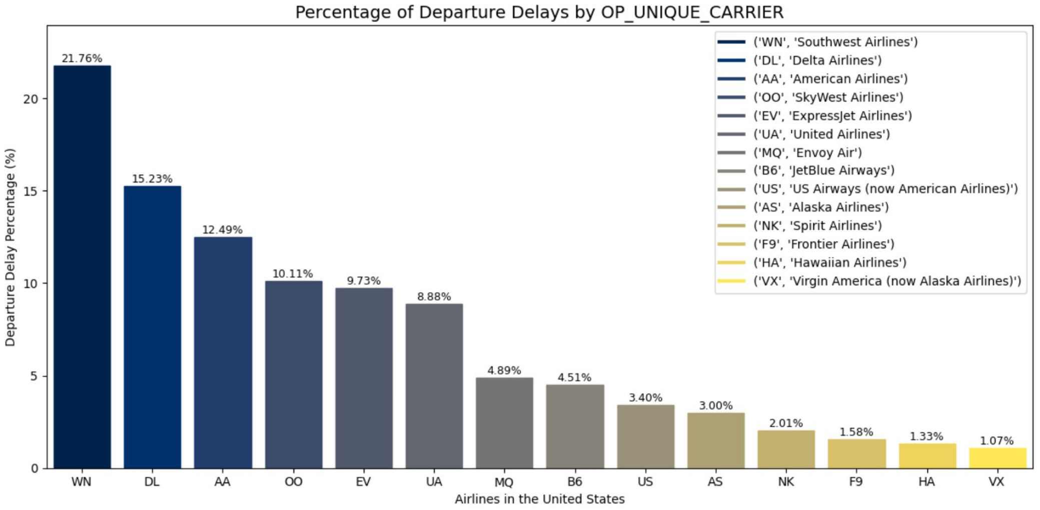


EDA

Operational-associated

- **Distance of the Flight:**
  Likelihood of delays due to factors like fueling time and air traffic.
  - In general, shorter distances are associated with higher departure delays.
- **Airline operation:**
  Often influence delay pattern due to varying operational efficiencies and policies.
  - Top three airlines with the highest departure delay:
    - Southwest airlines
    - Delta airlines
    - American airlines

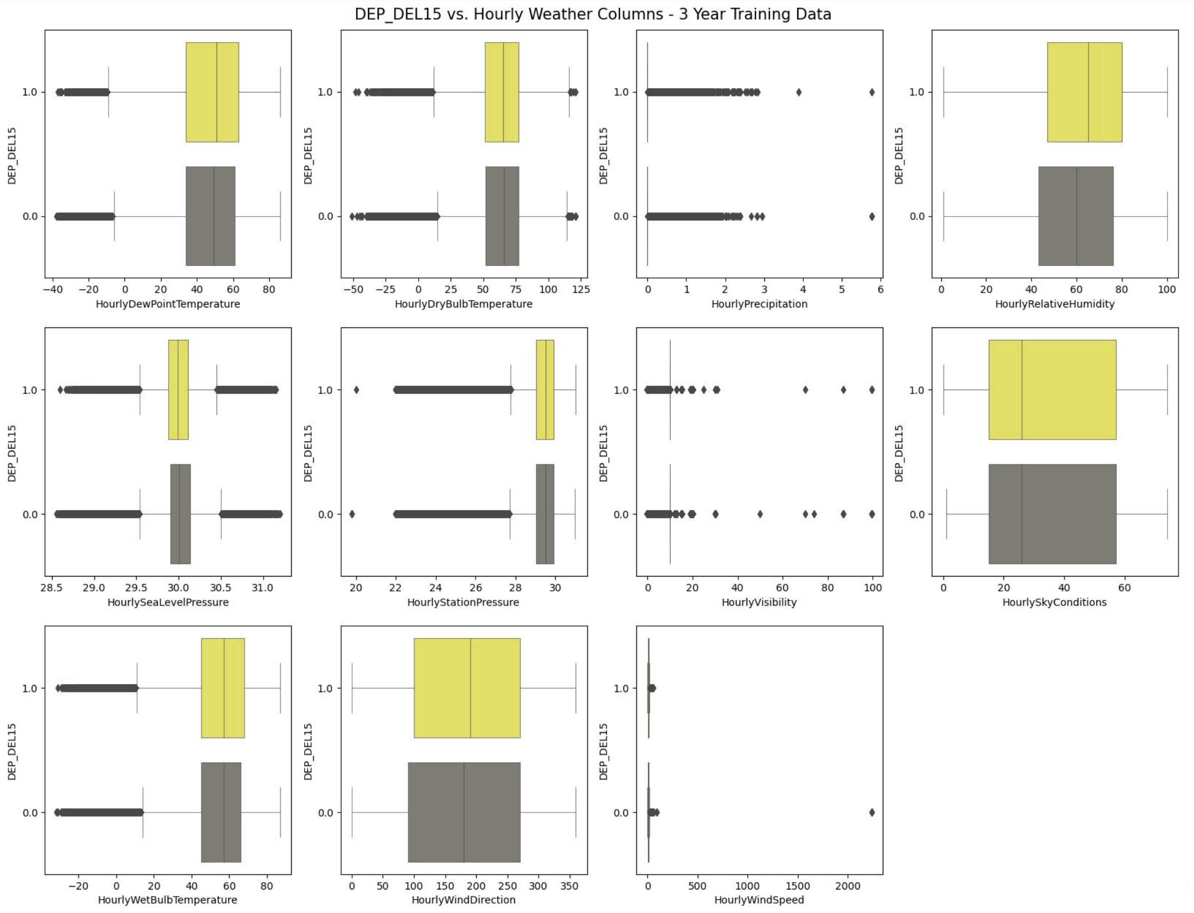




EDA

Weather-associated

- **HourlyWindSpeed:** The median wind speed for delayed flights is higher than that for non-delayed flights.
- **HourlyVisibility:** Notable difference - with delayed flights having a lower median.
- **HourlyPrecipitation:** Instances of delay at higher precipitation levels.
- **HourlyRelativeHumidity:** The median relative humidity is higher for delayed flights.



Feature Engineering

Seasonality-based Features:

- The number of days before/after the most impactful holidays

5_DAYS_DIST_FROM_Independence
7_DAYS_DIST_FROM_Christmas
7_DAYS_DIST_FROM_NewYear

Time-based Features

- Capture the cyclical hidden trends in flight delays

del15_2hr_before: *Delay 15 minutes or greater in the past 2 hours, not just the direct previous flight (1 = yes, 0 = no)*
TIME_INTERVAL_OF_DAY: *4 clusters of day using KMeans*
DAY_TYPE *(1 = weekend, 0 = weekdays)*
DAY_OF_WEEK_sin & **DAY_OF_WEEK_cos**
CRS_DEP_HOUR_sin & **CRS_DEP_HOUR_cos**
DAY_HOUR_interaction

Operational-based Features

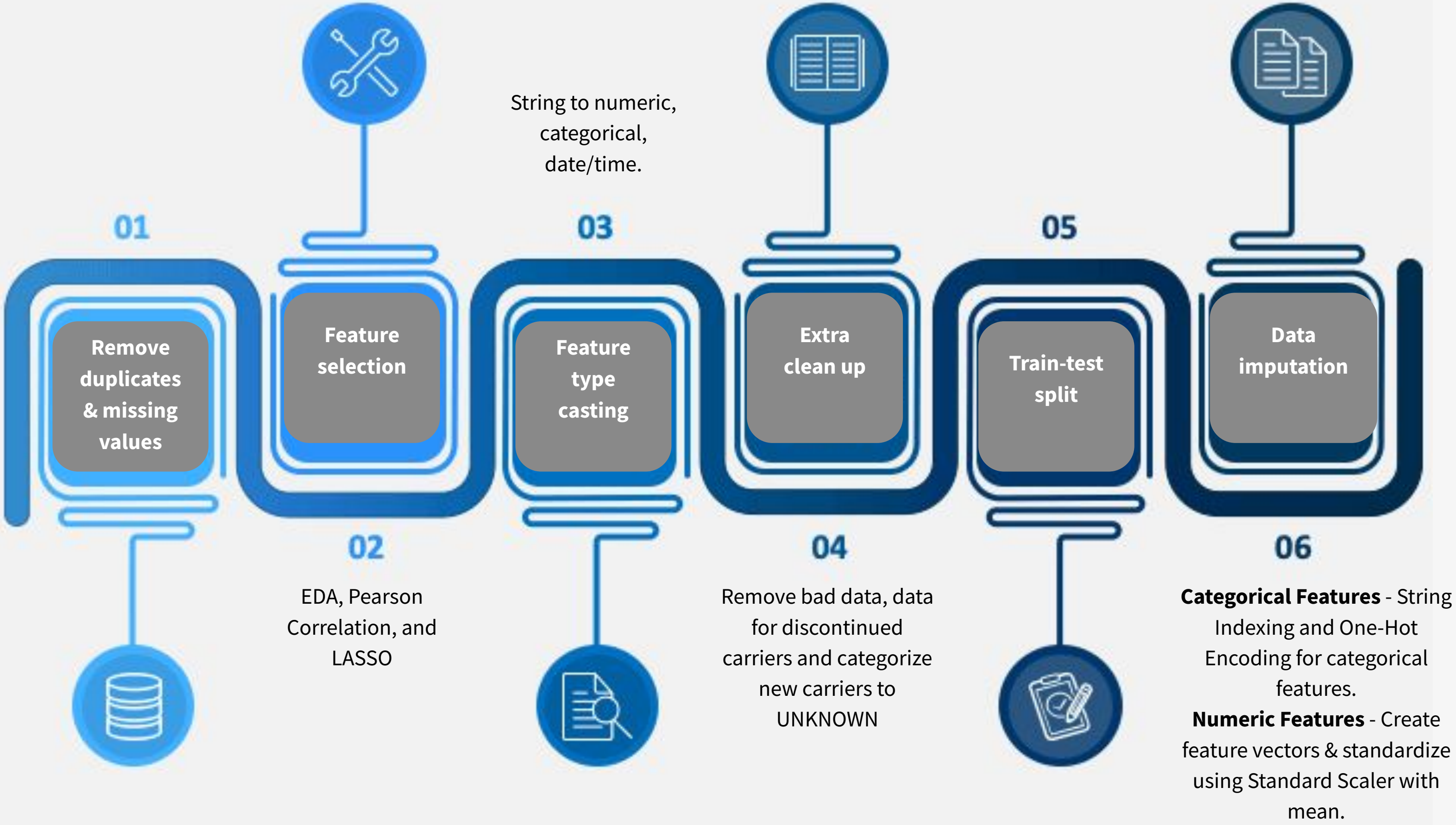
- Influence on the logistics and management of flight operations

FL_DISTANCE_GROUP: *Grouped by travel distance*
CARRIER_SIZE: *Carrier Size by the Number of Flights*

03

Modeling Pipeline

Data Pre-Processing



Data Pre-Processing

Pre-Processing Step	Further Details
Remove outliers or missing values	
Feature selection	• Use Pearson Correlation and LASSO to select relevant features.
Feature type casting	• String to numeric, categorical, date/time.
Extra clean up	• variable formatting and removal of canceled flights.
Train-test split	• Specified further in next slide.
Data imputation	• Categorical Features - String Indexing and One-Hot Encoding for categorical features. • Numeric Features - Create feature vectors & standardize using Standard Scaler with mean.

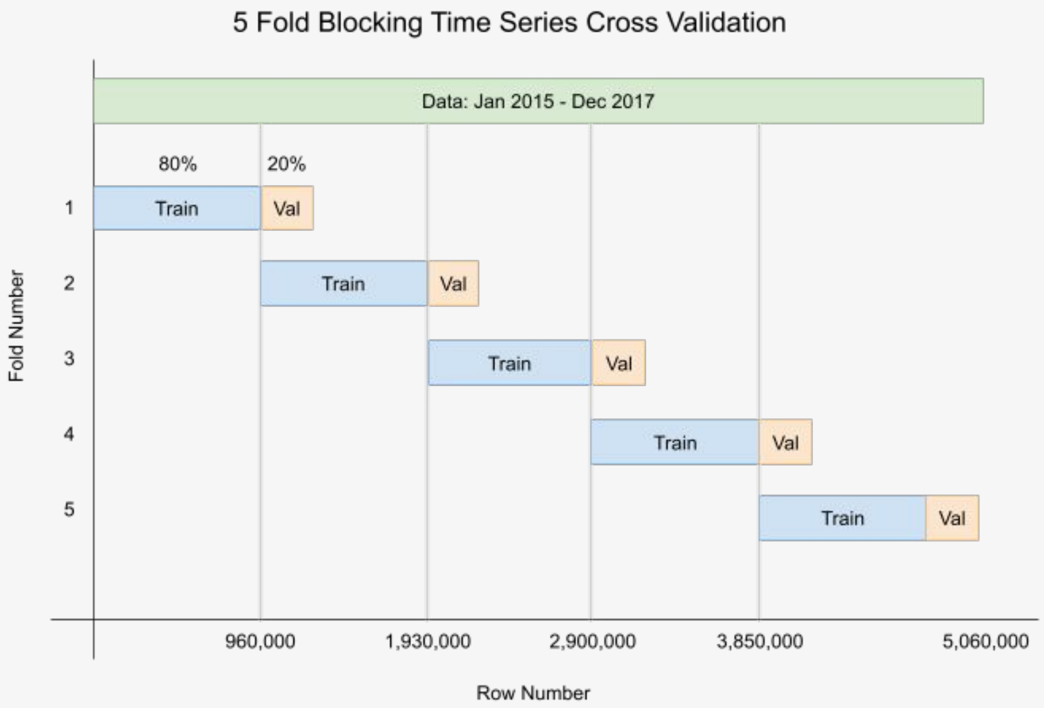
Cross Validation Split for Time Series

Data Splits

- Train set – 2015-2017 ~ 6 million rows (after undersampling)
- Val set – 2018 ~ 7 million rows
- Test held-out set – 2019 ~ 7.4 million rows

OTPW 5 Year Cross Validation

- 5 Folds – 80/20 train/val per fold
- Train size per fold ~ 960k rows
- Val size per fold ~ 240k rows



Evaluation Metrics

Primary Metric:

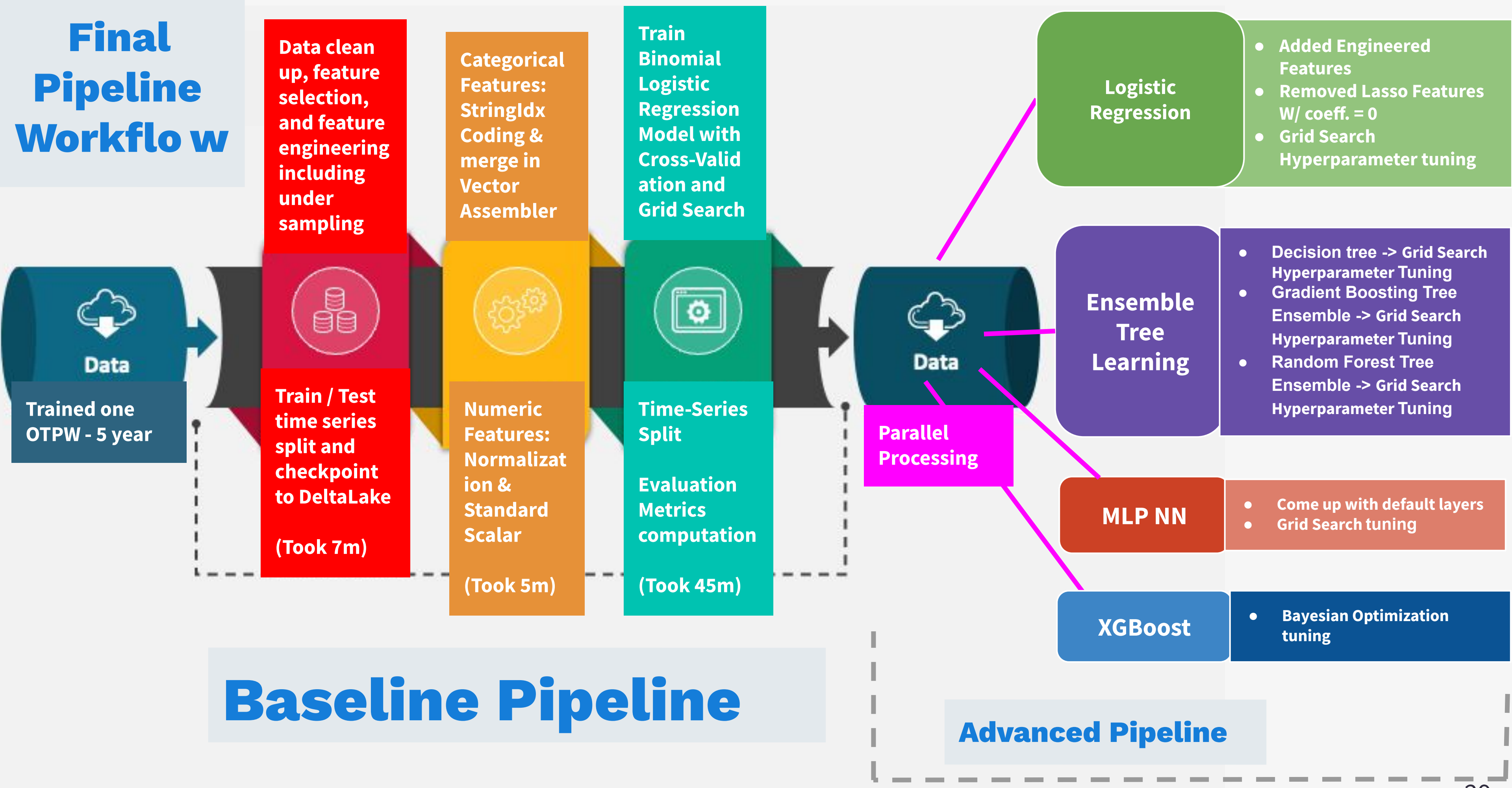
- F2 Score

Secondary Metrics for Results based analysis:

- Recall
- Precision

Beta Value	5-fold CV F-Beta Score
1.1	0.6054
1.2	0.6059
1.3	0.6065
1.4	0.6072
1.5	0.6079
1.6	0.6087
1.7	0.6094
1.8	0.6101
1.9	0.6108
2.0	0.6114

Final Pipeline Workflo w

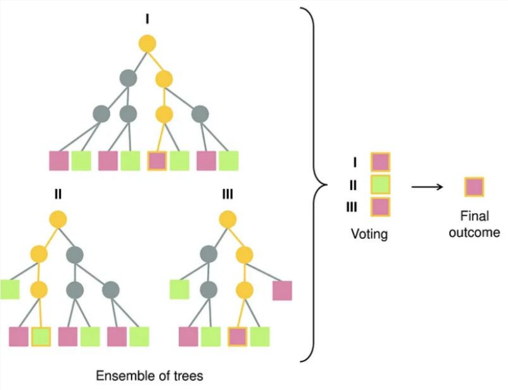


Decision Tree and Ensembles

Basic Decision Tree

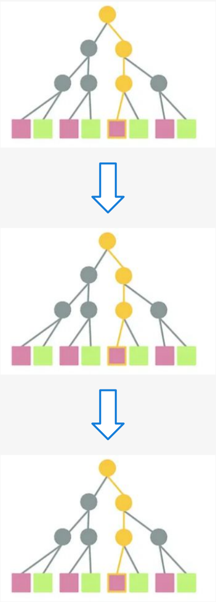
1. Start from top
2. Make a decision
3. Split into branches
4. Keep asking
5. Leaf nodes

Random Forest Tree Ensemble



1. Build Multiple Trees
2. Randomness
3. Voting

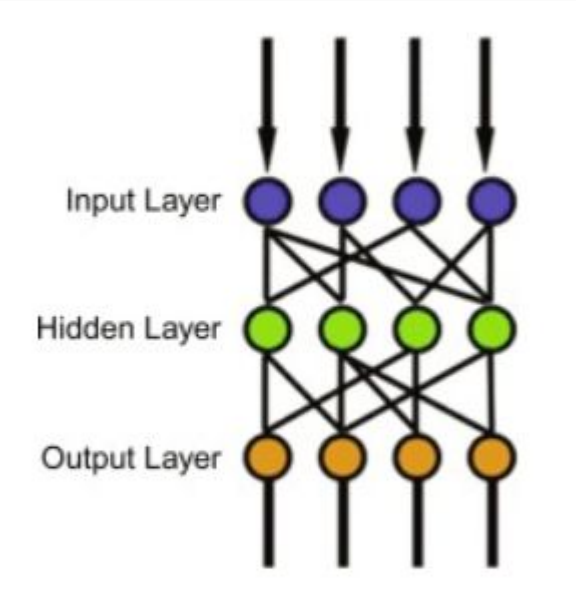
Gradient Boosting Tree Ensemble



1. Building Trees Sequentially
2. Learning from Errors
3. Combining Predictions

Multi-layer Perceptron Classifier

Multilayer perceptron classifier (MLPC) is a classifier based on the feedforward artificial neural network. MLPC consists of multiple layers of nodes. Each layer is fully connected to the next layer in the network.



- Layers = [31, 64, 32, 16, 8, 2]
- weights w and bias b with K+1 layers

$$\mathrm{y}(\mathbf{x}) = \mathrm{f}_K(\ldots \mathrm{f}_2(\mathbf{w}_2^T \mathrm{f}_1(\mathbf{w}_1^T \mathbf{x} + b_1) + b_2) \ldots + b_K)$$

- 4 hidden layers (Sigmoid)

$$\mathrm{f}(z_i) = \frac{1}{1 + e^{-z_i}}$$

- 1 output layer (Softmax)
- 2 Classes

$$\mathrm{f}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}}$$

04

Results & Next Steps

Results

Training:
F2 Score: 0.6351

Recall: 0.6357
Precision: 0.6351

Validation:
F2 Score: 0.6114

Recall: 0.6216
Precision: 0.6282

Confusion Matrix:

	TRUE	
PREDICTION	Not Delayed (0)	Delayed (1)
Not Delayed (0)	258,806 34.98%	113,751 15.37%
Delayed (1)	166,211 22.47%	201,086 27.18%

Results on Train vs Validation

Model Granularity	Best Hyperparameter	F2 Score	Recall	Precision	Training Time
Baseline	maxIter, [100], regParam, [0.01] elasticNetParam, [0.0]	Train: 0.6503 Val: 0.5476	Train: 0.6569 Val: 0.5497	Train: 0.6719 Val: 0.7939	9.495 min
Logistic w/ engineer features	maxIter, [100], regParam, [0.01] elasticNetParam, [0.0]	Train: 0.5488 Val: 0.5751	Train: 0.5181 Val: 0.5264	Train: 0.7196 Val: 0.9124	9.917 min
Basic Decision Tree	impurity, ['entropy'], maxDepth, [30] minWeightFractionPerNode, [0.08]	Train: 0.6215 Val: 0.6000	Train: 0.6130 Val: 0.5538	Train: 0.6583 Val: 0.9002	13.325 min
Gradient Boosting Tree Ensemble	maxIter, [100], maxDepth, [30] stepSize, [0.1] minWeightFractionPerNode, [0.08]	Train: 0.6204 Val: 0.6300	Train: 0.6009 Val: 0.5849	Train: 0.7128 Val: 0.9114	1.13 hrs
Random Forest Tree Ensemble	impurity, ['entropy'],numTrees, [50] maxDepth, [30], bootstrap, [True] minWeightFractionPerNode, [0.03]	Train: 0.5377 Val: 0.5399	Train: 0.5059 Val: 0.4894	Train: 0.7183 Val: 0.9194	21.6 min
Multilayer Perceptron NN	maxIter, [100], blockSize, [128] stepSize, [0.03]	Train: 0.5902 Val: 0.6075	Train: 0.5672 Val: 0.5614	Train: 0.7045 Val: 0.9047	3.11 hrs

Final Model Results on Test Set

- Re-train Gradient Boosting Tree Ensemble model on 2015-2018
- Held-Out Test on 2019

Model Granularity	Best Hyperparameter	F2 Score	Recall	Precision	Training Time
Gradient Boosting Tree Ensemble	maxIter, [100] maxDepth, [30] stepSize, [0.1] minWeightFractionPerNode, [0.08]	Train: 0.6203 Test: 0.6407	Train: 0.6019 Test: 0.5969	Train: 0.7069 Test: 0.9066	1.48 hrs

Leakage

Model Granularity	F2 Score	Leakage?
Baseline	Train: 0.6503 Val: 0.5476	
Logistic added engineered features	Train: 0.5488 Val: 0.5751	
Basic Decision Tree	Train: 0.6215 Val: 0.6000	
Gradient Boosting Tree Ensemble	Train: 0.6204 Val: 0.6300	
Random Forest Tree Ensemble	Train: 0.5377 Val: 0.5399	

Results

Area under ROC curve

F Beta score:
0.5808588926654742

Confusion Matrix:

	Predicted	
TRUE	Not Delayed (0)	Delayed (1)
Not Delayed (0)	1034100 99.90%	1189 0.60%
Delayed (1)	264704 99.30%	1753 0.10%

Discussion of Results

Performance & Scalability Concerns

Performance & Scalability

- Used Delta Lake which gave efficiency boost over parquet
- We learned that utilizing cache() at the right places is very critical!
- Many Variables act as proxies for the others

Feature Engineering (cont'd)

- **Operational Features**
  - ***Distance wise (in Miles):***
    - Short distance (1): < 749
    - Medium distance (2): >= 750 and < 1249
    - Long-haul distance (3): >= 1250

- **Weather-related Features**
  - High (0) or Low (1) ***hourly visibility*** with threshold of < 10

05

Conclusion & Next Steps

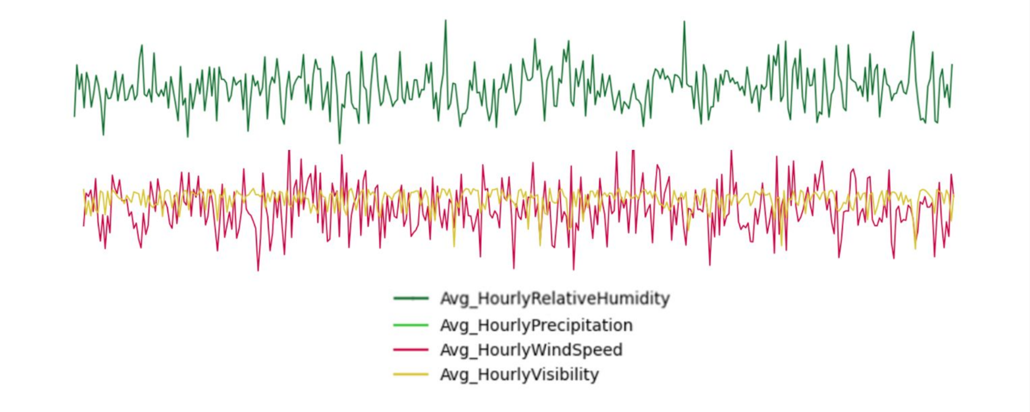
Concluding Remarks

- Best performing model
  - **Gradient Boosting Tree Ensemble -** sequential complexity
- Number of features - 31
  - Opportunity to slim down
- Next steps
  - Significant improvement on Precision
  - Less significant improvement on Recall

Top 10 Best Features - Based on LASSO Coeff.
CRS_DEP_TIME
HourlyRelativeHumudity_Imputed
HourlyWindSpeed_imputed
6_DAYS_DIST_FROM_NewYear_idx
CRS_ELAPSED_TIME_imputed
HourlyPrecipitation_imputed
6_DAYS_DIST_FROM_Christmas_idx
HourlyVisibility_imputed
TIME_INTERVAL_OF_DAY
CRS_DEP_HOUR_sin

Future Work

- Advanced imputation & future work
  - Seasonal ARIMA - more advanced approach
    - Currently debugging
- Anomaly outlier detection



Thanks!

Do you have any questions?

Contact:
Heesuk Jang – jheesuk@ischool.berkeley.edu
Karsyn Lee – karsyn@ischool.berkeley.edu
Stephanie Cabanela – scabanela@ischool.berkeley.edu
Raymond Tang – raymond.tang@ischool.berkeley.edu