# Flight Delay Prediction

Julia Rowe
Data Scientist

# Outline

Intro/Background

Problem Statement

Data

Data Preprocessing

Preliminary Modeling

Final Models

Conclusion

# Main reasons of flight delays

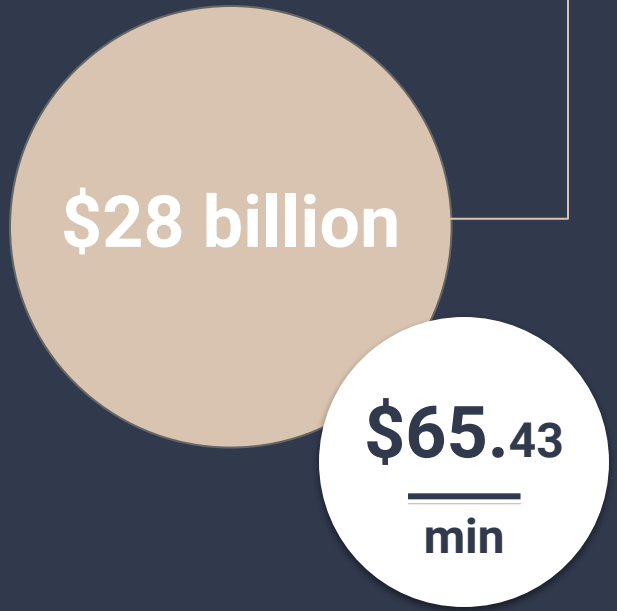
Based on data from US Bureau of Transportation Statistics flight delays (in 2015) are caused by...

| | |
|---|---|
| 39.8% | Late-arriving aircraft |
| 32.2% | Air Carrier |
| 22.9% | National Aviation System (NAS) delay |
| 5% | Extreme Weather |
| 0.1% | Security |

Airlines do not report the causes of the late-arriving aircraft. Weather contributed to 32.8% of total delay minutes in 2015.

("Understanding the Reporting of Causes of Flight Delays and Cancellations", 2021).

# Flight delays are costly for airline companies




FAA/Nextor estimated the annual costs of delays...to be $28 billion ("U.S. Passenger Carrier Delay Costs", 2020).

In 2015, the cost of aircraft block (taxi plus airborne) time for U.S. passenger airlines was $65.43 per minute ("Cost of Aircraft Delay to U.S. Passenger Carriers").

# Problem Statement:
# Can we predict severity of flight delays?

**For domestic flights in the U.S. minimum connection times range from 30 minutes to 2 hours. (“What is Minimum Connection Time?”, 2021).**

3 classes

| **On-time** | **Minor delay** | **Major delay** |
| --- | --- | --- |
| (early or no delay) | (0-30 min delay) | (30+ min delay) |

# Data

**‘2015 Flight Delays and Cancellations’**

## 5,819,079 flights

from Kaggle

**Daily climate data from the National Oceanic and Atmospheric Administration (NOAA) in 2015**

## Queens, NY

&

## Chicago, IL

# Data Preprocessing

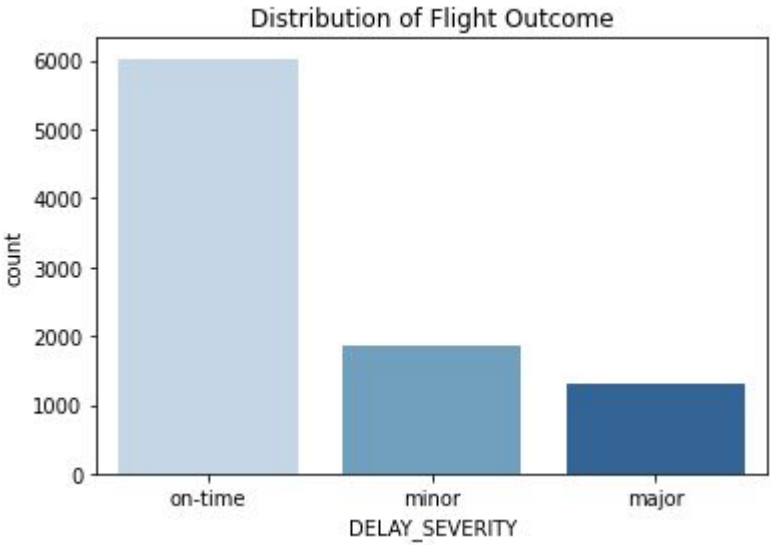
## Flight Data

- Scoped data to one flight plan:
  - Origin airport: LGA
  - Destination airport: ORD
- ~80% of delay type columns were null
- Included only features for data that is only known prior to the flight
- Extracted data from date and time columns into categorical data

## Climate Data

- Merged climate data from Queens & Chicago
- Almost all weather-type columns were null
- Included various features for daily temperature, wind speed, wind direction, snowfall, precipitation

# Baseline Model




**ratios:**

on-time: 0.656390

minor: 0.202201

major: 0.141410

**Baseline score is 65.6%**

# Preliminary modeling:

| Model | Train Scores | Test Scores | Recall Scores |
|---|---|---|---|
| **XGB Classifier** | 86.4% | 72.1% | 56.6% |
| **Random Forest Classifier** | 92.0% | 70.2% | 55.4% |
| **Extra Trees Classifier** | 92.0% | 69.0% | 54.4% |
| **Decision Tree Classifier** | 92.0% | 65.6% | 52.8% |
| **Gradient Boosting Classifier** | 73.2% | 71.2% | 48.6% |
| **KNeighbors Classifer** | 75.5% | 68.8% | 47.2% |
| **Bernoulli Naive Bayes** | 63.1% | 64.1% | 46.1% |
| **Ada Boost Classifier** | 69.2% | 68.4% | 43.7% |
| **SVC** | 71.7% | 69.0% | 43.2% |
| **Logistic Regression** | 67.8% | 68.2% | 42.1% |

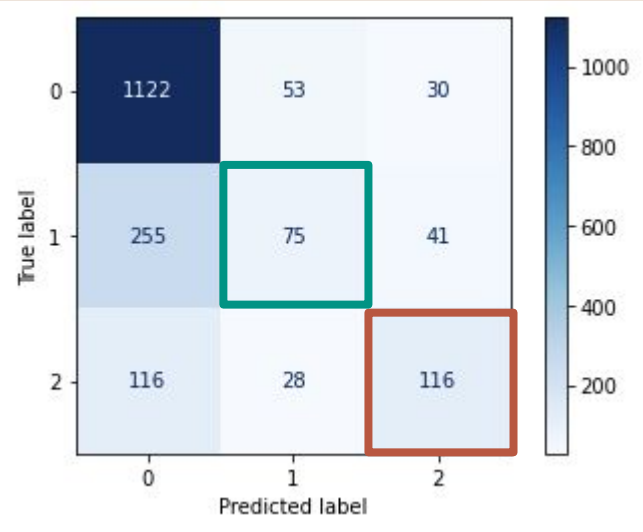
**Highest recall score**

**Train & Test scores are not overfit and has high recall scores**

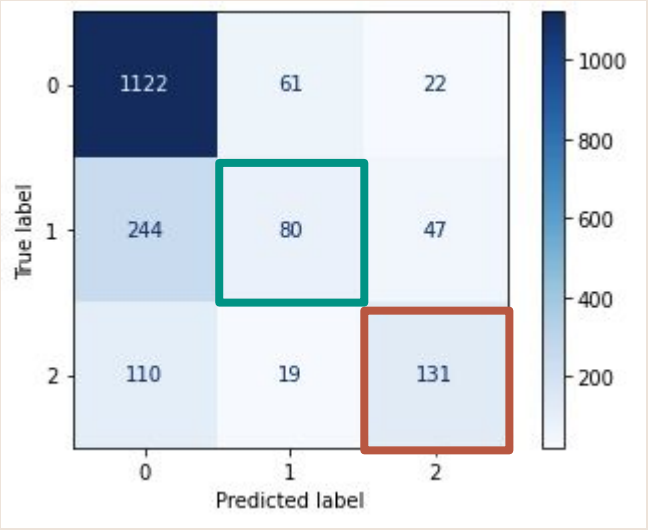
**Recall scores calculated with macro-average

# Final Models

Gradient Boosting Classification Matrix




XGBoost Classification Matrix




Final models had the best predictions for minor delay (green) and major delay (red).

# Final Models

**Gradient Boosting Classifier:**

Train score: 73.2% → 73.**9**%
Test score: 71.2% → 71.**5**%

Recall score: 48.6% → **52.6%**

**XGBoost Classifier:**

Train score: 86.4% → **76.2%**
Test score: 72.6% → **72.1%**

Recall score: 56.6% → **55.0%**

# Conclusions

- Model performed better than baseline
- Explore other methods to deal with unbalanced data
  - Oversampling
  - Decrease the threshold
- Need to collect more data:
  - Hourly climate data
  - Data with complete weather-type values
  - Data with complete delay-reason values
- Expand the scope of the project to include additional flight plans
- Future Project ideas:
  - Predict flight delay from weather before the flight