Flight Ahead - Advanced Predictive Models for Flight Delays

261 Machine Learning at Scale, December 14, 2023

Team 5-1: **Carolyn Dunlap - Phase III Leader**
Hamsini Sankaran, Jasmine Teo, Siva Thiyagarajan

Target Audience: Commercial Airlines

Outline

- Introduction
- Exploratory Data Analysis
- Modeling Pipeline
- Feature Engineering
- Key Model Results
- Discussion
- Conclusion

Team


Carolyn


Hamsini


Jasmine


Siva

Flight Delays

Problem:

- **20.8%** of flights delayed so far in 2023[1]
- FAA/Nexor estimated total cost of delays was **$28 billion** in 2018[2]
- Flights delays are **increasing** each year[1]

Solution:

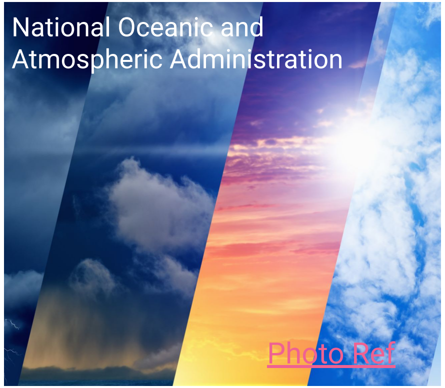
Large-scale, machine learning model to predict flight delays before they occur

[1] Bureau of Transportation Statistics
[2] https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/

Data

Flights

Weather



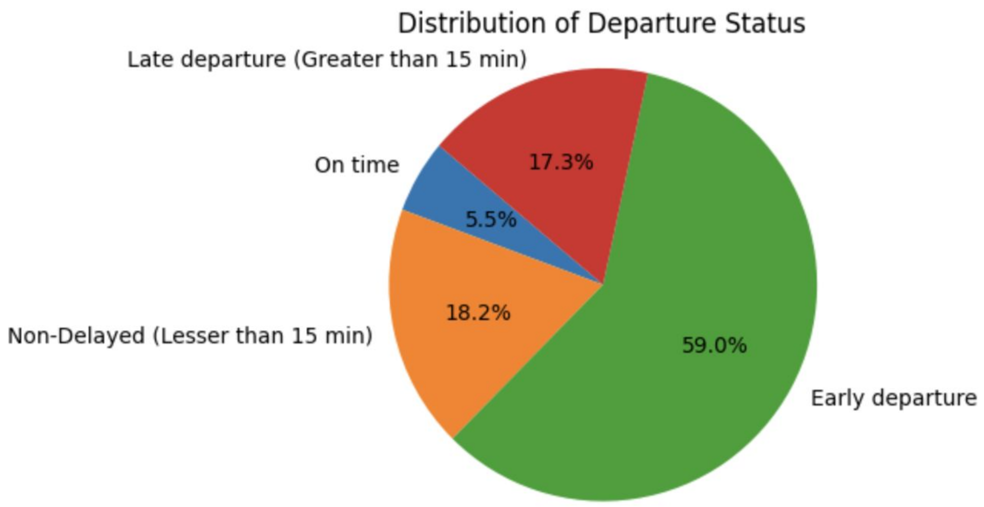
Airport Metadata

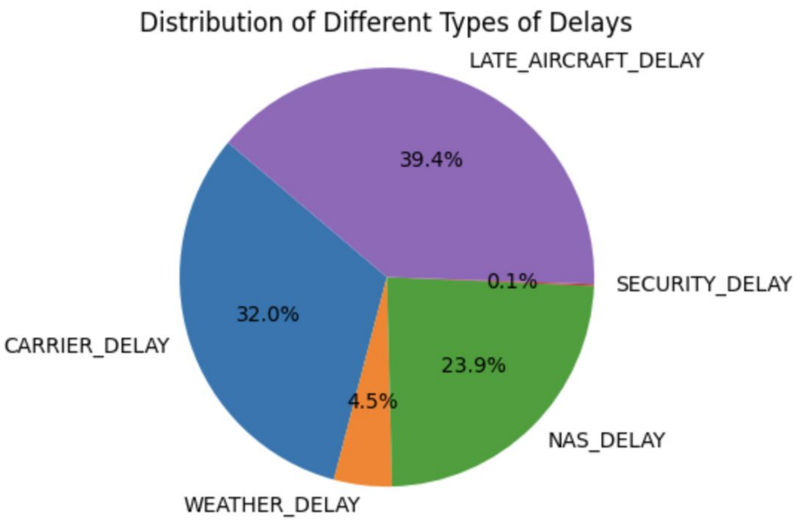
216 total features
5 years (2015 - 2019)
31,746,841 flights

Key target variable:
Departure delay of more than 15 minutes

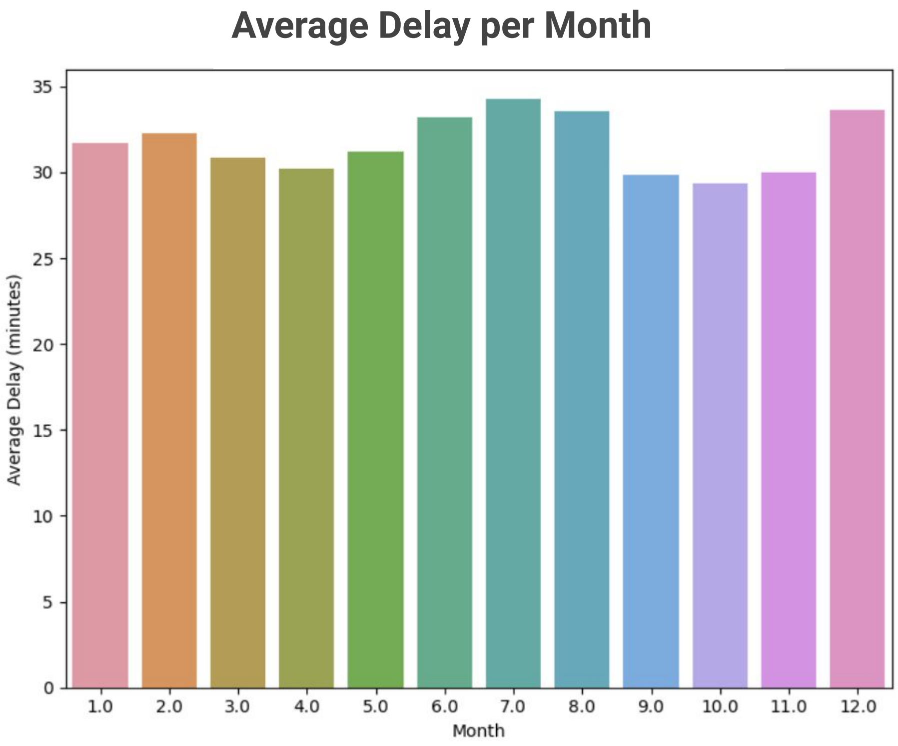
Flight Delay Metrics

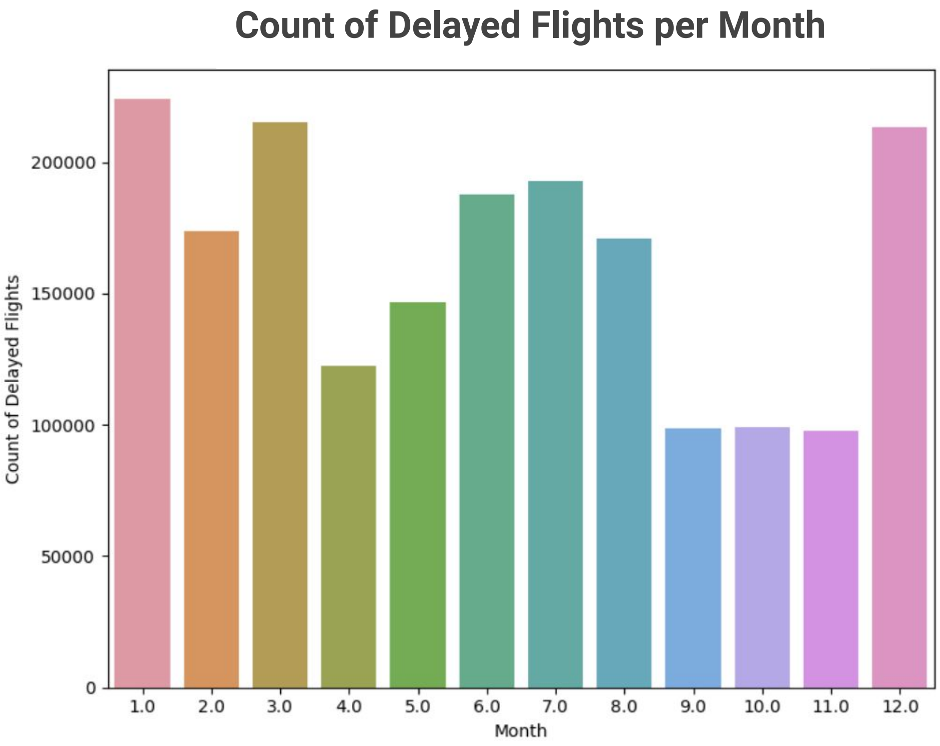
3-Year Data (2015-2018)





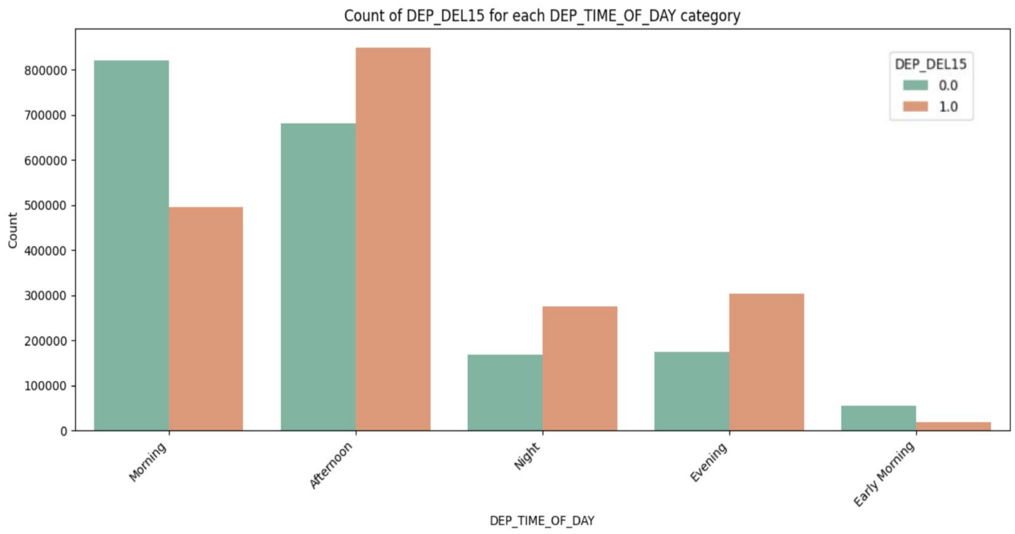
Flight Delay Metrics by Month



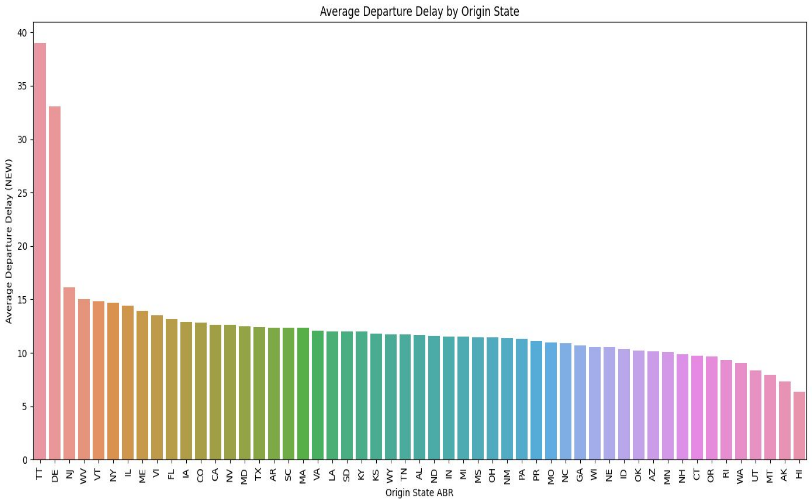


Flight Delays by Time of Day and Origin State

Number of Flights Delayed By Departure Time Block

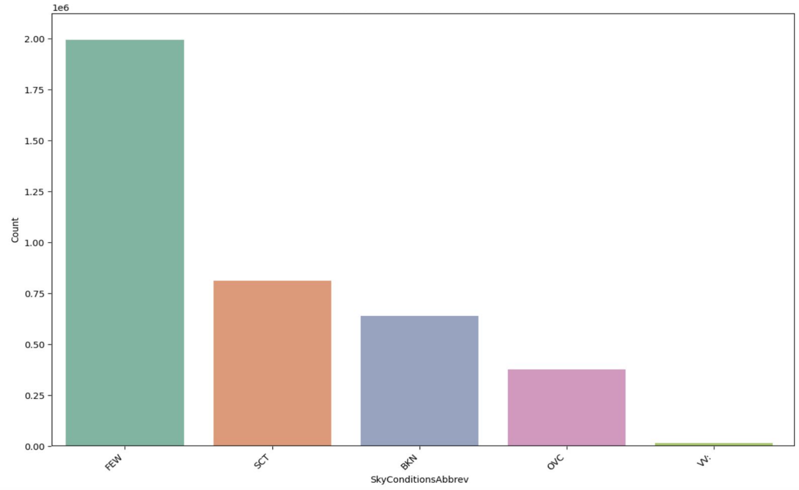


Average Departure Delay by Origin State



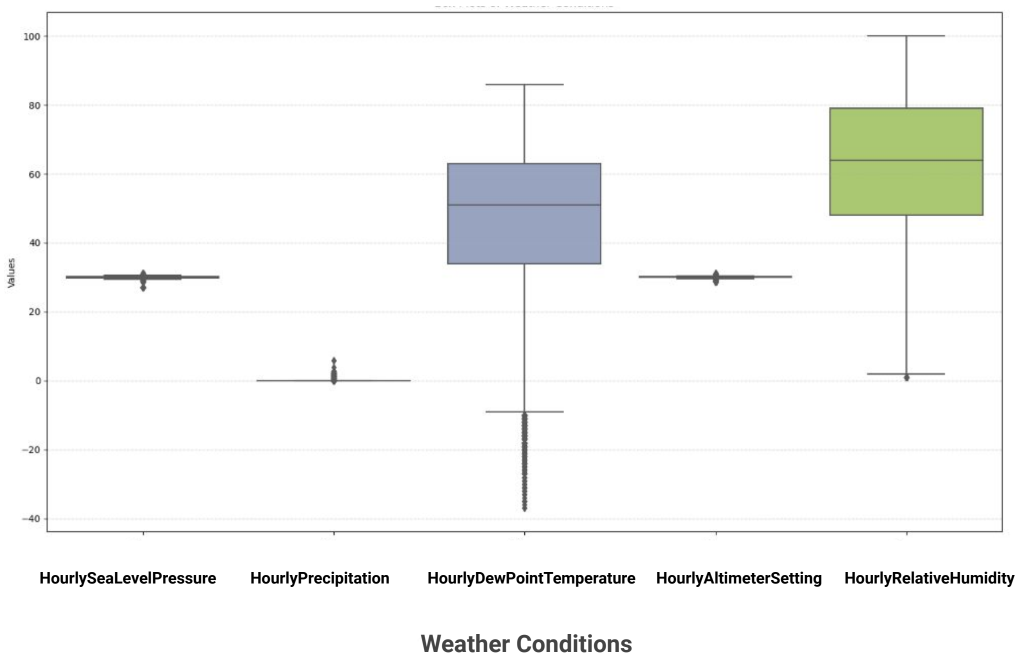
Weather & Sky Effects on Flight Delays

Distribution of Sky Conditions

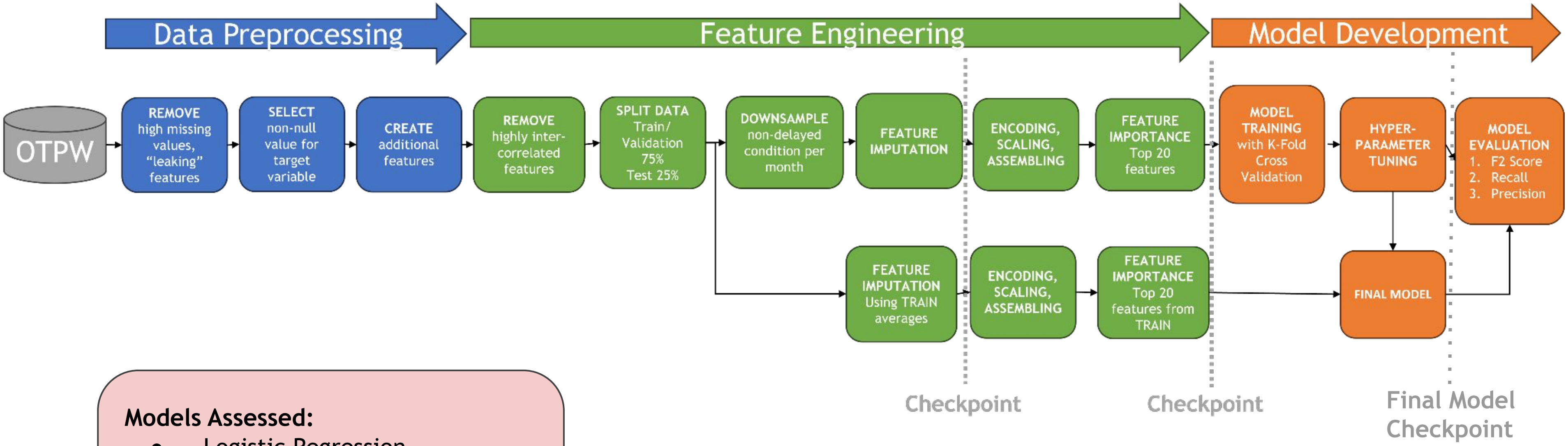


FEW: Few clouds (1/8 to 2/8 of the sky covered)
SCT: Scattered clouds (3/8 to 4/8 of the sky covered)
BKN: Broken clouds (5/8 to 7/8 of the sky covered)
OVC: Overcast clouds (8/8 of the sky covered)
VV: Vertical visibility (sky is completely obscured, often due to fog or heavy precipitation)

Boxplots of Weather Conditions



Our Modeling Pipeline



Models Assessed:

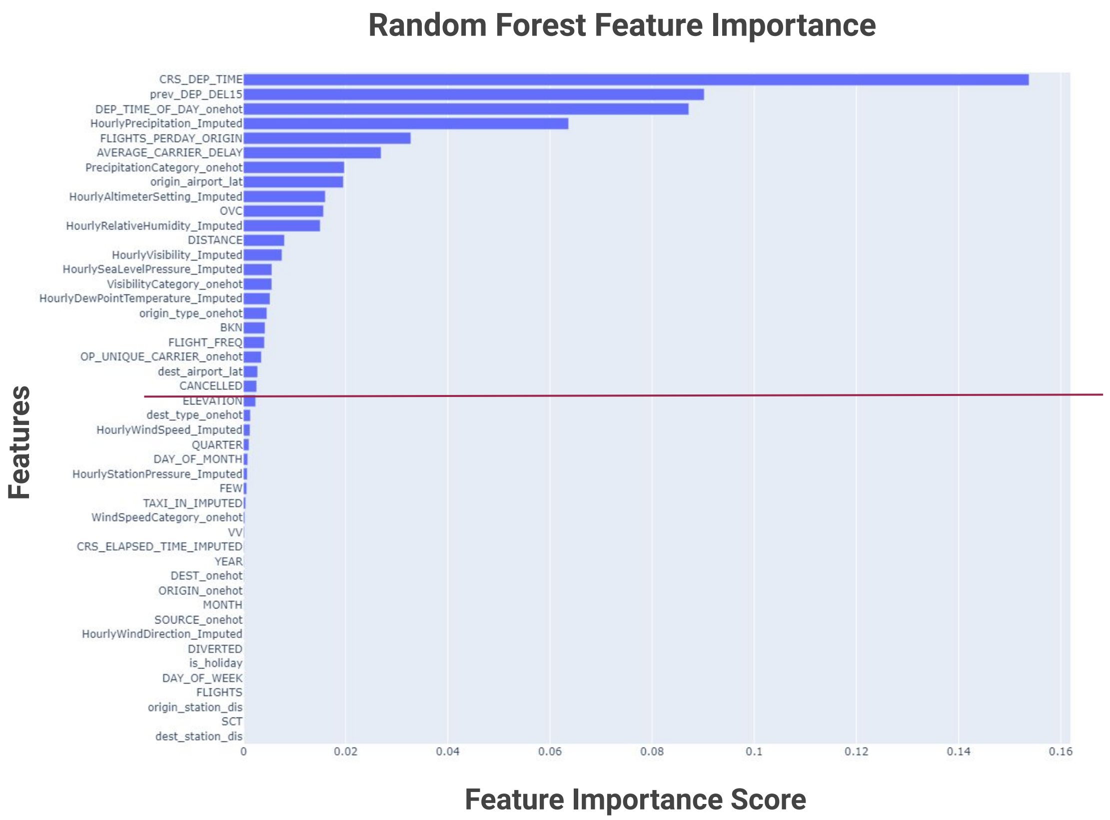
- Logistic Regression
- Random Forest
- Multi-layer Perceptron (MLP)
- Gradient-boosted Decision Tree (GBT)

Feature Engineering Decisions

- Exclude features with many missing values
- Exclude features for data leakage
- Design new features
- Impute features having missing values
- Exclude intercorrelated numerical features
- Select top features

Derived Feature	Raw Feature
Flight Frequency	Tail Number
Flights per day at Destination airport	Day of week, Carrier, Destination
Flights per day at Origin airport	Day of week, Carrier, Origin
Average Carrier Delay	Departure Delay, Carrier
Departure time of the day	Scheduled Departure Time block
Arrival time of the day	Scheduled Arrival Time Block
Prior Departure delay (> 2hrs)	Departure Delay
Holiday (-5/+2 days)	Flight Date
Few Clouds, Scattered Clouds, Broken Clouds, Over Cast, Vertical Visibility	Hourly Sky Conditions

Final Features



Flight

- Scheduled departure time
- Hourly relative humidity
- Origin airport latitude
- Flight distance
- Origin airport type (large/small)
- Airline carrier
- Cancelled flight
- Previous flight departure delay (yes/no)
- Average carrier delay
- Flight frequency
- Departure time of day
- Flights per day at the origin

Weather

- Hourly Precipitation
- Overcast sky conditions
- Hourly altimeter setting
- Hourly dew point temperature
- Hourly visibility
- Broken clouds sky condition
- Hourly sea level pressure

Results on 5 Years Data

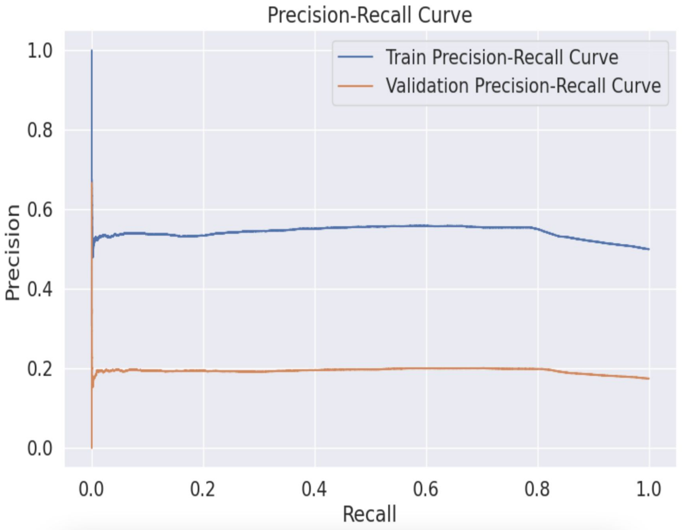
Model	Cross Fold Evaluation							Full Evaluation					
	Train				Validation			Full Training			Blind validation		
	F2 SCORE	RECALL	PRECISION	Training Time (min)	F2 SCORE	RECALL	PRECISION	F2 SCORE	RECALL	PRECISION	F2 SCORE	RECALL	PRECISION
Logistic Regression	**0.6360**	0.6411	0.6198	22.44	**0.6178**	0.6200	0.6152	**0.6476**	0.6411	0.6114	**0.4875**	0.6560	0.2405
Random Forest	**0.6308**	0.6264	0.6575	29.06	**0.6171**	0.6134	0.6418	**0.6504**	0.6264	0.6330	**0.5010**	0.6628	0.2534
GBT	**0.6111**	0.6008	0.6615	23.86	**0.5969**	0.5869	0.6483	**0.6335**	0.6008	0.6400	**0.4954**	0.6435	0.2580
MLP	**0.6225**	0.6549	0.5309	15.83	**0.6204**	0.6540	0.5287	**0.7645**	0.6549	0.5412	**0.5111**	0.8645	0.1939

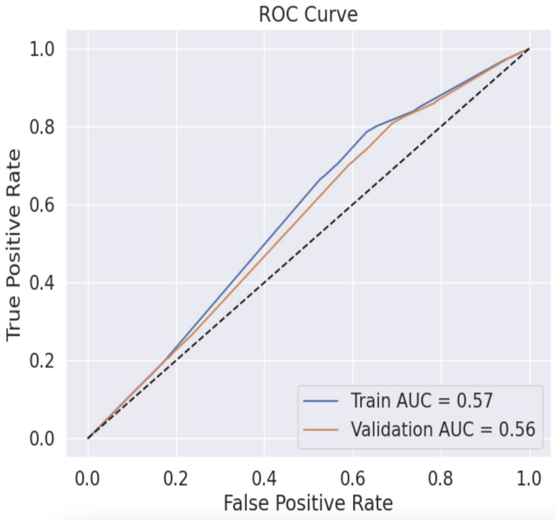
Hyper Parameters:

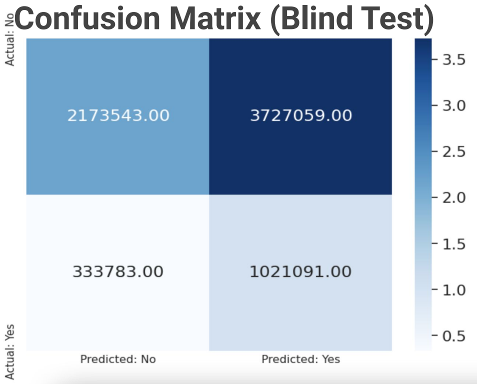
- Logistic Regression: {maxIter=10, regParam=0.01, elasticNetParam=0.7}
- Random Forest: {maxDepth=10, numTrees=20}
- GBT: {maxIter=10, maxDepth=5, maxBins=32}
- MLP: {maxIter=20, layers=[2 hidden layers], blockSize=128}

Best Model Results

Best Model	Parameters	F2 score	Recall	Precision	AUC
MLP	**Iteration- 20, Features - 40 Hidden layers - 2 Neurons - [10, 5]**	0.5021	0.7536	0.2151	0.5620







Discussion And Future Work

1. **Best Model:** MLP excels in predicting flight delays with F2 score of 50%, 75% recall and 21% precision

2.Advanced Feature Engineering

3. Scalability and Adaptability

4 . Deploy the model and integrate it on a flask web application

Cost Analysis

Average Delay per month (32 minutes)

→

Each minute delay cost for airline
$74.20 USD

→

Average number of delayed flights per month
150,000

→

Total Annual cost airline delays
$ 4,273,920,000

Model Development cost
4FTE development (150,000$ for 3 months) + 1 FTE for maintenance (150,000$/year)

→

Overall increase in customer retention

[1] https://www.
[2] https://www.

Conclusion

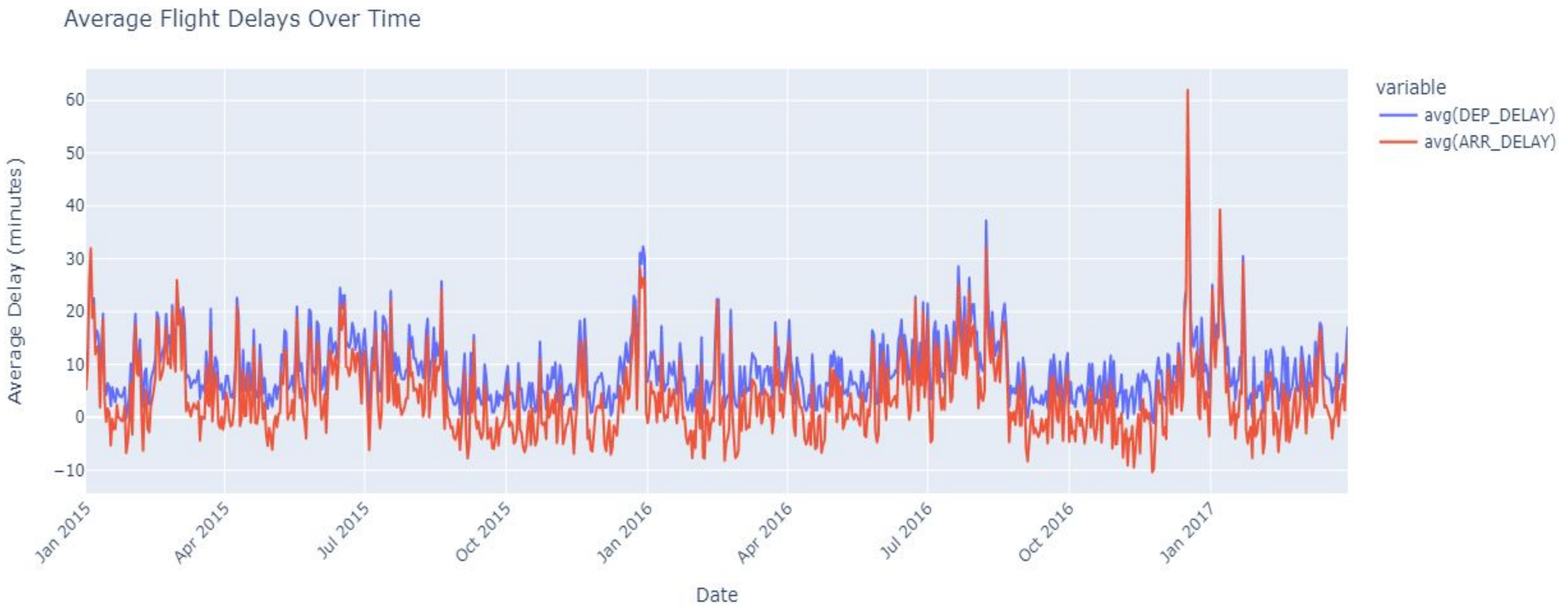
Key takeaways

- Predicted departure delays, enabling airlines to take early action and streamline operations.
- Leveraged advanced key features like weather conditions, carrier delays, flight frequency, previous departure delays, airport characteristics
- Utilized advanced machine learning models, including Multi-Layer Perceptron , Random Forest, Gradient Boosted Decision Trees
- Predicted departure delays, helping airlines to boost efficiency and enhance customer satisfaction, and reduce operational costs.

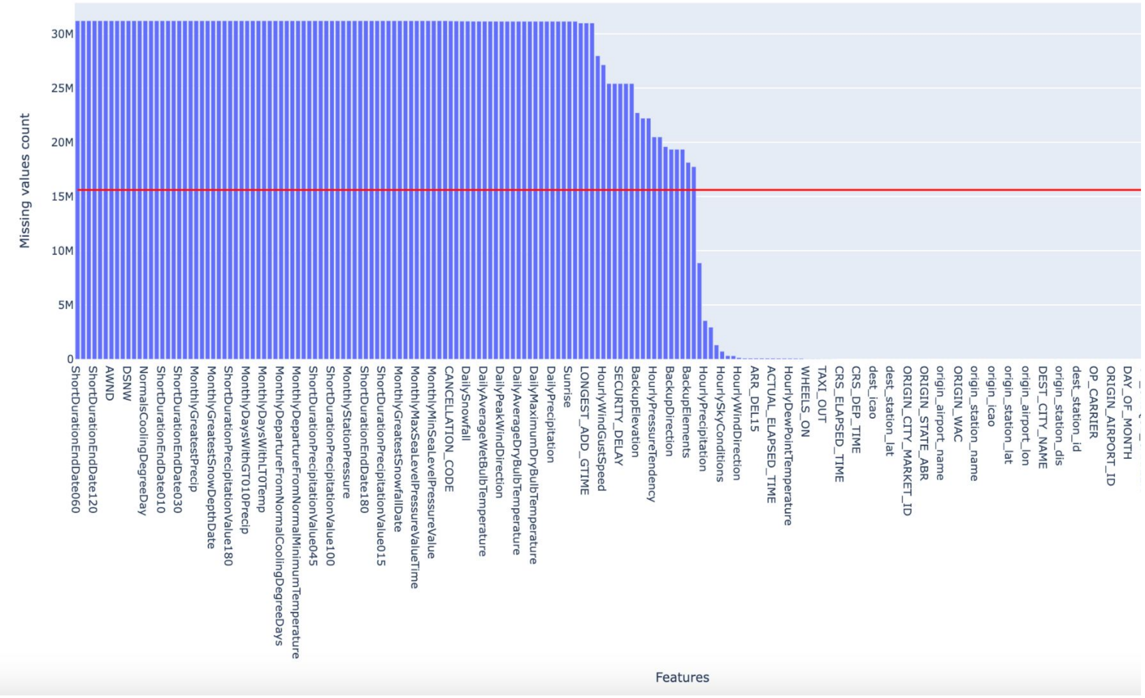
Thank you!

Questions? Comments? Concerns?

Average Flight Delays Over Time



Missing Values by Feature



Features with >50% missing values

- MonthlyStationPressure
- MonthlyGreatestSnowfallDate
- CANCELLATION_CODE
- DailySnowfall
- DailyAverageWetbulbTemperature
- DailyPeakWindDirection
- DailyMaximumDryBulbTEmperature
- DailyPrecipitation
- Sunrise
- HourlyWindGustSpeed
- SECURITY_DELAY
- BackupDirection
- ShortDurationEndDate
- etc.