Predicting Flight Delays

The US Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the performance of domestic flights operated by various air carriers. This summary information includes information on whether a flight was on-time or not, how long it was delayed by, what were the factors influencing a particular flight's delay, etc. According to data from the Bureau of Transportation Statistics (BTS) of the United States, over 20% of US flights were delayed during 2018, which resulted in a severe economic impact equivalent to 41 billion USD. These delays not only cause massive economic losses but also lead to unhappy customers, stressed flight attendants, etc. A particular flight's arrival delay can also be influenced by factors beyond its control - bad weather conditions, traffic on the runway, airline strikes, etc. Hence, building a model which can accurately predict flight arrival delays between different domestic hubs based upon basic flight information will help passengers plan better and will also help airlines allocate resources more efficiently.

I. Data

The U.S. Department of Transportation's Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report, and they also graciously released flight data for 2015 publicly on Kaggle (which is where I happened upon it). There are three separate data files:

(a) Airline information – contains the IATA code and the airline name.
(b) Airport information – contains the IATA code, names of the origin and destination airports, and their geographical information (city, state, latitude, longitude).

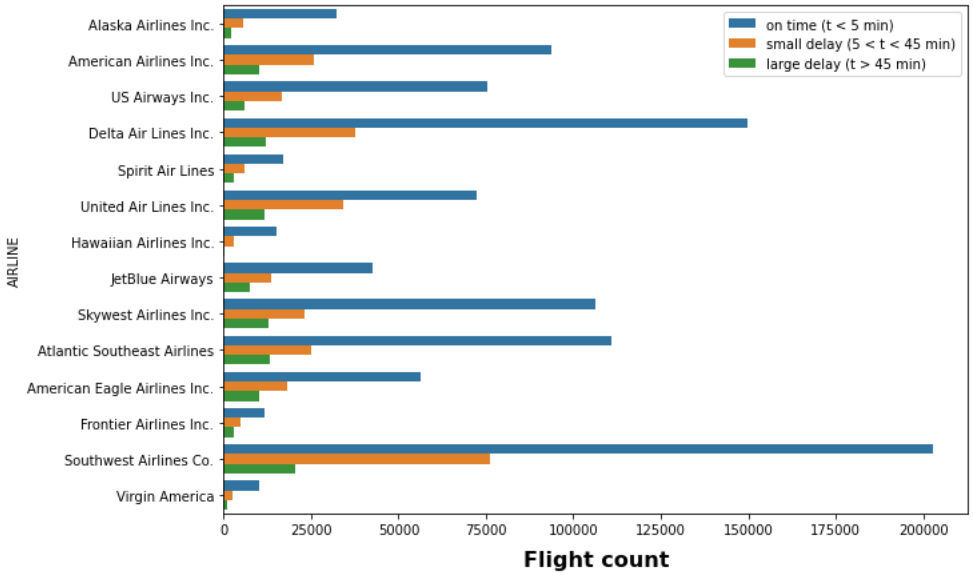
(c) Flight information – contains information about the actual trip, i.e. when the flight departed, where it departed from, was there a delay, how much of a delay was present, etc.

I wanted to make use of all of this information, and so I combined the datasets using the respective IATA codes. The dataset was very large, and so I decided to just focus on the flights that departed between Jan and March of 2015. The final raw dataset I worked with consisted of 1,403,471 rows (unique trips).

Feel free to examine the data here: https://www.kaggle.com/usdot/flight-delays

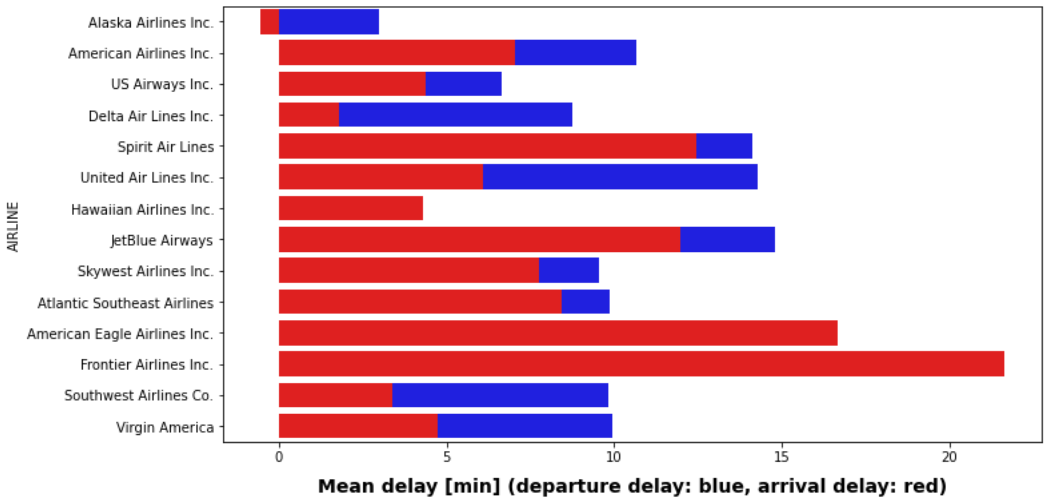
II. Exploratory Data Analysis

First, I examined some basic statistical information to get a feel for which airlines had the greatest delays. Frontier Airlines had the greatest average departure delay out of all the airlines (20 minutes), followed by Spirit Airlines (14 minutes). One possible reason for these airlines to have a greater delay than other airlines is that these are smaller/budget airlines and so do not have the same infrastructure as other, bigger, airlines. Hawaiian Airlines has the lowest average departure delay - might this have something to do with the general 'Aloha' nature of our Hawaiian friends? Below is a chart to examine the scale of delays for all the flights these airlines flew during the months in consideration:



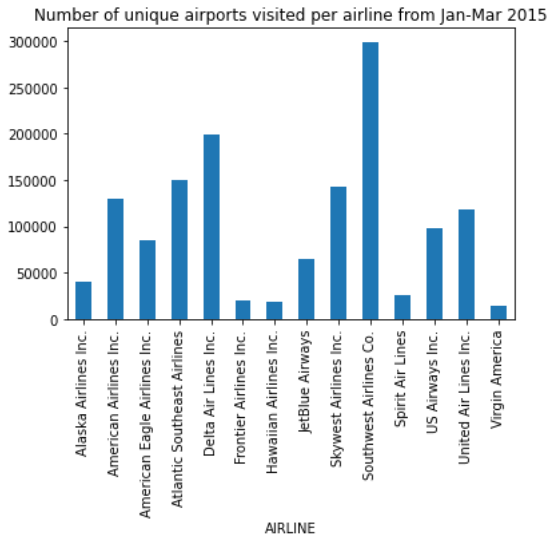
We see that although Frontier and Spirit had large average delays, they probably did not have as many flights as the other airlines, which could have skewed the distribution.

I was ultimately interested in predicting a flight's arrival delay since I believe that this is a more useful metric for delay (as compared to departure delay). Hence, I also wanted to understand the distribution of average arrival vs departure delays per airline:



Based on this chart, it looks like all the airlines have longer departure delays than arrival delays, except for Hawaiian Airlines. My intuition is that the flights can adjust speed when in air to make up for lost time, while departure delays are sometimes out of their control. It is worth noting that Alaska Airlines is the only airline among all of them which reaches the destination earlier than scheduled, on average.

I now want to examine how many unique airports were visited by each of these airlines from Jan - March 2015. This will help me understand the volume of each airline (in terms of number of destinations).



III. Feature Engineering/Data Pre-Processing

In order to determine which columns were the most useful to predict arrival delay with, I decided to explain the correlation of each variable with arrival delay. The most correlated features were departure delay and the breakdown of the airline delay (determining how much of the delay was

attributed to specific factors). The least correlated features were the latitude, longitude, and flight number (as expected, since these variables do not have a lot of important information encoded into them). I decided to only use the following columns in my prediction:

- Time of flight (month, day of week, etc)
- Airline
- Distance of flight
- Origin and destination airports

I decided to also predict the arrival delay (making it a regression problem) rather than binning the delay and predicting a categorical variable. In order to transform the object columns like the airline and the airport name, I used a label encoder to transform them into numerical values. Finally, I used the `get_dummies` function to categorically encode the month of the flight and the day of week that the flight departed.

IV. Modeling

First, I split the data into training and test sets. Then, I scaled the entire dataset to standardize all of the variables. This was to ensure that no particular variable would be unfairly weighted by the model. I wanted to test the performance of a few models to determine which would work the best on my dataset. These are the results I obtained:

```
          Random Forest Regressor
  Mean Absolute Error: 0.5711842999035724
   Mean Squared Error: 4.300947444672979
Root Mean Squared Error: 2.073872571946738
          R2 :  0.9972877439740886

        Gradient Boosting Regressor
  Mean Absolute Error: 5.861076559749258
  Mean Squared Error: 63.629400053702554
Root Mean Squared Error: 7.976803874591788
          R2 :  0.9598741379798696

          Decision Tree Regressor
  Mean Absolute Error: 1.0164757910137232
   Mean Squared Error: 7.38459583604486
Root Mean Squared Error: 2.717461285105063
          R2 :  0.9953431389681267

              Linear Regression
  Mean Absolute Error: 5.285331088876385
   Mean Squared Error: 60.11562382929442
Root Mean Squared Error: 7.753426586309722
          R2 :  0.9620899894546787
```

As we see, random forest seems to perform the best, with the lowest MAE score out of all of the models. Hence, I decided to go ahead and use this regressor to train on my data. I applied hyperparameter tuning by making use of randomized search to determine the best settings to

apply to the random forest. Finally, I trained this model on my data and determined a test score of 0.99.

V. Recommendations

- My insights can be used by airline officials to determine if there are specific times of the year when certain airlines/airports are more prone to delays than others. This will help them staff those locations better to manage their delays.
- Passengers can also use my model's results to understand the impact that arrival delays of the flight's previous trip can have on the current trip they are embarking on. It will also help them make more informed decisions when they book tickets.

VI. Next Steps

- In the future, I would love to build a UI that can take in latitude and longitude information and be able to output a score that travelers can use to understand a particular location's 'delay factor.' They can also search for specific markets/routes and specific times of the year to better predict the delay factor.
- It would also be amazing if weather information can be incorporated into this model – weather delays play a big and uncertain role in causing airline delays, and being able to account for it would help improve the model and make it more generalizable.