FNCE 237: PREDICTING FLIGHT DELAYS

Sophia Africk & Kaitlynn Soo
1. **Cause:** What is the primary cause of delay for flights in the United States? How does the cause of delay impact the length of the delay?

2. **Correlated Factors:** What other factors (airport size, airline, season, day of the week, time of day, etc.) are correlated with delays and the length of delays?

**Defining Delayed**
The Department of Transportation’s Bureau of Transportation Statistics defines a flight as delayed if it **arrives at or departs from the gate 15 minutes or more after the scheduled time**
Data

Original data set 2009-2018 with 60 million+ observations

- Narrowed scope to 2018 and top three origin airports → ~7 million flight observations and ~900,000 delayed flights

Notable findings (conditioned on delay):

- **Average delay time**: 61.87 minutes
- Delay times and fraction of flights delayed **highest in June** by month and **Summer** by season
- **Airlines with highest average delays**: Frontier, Spirit, ExpressJet
- **Airlines with highest average delay times**: Frontier, Jetblue, Comair
- **Destination airports with highest average delay times**: San Francisco International Airport (SFO), LaGuardia Airport (LGA), and Dallas/Fort Worth International Airport (DFW)
- **Destination airports with highest fraction of delayed flights**: Brownsville South Padre Island International Airport (BRO), Long Island MacArthur Airport (ISP), and Meridian Regional Airport (MEI)
  - EWR only large airport on the list
Visualizations with Mapbox

Airport destinations map visualization with Philadelphia International Airport hover box

Airport destinations with scaled bubbles and heat map indicating average delay minutes

Airport destinations visualization zoomed in on the Northeast region with JFK Airport hover box

Airport destinations with scaled bubbles and heat map indicating delay counts

Airport destinations with scaled bubbles for average delay minutes and color indicating airport type
Modeling

### Categorical: Will My Flight Be Delayed?

**Naive Model**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>pact</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80.43%</td>
<td>0.00%</td>
<td>80.43%</td>
</tr>
<tr>
<td>1</td>
<td>19.57%</td>
<td>0.00%</td>
<td>19.57%</td>
</tr>
<tr>
<td>ppred</td>
<td>100.00%</td>
<td>0.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Cost = \(c \times TP + 0 \times TN + 0.5c \times FP + 2.5c \times FN\)

### Continuous: How Long Will My Delay Be?

Random Forest gives best RMSE, but OLS is more interpretable without sacrificing significant accuracy

OLS: \(-41.830565\)
Lasso: \(-41.830478\)
Random Forest: \(-40.570830\)
Gradient Boosting: \(-40.708335\)

**OLS Coefficient Highlights**

- Distance: \(-0.0007\)
- Reasons: Weather (23.70) → NAS (2.44)
- Airlines: United (3.68), JetBlue (4.95)
- Late Flight: \(-3.70\)