PREDICTIVE Flight Delays

GROUP 2

Maddie Bowman, Jinlu Wang, Belinda Ho
Thomas Ngo, Nelson Lin, Meredith Jones

PROJECT OVERVIEW

PROPOSAL

Using historical flight and weather data, build a machine learning model that predicts probability of flight being delayed by 15 minutes or more.

The user interface also features real-time weather forecasts for both the departure and destination airports, enabling travelers to better prepare for their trips.

DATASET

Kaggle Dataset - **2019 Airline Delays w/Weather and Airport Details**

Classification dataset with detailed airline, weather, airport and employment information for 2019.

Sourced: ***Bureau of Transportation statistics***

National Centers for Environmental Information (NOAA)

KAGGLE DATA

26 Columns — 6,489,062 Rows

MONTH	DAY OF WEEK	DEPARTURE DELAY OVER 15 MIN	DEPARTING FLIGHT DISTANCE (IN TIME)	DEPARTING FLIGHT DISTANCE TO BE FLOWN	SEGMENT NUMBER	CONCURRENT FLIGHTS
NUMBER OF SEATS	CARRIER AIRLINE	AVG. # OF MONTHLY FLIGHTS (AIRPORT)	AVG. # OF MONTHLY FLIGHTS (AIRLINE)	MONTHLY FLIGHTS (AIRLINE/AIRPORT)	AVG. MONTHLY PASSENGERS (DEPARTING AIRPORT)	
AVG. MONTHLY PASSENGERS (AIRLINE)	FLIGHT ATTENDANTS PER PASSENGER	GROUND EMPLOYEES PER PASSENGER	PLANE AGE	DEPARTING AIRPORT	LATITUDE (DEPARTING AIRPORT)	
LONGITUDE (DEPARTING AIRPORT)	PREVIOUS AIRPORT	PRECIPITATION DAILY (INCHES)	SNOWFALL DAILY (INCHES)	SNOW ON GROUND PER DAY (INCHES)	MAX DAILY TEMP	MAX DAILY WIND SPEED

Source: https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

DATA PREPROCESSING

POTENTIAL FEATURES IMPACTING FLIGHT DELAYS

WEATHER CONDITIONS

- ✈ Flight Visibility
- ✈ Precipitation
- ✈ Extreme Temperature
- ✈ Wind and Speed Direction
- ✈ Snowfall/Snow on Ground
- ✈ Severe Weather Storms

INSTANCES OF PEAK TRAVEL

- ✈ Peak Travel Times
- ✈ Holiday Seasonality
- ✈ Busy Days for Travel (weekends, Mondays)
- ✈ Concurrent Flights (tarmac congestion)

AIRLINE & AIRCRAFT IMPACT

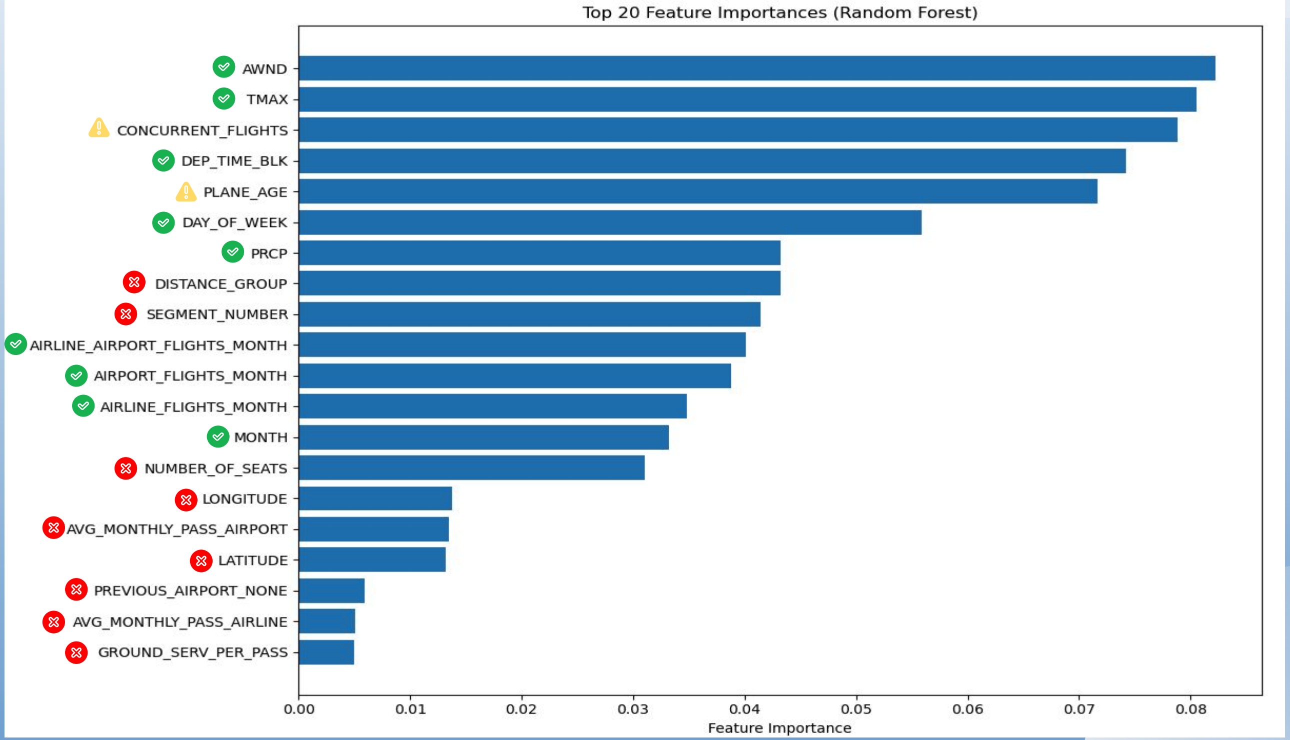
- ✈ Airline Carrier (are there specific airlines that see flight delays more often?)
- ✈ Age of Departing Aircraft (does this assumption of increased maintenance times cause higher levels of delay?)
- ✈ Prior Flights to Departure (subsequent flights for an aircraft allows for possibility of prior flight delay)
- ✈ Departing Airport (busier airports, with higher traffic levels can more commonly result in flight delays)
- ✈ Flight Distance for Departing Aircraft (will a flights' length impact likelihood of delay - *maintenance factors*?)

FEATURE IMPORTANCE

Feature Availability

Current database structure and access to necessary APIs for live data to user input prevented certain features from being utilized in our model to dashboard design.

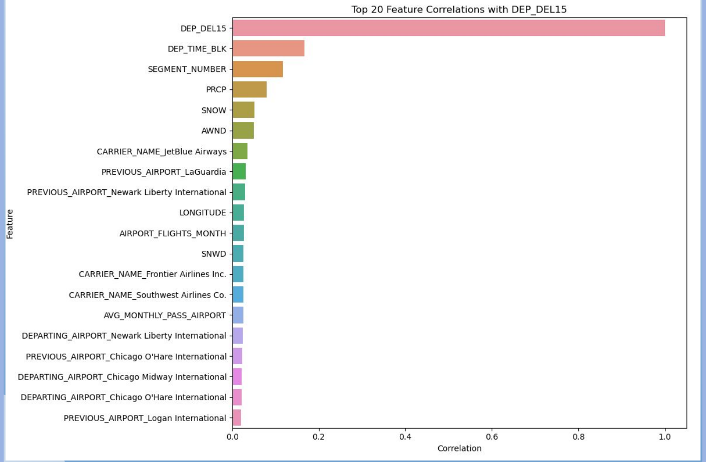
	Feature Used
	Use Restricted
	Not Included in Model Training

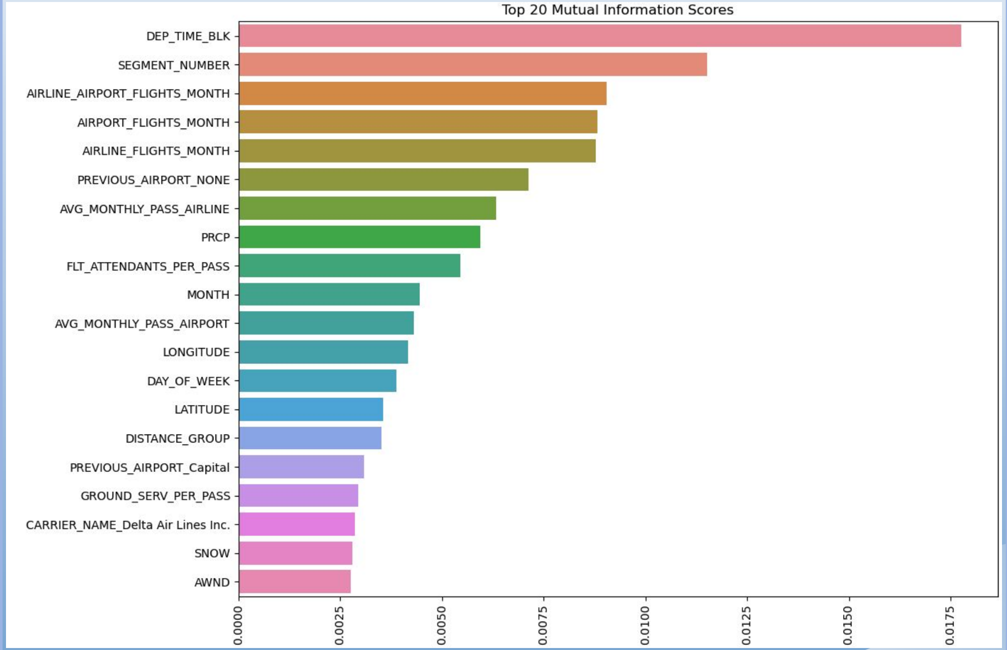


FEATURE IMPORTANCE

What can these charts tell us?

Further confirmation of feature selection, based on mutual dependence AND correlation between delays **[DEP_DEL15]** and each individual column in our dataset.





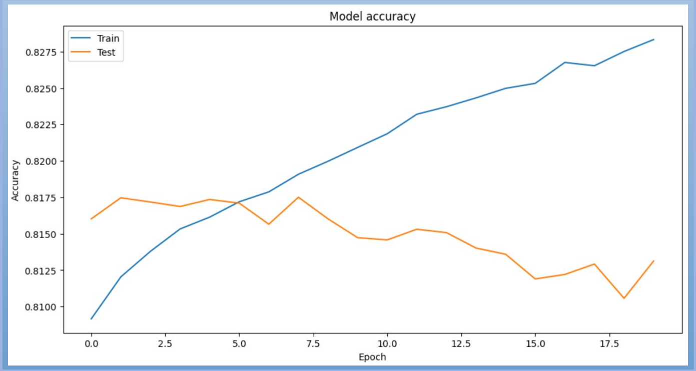
MODEL SELECTION

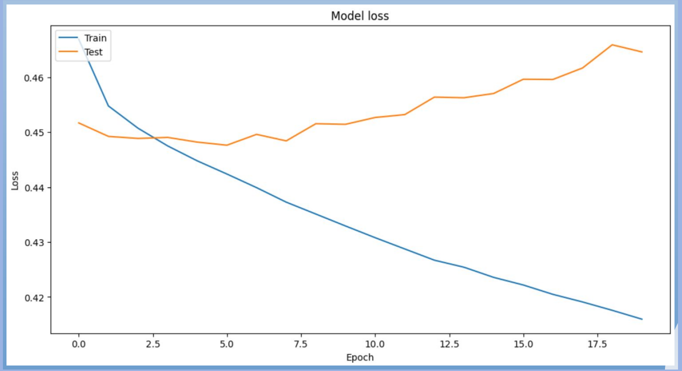
2 Model Types Tested → ***13 Total Optimizations***

XGBoost Model

Neural Network Model

- ✈ Original Neural Network Model + increased number of neurons
- ✈ Original Neural Network Model + add hidden layer
- ✈ Original Neural Network Model + feature engineering



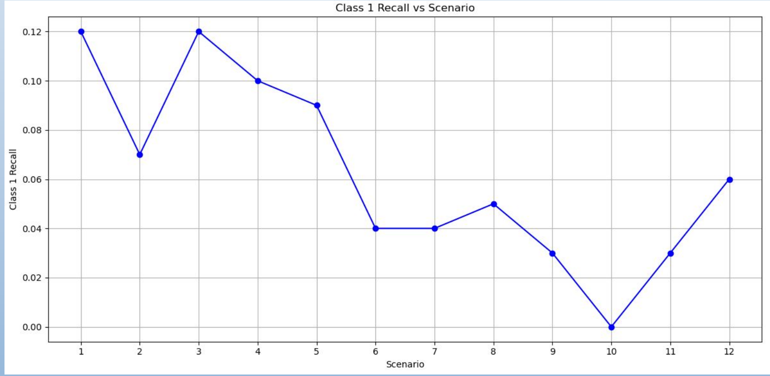


FEATURE ENGINEERING

Steps to Improve Model Accuracy

Scenario	Features
1	With all Features
2	All features without Plane Age
3	All features without Concurrent Flights
4	All features without Prcp
5	All features without Plane Age & Concurrent flight
6	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK', 'CONCURRENT_FLIGHTS', 'PLANE_AGE','PRCP', 'TMAX', 'AWND']
7	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK', 'CONCURRENT_FLIGHTS', 'PRCP', 'TMAX', 'AWND']
8	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK','PLANE_AGE', 'PRCP', 'TMAX', 'AWND']
9	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK', 'PRCP', 'TMAX', 'AWND']
10	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK', 'TMAX', 'AWND']
11	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK','TMAX', 'AWND' ,'AIRPORT_FLIGHTS_MONTH', 'AIRLINE_FLIGHTS_MONTH', 'AIRLINE_AIRPORT_FLIGHTS_MONTH']
12	['MONTH', 'DAY_OF_WEEK', 'DEP_TIME_BLK','TMAX', 'AWND', 'AIRPORT_FLIGHTS_MONTH', 'AIRLINE_FLIGHTS_MONTH', 'AIRLINE_AIRPORT_FLIGHTS_MONTH', 'PRCP']

Experimental Model - Outcome Plotted



MODEL OPTIMIZATION

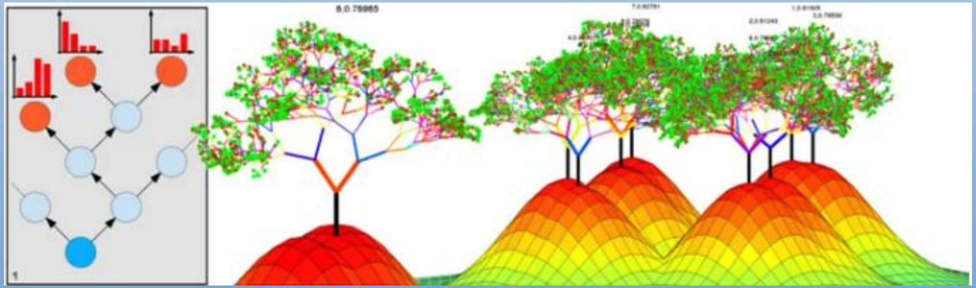
Continuous Improvement

Only 2 days until the presentation...

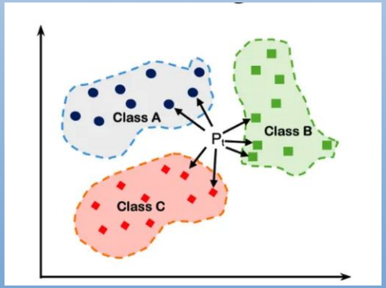
→ ***Not very happy with overall model performance, continue to optimize!***

Revert back to simpler models that work well with **classification** problems:

Random Forest



+ K Nearest Neighbors



MODEL PERFORMANCE

Random Forest

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.92	0.88	210437
1	0.42	0.24	0.30	49126
accuracy			0.79	259563
macro avg	0.63	0.58	0.59	259563
weighted avg	0.76	0.79	0.77	259563

RF produced more balanced results and provides much higher chance of predicting true delays.

NNW

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.99	0.90	42286
1	0.54	0.06	0.11	9627
accuracy			0.82	51913
macro avg	0.68	0.52	0.50	51913
weighted avg	0.77	0.82	0.75	51913

KNN

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.94	0.88	42286
1	0.39	0.17	0.24	9627
accuracy			0.80	51913
macro avg	0.61	0.55	0.56	51913
weighted avg	0.75	0.80	0.76	51913

MODEL TESTING

```
input = [10.00, 7.00, 100.00,10.00, 20000.00,31600,1500, 50.0, "1900-1959"]
```

```
# Make the prediction
prediction_result = model.predict_proba(prepped_data) * 100
delay_percentage = prediction_result[0][1]  # Probability of class 1 (delay)
```

	MONTH	DAY_OF_WEEK	TMAX	AWND	AIRPORT_FLIGHTS_MONTH	AIRLINE_FLIGHTS_MONTH	AIRLINE_AIRPORT_FLIGHTS_MONTH	PRCP	ENCODED_DEP_TIME_BLK
0	10.0	7.0	100.0	10.0	20000.0	31600	1500	50.0	14

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 102ms/step
Predicted delay percentage: 33.74%
```

✓ **Model passes testing → ready for flask connection and dashboard predictions**

FLASK APPLICATION

BACKEND

Flask Application

→ Parses Frontend User Inputs
→ Calculates and Returns Delay Probability
→ Database Queries for Multiple Endpoints

FRONTEND

HTML, CSS, JavaScript

→ Predictive Flight Delay Dashboard
→ Historical Delay Trends (Interactive Visuals)
→ Navigation to Available Routes & Data

DATABASE

PostgresSQL - SQLAlchemy

Import: *'full_data_flightdelay.csv.zip'*
Database: *flightpredict*

APPLICATION DESIGN

USER INPUT

User inputs available flight information; Origin and destination airport, date of flight, airline and flight number.

VISUAL OUTPUT

Output visualizations are connected to model predictions and live flight & weather data from API.

PROCESS DATA

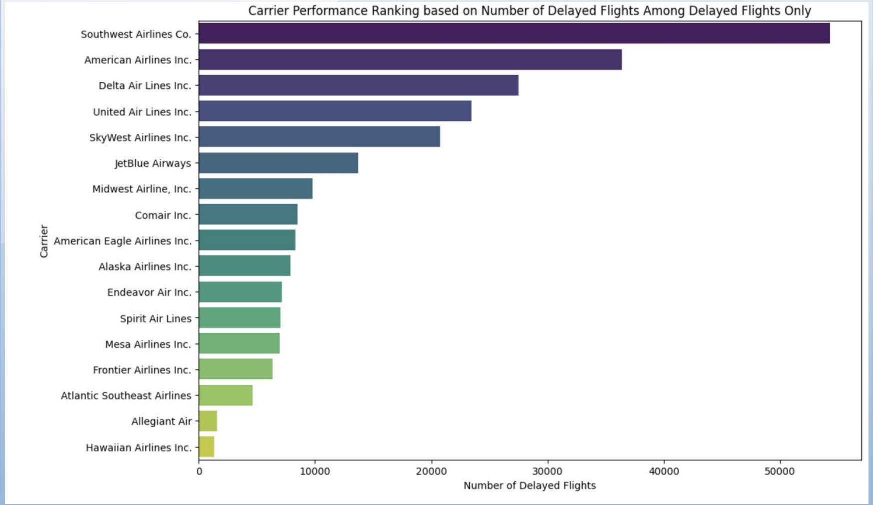
Machine learns from historical instances of flight delays, flagging for parameters that have produced past delays.

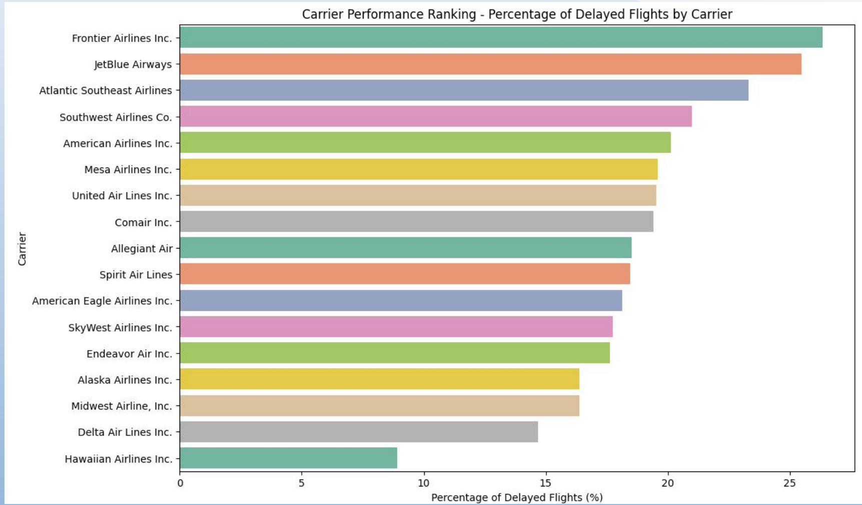
PREDICT

Live weather forecast API is compared to historical data, outputting the probability of a flight delay.


FLASK APP
DEMO

CARRIER PERFORMANCE





Frontier
jetBlue
Atlantic Southeast
Southwest
American

Hawaiian
Delta
Midwest
Alaska
Endeavor Air

CONCLUSION & LIMITATIONS

CONCLUSION

Flight delays are **highly unpredictable** due such wide variety of factors impacting the possibility of a flight being delayed *(think of the recent Crowdstrike outage).*

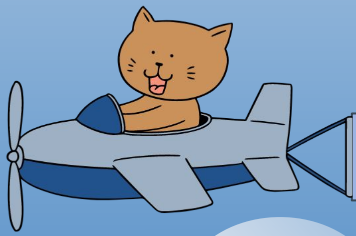
LIMITATIONS

- ✈ Some features are retrieved from the database rather than live data
  (This restricted our coverage to certain airports)
- ✈ Two features deemed important were NOT attainable through user input
- ✈ Large size of database hinders performance in parts of the application
- ✈ Time frame of a prediction must be within 5 days of the current date
  (Due to weather forecast availability through API)

FINAL THOUGHTS

NEXT STEPS

- ✈ Find other data/variables not provided in initial dataset, *potentially impacting delays* (airlines being short staffed, cap on pilot flying hours, outdated systems, etc.)
- ✈ Source more than 1 year's worth of historical data
- ✈ Regularly update flight database with live results to track detailed delay trends
- ✈ Access to higher computing power and memory resources for model fine tuning
- ✈ Additional time debugging JavaScript to improve dashboard functionality
- ✈ Hosting application online for public use



Surprising Discovery:
Spirit was NOT the worst airline carrier experiencing flight delays!

THANK YOU
HAPPY GRADUATION!

GROUP 2

Maddie Bowman, Jinlu Wang, Belinda Ho
Thomas Ngo, Nelson Lin, Meredith Jones

REPOSITORY:

github.com/maddiebowman/predictive_flight_delays

APPENDIX

WEATHER ANALYSIS

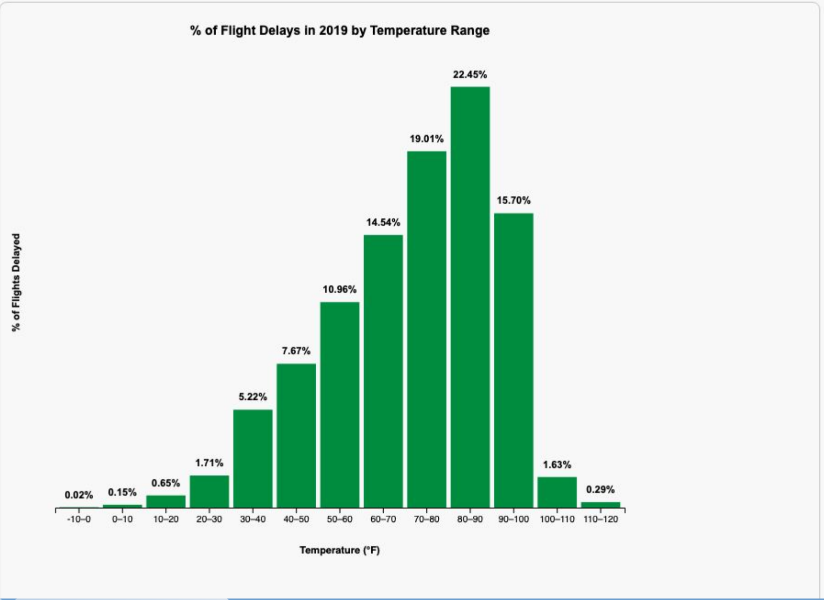
Demystifying Predictive Flight Delays with Historical and Real-Time Data

API sources: OpenWeatherMap + weather.gov

WEATHER CONDITIONS

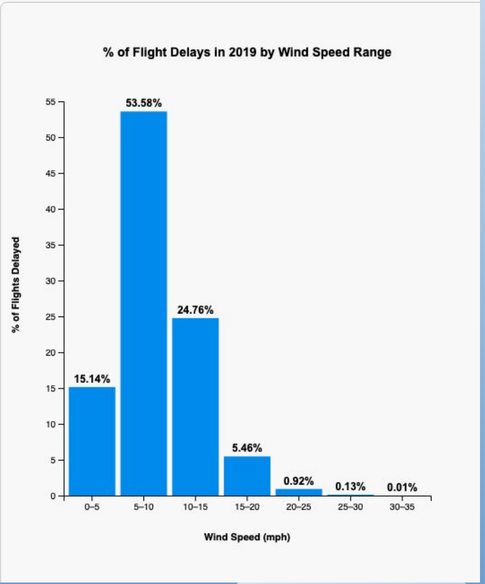
Temperature

Temperature range between 80 F and 90 F had the highest impact on flight delays in 2019, followed by 70 F and 80 F.



Wind Speed

Wind speed range between 5 and 10 mph had the highest impact on flight delays in 2019.



Note: this could be because most flights are within this temp and wind range, which is why it's important for us to look at all factors when assessing flight delays rather than just a single factor.

API sources: OpenWeatherMap + weather.gov

WEATHER CONDITIONS

PRECIPITATION

Precipitation did not appear to be a factor in 2019 dataset, despite being an important factor in model prediction.