Decision Fly-ence
Team 13

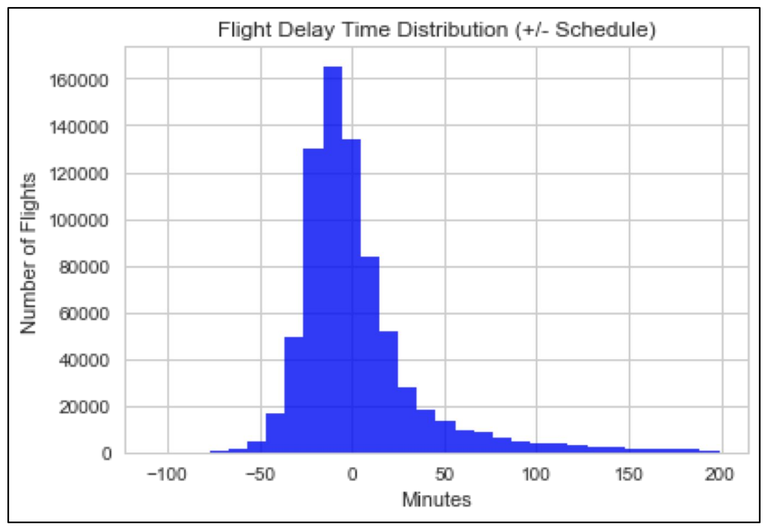
Predicting Flight Delays 5 Days In Advance

Tyler Shiovitz tshiovitz3 | Matthew Beaver mbeaver3 | Benjamin Confer bconfer3 | Samantha He sko62 | Avery Scott ascott90

Motivation

In a 2010 study, researchers estimated that domestic flight delays cost passengers **$16.7 billion** annually and have an overall economic impact of **$32.9 billion**. Additionally, the Bureau of Transportation Statistics reports that approximately **20%** of all domestic flights are delayed by 15 minutes or more.

The goal of this project is to provide airline travelers with better insight into potential disruption of an upcoming trip **up to five days** into the future using weather forecast data. With better planning around weather disruptions and other delays, we can help them take precautionary measures to minimize inconveniences and, perhaps more importantly, economic losses.



Data

We used free, public data from the BTS* as the basis for our model. We used an R script to send individual POST requests and download **120 months of flight performance data**. The initial data set was **~9M records and ~2GB**. We focused the model on flights between **4 target airports (ATL, LAX, JFK, ORD)**. This limited the flight training dataset down to **~765k records**.

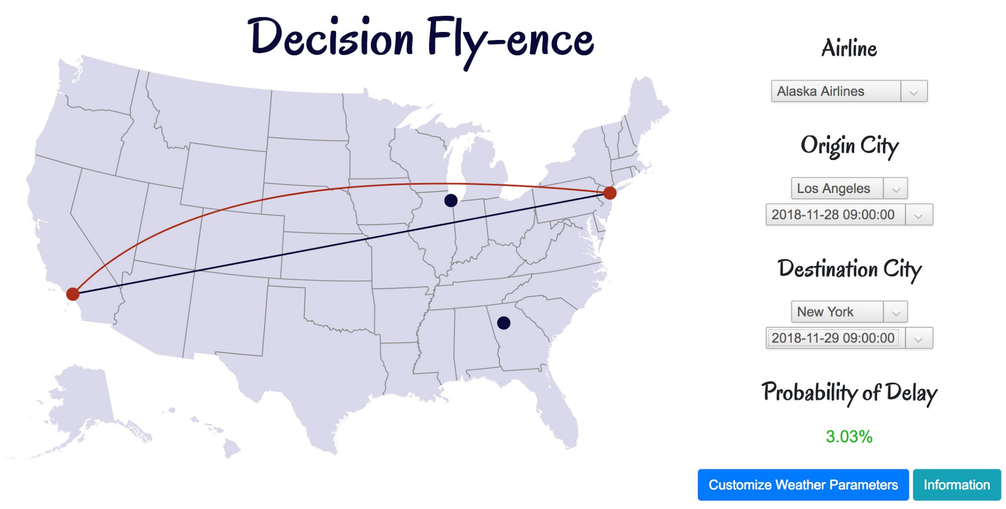
We merged the flight data with **daily historical weather data** downloaded from the NOAA** over the same time period to look for correlations between weather events and flight delays. Lastly, we **integrated a weather forecast API** from OpenWeatherMap.org to generate delay predictions within our visualization tool.

*https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
**https://bigquery.cloud.google.com/dataset/bigquery-public-data:noaa_gsod

Visualization

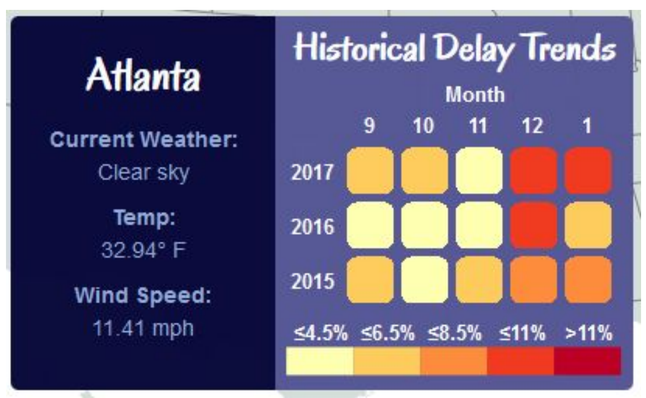
Predictions Up to 5 Days

User will input flight information and our model will use our logistic regression algorithm trained on historical flight and weather data and real-time weather forecast to determine the probability of delay.



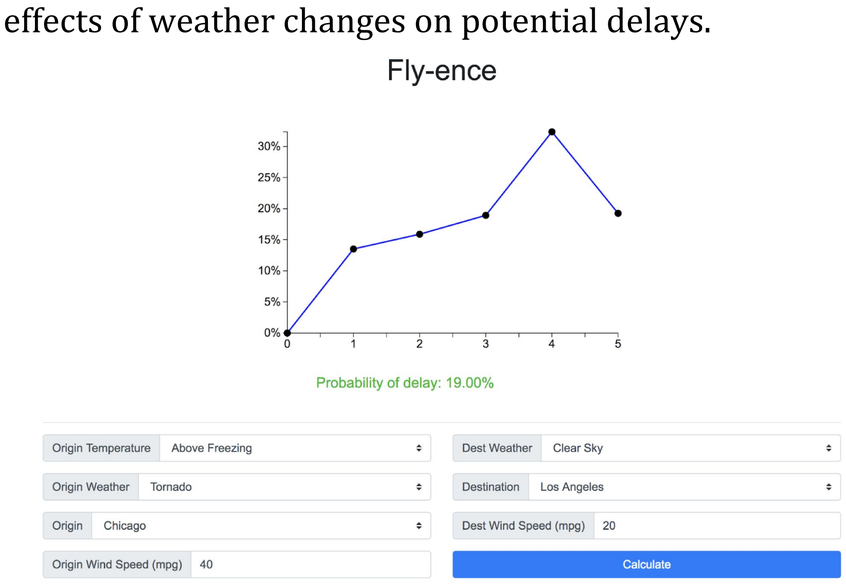
Historical Flight Metrics

Observe flight delays in the previous year for each city when hovering over our choropleth map. Representations in our heat map display a color scale which shows overall percentage of flight delay for the past three years.



Customizable Flight Parameters

Input various combinations of weather information to observe changes in the probability of delay to see the effects of weather changes on potential delays.



Algorithm

Our approach consists of accurately predicting the probability that a flight will arrive significantly later than scheduled. We tested various CART and ML models such as logistic and linear regression alongside those mentioned above, but we consistently leveraged a larger, more useful data source than predecessors in our research. Our approach incorporates a combination of weather forecasts, historical flight performance,and airline network effect. We limited our data to 4 airports of interest, ATL (Atlanta), JFK (New York), LAX (Los Angeles), and ORD (Chicago). We classify a flight as "delayed" if the projected arrival time is 60 minutes later than scheduled.

Experiments & Results

We aim to answer the following questions through our experiments:

1. What features were most important in affecting delays?
2. How much does weather contribute to flight delays and/or cancellations?
3. What is the right delay time cut-off for prediction?
4. How can we increase performance of our prediction model?
5. What is the best model to use in estimating flight delays?

In order to compare our models we will be using the following metrics:

1. Percent Accuracy
2. Confusion Matrix
3. Area Under ROC Curve
4. R^2 for Linear Regression

Experiment Summary

#	Model Type	Dependent Variable	Classification Threshold	Test Accuracy	FN*	FP*	AUC	Airline Included?
1	LR	30	0.3	85%	14%	1%	.677	No
2	MLP	30	n/a	85%	14%	14%	.677	No
3	RFC	30	n/a	86%	11%	3%	.738	No
4	LR	30	0.1	48%	3%	50%	.672	Yes
5	LR	30	0.2	80%	9%	11%	.672	Yes
6	LR	30	0.3	85%	12%	3%	.668	Yes
7	LR	30	0.4	86%	13%	1%	.668	Yes
8	LR	30	0.5	87%	13%	1%	.672	Yes
9	LR	60	0.1	81%	4%	14%	.699	Yes
10	LR	60	0.2	92%	6%	2%	.699	Yes
11	LR	60	0.3	93%	7%	1%	.700	Yes

Chosen Model
All percentages are rounded to the nearest whole percentage
*False Negatives/Positives reported as percentage of total test predictions

Logistic Model:

Variables: Year, Month, Airline, Flight Path, and Origin & Destination Weather Conditions.

$$\geq 60\ min\ Delay = \beta_0 + \beta_1 Year + \beta_2 Month + \beta_3 Airline + \beta_4 Destination\ Airport + \beta_0 Origin\ Airport + \beta_0 Destination\ Weather + \beta_7 Origin\ Weather + \varepsilon_i$$