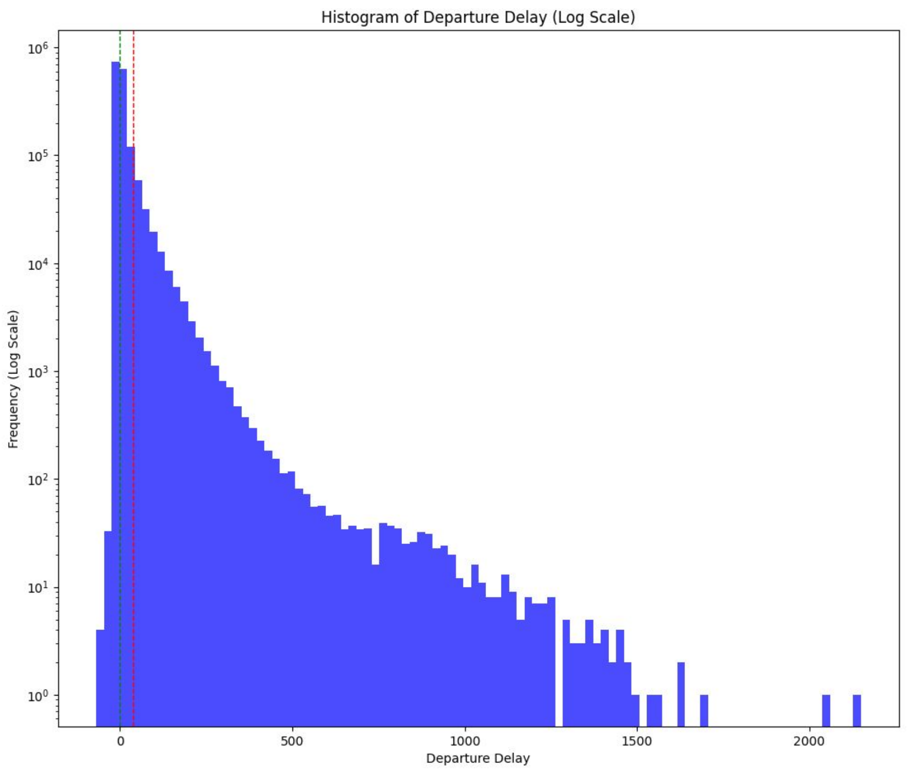
Flight Delay Prediction

Asher Erickson, Linqing Mo, Kyllan Wunder

Explaining the problem

- Impact on customers
- Impact on airlines
- Delaying other flights
- Economic impact
- Environmental impact



Data Description

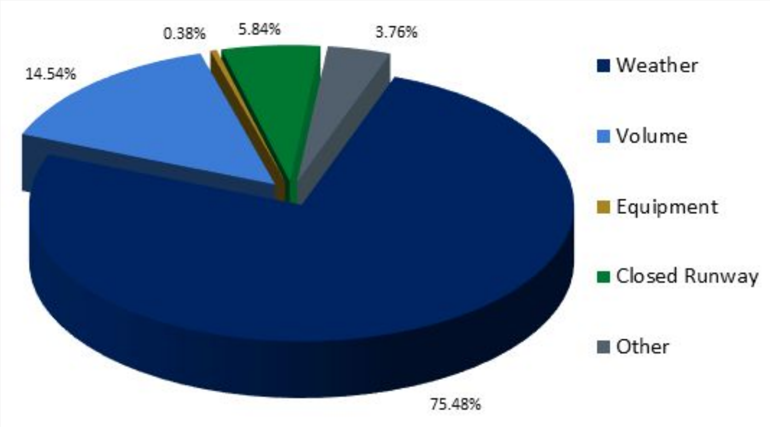
- Flights from 2014 to 2018 collected by the US Office of Airline Information and the Bureau of Transportation Statistics, including date, time, origin, destination, airline, and flight delay status.
- 60 files
- Over 30 millions rows of flights data
- 109 variables

- Focus on 9 main airports and 5 main airlines: 2 millions rows

Data Cleaning

- Remove columns with most missing values
- Remove arrival statistics(actual arrival time etc.): the goal is to predict if a flight will be delayed before it leaves
- Combine columns with repeating information(StateID, StateName etc.)
- Remove rows with missing value: about 1%

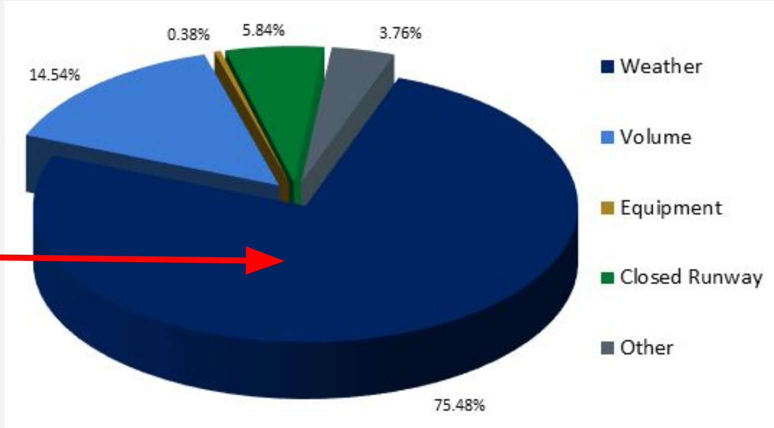
- 24 variables remaining



Causes of air traffic delay in the National Airspace System - FAA

Combine Weather Data

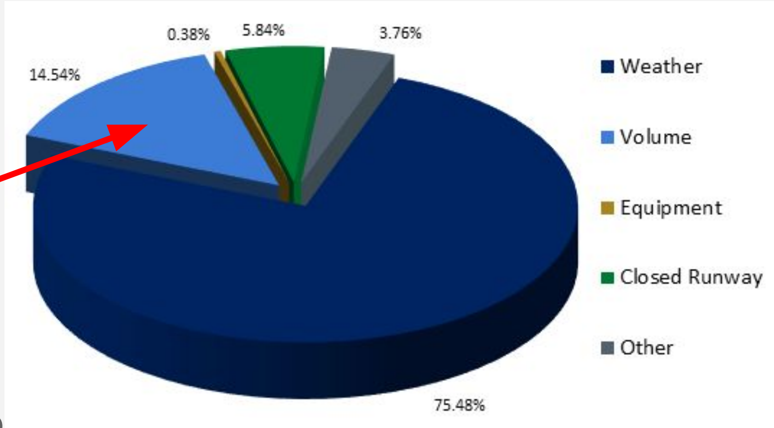
- About 75% of delays come from weather



- Request Climatological Data Station Details Data from NOAA(National Centers for Environmental Information) for each airport station
- Partition hourly weather type descriptions and group them into different levels based on FAA records
- Combine hourly wind speed, hourly visibility and hourly weather type into Flight dataframe

Visibility	WindSpeed	SevereWeather	BadWeather	DepDelay
10.0	5.0	0	0	-3.0

Create More Variable and Targets



- About 15% of delays come from volume of airport

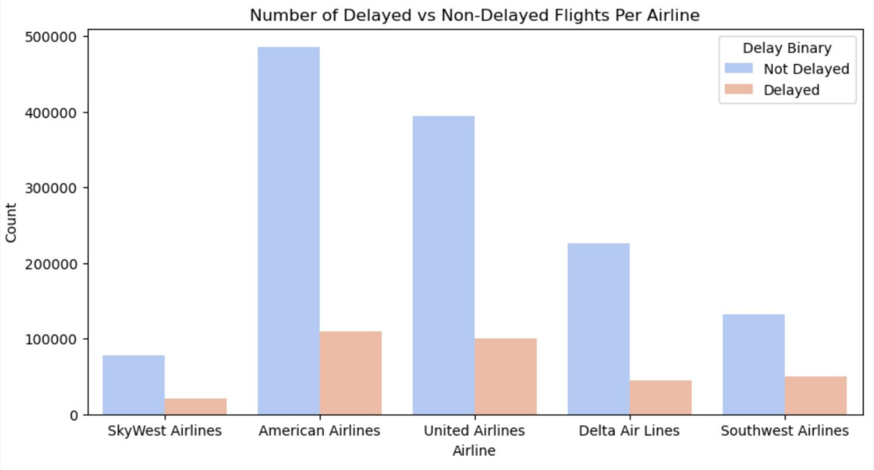
- Flight Density is created based on counts of flight departure in a 3 hours interval for each flight
- Holiday variable is created based on if the flight is on the week of federal holidays

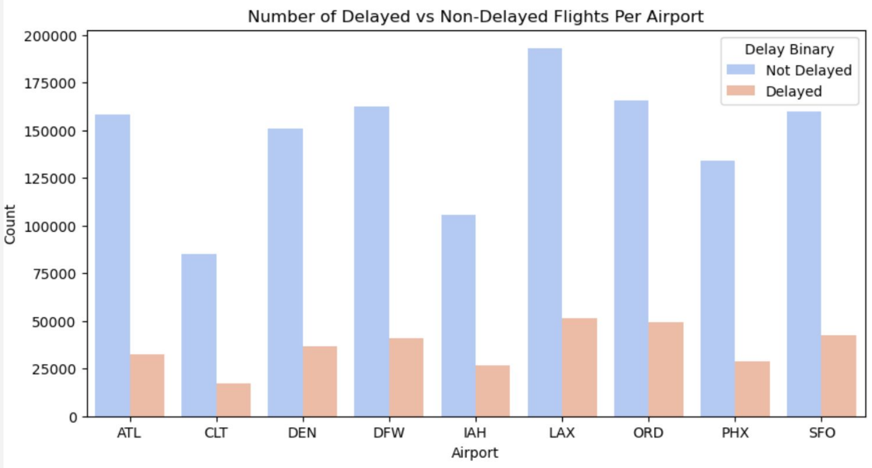
- Two targets for classification problem: delay_binary(with threshold of 15 min)

  delay_interval(multiclass classification)

is_holiday_week	OriginFlightDensity	DepDelay	delay_binary	delay_interval
1	6.0	21.0	1	3

How to Choose Airlines/Airports





Overall: 19.87%

Delta Airlines (DL): 16.50%
American Airlines (AA): 18.42%
United Airlines (UA): 20.30%
SkyWest Airlines (OO): 21.32%
Southwest Airlines (WN): 27.68%

Charlotte Douglas (CLT): 16.64%
Atlanta (ATL): 17.07%
Phoenix (PHX): 17.73%
Denver (DEN): 19.59%
Dallas/Fort Worth (DFW): 20.08%
Houston (IAH): 20.13%
San Francisco (SFO): 21.03%
Los Angeles (LAX): 21.09%
Chicago O'Hare (ORD): 22.92%

Airline Airport dependency

Independent if P(A) * P(B) = P(A ∩ B)

P(Delta) * P(Phoenix) = P(Delta ∩ Phoenix)

P(.165) * P(.1773) = P(.785)

No equal thus dependant

All combinations of airlines and airports are dependent

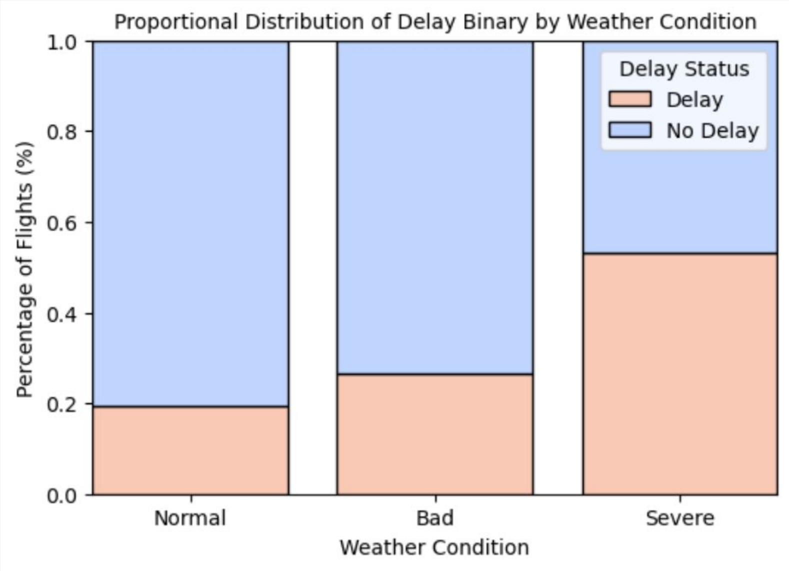
Best combinations:

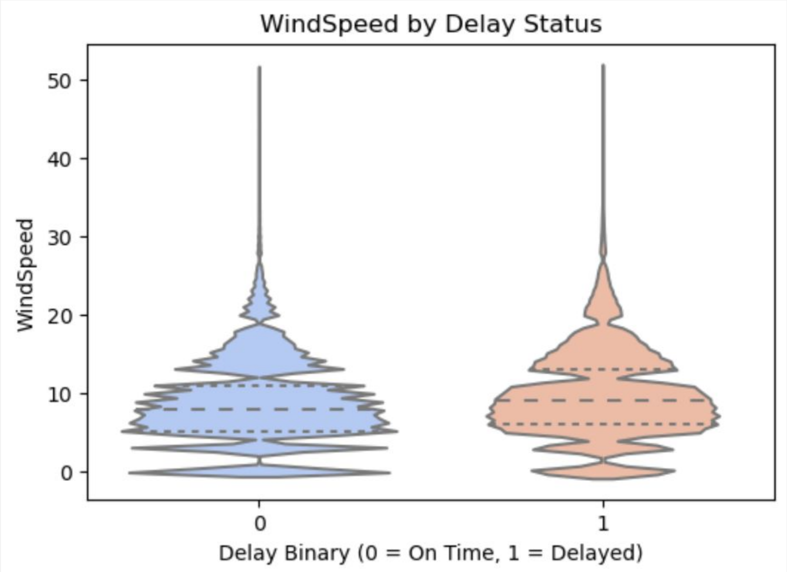
PHX	DL	7.85%
CLT	DL	13.01%
PHX	UA	13.35%
DEN	DL	13.50%
DFW	UA	13.90%

Worst combinations:

DEN	WN	24.81%
ORD	UA	24.86%
PHX	WN	26.24%
SFO	WN	28.66%
LAX	WN	32.24%

How Weather Affects Flight Delay



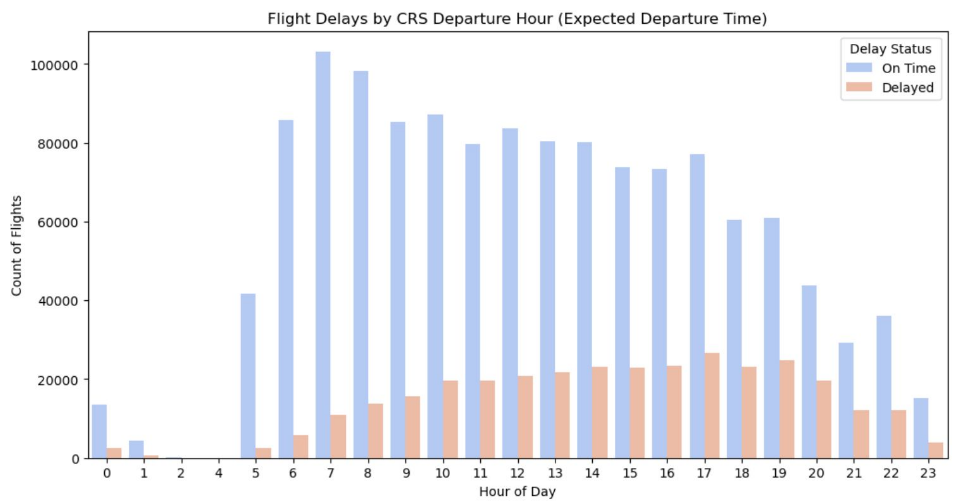


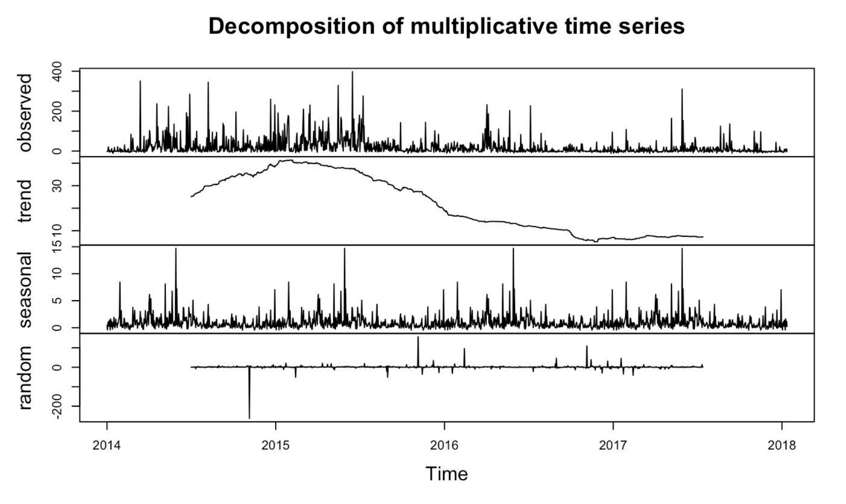
Ratio of flights under bad weather: 0.0271
Ratio of flights under severe weather: 0.008

Ratio of delay under normal weather: 0.194
Ratio of delay under bad weather: 0.264
Ratio of delay under severe weather: 0.530

Delay	Yes	No
Mean	9.7	8.8
25% percentile	6.0	5.0
75% percentile	13.0	11.0

Delay Over Time

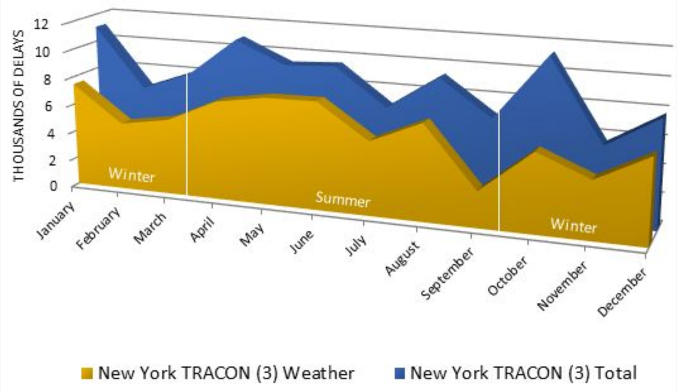




Ratio of Delayed Flights by Hour:

3-5	0.058	6-8	0.097
9-11	0.179	12-14	0.212
15-17	0.245	18-20	0.290
21-23	0.259		

From FAA (shows similar seasonality)



Naive approach

First predict if a flight will be delayed

Accuracy: 79.3%

Feature importance:

DestAirportSeqID	50.36
OriginAirportSeqID	49.64

Then predict how long the delay will be

Accuracy: 52%

Feature importance:

DestAirportSeqID	50.03
OriginAirportSeqID	49.72
Year	0.08
CRSArrTime	0.06
CRSDepTime	0.05
Distance	0.04
CRSElapsedTime	0.01

Comparing models (binary)

Logistic Regression: 79.4%

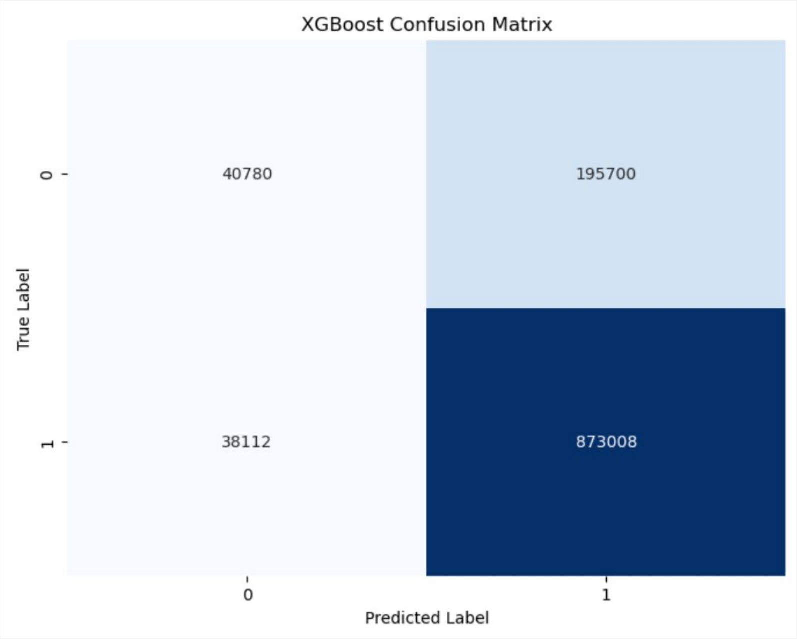
Random Forest: 79.6%

XGBoost Classification: 79.8%

Neural Network: 80.3%

Feature importance for XGBoost:

```
Top features:
CRSDepTime: 0.13519437611103058
Reporting_Airline_WN: 0.11107633262872696
SevereWeather: 0.08239080011844635
CRSArrTime: 0.07548076659440994
Visibility: 0.06928591430187225
DestAirportSeqID: 0.05322586745023727
Month: 0.041327688843011856
Origin_ORD: 0.032815366983413696
```



Comparing models (multi)

Neural Network: 58.25%

Xgboost: 51.45%

Random Forest: 51.43%

KNN: 49.96%

Feature importance:

SevereWeather (12.0%)

CRSDepTime (8.07%)

Reporting_Airline (7.44%)

DestAirportSeqID (6.5%)

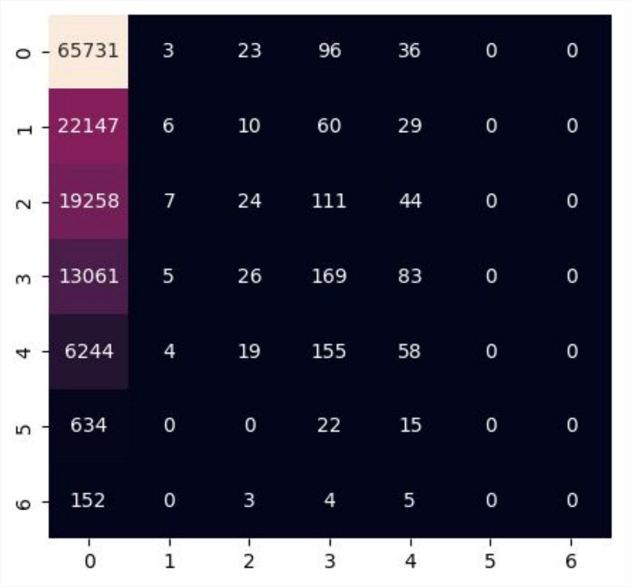
Distance (5.86%)

Month (5.67%)

OriginAirportSeqID (5.67%)

Visibility (4.9%)

BadWeather (4.34%)



Xgboost results

Future Work

- Build more detailed model on how weather affects flights delay
- Research on relationship between Hour of day and flights delay
- Calculate real airports volume
- One-hot ensemble for some categorical variables