Predicting Flight Delays

Alaska Airlines Data Science Interview
Practical Assessment

Emily Yamauchi

Background and Objective

- Airlines are motivated to keep flights on-time, and crews may be asked to fly faster in order to mitigate flight delays
- But flying faster comes at a cost, so it does not make business sense to increase the speed if it does not mitigate delays
- Therefore, the aim is to predict whether a flight would be delayed or not based on the given variables
- Flights are classified as being delayed if they arrive at **14 minutes or more** after the scheduled arrival time

Data Sources

- Flights table- given
  The given dataset contained nearly 300,000 rows of data, with each row representing a flight from March 2020 to Feb 2021. The schema is shown on the right
- FAA Aircraft Data
  Using the Tail Number given in the dataset described above, aircraft age was found from FAA archives

Schema of the data given

Data:

The attached .csv file contains flight data from March 2020 to Feb 2021. Each row is a flight.

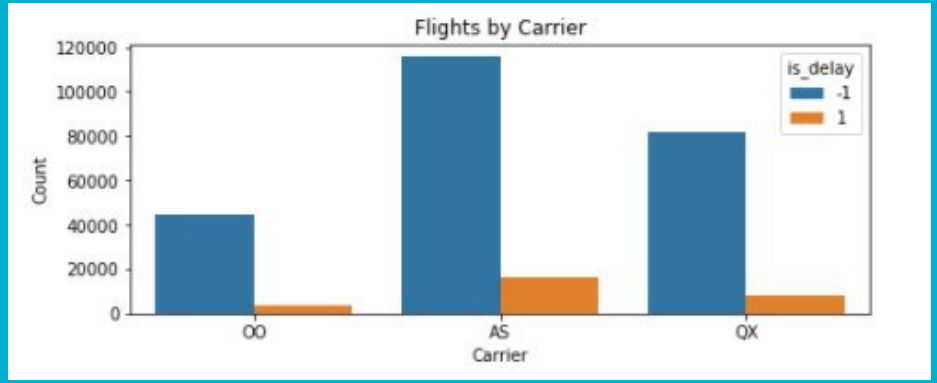
- `OPER_CARR_CD` —Operating airline – AS = Alaska Airlines, QX = Horizon Airlines, OO = Skywest Airlines
- `ACT_ORIG` —Origin airport code
- `ACT_DEST` —Destination airport code
- `FLT_NBR` —Flight number
- `ACT_FULL_TAIL_NBR` —Unique identifier for each individual aircraft
- `ACT_AC_TYPE` —Aircraft types with different performance and seat configuration – Boeing 737, Airbus A320/321, De Havilland DH8-400, Embraer E175
- `SKD_MILES` —Number of flight miles between origin and destination
- `ACT_TAXI_OUT_MIN` —Minutes between gate departure and take off from runway at the origin
- `ACT_TAXI_IN_MIN` —Minutes between touch down onto runway and gate arrival at the destination
- `SKD_ZULU_DPTR_DTTM` —Scheduled departure datetime in UTC time
- `ACT_ZULU_DPTR_DTTM` —Actual departure datetime in UTC time
- `SKD_ZULU_ARRV_DTTM` —Scheduled arrival datetime in UTC time
- `ACT_ZULU_ARRV_DTTM` —Actual arrival datetime in UTC time
- `ACT_ORIG_GMT_OFFSET` —Time offset to convert from UTC time to the origin airport's local time
- `ACT_DEST_GMT_OFFSET` —Time offset to convert from UTC time to the destination airport's local time

Assumptions Made

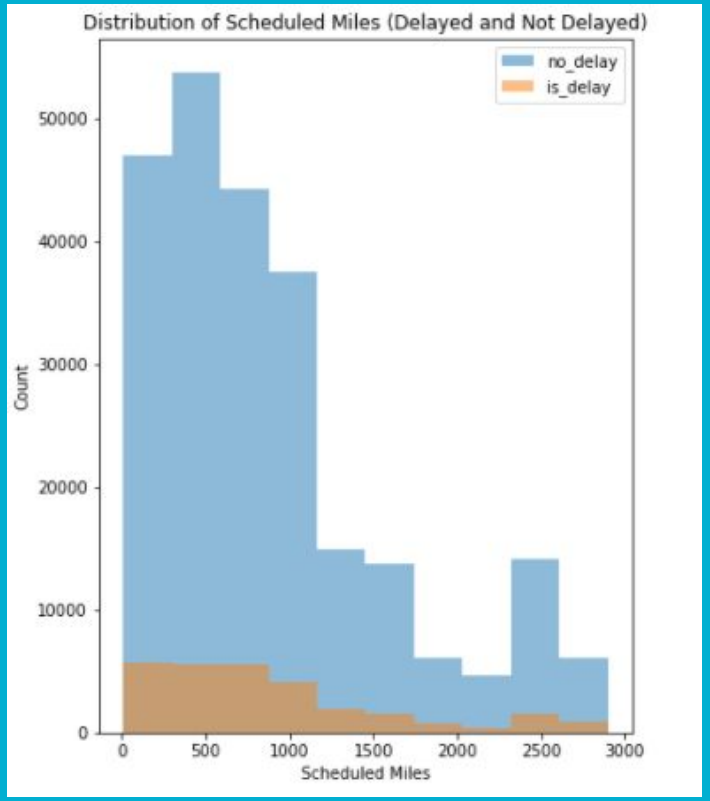
- Scheduled flight time:
  Some flights are missing data for scheduled arrivals which makes it difficult to determine whether the flight was delayed or not. Assumed then that scheduled flight times are consistent for given origin, destination, and flight number combination
- Aircraft age:
  Aircraft age was retrieved from FAA archives, but some aircrafts did not have the manufactured date- date of certification was used as a proxy in that case

Descriptive Statistics/Visualizations

- Approximately 10% of total flights were classified as delayed per the definition
- Breakdown per airline is as below:

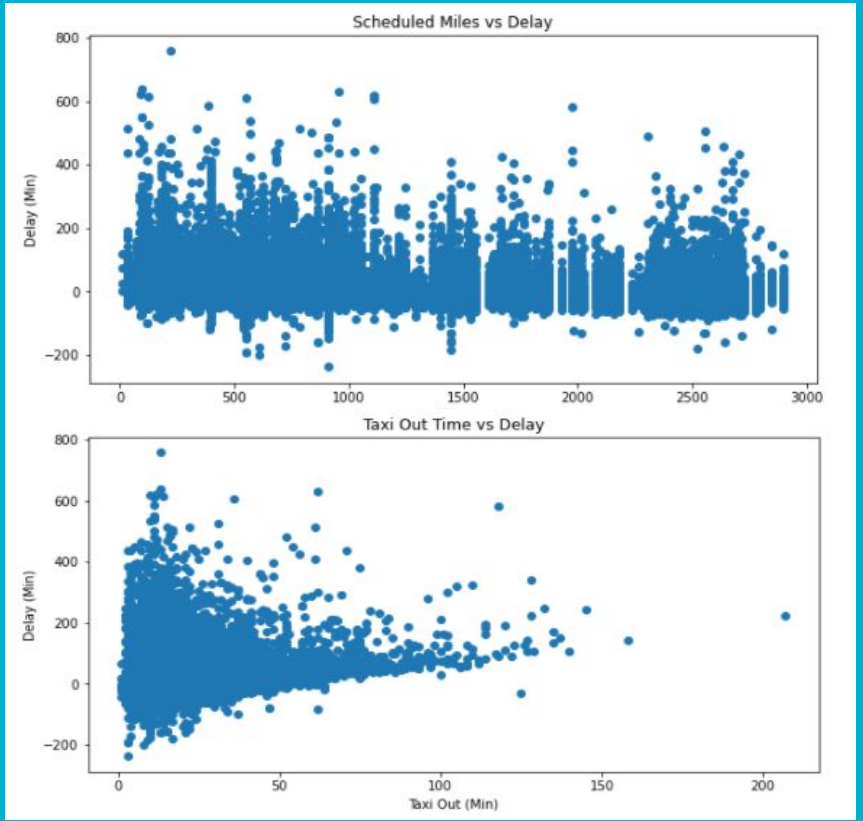


Percent of Flights Delayed by Carrier	
AS	12.36%
OO	7.00%
QX	9.09%



Descriptive Statistics/Visualizations

- The relationship between scheduled miles and delays is unclear
- However, there does appear to be some relationship between taxi out time and delays



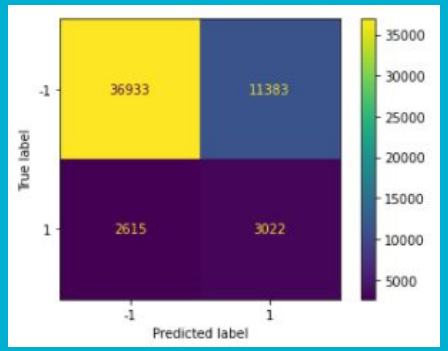
Model Selection and Iterations

- As this is a classification problem, logistic regression was used to fit and predict the likelihood of a delay
  - Used the LogisticRegressionCV model from sci-kit learn, with validation folds = 5
  - Used `balanced` method for response weights, as delayed class was roughly 10% of the training set
- Variables used-
  - Categorical variables: Operating airline, aircraft type, origin airport*, destination airport*, month of departure (*see below for various iteration)
  - Numerical variables: Scheduled miles, taxi out minutes, aircraft age
    - Taxi in minutes was excluded, as the purpose of this study was to determine whether a flight would be delayed or not midflight. Including taxi in minutes would be 'cheating', as at that point the flight would already have landed
- Iteration 1:
  - Excluded high cardinality airport codes- underlying data such as weather, wind would be meaningful, not the actual locations
- Iteration 2:
  - Included airport codes, higher dimension

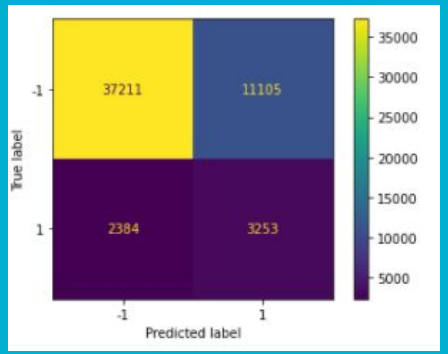
Model Performance

- What's important for the business decision?
  - High precision: flights classified as delay are actually delayed → low false positives means saving fuel costs
  - High recall: captured actual flight delays → low false negatives mean that potentially delayed flights are mitigated, improving flight performance
- Including the high cardinality airport codes did not improve performance
- As a next step, would like to include underlying variables such as weather data (can be binary for is clear, not clear, or ordinal/categorical for more granularity), or wind speed/direction

```
iteration 1:
---------------
X_train.shape:  (215810, 32)
y_train.shape:  (215810,)
X_test.shape:  (53953, 32)
y_test.shape:  (53953,)
---------------
misclassification:  0.2594480381072415
---------------
precision:  0.9243649105243399
recall:  0.7644051659905622
f1:  0.836809371141145
---------------
time took:  8.820134401321411
```




```
iteration 2:
---------------
X_train.shape:  (215810, 278)
y_train.shape:  (215810,)
X_test.shape:  (53953, 278)
y_test.shape:  (53953,)
---------------
misclassification:  0.2500139009878969
---------------
precision:  0.9196075523922499
recall:  0.7701589535557579
f1:  0.8382743861230008
---------------
time took:  44.11552453041077
```



What's Next?

- More data?
  - The origin/destination airport codes do not provide much meaning to the model without more underlying data- weather, windspeed, etc.
- Production level use?
  - Data pipeline- access underlying flight data as well as additional weather/wind data, aircraft age
  - Framework and definitions- What about weather during the flight path? Would need to develop methodology to define what weather data points would be relevant- but anecdotally have experienced delays due to weather not at origin/ destination, but during flight
- Business integration?
  - The classification model aids decision making when faced with potential delays
    - Case 1: potential delay, correct classification → model suggests to increase flight speed to mitigate potential delay, improving on-time performance
    - Case 2: no potential delay, correct classification → model suggests no delay, no additional actions taken, saving fuel costs from not increasing flight speed
    - Case 3: potential delay, incorrect classification → model incorrectly suggests no delay, no additional actions taken, saves fuel costs but decreases on-time performance
    - Case 4: no potential delay, incorrect classification → model incorrectly suggests to increase flight speed, on-time performance is improved/arrives early, but increases fuel costs
  - Which is worse error from business perspective? Case 3 or Case 4?

Thank you!

https://www.linkedin.com/in/eyamauchi/

https://github.com/emi90