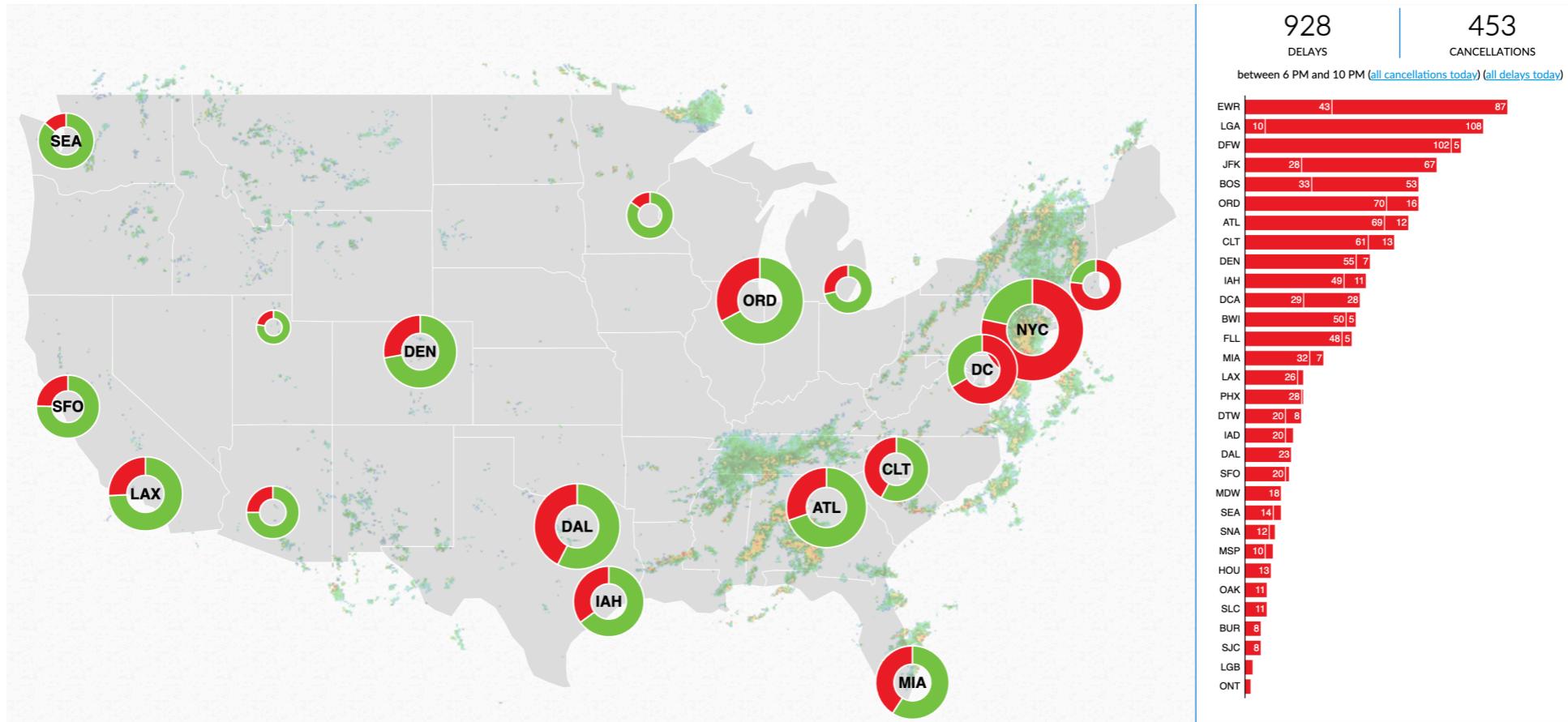


Predicting Flight Delays from Weather

Cameron Czerpak

Background



Flight Aware Misery Map

- Over 220,000 US flights were delayed in 2019 (Bureau of Transportation Statistics)
- Delays cost airlines \$30 billion a year in lost revenue (Cirium) and hamper roughly 20% of the 4.5 billion yearly airline passengers worldwide (ICAO)
- Frequent fliers rely on 3rd party flight tracking apps like Flight Aware to get updates on their travel plans

Current Methods Flight Aware

Flight Details

updated 44 seconds ago

[View track log](#)

[Track inbound plane](#)

[All flights between BWI and STL](#)

Departure Times	
Gate Departure	Takeoff
06:22AM EDT	06:39AM EDT
Scheduled 06:20AM EDT	Scheduled 06:30AM EDT
Taxi Time: 17 minutes	
Average Delay: Less than 10 minutes	
Arrival Times	
Landing	Gate Arrival
07:31AM CDT	07:39AM CDT
Scheduled 07:26AM CDT	Scheduled 07:30AM CDT
Taxi Time: 8 minutes	
Average Delay: Less than 10 minutes	

Southwest WN 579

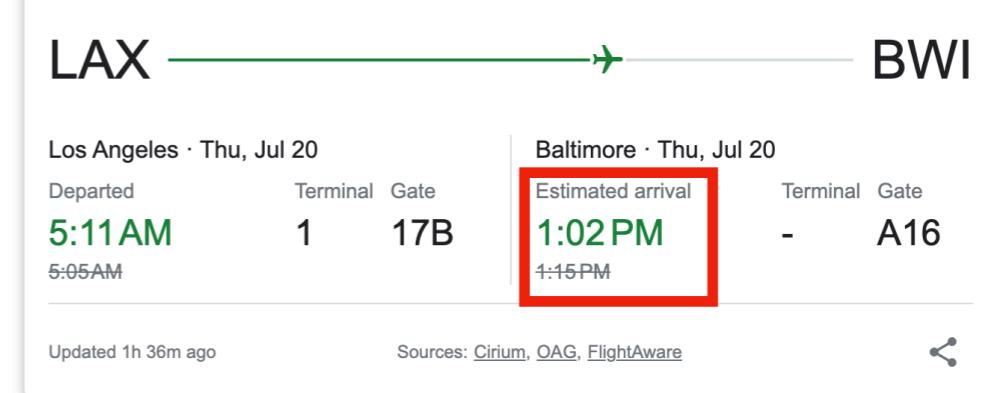
6 flights found

Wed, Jul 19 Thu, Jul 20 Fri, Jul 21 Sat, Jul 22

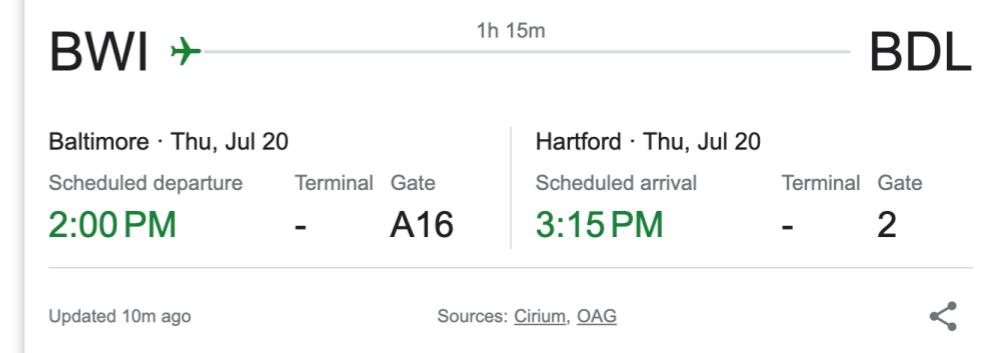
5:05AM

WN 579

Baltimore BWI



2:00 PM WN 579
ON TIME Hartford BDL



When Booking

View average departure and arrival delay on FlightAware

Day of flight

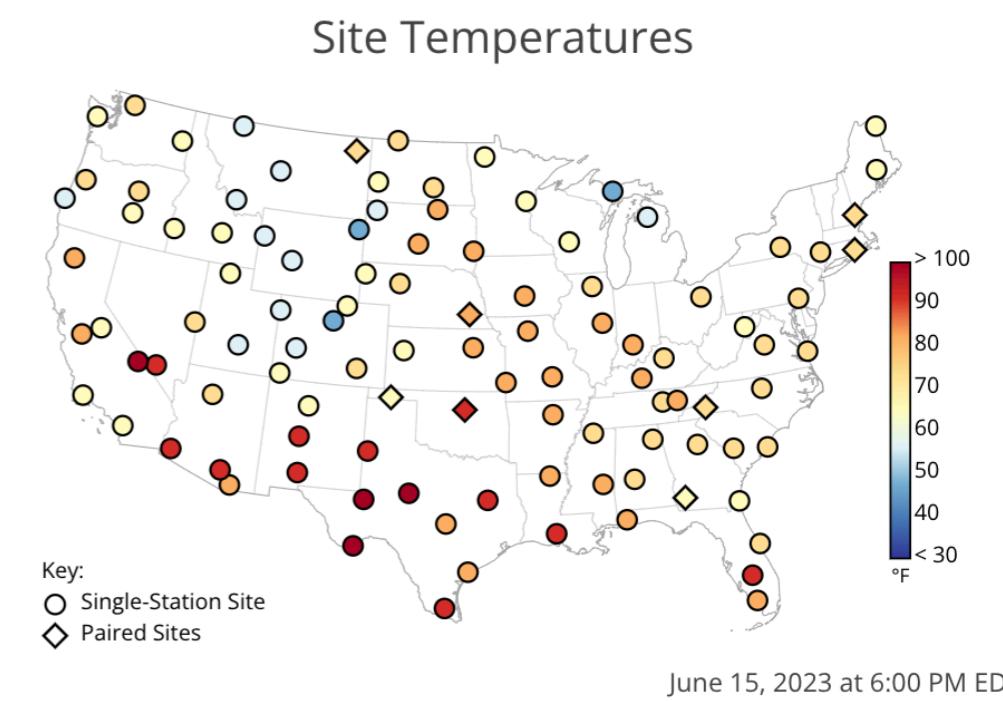
Check if the inbound flight is delayed Google/FlightAware

Objective

- Problem: Require guessing
- **Aim: To build a tool that predicts flight delays from historical flight and weather data**
 - Schedule extra travel days before important events
 - Fly into a different airport hub to make connection

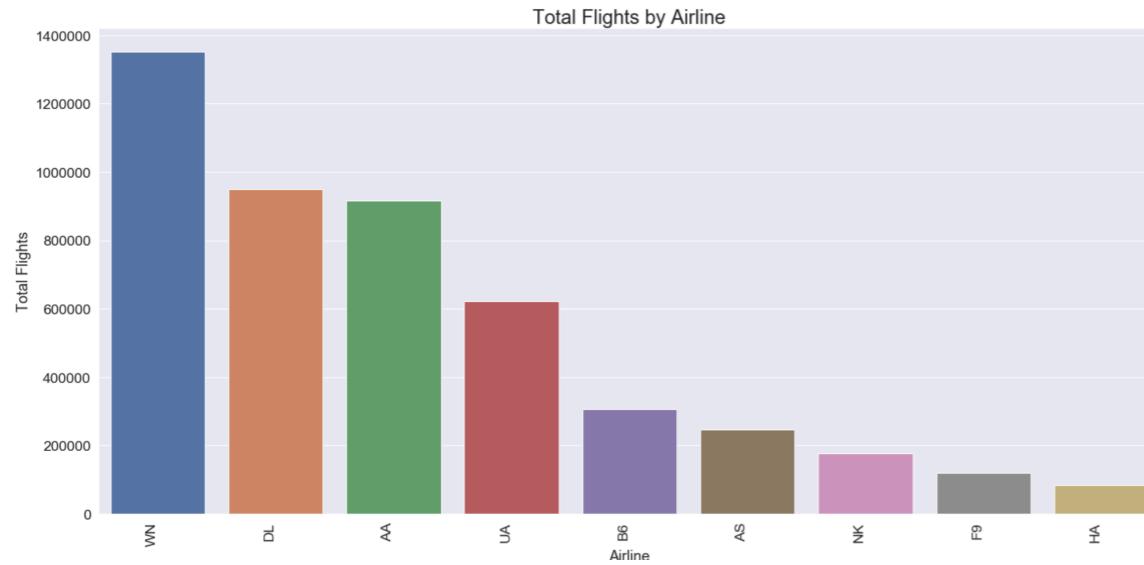
Methods

- Data: Kaggle (Flights 2014-2018), National Centers for Environmental Information (Weather 2014-2018)
- Outcome: Arrival - On time/Delayed
- Model: Classification (XGBoost, Random Forest) (80% Training, 10% Validation, 10% Testing)
- Avoiding data leakage: 2014-17 Train, 2018 Half Test Half Validation

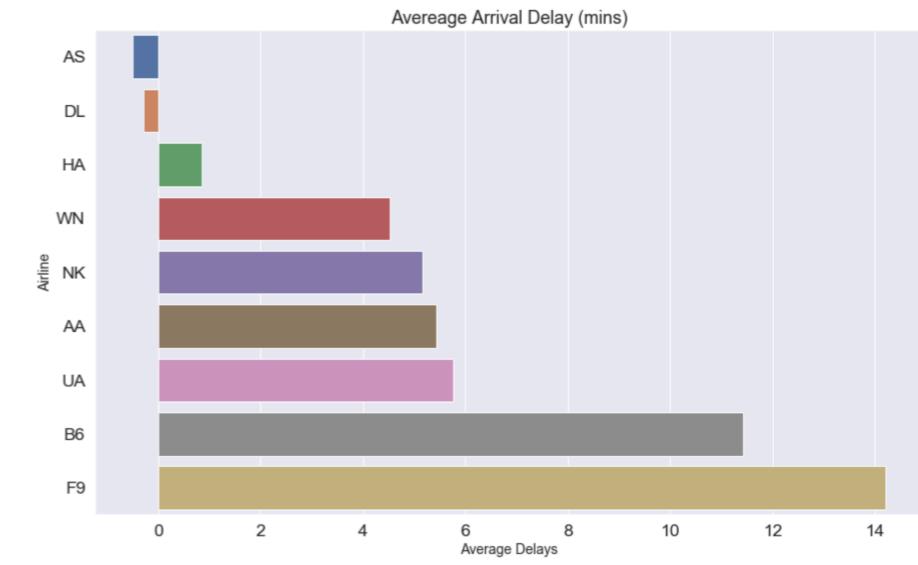


Climate Monitoring Stations

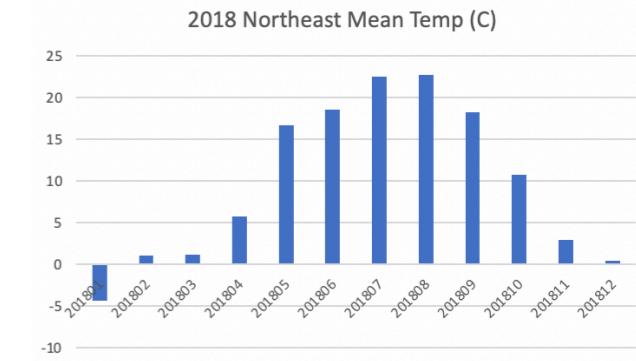
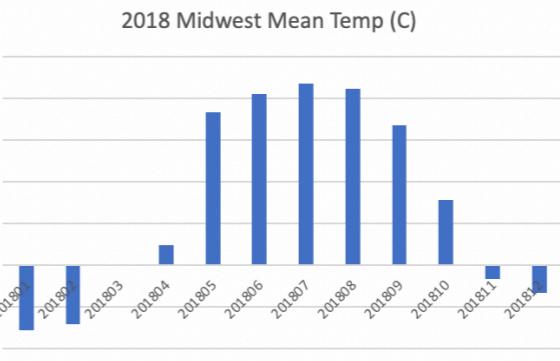
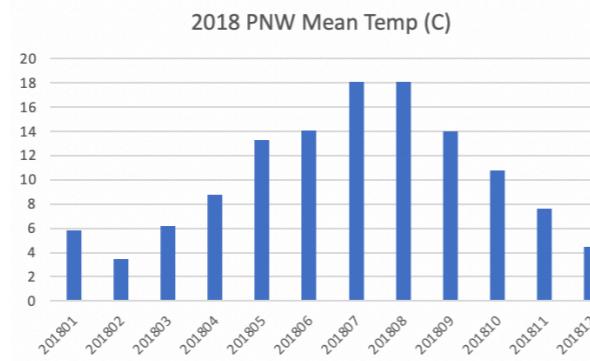
Exploration



Southwest overrepresented in the dataset

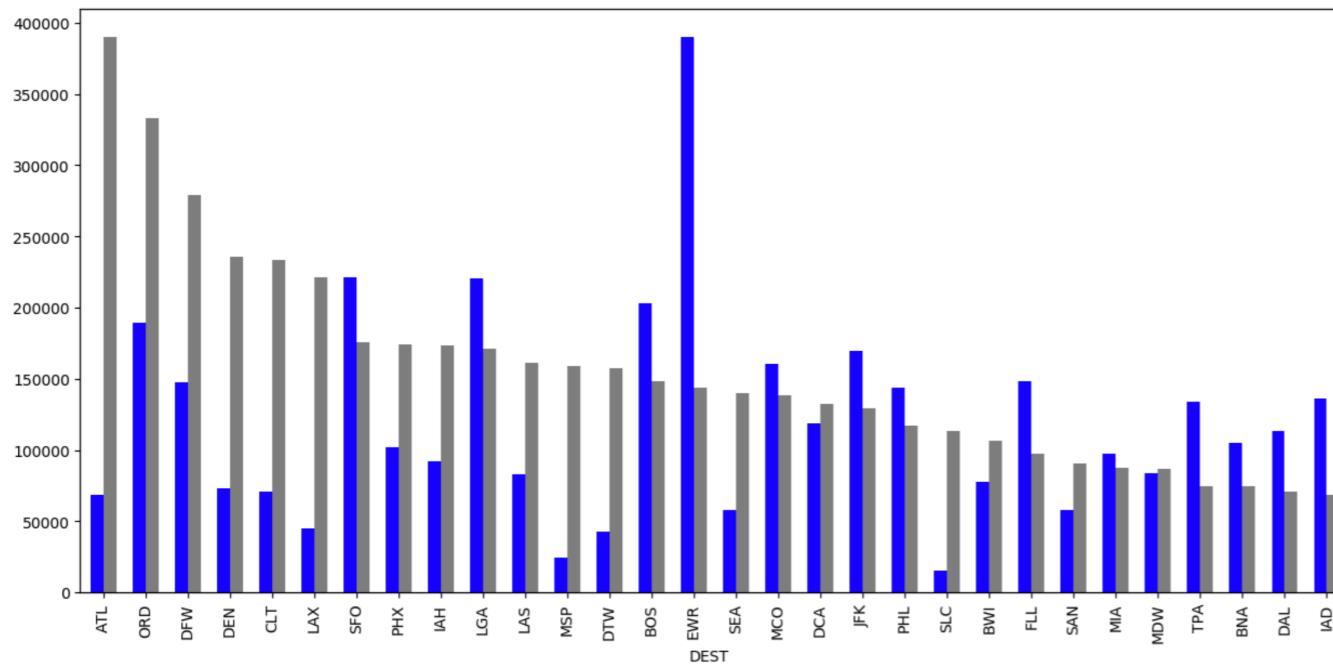


Average delay under 15 minutes



Temperature and Rainfall based on 7 US regions

Features and Cleaning



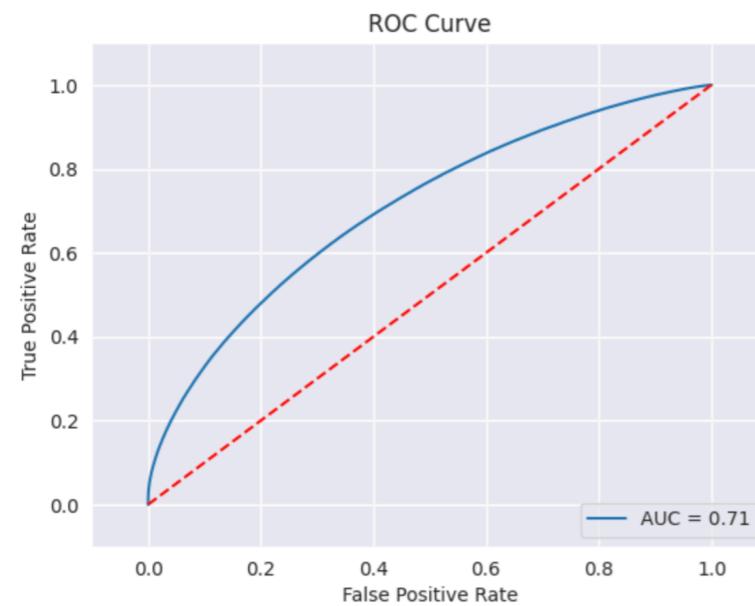
Airline (4x)
Departure Time
Arrival Time
Distance
Regional Temperature (7x)
Regional Precipitation (7x)
Day
Month
Holiday
Origin airport
Arrival airport

Features

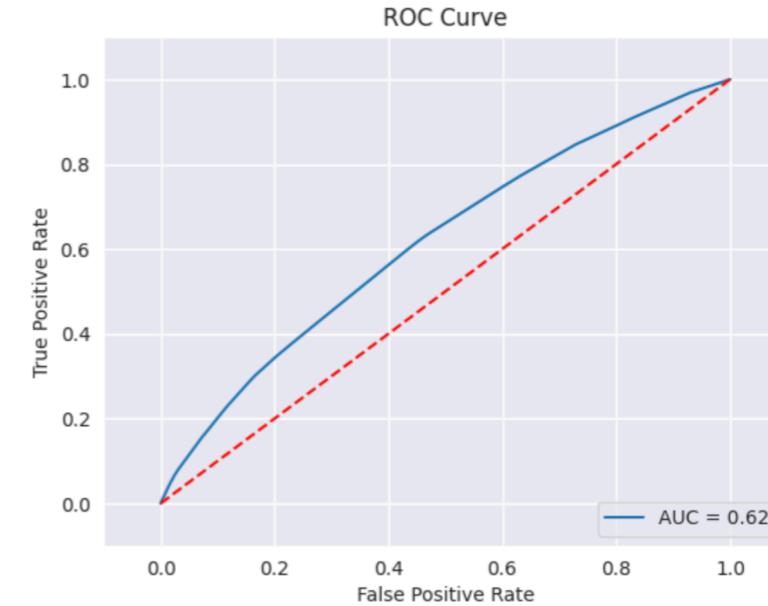
- Airline, Origin, and Destination - categories most valuable
 - `pandas.get_dummies` - 30 columns to 444 columns - each year >16GB
- Reduced to Major 4 US carriers and top 30 US airports

Outcomes

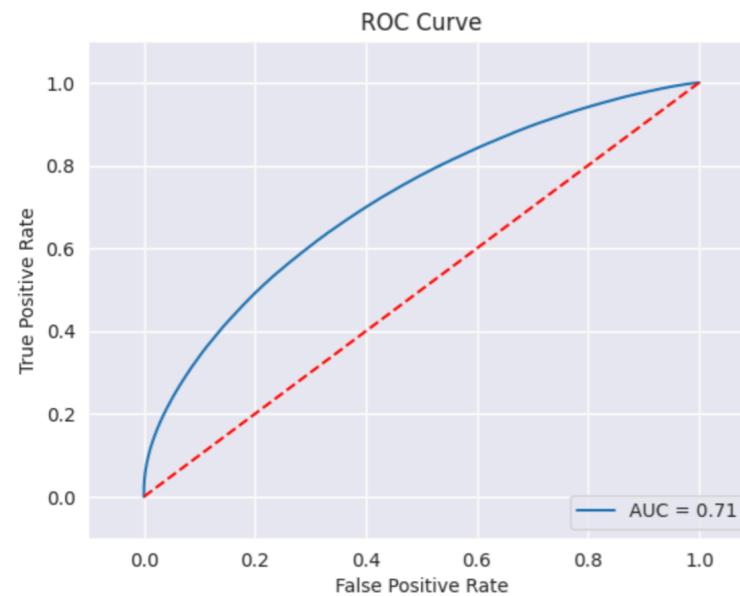
2014-17 Train, 2018 Test/Validation



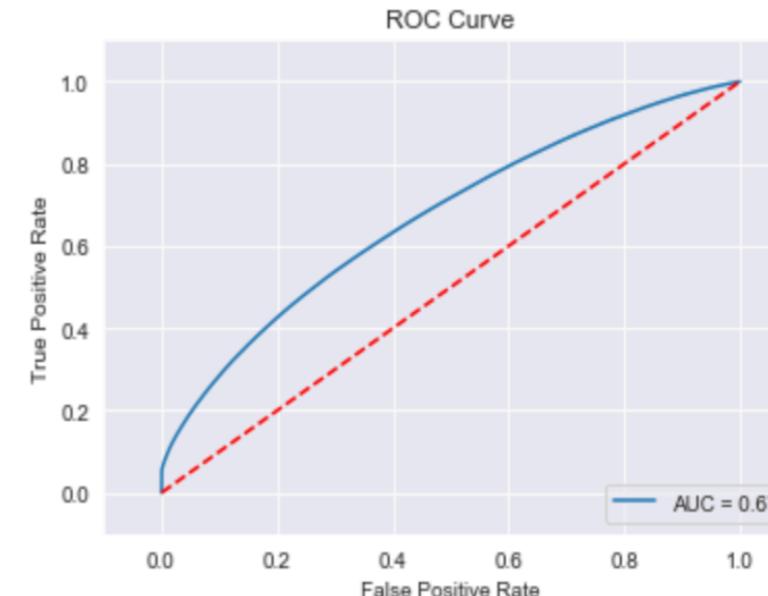
XGBoost
Training Accuracy 70.9%
Testing Accuracy 68.5%



Random Forest
Training Accuracy 57.5%
Testing Accuracy 57.4%



XGBoost with optimization
Training Accuracy 72.8%
Testing Accuracy 68.7%



XGBoost All Airlines w/o Airport
Training Accuracy 65.7%
Testing Accuracy 65.7%

Day of Prediction

Southwest WN 579

6 flights found

Wed, Jul 19 Thu, Jul 20 Fri, Jul 21 Sat, Jul 22

5:05AM WN 579 Baltimore BWI

LAX → **BWI**

Los Angeles · Thu, Jul 20 Baltimore · Thu, Jul 20

Departed Terminal Gate Estimated arrival Terminal Gate

5:11AM 1 17B **1:02PM** - A16

5:05AM 4:15PM

Updated 1h 36m ago Sources: [Cirium](#), [OAG](#), [FlightAware](#)

2:00PM WN 579 Hartford BDL

BWI → **BDL**

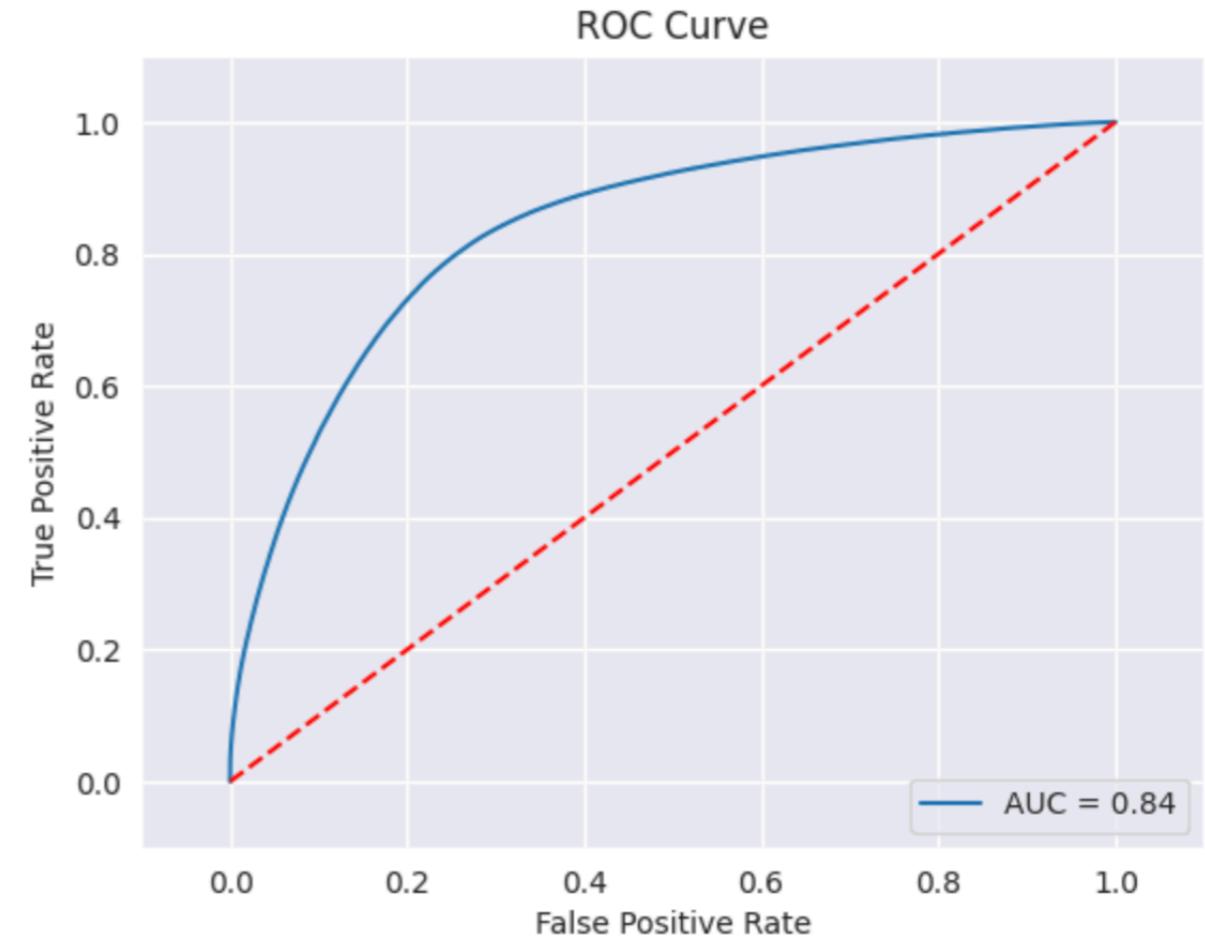
1h 15m

Baltimore · Thu, Jul 20 Hartford · Thu, Jul 20

Scheduled departure Terminal Gate Scheduled arrival Terminal Gate

2:00PM - A16 **3:15PM** - 2

Updated 10m ago Sources: [Cirium](#), [OAG](#)



	[[473209 107357] [102879 248257]]		precision	recall	f1-score	support
0			0.82	0.82	0.82	580566
1			0.70	0.71	0.70	351136
accuracy					0.77	931702
macro avg			0.76	0.76	0.76	931702
weighted avg			0.77	0.77	0.77	931702

Day of flight
If departure flight is delayed

XGBoost
Training Accuracy 78.0%
Testing Accuracy 77.4%

Conclusions

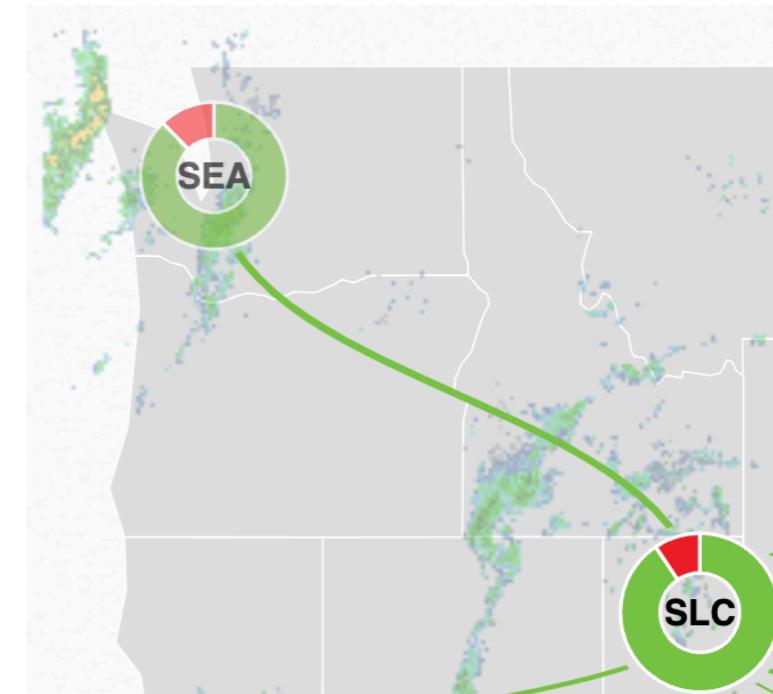
- XGBoost best model 68.5% accuracy compared to Random Forest 57.4% accuracy
- Optimization with XGBoost <1%, but 3x training time (Google Colab)
- Dropping smaller airports and airlines +3% accuracy
- +8% higher prediction rate the day of the flight (when departure status is known)

Future Work

- Improve model:

Arrival Times	
Landing	Gate Arrival
07:31AM CDT	07:39AM CDT
Scheduled 07:26AM CDT	
Scheduled 07:30AM CDT	
Taxi Time: 8 minutes	
Average Delay: Less than 10 minutes	

Average delay



Weather on flight path

- Convert the model from binary classification to on time, small <15min, medium 15-30min, and large delay >30 min

Thank you