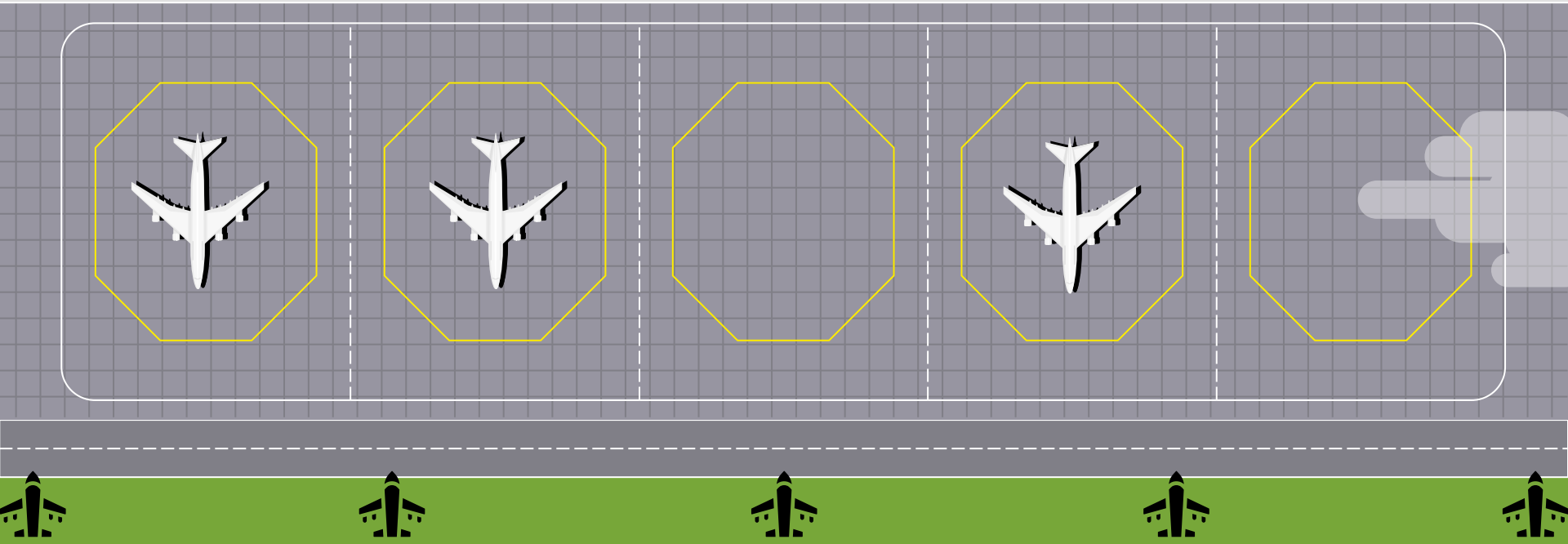
Will My Flight Be Late?

Simon Guichandut, Tim Hallatt, Ketan Sand

Erdos Datascience Bootcamp - Fall 2023



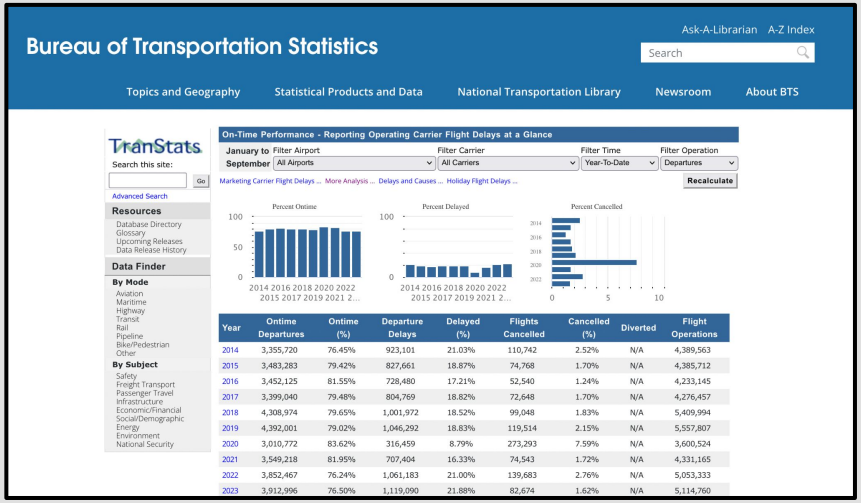
Problem, Stakeholders and KPIs

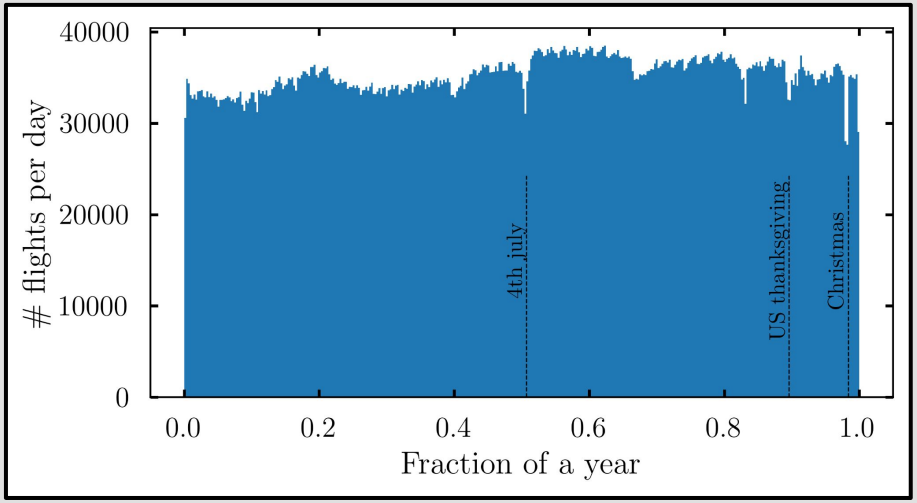
In a single year, there are 200,000 hours of combined flight delays in just the 20 of busiest airports of the United states. Due to all this the US economy suffers a $32.9 billion annual loss.

Stakeholders - Travelers, Airlines, Airport Authorities, Government (Tourism Sectors), Insurance companies.

KPIs - Precision/Recall: False Positive vs. False Negative rates, Receiver Operating Characteristic (ROC), Detection Error Tradeoff (DET)

Data Collection

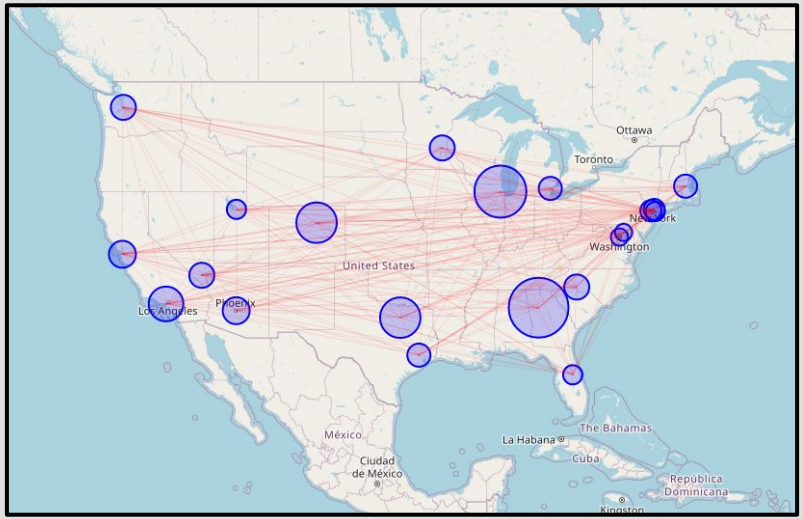




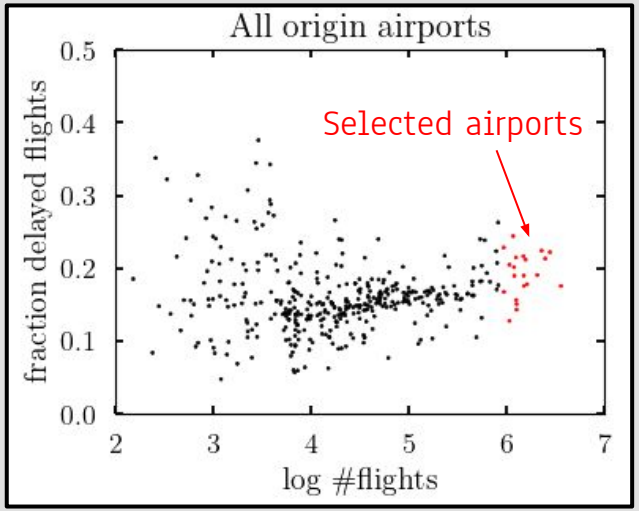
US Bureau of transportation statistics
Data - 2012 - 2023
Automated using Selenium

More than 30,000 flights per day!
Dips on the Holidays

Cleaning

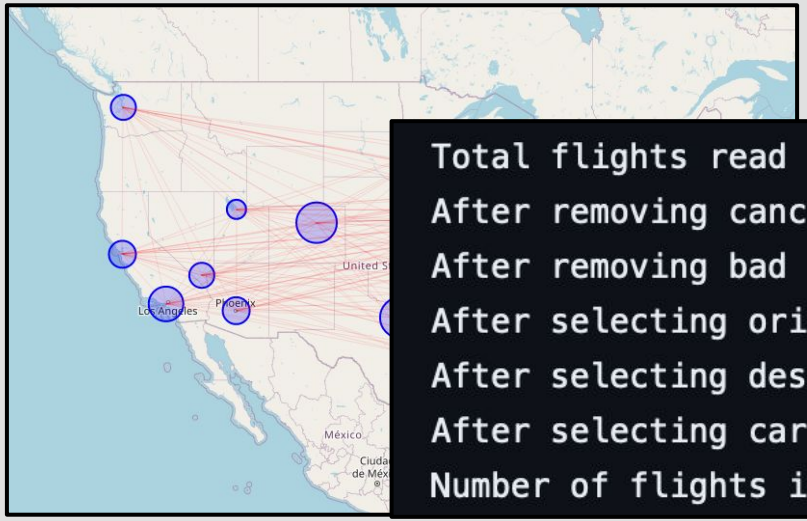


20 Busiest Airports
Removed Cancelled Flight
Top 8 Aircraft Carriers

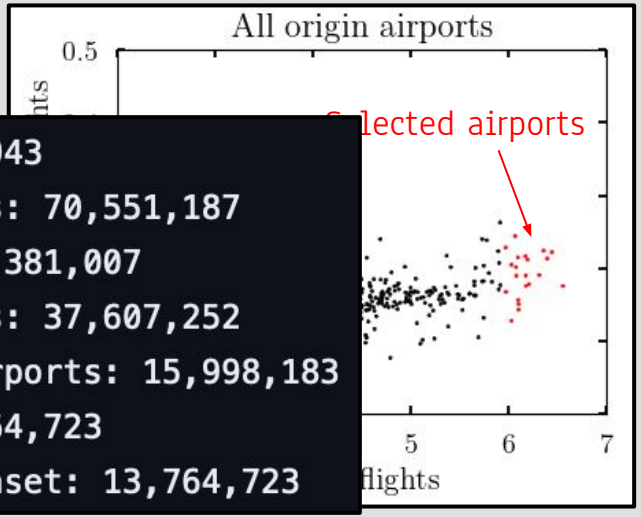


The smallest airports have the most delays: to avoid bias, we take the 20 largest airports to balance number of flights (20% of total) with delays

Cleaning




```
Total flights read in: 71,983,943
After removing canceled flights: 70,551,187
After removing bad columns: 70,381,007
After selecting origin airports: 37,607,252
After selecting destination airports: 15,998,183
After selecting carriers: 13,764,723
Number of flights in final dataset: 13,764,723
```

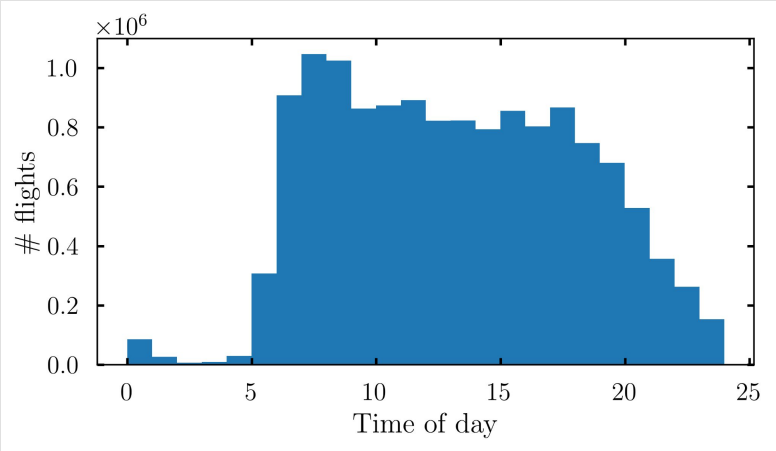


20 Busiest Airports
Removed Cancelled Flight
Top 8 Aircraft Carriers

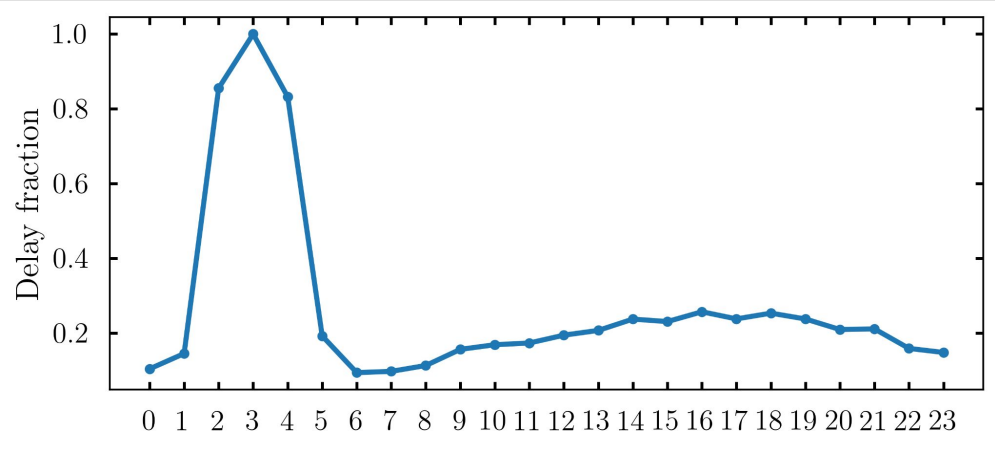
The smallest airports have the most delays: to avoid bias, we take the 20 largest airports to balance number of flights (20% of total) with delays

Exploratory Plots

Number of flights

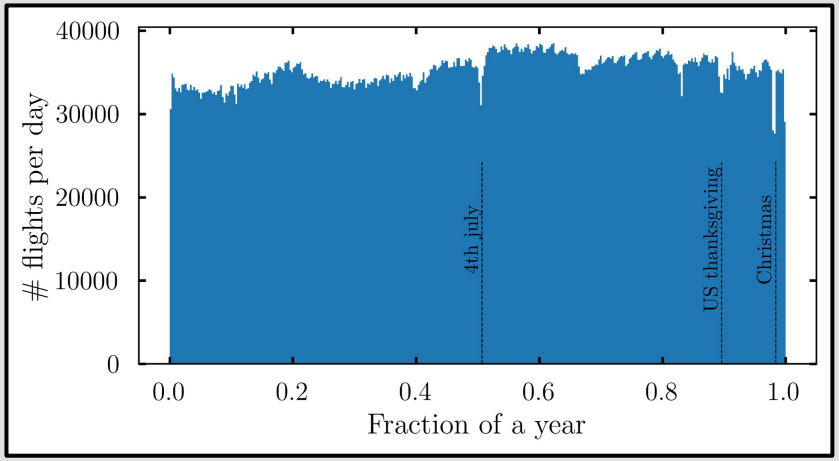


Delay fraction

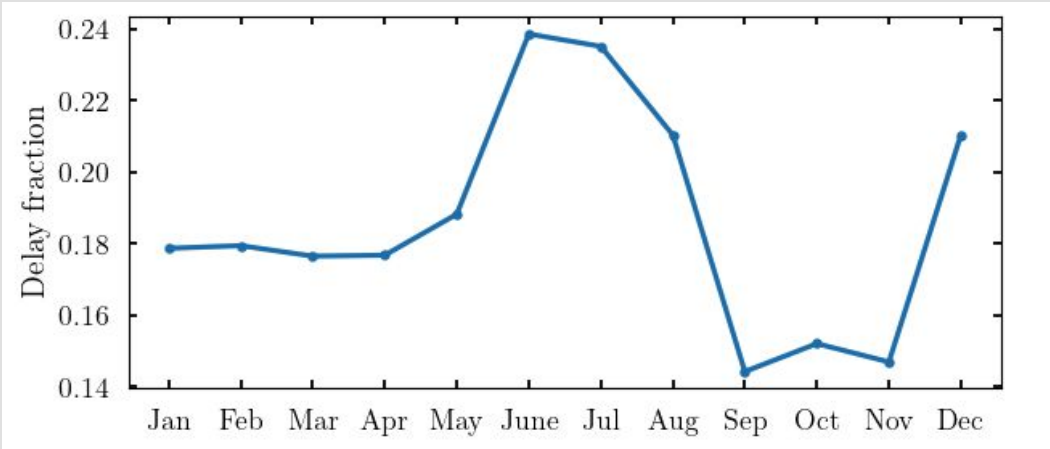


Exploratory Plots

Number of flights

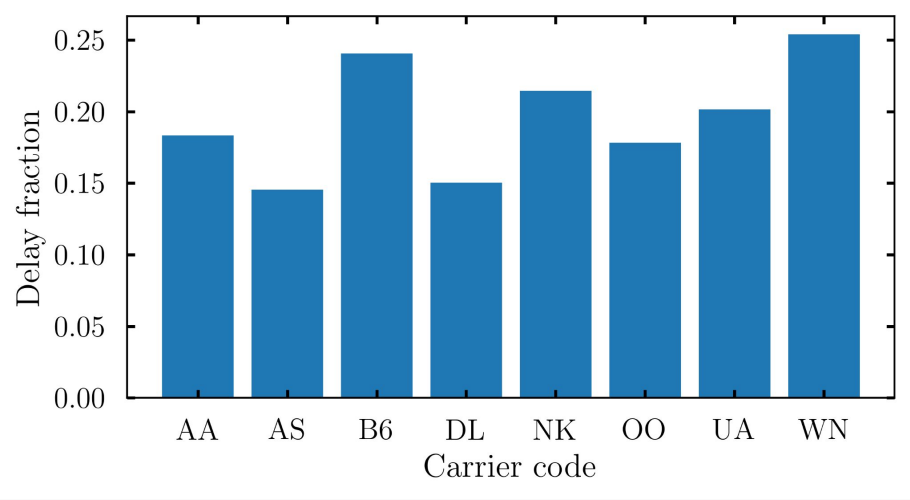


Delay fraction

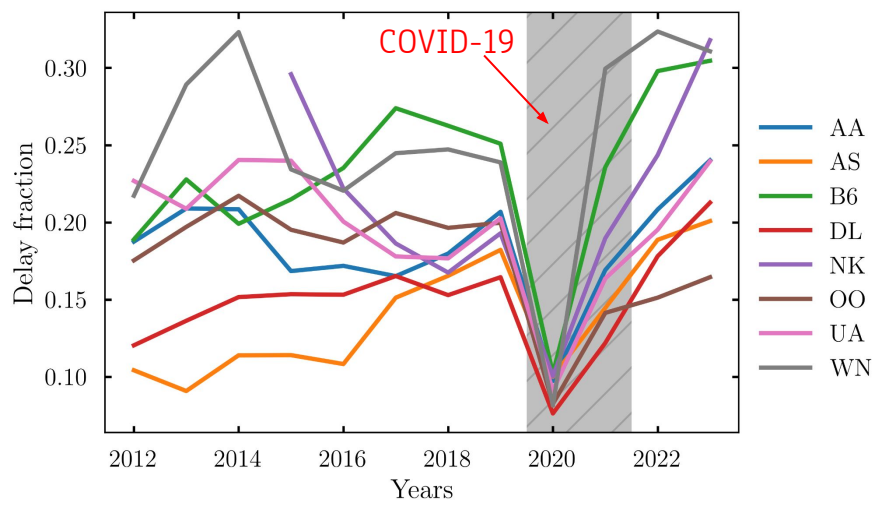


Exploratory Plots

Carrier matters: Southwest (WN) worst, Alaskan Airlines best (AS)

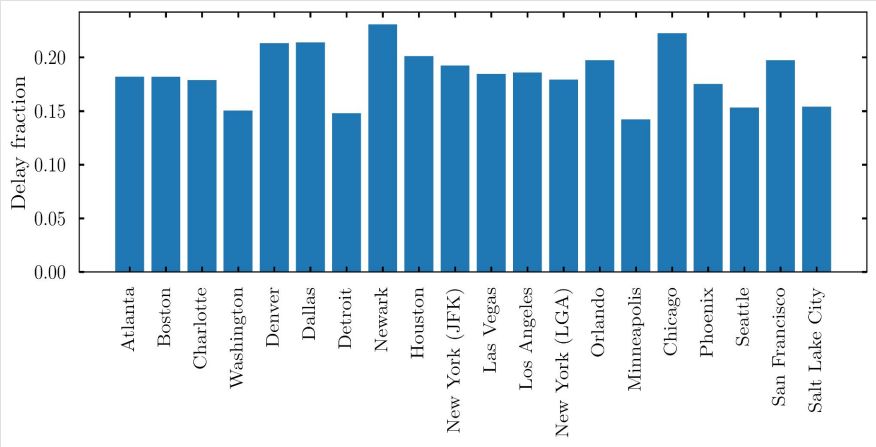


Jetblue (B6) + Southwest (WN) grow worse post-COVID

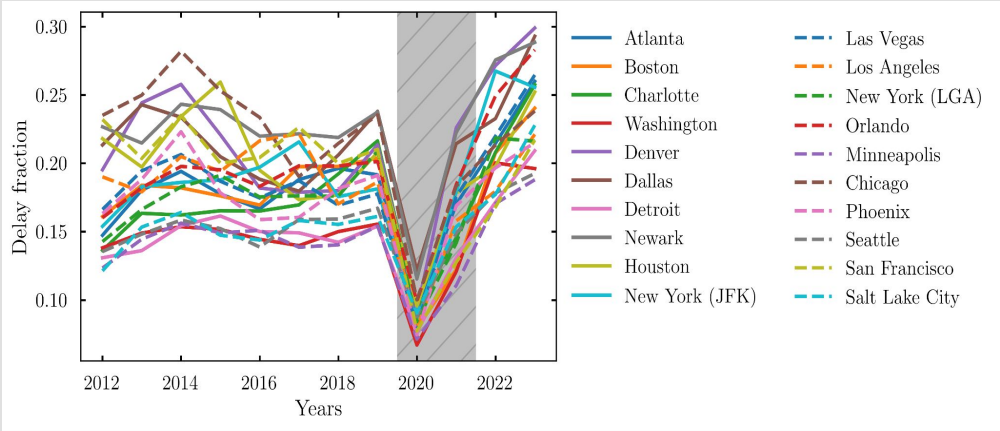


Exploratory Plots

Airport matters, weakly

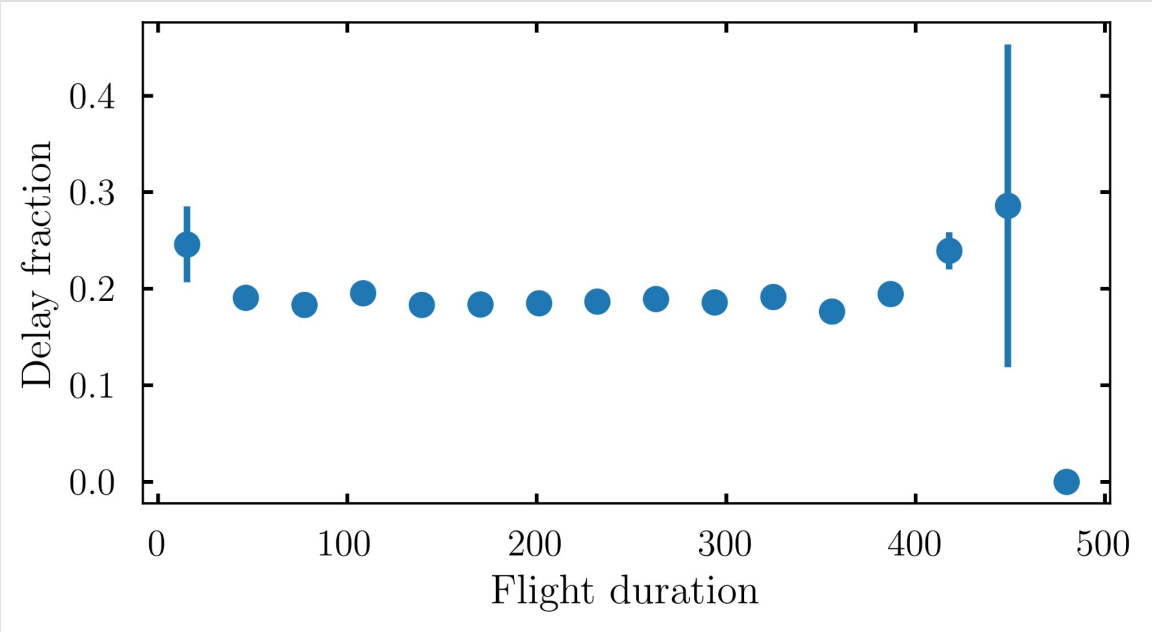


Dallas, Newark, Orlando are bad post-COVID



Exploratory Plots

Flight duration correlates weakly with delay fraction



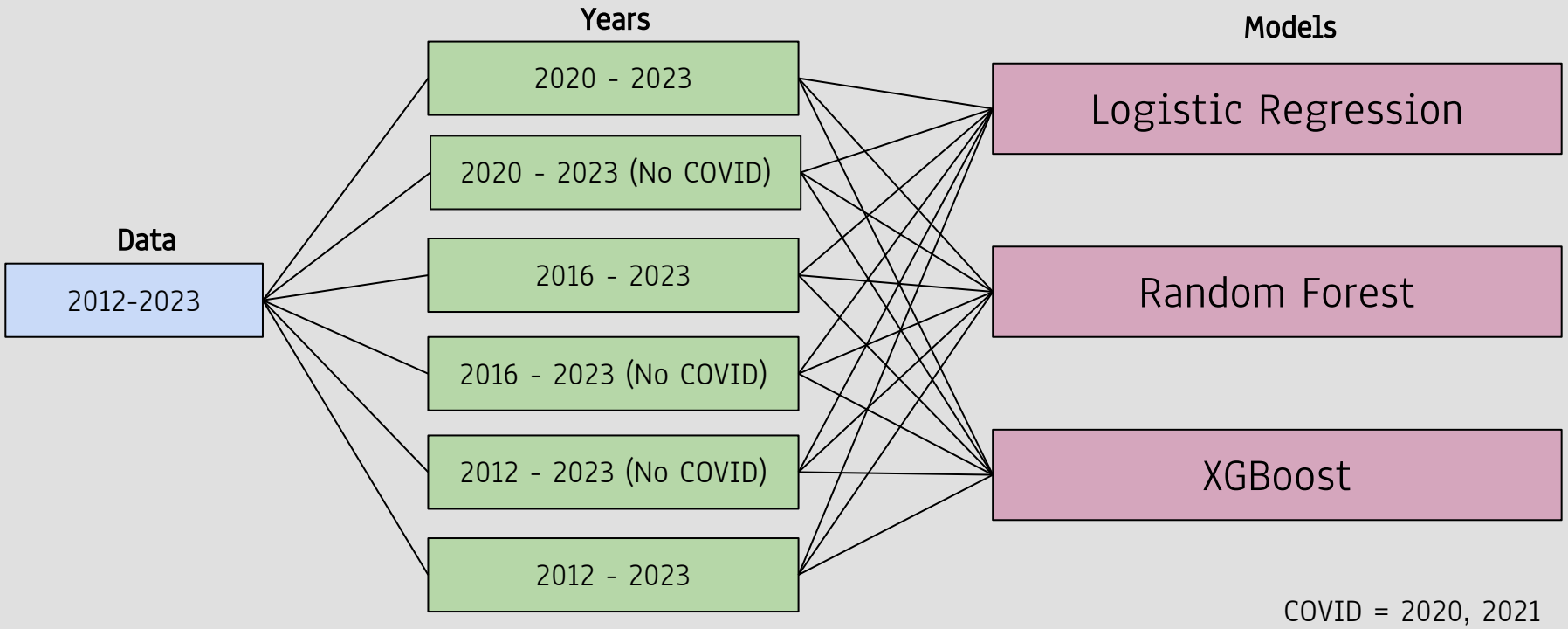
Workflow

Preparing the data

- **Predicted variable**: Delay times >15min are categorized as TRUE
- **Scaling**: Normalize flight duration to 0-1
- **Categorical variables**: One-hot encore carrier and airport variables
- **Cyclic variables**: Convert day & year fraction to sin & cos components

$$\text{day\_frac\_sin} = \sin(2\pi \cdot \text{day\_frac}) \qquad \text{year\_frac\_sin} = \sin(2\pi \cdot \text{year\_frac})$$
$$\text{day\_frac\_cos} = \cos(2\pi \cdot \text{day\_frac}) \qquad \text{year\_frac\_cos} = \cos(2\pi \cdot \text{year\_frac})$$

Workflow

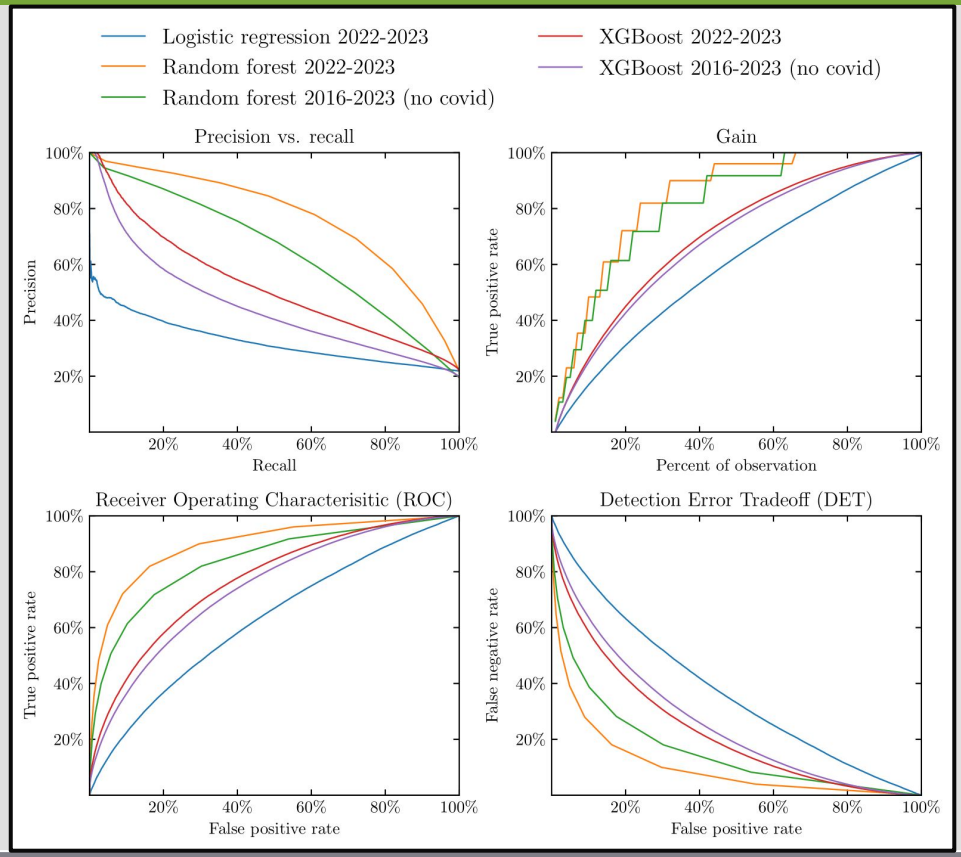

Data
2012-2023
Years
2020 - 2023
2020 - 2023 (No COVID)
2016 - 2023
2016 - 2023 (No COVID)
2012 - 2023 (No COVID)
2012 - 2023
Models
Logistic Regression
Random Forest
XGBoost
COVID = 2020, 2021

Best Result

Train on 2022 - June 2023
Model - Random Forest

Testing On
July and August 2023

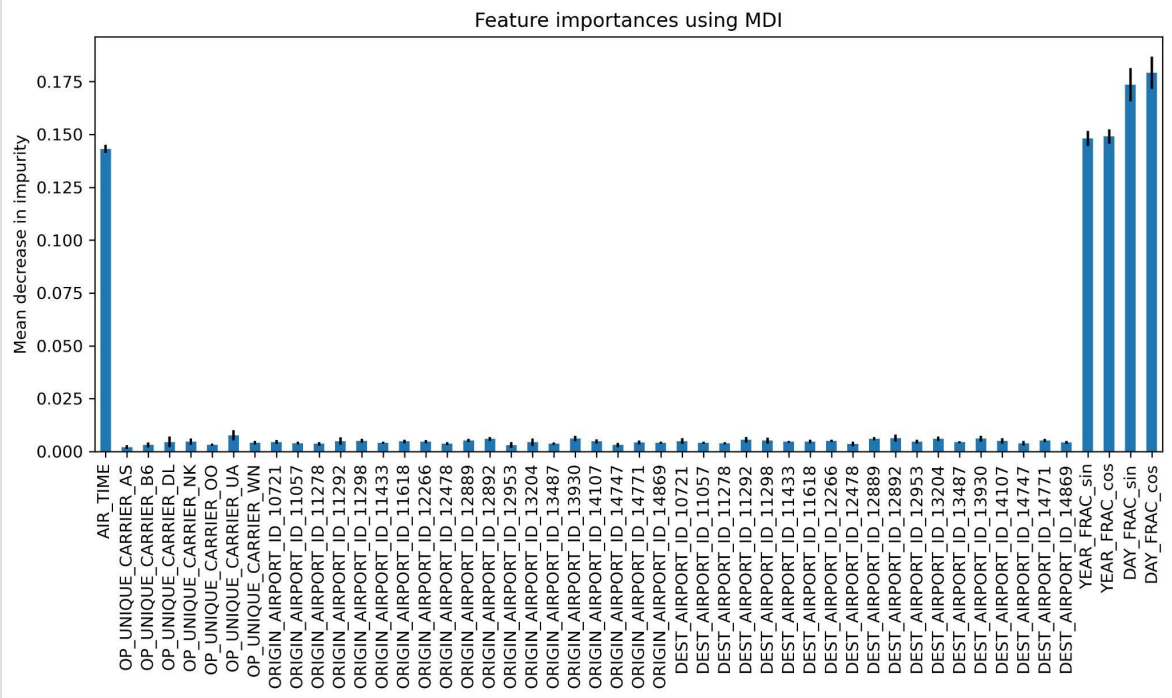
On delayed flight category	Precision	Recall	F1-score	Accuracy
Validation set	0.69	0.72	0.71	0.87
Testing set	0.5	0.7	0.58	0.72



Feature Importance

What determines if your flight will be late?

1. Time of day
2. Time of year
3. Flight duration



Delay Time Range*

Logistic Regression

Accuracy: 0.41
Classification Report:

	precision	recall	f1-score	support
1.0	0.40	0.50	0.44	28615
2.0	0.37	0.00	0.01	24765
3.0	0.41	0.65	0.51	29892

Random Forest

Accuracy: 0.64
Classification Report:

	precision	recall	f1-score	support
1.0	0.65	0.71	0.68	28615
2.0	0.60	0.51	0.55	24765
3.0	0.67	0.70	0.68	29892

XGBoost

Accuracy: 0.57
Classification Report:

	precision	recall	f1-score	support
1.0	0.57	0.65	0.61	28615
2.0	0.53	0.37	0.43	24765
3.0	0.60	0.67	0.63	29892

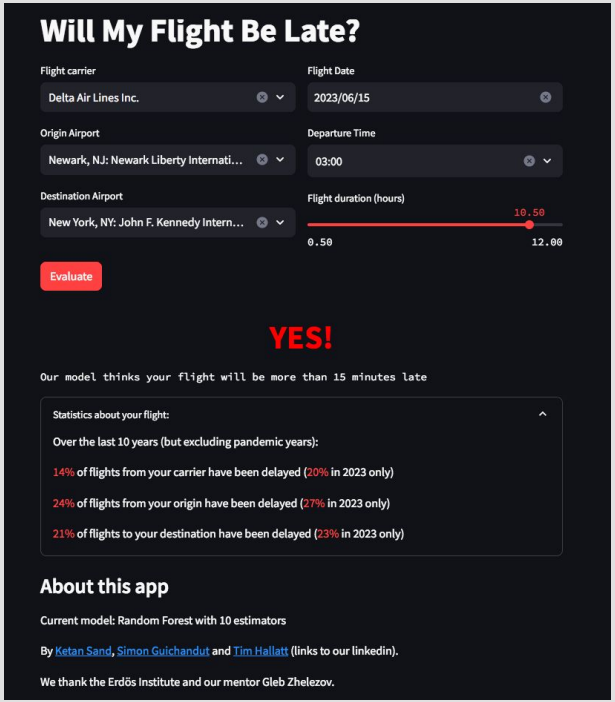
1 -> 15 mins < Delay time <= 30 mins

2 -> 30 mins < Delay time <= 60 mins

3 -> Delay time >= 60 mins

*OPTIMIZATION NEEDED!

Web-App



https://willmyflightbelate.streamlit.app/

SCAN ME

Future Work

- Calibration of model: predict probability, rather than Yes/No
- Delay range prediction: Optimize and Add to the Website
- Hyperparameter tuning: RandomForest takes too long
  - n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, bootstrap
  - Using RandomForestRegressor and RandomizedSearchCV
  - Notebook in repo
- Train Model that also provides the cause of the delay
- Get more parameters and apply neural networks

Special thanks to our mentor Gleb Zhelezov!!